SPEECH PRODUCTION MODELLING
WITH PARTICULAR REFERENCE TO ENGLISH

by
Celia Scully
University College London
Department of Phonetics and Linguistics

A thesis submitted for the degree of Doctor of Philosophy
in the University of London

1990
ABSTRACT

Many of the complexities of structure in speech signals are related to the processes of speech production. The aim of this study is to develop a better signal model in the form of a computer-implemented composite model of speech production and to apply it to some allophone sequences for British English.

The stages of speech production included in the model are: articulation, aerodynamics, derivation of acoustic sources, filtering by the vocal tract acoustic tube and radiation of a sound pressure wave. The aerodynamic processes give interactions between the various acoustic sources and between the sources and filter shapes. As a result, covarying bundles of acoustic pattern features were found in the model's outputs; these were qualitatively and, in some cases, quantitatively in agreement with corresponding patterns in natural speech.

The linguistic, anatomical and acoustic frameworks of the study are set out. Speech production processes are discussed as theory and data in relation to models. The data are drawn from natural speech production and other sensori-motor skills. The actions of speech are described kinematically. The basic physical principles and equations needed to simulate aerodynamic processes are set out. Different approaches to the acoustic processes of sources and filtering are considered. The conditions needed for the sources are described.

The composite model used in this study is described in terms of the basic principles, implementation methods and assessment.

The modelling of some phonetic classes relevant to English speech is described. Simulations of some minimal and non-minimal articulatory
contrasts, including eight published papers, are presented. A quantitative but flexible time scheme planning framework was developed and used as an input stage for the model.

The general conclusions from and limitations of the study are discussed. Future development and work are suggested.
ACKNOWLEDGEMENTS

To undertake the modelling of speech production unaided would be a futile task and I owe debts of gratitude to many colleagues. The development of the computer-implemented model from a few pages of Algol to a large number of interactive graphically oriented programs with many parameters and options was a joint effort with Mr. Ted Allwood. Ted produced good ideas for the forms of representation of the model, implemented them accurately and quickly, contributed greatly to the modelling experiments themselves and assisted with the writing of reports and papers; my debt to him is very great.

I should like to thank Colleagues in the Department of Linguistics & Phonetics, University of Leeds, especially Miss Marion Shirt, Dr. Peter Roach, Mr. David Barber and Mrs. Helen Roach, who spoke into masks, made studio recordings and applied their auditory skills to synthetic sounds, often very difficult to describe. The unfailing technical help and interest of Mr. Eric Brearley, Chief Technician of the Department has been particularly important; also that of members of the Department of Mechanical Engineering and other colleagues in the University of Leeds. I should like to thank particularly Professor Alan De Pennington and Dr. Susan Bloor who saved me from batch processing and gave much encouragement and help in preparing project applications, Dr. Malcolm Bloor who advised us on aerodynamics, especially the mixed model approach, Mr. Jim Swift the Systems Manager for the VAX, Mr. Stuart Allen and Mr. Stan Cail who designed and constructed the digital hardware clock for the model's output, Dr. Gordon Lockhart and Dr. Muhammed Zaid who helped us to design and construct the anti-aliasing filter; and from outside Leeds, Dr. Steve Terepin who advised us on the program for low-pass filtering with down-sampling and Mr. Nicholas Husband who gave us a
copy of his computer program for the reflected pressure wave method of filtering.

Of the many scholars outside Leeds from whose advice I have benefitted, I should like to thank, in particular, Professor Gunnar Fant whose book provides so many answers, but who still patiently answered my questions, Professor Ken Stevens who advised me on aerodynamics, Dr. John Holmes who advised and instructed me on many aspects of speech signals and production processes, and Professor Adrian Fourcin, the Supervisor of this work, for his wise advice and profound insights into the nature of speech communication.

The Science and Engineering Research Council (SERC, formerly SRC) supported the model development and modelling, through a series of projects.

This work describes attempts to capture quantitatively and in highly simplified form the behaviour of natural speech. The researchers who have published papers, sent me copies of their data and discussed the issues at conferences and in the journals are far too many to name individually, but I am grateful to them all. The inadequacies of the model, failures of understanding and actual errors, in this text and in the model, are entirely my responsibility.

--------------------------------------------------------------------------------
Professor Ingo Titze (almost verbatim): "If you are trying to model speech processes you must work as hard as you possibly can; then you may perhaps achieve a fairly good qualitative match with natural speech."

--------------------------------------------------------------------------------

Dedicated to Dr. John C. Scully who helped with his careful proof reading and found phonetics for me in the first place.
CONTENTS

ABSTRACT 2
ACKNOWLEDGEMENTS 4
CONTENTS: CHAPTERS I, II, III AND IV 6
LIST OF FIGURES 10
LIST OF TABLES 11
LIST OF APPENDICES: PUBLICATIONS 12
LIST OF UNITS AND ABBREVIATIONS 13
LIST OF SYMBOLS AND NOMENCLATURE IN THE MODEL
WITH REPRESENTATIVE VALUES FOR PARAMETERS 13
SPECTROGRAM AXES 16
VALUES FOR PHYSICAL CONSTANTS 16
SYMBOLS AND KEY WORDS FOR THE PHONEMES OF AN
RP ACCENT OF BRITISH ENGLISH 16

CHAPTER I: INTRODUCTION AND THE FRAMEWORK FOR THE STUDY 19

Introduction: aims and objectives of the study; productive and receptive aspects of speech communication 19
I.1 Linguistic framework 23
I.1.1 Accent and style of English modelled 24
I.1.2 Linguistic background of the real speakers 24
I.1.3 Syllable and word structure in RP English 24
I.2 Anatomical and physiological framework 26
I.3 Acoustic framework 27
I.4 Stages of speech production 33

CHAPTER II: SPEECH PRODUCTION PROCESSES 39

Introduction 39
II.1 Speech production as skilled sensori-motor activity 39
Introduction 39
II.1.1 Degrees of freedom, functions and the concept of a schema 40
II.1.2 Path dynamics, costs and constraints 45
II.1.3 Accuracy and variability 49
II.1.4 Concepts applied to speech 55
II.1.5 Feedback and feedforward 57

II.2 Mechanical properties of the structures 61

II.3 Stages of speech production: theory and data in relation to models 65
II.3.1 Neural and muscular processes 65
II.3.2 Articulation 69
   Introduction 69
   II.3.2.1 Configurations of the vocal tract for vowels and diphthongs 73
   II.3.2.2 Configurations of the vocal tract for consonants 82
   II.3.2.3 Cross-dimension to cross-section area mapping 85
   II.3.2.4 Shapes and general properties of movement paths 88
   II.3.2.5 The articulators and their movement path durations 93
   II.3.2.6 Timing, coordination and variability 111

II.3.3 Aerodynamics 112
   Introduction 112
   II.3.3.1 Lung volumes 113
   II.3.3.2 Use of the system in breathing and speech 114
   II.3.3.3 Distribution and aerodynamic properties for the air in the respiratory tract 117
   II.3.3.4 Resistance 124
   II.3.3.5 Compliance 132
   II.3.3.6 Inertance 133
   II.3.3.7 Aerodynamic data for natural speech 135

II.3.4 Acoustic sources 137
   Introduction 137
   II.3.4.1 Voice: introduction 138
   II.3.4.2 Theory of voicing mechanisms 141
   II.3.4.3 Phonation types and registers 145
   II.3.4.4 Glottal area, glottal volume flowrate and vocal fold contact area during voicing 146
   II.3.4.5 Source-filter interactions 149
   II.3.4.6 Flow in the closed phase 151
   II.3.4.7 Choice of a voice source model 151
   II.3.4.8 The range of voice waveshapes 153
   II.3.4.9 Control of sound pressure level 155
   II.3.4.10 Control of fundamental frequency 157
   II.3.4.11 Jitter and shimmer 162
   II.3.4.12 Turbulence noise 163
   II.3.4.13 Transient 167

II.3.5 Filtering 168
II.3.6 Radiation 170

II.4 Mapping across stages: stability and sensitivity 172
II.5 Specific models of speech production

II.5.1 Neural and muscular models
II.5.2 Articulatory models
II.5.3 Aerodynamic models
II.5.4 Acoustic models and electrical analogues
II.5.5 Composite models and the need to include an aerodynamic stage

CHAPTER III: A COMPOSITE MODEL OF SPEECH PRODUCTION

III.1 Principles

III.1.1 Inputs
III.1.2 Articulation
III.1.3 Aerodynamics

III.1.3.1 Physical principles applied and assumptions made

III.1.3.2 Respiratory control
III.1.3.3 The simultaneous differential equations

III.1.3.4 Composite constrictions

III.1.4 Acoustic sources

III.1.4.1 Voice
III.1.4.2 Turbulence noise
III.1.4.3 Transient

III.1.5 Acoustic filtering

III.1.6 Radiation

III.1.7 Outputs

III.2 Implementation

III.2.1 Inputs
III.2.2 Articulation
III.2.3 Aerodynamics
III.2.4 Acoustic sources

III.2.4.1 Voice
III.2.4.2 Turbulence noise
III.2.4.3 Transient

III.2.5 Acoustic filtering
III.2.6 Radiation
III.2.7 Outputs

III.2.8 Feedback and computation times

III.3 Methods of assessment

III.3.1 Comparisons with traces from natural speech
III.3.2 Auditory assessment
III.3.3 Perception experiments
CHAPTER IV: INTRODUCTION TO THE APPENDICES AND SOME RESULTS OF THE MODELLING
CONCLUSIONS FROM THE STUDY
LIMITATIONS OF THE STUDY
SUGGESTIONS FOR FUTURE WORK

Introduction 288
IV.1 Modelling of some phonetic classes 289
 IV.1.1 Static vowels 289
 IV.1.2 Diphthongs 290
 IV.1.3 Approximants 293
 IV.1.4 Fricatives 297
 IV.1.5 Plosives 297
 IV.2 Implications of some minimal and non-minimal articulatory contrasts
 Introduction 297
 IV.2.1 Coordination in vocal attack 298
 IV.2.2 Fricatives: voiced and voiceless; in different contexts 298
 IV.2.3 Devoiced [r] 299
 IV.2.4 Sequences of the form [CCV] and [CəCV] 299
 IV.2.5 Articulatory perturbations 300
 IV.3 Conclusions from the study 301
 IV.4 Limitations of the study 301
 IV.5 Suggestions for future work 306

FIGURES 307
REFERENCES 345
APPENDICES 1 TO 8 367
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig.1</td>
<td>Some of the anatomical structures used in speech</td>
<td>307</td>
</tr>
<tr>
<td>Fig.2</td>
<td>Formant frequencies as functions of vocal tract shape</td>
<td>309</td>
</tr>
<tr>
<td>Fig.3</td>
<td>Terminology and segmentation criteria for spectrograms</td>
<td>311</td>
</tr>
<tr>
<td>Fig.4</td>
<td>Simplified one-dimensional movement path</td>
<td>312</td>
</tr>
<tr>
<td>Fig.5</td>
<td>Stages of speech production, excluding feedbacks</td>
<td>313</td>
</tr>
<tr>
<td>Fig.6</td>
<td>Tongue body movement paths in natural speech</td>
<td>314</td>
</tr>
<tr>
<td>Fig.7</td>
<td>Movement paths in natural speech and the representation of tight closure configuration</td>
<td>315</td>
</tr>
<tr>
<td>Fig.8</td>
<td>Nomenclature for the lungs</td>
<td>316</td>
</tr>
<tr>
<td>Fig.9</td>
<td>Forces for the control of lung volume</td>
<td>317</td>
</tr>
<tr>
<td>Fig.10</td>
<td>Distribution of volume and flow resistance in the subglottal airways</td>
<td>318</td>
</tr>
<tr>
<td>Fig.11</td>
<td>Data for subglottal flow resistance as a function of lung volume</td>
<td>319</td>
</tr>
<tr>
<td>Fig.12</td>
<td>Sketches of the range of voice source waveshapes for natural speech</td>
<td>320</td>
</tr>
<tr>
<td>Fig.13</td>
<td>Parametric representation of the voice waveform</td>
<td>321</td>
</tr>
<tr>
<td>Fig.14</td>
<td>Variation of voice waveshape parameters with controlling parameters in natural speech</td>
<td>322</td>
</tr>
<tr>
<td>Fig.15</td>
<td>The composite model: the articulatory model</td>
<td>323</td>
</tr>
<tr>
<td>Fig.16</td>
<td>The composite model: vocal tract shapes approximating to tongue factors 1 and 2; articulatory transitions</td>
<td>324</td>
</tr>
<tr>
<td>Fig.17</td>
<td>The composite model: the aerodynamic model</td>
<td>325</td>
</tr>
<tr>
<td>Fig.18</td>
<td>The composite model: subglottal airways conductance</td>
<td>326</td>
</tr>
<tr>
<td>Fig.19</td>
<td>The composite model: composite constrictions</td>
<td>326</td>
</tr>
<tr>
<td>Fig.20</td>
<td>The composite model: three representations of the voice source (a) pulse; (b) Rosenberg (c) Fant</td>
<td>327</td>
</tr>
<tr>
<td>Fig.21</td>
<td>The composite model: block structure</td>
<td>328</td>
</tr>
<tr>
<td>Fig.22</td>
<td>The composite model: variation of K with AG in the Fant representation of the voice source</td>
<td>329</td>
</tr>
<tr>
<td>Fig.23</td>
<td>The composite model: voice table for linear mapping</td>
<td>330</td>
</tr>
<tr>
<td>Fig.24</td>
<td>The composite model: voice amplitude VOIA and F&lt;sub&gt;0&lt;/sub&gt; as functions of glottal area AG</td>
<td>331</td>
</tr>
<tr>
<td>Fig.25</td>
<td>The composite model: modulation of a noise source</td>
<td>332</td>
</tr>
<tr>
<td>Fig.26</td>
<td>Assessment of the model: radiation transform and low-pass filtering with down-sampling</td>
<td>333</td>
</tr>
<tr>
<td>Fig.27</td>
<td>An example of the modelling of different phonation types with an [i:] vowel</td>
<td>334</td>
</tr>
<tr>
<td>Fig.28</td>
<td>An example of the simulation of women's speech: [i:] vowels from different speaker types</td>
<td>335</td>
</tr>
<tr>
<td>Fig.29</td>
<td>Unsuccessful attempts at modelling [r]</td>
<td>336</td>
</tr>
<tr>
<td>Fig.30</td>
<td>A successful [r] from the model compared with natural speech</td>
<td>337</td>
</tr>
<tr>
<td>Fig.31</td>
<td>Modelling of [t] compared with natural speech</td>
<td>338</td>
</tr>
<tr>
<td>Fig.32</td>
<td>Modelling of [s] in open and close vowel contexts</td>
<td>339</td>
</tr>
<tr>
<td>Fig.33</td>
<td>Modelling &quot;A hiss it said&quot; and &quot;A his it said&quot;</td>
<td>340</td>
</tr>
<tr>
<td>Fig.34</td>
<td>Modelling of a devoiced [s]</td>
<td>341</td>
</tr>
</tbody>
</table>
Fig. 35 Modelling of [ɑtɪr'ɛɪ] and [ɑtɑr'ɛɪ] 342
Fig. 36 Modelling timing perturbations for [pə:s] and [pəːz] 343

LIST OF TABLES

Table I Geometry of the subglottal airways 120
LIST OF APPENDICES: PUBLICATIONS


LIST OF UNITS AND ABBREVIATIONS

kHz kilohertz \((10^3 \text{ Hz})\)
ms millisecond \((10^{-3} \text{ s})\)
µs microsecond \((10^{-6} \text{ s})\)
cm centimeter \((10^{-2} \text{ m})\)
µm micrometer \((10^{-6} \text{ m})\)
L litre \((10^3 \text{ cm}^3)\)
gm gram \((10^{-3} \text{ kg})\)
cmH$_2$O \((98 \text{ Pa})\) (cm of water)

Other abbreviations are defined in the text.

Multiplication is shown as \(x \times y\) or \(x.y\) or \(xy\) depending on the layout of the expression.

LIST OF SYMBOLS AND NOMENCLATURE IN THE MODEL
WITH REPRESENTATIVE VALUES FOR PARAMETERS

Programs of the model

ARTIC Articulation
AERO Aerodynamics, acoustic sources
FILT Acoustic filtering
OUT Output of signals via microprocessor
SIGNAL Construction of waveforms and signal analysis

Commands in the model

TMOD ACAL FLOW SORV etc.
Symbols for some of the physical variables and the equivalent symbols in the model

- $F_0$: fundamental frequency of the voice source
- $V_c$: vocal tract cavity volume as a function of time
- $V_L$: lung volume as a function of time
- $\frac{dV_L}{dt}$: rate of change of lung volume with time
- $P_L$: air pressure in the lungs with atmospheric pressure as reference
- $P_{exp}$: nett expiratory pressure
- $\Delta P$: air pressure drop across a constriction
- $P_{sg}$: subglottal air pressure with atmospheric pressure as reference
- $P_C$: air pressure in the vocal tract cavity with atmospheric pressure as reference
- $\frac{dP_C}{dt}$: rate of change of oral air pressure with time
- $P_{sg} - P_C$: air pressure drop across the glottis
- $P_a$: atmospheric air pressure
- $U$: volume flowrate of air
- $U_L$: volume flowrate of air out of the lungs
- $U_g$: volume flowrate of air through the glottis
- $U_c$: volume flowrate of air through the vocal tract constriction
- $U_v$: volume flowrate of air through the velopharyngeal port
- $A$: minimum cross-section area of a constriction
- $A_g$: glottal area
- $A_c$: minimum cross-section area of a vocal tract constriction
- $A_v$: cross-section area of the velopharyngeal port
- $l_g, l_c, l_v$: length of the glottal, vocal tract and velopharyngeal constriction, respectively
Representative values for some of the mechanical and aerodynamic parameters and the symbols used in the model

**Lung Volumes (see Figure 8)**

- **Total Lung Capacity (TLC)** defined as (RV + VC)
  - Vital capacity (VC) VCLU 4500 (3300 to 7000) cm$^3$
  - Residual volume (RV) RVLU 2000 (1000 to 2800) cm$^3$
  - Functional residual capacity (FRC) 3500 (0.2*VC to 0.45*VC) cm$^3$

**Subglottal Airways**

- Maximum value of flow conductance $G_{S_{G_{MAX}}}$ 2000 cm$^3$s$^{-1}$(cmH$_2$O)$^{-1}$
- Minimum value of flow conductance $G_{S_{G_{MIN}}}$

**Lung Walls and Rib Cage Combined**

- Compliance $C_{LW}$ CLW 100 (50 to 150) cm$^3$/cmH$_2$O
- Inertance $I$ 0.01 cmH$_2$O/Ls$^{-2}$
- Viscous resistance $R$ 1.2 cmH$_2$O/Ls$^{-1}$

**Vocal Tract Walls**

- Compliance $C_{C_{CW}}$ CCW 0.6 cm$^3$/cmH$_2$O

**Initial Conditions: Symbols and Representative Values**

- Initial value of lung volume VLU0 (3000 cm$^3$/s)
- Initial value of air pressure in the lungs PLU0 (0)
- Initial value of air pressure in the vocal tract PC0 (0)
SPECEROGRAM AXES

The abscissa shows time, with a time marker for 100 ms
The ordinate shows frequency, with calibration lines
at 1 kHz intervals
The degree of darkening shows sound pressure level S.P.L.
(intensity level I.L.)

VALUES FOR PHYSICAL CONSTANTS

Acceleration due to gravity $g \quad G \quad 980 \text{ cm/s}^2$
Atmospheric air pressure $P_a \quad \text{PAT} \quad 1030 \text{ cmH}_2\text{O}$
Kinematic coefficient of viscosity $\nu = \mu / \rho_a \quad 0.14 \text{ cm}^2\text{s}^{-1}$
Dynamic coefficient of viscosity $\mu \quad \text{MU} \quad 1.84 \times 10^{-4} \text{ gm/cm.s}$
(both for air at 20°C and 1030 cmH$_2$O)
The values in the vocal tract may be higher than this, since $\nu$ and $\mu$
increase with temperature for a gas (Meyler and Sutton, 1958, p.308).
Velocity of sound in the vocal tract $c \quad 350 \text{ m/s}$
(moist air at body temperature 37°C)
Ratio of specific heats for air $\gamma = c_p / c_v \quad 1.4$
($c_p$: specific heat at constant pressure)
($c_v$: specific heat at constant volume)
Density of air in the vocal tract $\rho_a \quad \text{RHOAT} \quad 0.0013 \text{ gm/cm}^3$
(moist air at body temperature 37°C)
Empirical constant in the orifice $\# K \quad K \quad 0.875$
equation

$\#$ These values for $\rho_a$ and for $K$, the values used in the model,
were based on van den Berg et al. (1957). Combined, these give an
orifice equation, Equations (II.12) and (II.13), consistent with a
correct value for $\rho_a$ of 0.00113 gm/cm$^3$ and a $K$ value of 1 (see
Section II.3.3.4).
SYMBOLS AND KEY WORDS FOR THE PHONEMES OF
AN RP ACCENT OF BRITISH ENGLISH

[ / ] Phonological structure [ ] Phonetic description

V Vowel C Consonant

['V] A stressed vowel ["V] An unstressed vowel

Long vowels

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i:/</td>
<td>&quot;eat&quot;</td>
</tr>
<tr>
<td>/a:/</td>
<td>&quot;card&quot;</td>
</tr>
<tr>
<td>/u:/</td>
<td>&quot;boot&quot;</td>
</tr>
<tr>
<td>/ɔ:/</td>
<td>&quot;fall&quot;</td>
</tr>
<tr>
<td>/ɜ:/</td>
<td>&quot;serve&quot;</td>
</tr>
</tbody>
</table>

Short vowels

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/I/</td>
<td>&quot;hit&quot;</td>
</tr>
<tr>
<td>/e/</td>
<td>&quot;fetch&quot;</td>
</tr>
<tr>
<td>/æ/</td>
<td>&quot;sat&quot;</td>
</tr>
<tr>
<td>/e/</td>
<td>&quot;again&quot;</td>
</tr>
<tr>
<td>/ɜ/</td>
<td>&quot;foot&quot;</td>
</tr>
</tbody>
</table>

Diphthongs

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ei/</td>
<td>&quot;play&quot;</td>
</tr>
<tr>
<td>/æi/</td>
<td>&quot;ear&quot;</td>
</tr>
<tr>
<td>/eə/</td>
<td>&quot;air&quot;</td>
</tr>
<tr>
<td>/ʌə/</td>
<td>&quot;poor&quot;</td>
</tr>
</tbody>
</table>

/ɔə/ "four" is sometimes included as a diphthong in the system, but Gimson (1980, p.117) states that this is found in conservative RP and that "/ɔə/ increasingly replaces earlier /ɔə/ forms...".

Consonants

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>&quot;pin&quot;</td>
</tr>
<tr>
<td>/t/</td>
<td>&quot;till&quot;</td>
</tr>
<tr>
<td>/k/</td>
<td>&quot;cat&quot;</td>
</tr>
<tr>
<td>/b/</td>
<td>&quot;bit&quot;</td>
</tr>
<tr>
<td>/d/</td>
<td>&quot;dog&quot;</td>
</tr>
<tr>
<td>/ɡ/</td>
<td>&quot;gate&quot;</td>
</tr>
<tr>
<td>/ʃ/</td>
<td>&quot;cheap&quot;</td>
</tr>
<tr>
<td>/ʃ/</td>
<td>&quot;shoe&quot;</td>
</tr>
<tr>
<td>/h/</td>
<td>&quot;head&quot;</td>
</tr>
<tr>
<td>/dʒ/</td>
<td>&quot;joke&quot;</td>
</tr>
<tr>
<td>/z/</td>
<td>&quot;azure&quot;</td>
</tr>
<tr>
<td>/f/</td>
<td>&quot;fat&quot;</td>
</tr>
<tr>
<td>/θ/</td>
<td>&quot;thin&quot;</td>
</tr>
<tr>
<td>/s/</td>
<td>&quot;son&quot;</td>
</tr>
<tr>
<td>/v/</td>
<td>&quot;vain&quot;</td>
</tr>
<tr>
<td>/ð/</td>
<td>&quot;this&quot;</td>
</tr>
<tr>
<td>/z/</td>
<td>&quot;zoo&quot;</td>
</tr>
<tr>
<td>/m/</td>
<td>&quot;man&quot;</td>
</tr>
<tr>
<td>/n/</td>
<td>&quot;nose&quot;</td>
</tr>
<tr>
<td>/ŋ/</td>
<td>&quot;sing&quot;</td>
</tr>
<tr>
<td>/l/</td>
<td>&quot;lid&quot;</td>
</tr>
<tr>
<td>/r/</td>
<td>&quot;red&quot;</td>
</tr>
<tr>
<td>/j/</td>
<td>&quot;yet&quot;</td>
</tr>
<tr>
<td>/w/</td>
<td>&quot;woman&quot;</td>
</tr>
</tbody>
</table>
Where languages other than English are included in the discussion the phonetic symbols used conform to the International Phonetic Alphabet, revised to 1979 (IPA, 1979).
CHAPTER I: INTRODUCTION AND FRAMEWORK FOR THE STUDY

Introduction: Aims and objectives of the study; productive and receptive aspects of speech communication

This study is concerned with the establishment of algorithmic relations between configurations of the vocal tract and the other air-filled tubes of the respiratory tract and the acoustic pattern features which are basic to speech. Traditionally, speech sounds are described in terms of the way they are produced. With the advent of quantitative methods of acoustic description, notably the sound spectrograph, it might be considered that for most practical purposes, such as the synthesis of speech, automatic recognition of natural speech, coding and transmission of speech signals and aids for the speech-disabled, understanding could be increased by purely acoustic analysis, since this is what is received by listeners. The many complexities and regularities of the signal as generated by the huge number of different speakers and different conditions encountered in everyday life are well known to ordinary listeners. Perceptually equivalent patterns can be recognised as such. A perceptual processing level, called normalisation, has been proposed in which listeners operate upon the auditory correlates of acoustic signals alone, without direct knowledge of speech production processes in general or in specific instances, and even without the ability to produce speech themselves (Fourcin, 1972, 1975).

Speech signals have special properties: not only because their contrasting acoustic patterns must be perceptible as such to listeners, but also because these patterns are generated by particular sets of mechanisms and processes - those of speakers. Thus the range of possible acoustic patterns in speech is in part determined by the constraints necessarily imposed by the basic
mechanisms of speech production. Within these basic limits, normal adult speakers have much freedom to combine and sequence actions, but there are limitations on the speed with which actions can be performed. Speakers are bound by physical laws of neural transmission, mechanics, aerodynamics and acoustics and by the dimensions and mechanical properties of their own speech producing apparatus.

It should be possible to relate knowledge of the physical processes of speech production to linguistic analyses of any language. The phonological systems of the language impose limits on the permissible forms of acoustic patterns used for their expression; in phonetic descriptions the domain may be auditory, acoustic or articulatory, but an articulatory view has been found useful for bringing structure and organisation into the description and understanding of spectrograms or other acoustic analysis displays. The immense value to researchers of the source-filter theory of speech production as set out from 1941 to the early 1960s by, for example, Chiba and Kajiyama (1941), Dunn (1950), Stevens et al. (1953), Stevens and House (1955, 1956), Fant (1960, 1970) is taken for granted in most quantitative analyses of speech, including the present study. Within that conceptual framework, acoustic sources, or excitations, and the filters or resonators which modify the sources to give outputs are independent of each other. To a first approximation, a linear system is assumed: at any stage, only those frequencies present in the input are found in the output; in a frequency analysis it is permissible to conflate source and filter characteristics.

A repeated finding in speech perception studies is that there is a multiplicity of cues potentially involved in the perception of speech. The possibility for this is found in the richness and complexity of the acoustic patterns of speech. There is a parallel
in speech production: a single change of action in production yields a multiplicity of acoustic pattern changes. Used in a straightforward way, the speech production mechanisms automatically generate richness and complexity of acoustic structure. Aerodynamic conditions link different portions of the mechanical system to each other and ensure that neither the different sources nor the sources and filters are completely independent of each other. As a result, covarying bundles of acoustic pattern features are to be expected in speech signals. Thus multiplicity of cues for the perception of a single phonological contrast can be traced back to production processes and constraints. Without having, in any literal sense, access to an immediate knowledge of these processes, listeners understand their auditory manifestations. If, however, the implications for acoustic structure of all of the processes of speech production could be made explicit, a deeper understanding of speech as a communication system would be gained. Disordered, as well as normal speech, would be both better understood and, potentially, better managed clinically.

What is generally called articulatory modelling is the representation of speech production, based on whatever data are available from fields such as human motor skills, observations on natural speech production, inference from acoustic structures and mathematical analysis based on the relevant physical laws such as continuity of mass flow.

It has been argued above that articulatory descriptions can shed light on acoustic structure, but these descriptions must be quantitative and sufficiently detailed. Traditional articulatory descriptions contain powerful insights, but the 'place', 'manner' and 'voicing' categories of the IPA chart (IPA, 1979) are not sufficiently precise and explicit to provide a prescription for the
production of consonants; and in the case of vowels the apparently articulatory dimensions of front-back, close-open and rounded-unrounded-spread are agreed by many phoneticians to be convenient labels for an auditory space.

A valid, quantitative framework needs to be established for the traditional phonetic categories, as a contribution to phonetic theory. There is a gulf between existing linguistic and phonetic analyses of spoken language on the one hand and, on the other hand, the observed facts of speech production as seen on X-ray films and the acoustic characteristics of speech as seen on spectrograms.

Even at a detailed phonetic level, symbols are generally discrete and ordered, without reference to a time domain. Acoustically, the speech signal is analysed as changing frequency components along a time axis, while speech production is characterised by continuous movement and overlapping gestures. Boundaries between successive phonetic units such as allophones in a symbol string are not often readily found in either the acoustic or the articulatory domains. Segmentation is recognised as necessary though problematical for acoustic analysis. Many studies have explored the relationships between these segments and phonetic units. Segmentation must be tackled for articulation also; units of articulation need to be related to phonetic units represented by symbols. Important underlying articulatory correlates of these segmental levels of representation are provided by studies of speech production processes.

The aims of the present study are:
1) to review the theory and data on which the modelling of speech production is founded;
2) to describe a composite model of speech production in which an
attempt is made to represent in highly simplified form the complex properties and observed behaviours of the real human systems;
3) to describe some experiments in which the composite model was manipulated so as to generate a variety of output acoustic patterns, with attempted explanations of the relevance of the results for speech communication with British English;
4) to work towards a quantitative but flexible descriptive framework for relating phonetic units to units of speech production.

In a work such as this it will not be possible to cover all aspects of the systems fully. Some aspects will be highlighted, others will be described only in outline. Many of the topics addressed here have been discussed by the author elsewhere; reference will be made both to published works and to one publication in press. Copies of eight published papers are included as Appendices to this study.

I.1 Linguistic framework

I.1.1 Accent and style of English to be modelled

The overall aim of this study is to explore the mapping from articulation to acoustic signal within the framework of a standard form of the language of British English. Accents may differ in terms of system, distribution and incidence of phonological units and in their phonetic realisation. The accent to be modelled is one close to that of a wide-based version of Received Pronunciation, general RP (Gimson, 1980, pp.91 & 302); it will be called Near-RP. Variations in phonetic realisations will be kept, as far as possible within the capabilities of the modelling, to within the limits of what is judged by phonetically naive listeners to be Near-RP, using recordings from a few English speakers as a guide to the accent. In much of the modelling these auditory goals are by no means achieved and
phoneticians have analysed the auditory phonetic properties of outputs from the model.

A simple, declarative style is to be modelled, with medium loudness and moderate speaking rate. The phonological contrasts to be examined are manifest in small sets of words of English, often two monosyllabic words. The words are taken to be produced in isolation or, in some cases, to carry the main sentence stress, but not in an extremely emphatic way with contrastive stress. Relative to spontaneous conversational speech, more precision is required for the synthetic utterances (O'Donnell and Todd, 1980, p.66).

1.1.2 Linguistic background of the real speakers

The aerodynamic and acoustic patterns generated by the modelling were compared with equivalent patterns from a few British speakers of English: generally two women and two men. Where auditory assessments of the naturalness of the synthetic speech were elicited some additional speakers made recordings for comparison.

All the speakers were members of the academic staff of the Department of Linguistics & Phonetics, University of Leeds. They come from different areas in the U.K. but all from England. None had a marked regional accent and they were judged to speak with accents close to the general RP in widest use.

1.1.3 Syllable and word structure in RP English

The phonological and phonetic systems in this accent of British English are described in Gimson (1980) and Roach (1983). The symbols used for the vowel and consonant phonemes and other linguistic labels such as stress markers are listed, with key words, on page 17.
The syllable in RP may be shown as (CCC)V(CCCC). There is one vowel or diphthong V which may be preceded by up to three consonants C and followed by up to four consonants. Within a polysyllabic word it is not necessary, for the purposes of this study, to assign consonants to one or other syllable of the word, since the auditory acceptability or otherwise of the whole word or longer utterance is the criterion for judging the success of the modelling.

For these judgements, the phonological feature of stress is likely to play an important part: the location of stress on one or another syllable of a word of English can, in certain cases, change the meaning of the word. Examples of this include noun-verb pairs such as "import" (noun) /ɪmˈpɔːt/ versus "import" (verb) /ɪmpˈɔːt/. The symbol ' before a vowel indicates that the syllable containing that vowel carries the word stress. Even when the meaning of a word is not altered by it, inappropriately placed stress, as judged auditorily, causes the sound sequence to be rejected as an example of normal English speech. Acoustic correlates for the perception of stress in English bisyllabic words have been shown to include the following: vocoid durations and duration ratios (see Section I.3 for the definition of vocoid in this work), pattern of fundamental frequency and frequency spectrum for the vocoids (Fry, 1955, 1958, 1964). In the experiments to be described the focus is on details of acoustic structure, with the signal segmented as described in Section I.3; the stressed syllable is given a falling pitch, by means of actions which generate a falling fundamental frequency over a portion of the utterance. It is difficult to make all the acoustic features appropriate in the modelling and the auditory balance between the two syllables in a word is often only a rough approximation to what is found in natural RP speech.
Very short voiced vocoid segments are important in English. A few word pairs are distinguished phonologically by the presence or absence of a short vowel /a/ between consonants. The presence or absence of a segment with [ə] or schwa-type acoustic formants and phonetic quality seems to be important for listeners' judgements of the word heard in pairs such as: "plight"-"polite", "blade"-"belayed", "train"-"terrain", "crowed"-"corrode" and "Clyde"-"collide" (Scully, 1973a). The actions needed to generate this kind of phonological contrast are discussed in Sections IV.2.4 and Appendix 3, Section IV.7 (Scully, 1987). Besides the short vocoid, other acoustic pattern features distinguish the pairs of words. This is an example of the multiplicity of acoustic changes sometimes found when different allophones of a phoneme are analysed, by placing the phoneme in different phonetic contexts.

I.2 Anatomical and physiological framework

The main anatomical structures involved in speech production are shown in Figures 1(a), 1(b) and 1(c). Although the structures are linked together through skeletal material, passive tissues and muscles, anatomy texts treat organs such as the larynx as entities. The organs of speech can be identified as the following: the lungs and conducting airways within the thorax or chest, the abdomen with the diaphragm separating it from the chest, the larynx, the tongue, the lips, the jaw or mandible, the soft palate or velum, the pharynx, and the nose. The whole set of air-filled passages from the air sacs of the lungs to the mouth and nose outlets is called the respiratory tract; the air-filled tube above the larynx and comprising the pharynx, the nasal cavities and the oral cavity is called the vocal tract.
The various portions of the vocal tract, as usually labelled in phonetic descriptions of articulation (for example Gimson, 1980; Roach, 1983) are shown in Figure 1(c). As discussed in Section II.3.2, it is not possible to define exactly where the boundaries between these regions lie. The labels will be used both for the solid structures and for nearby portions of the air-filled vocal tract tube.

The ear and hearing mechanisms are important for a speaker as well as for another listener. Auditory and other modalities of sensory feedback are of fundamental importance in the acquisition and skilled performance of speech. Speech is discussed as a sensori-motor skill in Section II.1. Details of the anatomical structures and their physiological functions in speech, insofar as these are known at present, may be found in, for example, Hardcastle (1976), Williams and Warwick (1980), Dickson and Maue-Dickson (1982). The arrangements and functions of the many muscles used are highly relevant for the modelling of the processes of speech production, since the muscles provide the forces which move the structures, but this study will not consider the muscles in detail, nor cite electromyographic data associated with their innervation. Articulation will not be considered here as dynamics, with analysis in terms of forces and mechanical properties; instead the emphasis will be on the movement paths for the structures treated kinematically. Since the data on which the articulatory block of the model is based are for natural speech, the properties and constraints of the neuromuscular system are implicit in the modelling.

I.3 Acoustic framework

As was stated in the Introduction above, the source-filter theory of the acoustic processes of speech production is assumed to provide an
appropriate basis for analysis and synthesis (Fant, 1960, 1970; Holmes, 1988). There are three types of sound source: voice, which originates in the larynx, turbulence noise, which originates where a jet of air emerges from a narrow orifice or constriction and becomes turbulent, and transient, a non-sustainable source which arises when air pressure at some point changes suddenly. The voice source is the quasi-periodic acoustic component of volume flowrate of air through the glottis $U_g$, which is related to the total and mean volume flowrates as discussed in Section II.3.4.1. Turbulence noise generated in the larynx just above the glottis will be called aspiration noise; turbulence noise originating from a constriction anywhere in the vocal tract between the larynx and the mouth or nose outlets will be called frication noise. The transient source is associated with the closing up or opening out of a severe vocal tract constriction, or complete closure; it will be taken to be located at or in front of the constriction.

The processes of filtering modify the soundwaves of the acoustic sources. To a first order of approximation, sources and filters are assumed to be acoustically independent of each other. It is recognised that, in a more detailed analysis, the sources and filters influence each other, as discussed in Section II.3.4.5. In the modelling, interactions between the voice source and the vocal tract filter will be introduced through the influence of aerodynamic effects which link them.

The resonances or poles of the filter function will be called $R_1$, $R_2$, $R_3$ and so on, labelled in ascending frequency order. Corresponding formants, spectral peaks in the radiated sound pressure wave, and their frequency values also, will be labelled $F_1$, $F_2$, $F_3$ and so on. Where the transfer function contains zeros these will be labelled $Z_1$, $Z_2$, $Z_3$ and so on.
When the acoustic source is frication noise located between a constriction of the non-nasal vocal tract and the mouth outlet, the poles of the back cavity remove energy and so introduce zeros into the transfer function from the source to the mouth outlet. If the constriction is of very small cross-section, the portions of the vocal tract filter behind and in front of the source act almost independently of each other and the poles of the whole filter almost coincide with the poles of the front and back cavities considered separately. In that case, the poles and zeros associated with the back cavity portion of the filter nearly cancel out and the radiated sound pressure wave contains mainly poles or resonances of the front cavity only, between the constriction and the mouth outlet. Their associated formants are seen on spectrograms as darker regions of noise; they will be labelled $F_a$, $F_b$, $F_c$ and so on, as shown in Figure 3, to distinguish them from formants arising from poles of the whole vocal tract in vowel-like filter configurations.

The most important aspect of the vocal tract filtering for the modelling described in this study is the mapping from vocal tract shape to formant pattern. The acoustically relevant property of the vocal tract air-filled resonator is its whole area function: the cross-section area of the tube along its whole length from the glottis to the lips and along the nasal cavity to the nostrils in addition, for articulatorily nasalised configurations.

Even if the vocal tract tube is quantised into about thirty to thirty-five sections each 0.5 cm in length, as is done for the filtering implementation described in Section III.2.5, the whole area function has too many degrees of freedom for convenience, quite apart from questions of naturalness. Parametric representation of the filter shape supplements traditional phonetic descriptions of tongue shapes and gives insight into the kinds of tube shapes that are
needed in order to obtain particular formant patterns in the synthetic speech signal (Stevens and House, 1955, 1956; Fant, 1960, 1970). In Fant's (1960, 1970) model the vocal tract filter is represented either as four abutting cylindrical tubes: for the back cavity, the constricted portion, the front cavity and the lip outlet shape (op. cit. Figure 1.4-9, pp.76-77), or with a horn-shaped constriction (op. cit. Figure 1.4-11 a),b),c), pp.82-84). Two of these Figures for the horn-shaped constriction are reproduced here as Figures 2(a) and 2(b).

In these Figures, the three parameters for a given vocal tract length, which is 16.5 cm if there is no lip section and 17.5 cm when a 1 cm lip section is added, are: position of the vocal tract constriction dfg, minimum cross-section area of the constriction $A_{\min}$, and lip outlet cross-section area $A_1$ with length $l_1$. A smaller value of $A_1/l_1$ for a narrower longer outlet tube simulates increased rounding and protrusion of the lips. Figure 2(a) shows the effects on formant frequencies of different degrees of lip rounding and protrusion when the vocal tract constriction is severe, with $A_{\min} = 0.65 \text{ cm}^2$. Figure 2(b) shows the acoustic effects of different degrees of vocal tract narrowing combined with a wide open mouth shape $l_1 = 0 \text{ cm}$ and $A_1 = 8 \text{ cm}^2$.

Particular acoustic pattern features are predicted from the tube and horn models, as discussed by Fant (1960, 1970, pp.63-90). Two examples are: first, the close approach of $F_2$ and $F_3$ if there is a severe palatal constriction at a distance from the glottis varying between about 11 and 13.5 cm depending upon the amount of lip outlet narrowing; second, starting with a palatal place for the constriction, dfg about 11 cm as for an [i:]-type vowel, the similar effects upon $F_2$ of either narrowing the lip outlet or retracting the constriction to a place nearer the glottis.
Cavity affiliation is an important consideration in the modelling. Where a formant frequency rises as the constriction moves forward from the glottis towards the lips, that resonance is affiliated mainly with the cavity in front of the constriction and its frequency is very sensitive to lip shape. To the left of the formant frequency maximum, where the formant frequency falls as the constriction moves forward, there is mixed cavity affiliation for that formant. Where the formant frequency in Figure 2(a) is essentially constant as lip shape changes that formant is affiliated mainly with the back cavity. Knowledge of this mapping is especially valuable for vowels, but is relevant for some consonants also. As discussed in Section IV.1.3, the vocal tract shapes needed for English [r] can be understood with reference to the nomograms of Figure 2(a).

Vocal tract shapes needed for fricative and plosive consonants can be approximately predicted from these nomograms. Formant transitions, which provide important acoustic cues for the perception of place and manner of articulation for consonants, have been charted with a similar acoustic vocal tract model by Stevens and House (1956).

Segmentation of spectrograms for the acoustic description of speech signals, whether for natural speech or for outputs from a model, is performed in this study on three distinct bases. The first of these is in terms of the source type: voice, turbulence noise, transient or silence. The second uses relative labels vocoid and contoid. Vowel-like segments, called vocoids, have voice as the source and stronger formants visible in the frequency band 500 to 4000 Hz approximately. Dark bands at very low frequencies, associated with closed vocal tract resonances, are excluded. Consonant-like segments or contoids are weaker in this same frequency region; their source type may be either voice or noise or both. The terms vocoid and contoid will be used as purely acoustic signal descriptions; they
are not taken to imply specific articulatory contrasts. Weakening of higher formants with a voice source is likely to occur when a portion of the vocal tract is severely narrowed, but the acoustic effects seen on spectrograms of the output may come from changes in the voice source as well as from filtering effects. The third criterion for segmentation is a sudden change of formant pattern, for example seen in noise if the place of articulation for a fricative changes suddenly.

Figure 3(a) shows the spectrogram segmentation criteria and the notation to be used for the acoustic description of speech-like sequences containing plosives. Voice onset time (VOT) defines the time interval from the transient at the release of the plosive to the first cycle of voicing seen on a spectrogram. VOT is positive if voicing begins after the release transient and is negative if voice is present before the release; if the acoustic closure is completely voiced, or with voice disappearing only just before the release, VOT will be negative and equal to acoustic closure duration. Voice persistence time (VPT) is the duration for voicing within the acoustic closure, measured from the offset of the preceding vocoid. Vocoid onset time (VCT) is the time interval from the release transient to the onset of a following vocoid. A similar set of labels is used for the segmentation of spectrograms for speech-like sequences containing fricatives, as shown in Figure 3(b).

Most of the acoustic energy for speech lies between about 50 Hz and about 10 kHz. Sounds between about 3 and 5 kHz are auditorily more prominent than might be supposed from their sound pressure levels because normal human hearing is particularly sensitive to this range of frequencies (Moore, 1982, p.45). Spectrograms are usually given a high frequency emphasis which provides a visual effect somewhat
comparable to the auditory one; the spectrograms in this work are of that kind.

1.4 Stages of speech production

Between a linguistic message, a meaning intended by a speaker, and an acoustic signal which transmits the encoded message to listeners who decode it, lie the successive stages of speech production, through which information is transmitted. The number of stages identified depends on the focus of interest. In this study the emphasis is upon the ways in which skilled, coordinated actions and an airstream, moving from a region of higher air pressure to a region of lower pressure, combine so as to generate acoustic disturbances in the airstream.

The study takes as a starting point for the description of speech production the positions, mechanical states and movement paths of all the solid structures used and the corresponding configurations of the air-filled tubes of the respiratory tract, comprising the lower or subglottal airways from the air sacs in the lungs to just below the glottis in the larynx, the glottis, and the vocal tract or supraglottal airways from the glottis to the mouth and nose outlets. This will be called the articulatory stage of speech production. In contrast to some usage elsewhere (see, for example, Abercrombie, 1967, pp.21-22), the articulatory system will not be confined to structures above the larynx. Articulation will not be considered to be functionally opposed to initiation and to phonation, as proposed by Catford (1977, p.15). It was not found possible to construct an internally consistent descriptive framework on these bases. Instead, all of the structures, whether in the larynx or below it or above it, are given equal status as articulators. A change to one of the
articulators is likely to alter the conditions in more than one subsequent stage of speech production.

Preceding stages may be said to include the ideation and neurolinguistic stages: selection of semantic content and of grammatical, lexical, phonological and phonetic characteristics, Laver (1980a); and the neural and neuromuscular stages: activity within the central nervous system for skilled control and innervation of muscles and the resultant patterns of muscle forces. Information about the linguistic planning and selection operations in the central nervous system can be inferred from observable linguistic behaviour, for example, slips of the tongue (for example, Boomer and Laver, 1968). The operations performed in the ideation and neurolinguistic stages will not be considered here. The linguistic message will be represented as a sequence of appropriate allophones of English phonemes; this forms one of the inputs to the model to be described here. Neuromuscular processes will be mentioned only briefly, in connection with skilled sensori-motor actions, Section II.1, and in connection with specific models of speech production, Section II.5.1.

The mean values of volume flowrate of air through different portions of the respiratory tract and of air pressure at different places in the respiratory tract changing at frequencies less than about 50 Hz will be considered to be aerodynamic parameters. The term aerodynamics, as used in this work, will exclude acoustic components of speech: rapid air movements and air pressure changes falling within the band of frequencies audible to human listeners with normal hearing, from about 20 Hz to 20 kHz, but for speech sounds mainly between 50 Hz and 10 kHz, as stated above in Section I.3.

Normal adult speech is produced almost entirely during the expiratory phase of breathing. The inspiratory phases for quiet breathing and
for speech are probably similar, although often of shorter duration and with higher volume flowrates during speech (unpublished laboratory data). The expiratory phase of speech differs from that of quiet respiration in two respects: for speech, the rate of decrease of lung volume and of volume flowrate of air out from the lungs is controlled and there is at least one major obstruction to the flow of air.

The acoustic sources of speech all arise at or near constricted portions of the respiratory tract: generally, in a healthy speaker, in the larynx or the vocal tract but not in the subglottal airways. Here a slowly moving and slowly changing airstream is partially converted into air movements and air pressure changes that are rapid enough to be perceived as audible sound by normal hearing people. The slowly changing airstream contributes, in combination with the states and positions of the movable solid structures, to the source-generating mechanisms. Therefore the aerodynamic processes can be considered to precede the processes of source generation; that is the view adopted in this study. Aerodynamic processes will be assumed to follow ideal gas laws under isothermal conditions, while the usual adiabatic conditions are assumed for acoustic disturbances (see, for example, Flanagan, 1972a, p.30).

The terms voice, voiced and voicing can have many different meanings. In this study, voiced and voiceless are used as general phonetic labels for phonologically contrasting allophones. Those terms do not define the domain of the voice source, for which the terms voice and voicing are reserved. The interactions between articulation and aerodynamics in the generation of acoustic sources are complex. Rothenberg (1968, p.69) emphasises the importance of making a distinction between an adjustment of the larynx such that at the speaker's normal transglottal pressure the vocal folds vibrate with
closure or near closure in each cycle and the vocal fold vibration itself. The former may be called a phonation state of the vocal folds and is an aspect of articulation; the latter is neither articulation nor acoustic source, since the voice source is the waveform of the periodically interrupted airflow. The myoelastic-aerodynamic theory of voicing (see, for example, Titze, 1980) is accepted as a basis for analysis and synthesis in this study. Vocal fold articulation in the model is matched to glottal area changes observed in natural speech, but with any oscillatory component omitted. Rapid oscillations of glottal area at the fundamental frequency of voicing are included only in the derivation of an optional within-cycle variation in glottal losses for the filtering stage, described in Appendix 4 (Scully and Allwood, 1982).

The final stage of speech production will be taken to be the radiation of a sound pressure wave, mainly from the mouth and nose. Auditory and other kinds of sensory feedback available to a speaker will not be considered in detail, nor will feedback loops be included as part of the model structure. The view taken here is that trial and error with monitoring is an important aspect of speech acquisition. Suitable patterns of articulation need to be discovered for the model also in order to approach a specified auditory sound pattern goal. The method employed is partly knowledge based on published theory and data and partly trial and error, with the experimenters providing the required auditory monitoring and feedback.
The stages of speech production will be identified in this study as follows:

1) **Neurolinguistic stage:** the ideas and the intended meaning of the message; the selection of appropriate sentence forms, syntactical structures, lexical items, and phonological and phonetic characteristics;

2) **Neural stage:** the planning and realisation of appropriately sequenced and coordinated neural signals to the muscles;

3) **Neuromuscular stage:** the patterns of innervation of the muscles and the resulting forces operating upon the movable structures;

4) **Articulatory stage:** the positions, adjustments, mechanical states and movement paths of the movable structures, the articulators; their shaping of the respiratory tract, from the lung walls to the mouth and nose outlets;

5) **Aerodynamic stage:** the low frequency components of time-varying patterns of airflow and air pressure throughout the respiratory tract;

6) **Acoustic sources stage:** the generation of acoustic components of airflow and air pressure;

7) **Acoustic filtering stage:** the modification of acoustic sources by processes of filtering, with a time-varying transfer function;

8) **Radiation stage:** the radiation of a sound pressure wave, primarily from the mouth and nose, but also from the throat walls and other surfaces.
Figure 5 shows these eight stages in a slightly simplified form. Stages 4) to 8) above are simulated in the composite model to be described here. Some aspects of neural timing organisation, Stage 2), might be said to be represented in the articulatory time scheme described in Appendix 3 (Scully, 1987). Some of the theory and data on which a model of this kind needs to be based are considered next, in Chapter II.
Chapter II SPEECH PRODUCTION PROCESSES

Introduction

This chapter considers processes for normal, not disordered, speech production by adults. Data from a number of different speakers and languages are considered here, so that the phonological system of contrasts and the phonetic patterning will not always be that of an RP accent of British English. Nevertheless, some trends in the data for items that are phonetically similar to the accent to be modelled will be considered here and used as a basis for modelling. Speech production is taken to consist of a number of different stages as discussed in Section I.4. The stages are considered in turn here. Variability, which pervades all of speech communication, is considered and the mapping between the stages, since they cannot be considered in isolation from each other. Normal adult speech production is probably the most widespread human skill and this aspect of speech production will be considered first.

II.1 Speech production as skilled sensori-motor activity

Introduction

Speech production will be considered here as an example of a goal-directed motor skill; the goals are taken to be a broadly defined sequence of auditory sound patterns following in the intended order, without extra, intrusive sounds. The processes must be robust so that the auditory goal may be reliably achieved in the presence of perturbations in the magnitude, direction and timing of muscle forces and in the resulting articulatory paths.
When considering normal adult speech it seems reasonable to suppose that the level of skill achieved is very high. A distinction needs to be made between skilled performance and the processes of skill acquisition which precede it. Some insight into speech production skills may be gained by consideration of other human motor skills, or even those of other animals. The implications suggested here for speech are speculative.

II.1.1 Degrees of freedom, functions and the concept of a schema

A view adopted by some neurophysiologists, following Hughlings Jackson, is that movements, rather than muscles, are represented in the motor area of the central nervous system (Morton, 1982).

If each of the muscles relevant to speech production is considered separately (see, for example, Hardcastle, 1976), then the number of controlling parameters could be said to be as high as forty; if the number of degrees of freedom is equated to the number of planes of motion possible at a joint, a usage adapted from physics and mechanics according to Kelso (1982, pp.23-24), then the figure might be small. The joints involved in speech production comprise the temporo-mandibular joint, the cricothyroid and cricoarytenoid joints in the larynx, and the joints between ribs and vertebrae; on this basis there might be said to be only one to two degrees of freedom for vocal tract shapes.

Neither of these approaches appears appropriate: the positions and movements of some of the structures must be controlled, but not necessarily with respect to a joint. One problem in accounting for patterns of tongue movement under the influence of muscle forces is that not all the tongue muscles are attached to a rigid skeletal structure. The velopharyngeal port must be opened or closed by a
combination of velum movement and nasopharynx constriction; the lips must be rounded, protruded, spread, retracted; the tongue must move within the vocal tract cavity partly independently of jaw actions. The concept of a number of articulators, structures which move more or less independently of each other, seems more appropriate. The solid structures correspond to the organs of speech of traditional phonetics nomenclature.

A precise figure, or even an approximate one, for the likely number of degrees of freedom in skills equivalent in complexity to speech production does not seem to be readily available. Approximate figures have been suggested, however, for the number of articulatory parameters that need to be controlled. Ladefoged (1979) for example suggests sixteen as necessary and sufficient, as follows:

1) tongue front raising and lowering;
2) tongue back raising and lowering;
3) tongue tip raising and lowering;
4) tip advancing and retracting;
5) pharynx width;
6) tongue bunching;
7) lateral tongue contraction;
8) lip height;
9) lip width;
10) lip protrusion;
11) velic opening;
12) larynx lowering;
13) glottal aperture;
14) phonation tension;
15) glottal length;
16) lung volume decrement.
Following Kelso (1982, Chap.2, p.28), it seems helpful to distinguish between coordination, control and skill. Within the overall auditory goal, the aim in speech production is to shape appropriately the vocal tract cavities and the orifices of the whole respiratory tract, as functions of time. The means used by normal adult speakers are controlled movement of about ten structures. Coordination concerns the discovery of functions which constrain some free variables so that they act as a behavioural unit, a synergy or coordinative structure. This use of the term coordination seems to be compatible with the notion of inter-articulator coordination, an essential concept for speech production. Articulators are likely to be combined into different sets for different portions of speech.

Values must be assigned to the variables in the functions; this is the process of control. As the values become more nearly optimised for the particular task the level of skill increases. Some articulators may need to be precisely coordinated in specific ways for a particular speech sound or class of sounds, while other articulators may be relatively free. More crucial and less crucial articulators, primary and secondary in Bothorel's (1983) terms, are likely to show respectively small and large dispersions for that particular class of sounds.

Two statements about degrees of freedom appear at first sight to be incompatible. Increased skill means an increase in the number of degrees of freedom to be controlled (Kelso, 1982, p.24), yet the process of coordination, which is a prerequisite for skill, constrains variables, thus reducing the number of degrees of freedom. It seems that these statements can be reconciled when the concept of the discovery of a function f for coordination is incorporated. Then the measures of goodness or skill of motor performance for a particular goal are taken to be:
(a) the number \((n)\) of independent variables \((V_x)\) included in the function \(f(V_1, V_2, \ldots, V_n)\); the larger the value of \(n\), the greater the skill;

(b) the distance from optimum of the values assigned to each of the variables \(V_1, V_2, \ldots, V_n\); the smaller the distances the greater the skill. What is meant by optimum needs to be defined.

Schmidt's (1982) concept of a schema provides a similar analysis. It is supposed that there is a generalised program or schema through which a relationship or rule is learned; the schema relates the movement outcome to the values of perhaps as many as four movement parameters:

1. overall duration for a single trajectory;
2. overall force used for a single trajectory;
3. spatial orientation for a single trajectory;
4. pattern size for sequenced trajectories, as in writing one's signature small or large (Schmidt, 1982, pp.222-224).

It is suggested that a recall schema is gradually built up from both successful and unsuccessful attempts at a goal-directed action. Knowledge of the results is stored, along with the sensory patterns accompanying the action. The values of the movement parameters are adjusted to take account of factors which influence the outcome, for example, weight and wind speed and direction when aiming to throw an object a specified distance. Gradually, as skill increases, more factors are taken into account and more dimensions are added to the multidimensional relationships. Goals become more finely differentiated. Not one but many functions of \(V_1, V_2, \ldots, V_n\) are implied by a schema of this kind.

Both errors and successes help to build up a stronger, better defined and more extensive schema relationship, in Schmidt's view (1982,
It is necessary to learn which combinations of parameter values to avoid as well as which to choose. In modelling speech production, for example, it is necessary to avoid extreme palatal narrowing of the vocal tract by the tongue when producing a bilabial or alveolar plosive; auditory feedback shows that some combinations of parameter values for the tongue give inappropriate results. In the experience of some speech therapists and teachers of the deaf (personal communication) large numbers of errors must be avoided because they cause discouragement. In these cases, it seems appropriate to aim at a simpler schema initially.

In trying to apply these motor skill concepts to speech, it seems reasonable to equate the variables $V_1, V_2, \ldots, V_n$ in (a) and (b) above with the approximately sixteen articulatory parameters that seem to be needed. The functions of $V_1, V_2, \ldots, V_n$ describe inter-articulator coordinations and the paths of individual articulators, for particular allophones in particular phonetic contexts. The totality of functions defines an articulatory time scheme for a particular speech sound sequence: a succession of parallel coordinated gestures.

A young child might be unskilled in the sense of using few articulators, yet highly skilled as regards inter-articulator coordination for those that are used. There is evidence for example that new-born babies acquire, generally within about three days, the skill of coordinating their actions so as to allow sucking and breathing to alternate, with low variability in the timings of events (Selley, 1980).

External factors such as wind speed for throwing a ball are not likely to influence articulation very much, but speakers need to learn to allow for factors such as a change of rate or style of
II.1

Speaking; they learn to speak while eating, while smoking a pipe, from different body postures and so on; they need to learn to make an articulator arrive at a required state or position from many different starting states, perhaps under increasingly severe timing constraints. The auditory goals become more differentiated as ordinary speech develops; further differentiation surely takes place for professional acting and singing. Timing variability in speech produced by young children is gradually reduced and more complex gestural sequences are gradually mastered (Sarah Hawkins, personal communication).

II.1.2 Path dynamics, costs and constraints

Some aspects of the basic principles and modes of operation which speech production may have in common with other motor skills have been touched on elsewhere (Scully, 1984b). Here the objectives, dynamics, costs and constraints of skilled movement will be discussed further and an attempt will be made to consider their relevance for speech production.

The concept of a series of target states for the articulators, with transitional movements towards but not necessarily reaching them (Lindblom, 1963) has proved fruitful for speech research. The term end point will be used here for the position or state reached by an articulatory parameter when it has zero rate of change with time at the end of a transition, regardless of whether that state was aimed at and reached or whether some further point was aimed at but not reached.

The moving structures of speech production have been analysed as second order mechanical systems with stiffness and viscous losses as well as mass (Öhman, 1967). With the applied forces modelled as
rectangular pulses, increasing force magnitude resulted in increased jaw displacement, accompanied by a rather small increase in the time to reach maximum displacement (Lindblom, 1967). In Nelson's (1983) analysis it was assumed instead that recoil forces due to stretched passive tissues could be neglected and the stiffness of activated muscles providing the forces was considered as part of the input. An upper limit for the applied force per unit mass implied an upper limit for acceleration if frictional losses were neglected. These assumptions, combined with the idealised force step function of Lindblom's (1967) analysis, give a very simple set of time functions for the path parameters displacement d and velocity v, Newton's laws of motion (Meyler and Sutton, 1958, p.275), as shown in Figure 4.

In this simplified dynamics analysis, displacements in the accelerating and decelerating phases are parabolic functions of time. If the applied forces F are increased but their application times remain constant, peak velocity \( v_{\text{max}} \) increases in proportion and so does total distance moved D. As long as path distance is increased mainly by this method there will be a very strong correlation between distance D and peak velocity \( v_{\text{max}} \). The process is constrained by the costs of increasing the forces. Introduction of a time gap between the end of the accelerating force and the start of the decelerating force could increase the distance moved during the transition once maximum force had been reached but would increase the time taken.

Movement paths for the dorsum of the tongue have been analysed with more emphasis on the changes in stiffness of the structures when muscles which form them change their state to give the forces needed for movement, (Ostry and Munhall, 1985, described in Section II.1.4).

There must be upper limits for the forces applied, for the total energy costs, for peak acceleration and for the maximum jerk value,
the rate of change of acceleration. Trade-offs between objectives, expressed as distances and durations, and the performance constraints listed have been considered for the dynamics of one-dimensional displacement-versus-time paths by Nelson (1983). He analysed velocity-time patterns having a single acceleration phase followed by a single deceleration phase, since these unimodal paths are inherently efficient. He derived the path shapes for the cases in which minimum duration for a given path distance or maximum distance for a given path duration is achieved by the use of maximum force per unit mass throughout, called "bang-bang" control. In the case where path duration and distance are specified, the minimum force needed was derived, called "bing-bing" control. Knee regions were found in graphs of cost versus path duration, for all four cost factors, where there is a good trade-off between path duration and costs. For shorter path durations all the costs increase very rapidly. This study gives quantitative expression to the notion of ease of articulation.

As an activity, speech production needs to be not too tiring and sustainable over long time spans. Ordinary speech, unlike trained singing or acting, seems likely to operate in low effort regions of muscle forces and with moderate movements about joints. Extremely open jaw positions, for example, seem unlikely to be found in conversational speech or ordinary text reading, although these are required for Cardinal Vowel 5! An upper limit for more or less continuous speech might be suggested: after sixteen hours of talking, one person consulted a speech therapist for his voice problem; the therapist's advice was to stop talking (Rolnick, 1980, personal communication). Presumably the rate of tiring increases and the time span for talking is reduced if a fast rate of speech is used.
The maximum forces in speech may be related to the maximum permissible forces applied to the muscles, tendons and bones. In a small animal where the masses are low acceleration can be high for a given force (Buller and Buller, 1980; Stålberg and Young, 1981), and this might be a factor in speech. Muscles can be classified as fast-twitch or slow-twitch. Fast-twitch muscles shorten about 2 to 2.5 times faster than slow-twitch muscles of the same mammal. Slow-twitch muscles are better suited to maintaining a fixed posture, since they use less energy when maintaining a constant force, and are less susceptible to fatigue. Physiologists consider actions in terms of speed of muscle movement, defined as rate of change in muscle length divided by the muscle's resting length. A value of 1/s is considered to be quite fast. During chewing, for example, the fastest muscle speed found is 5/s (K. Appenteng, 1983, personal communication). It seems that many land mammals have a maximum shortening velocity for their fast-twitch muscles that is just low enough to avoid self-injury (Buller and Buller, 1980).

It is necessary to avoid damage to the structures in other ways also. For example, the jaw must not be raised so rapidly that there is a danger of the upper and lower teeth clashing together. Consistent with this supposition, there are data which indicate that for some speakers jaw raising is slower than jaw lowering when both actions are performed rapidly (Kiritani et al., 1982).

The tongue can alter its shape greatly, but it is constrained by the essentially incompressible nature of its tissues to retain a nearly constant volume. There may be practical limitations on the stretching of its surface: Houde's (1968) cineradiographic study indicated a maximum stretch between marked tongue surface points of about 25% in the root region, but less than 10% elsewhere, during speech-like sequences. Kent (1972) found that a straight line drawn
between two tongue points, one in the dorsum, the other in the root region, could vary by 38% during an utterance. Although the tongue tip or tip-blade region needs to be considered as a structure in its own right, it is, of course attached to the main body of the tongue and is not infinitely extensible. It may be for this reason, among others, that the body of the tongue moves forward in the oral cavity for the production of consonants requiring a raised tongue tip or blade near the front of the hard palate. Since the underneath of the tip is attached to the floor of the mouth by the frenulum some cooperation between tongue and jaw is likely to be needed when a severe constriction of the vocal tract is to be made.

II.1.3 Accuracy and variability

Fitts (1954) quantified the interactions between movement path properties of distance D, duration T and accuracy defined by the width of the target W. The results of three different tasks studied could be expressed by a single binary index of performance $I_p$ which remained at an approximately constant value across a wide range of task conditions, viz.

$$I_p = - \frac{1}{T} \log_2 \left( \frac{W}{2D} \right) \quad \text{bit/s}$$  \hspace{1cm} (II.1)

$W$ is an expression of tolerance, in the same units as $D$; a large value of $W$ implies low accuracy requirements for the on-target end of the movement. Constancy of $I_p$ is known as Fitts' law. It is sometimes expressed as:

$$T = a + b \log_2 \left( \frac{2D}{W} \right)$$  \hspace{1cm} (II.2)

where $a$ and $b$ are constants (Schmidt et al., 1979).

Fitts' (1954) binary index of difficulty was:

$$I_d = - \log_2 \left( \frac{W}{2D} \right) \quad \text{bit/response}$$  \hspace{1cm} (II.3)
$I_d$ was insensitive to the work required to perform the task and to the direction of movement. It omitted from consideration the small finger movements involved in the tasks. Most of the magnitudes in Fitt's experiments lie well outside those for speech production: his distance $D$ ranged from 1 up to 32 inches and duration $T$ went as high as 1 second. Where a small $D$ value of 2 inches was combined with a very low difficulty index value of 1 or 2, $W$ equal to $D$ or $D/2$ respectively, movement durations were only about 200 ms, not much beyond articulatory transition durations in speech. The distances moved in speech are almost certainly less than 2 inches generally, so the tolerance for target width would need to be large in speech to be consistent with predictions extrapolated from Fitt's findings.

Motor tasks performed by children appear to conform to Fitt's findings, but with larger movement durations at the same index of task difficulty in younger children. Capacity, defined as index of difficulty divided by movement duration (Fitts and Peterson, 1964), increases with age (Sugden, 1980; Hay, 1981). It may be helpful to postulate two aspects of movement control: pre-programmed force-time plans for the muscles and feedback-mediated approaches to a target; the integration of the two may improve with age (Hay, 1981).

Where accuracy of transition duration was emphasised as the goal set, a tendency was found in arm movements by adults for the fractional duration error to decrease sharply if average velocity increased from 5 to 25 cm/s; it continued to decrease slightly as average velocity increased still further up to about 300 cm/s. These results were relatively independent of movement duration and distance taken separately (Newell et al., 1980). Average velocities for articulatory transitions in speech have been cited as 5 to 20 cm/s, or up to 30 cm/s in fast movements (Netsell, 1982).
Movements having durations between 80 and 250 ms have been found to be better retained under conditions of competing secondary tasks than those with longer durations of 600 to 2500 ms (Carlton, 1978). Distances, durations and velocities found in speech are considered in Sections II.3.2.4 and II.3.2.5. The findings outlined here suggest that there may be motor control advantages in avoiding very low velocities and very long durations for articulatory transitions.

An additional complicating factor for Fitt's law of approximately constant performance index has been demonstrated by Kelso et al. (1979). Both difficult and easy tasks were presented to subjects who moved their arms to reach a smaller or larger target, with smaller or larger tolerance respectively, at different distances. When either left or right arm was used alone or if both arms together performed tasks of equal movement difficulty, durations conformed to the trade-offs in Fitt's law: movement time decreased as distance decreased and/or tolerance increased. But when the two hands were given tasks of unequal difficulty, both arms moved together with virtually synchronised patterns of displacement, velocity and acceleration, such that the time taken was more than that for the easier task if performed on its own and a little less than that for the more difficult task on its own. The two arms appear to be controlled as a single unit. The motor control task is presumably simplified by the use of a single timing pattern; perhaps this is the simplest possible synergistic function through which two actions may be linked. The authors make it clear that they do not view this kind of control simplification as inevitable: the subject could probably learn to make independently timed two-arm movements if required to do so by a specific goal set.

There may be implications in these findings for movement control in speech. Houde's (1968) cineradiographic study showed different
portions of the tongue body moving in just this kind of synchronised way over paths of different distances. It may be a simple and convenient solution for a speaker to synchronise muscle force patterns for the whole tongue where the auditory goal does not require any particular coordination between different parts of the tongue. Perhaps young children do synchronise movement paths for other sets of articulators and only gradually learn to dissociate their timings and achieve more subtle auditory goals.

Like Nelson (1983), Schmidt (1982, pp.208 ff.) has considered the parameters of basic unimodal transitions. He focussed on the accuracies achieved and offered plausible explanations for movement parameter interconnections. His conceptual framework is that of a motor schema which has been built up for a particular motor skill, as outlined above in Section II.1.1. Operation at one point along a function of several independent task-related variables is equivalent to the selection of a particular, and appropriate, motor programme. Pairs of muscles acting reciprocally are assumed by Schmidt to provide the forces needed, in the following sequence: agonist force to accelerate the structure, antagonist force to decelerate it, agonist force again to clamp the movement at the end point. Timing relationships are part of the specification and the whole force-time pattern determines the distance moved and the time taken to move.

Let us suppose that the values of the agonist and antagonist forces are increased, while the timing pattern remains constant; then the total transition time is unaltered while its distance increases in proportion to the increase in force level. This appears to be a convenient and simple way of increasing distance traversed, since only one of the two aspects of the motor force-time program has been modified. Caution is needed, however, in speculating as to what constitutes ease, convenience and simplicity for the central nervous
system. It is perhaps relevant to note that larger muscles forces do not require more time to reach their peaks than smaller ones (Ghez, 1981).

Evart's monkey apparently programmed its muscles for the force required under different load conditions so as to achieve an imposed goal of approximately constant transition duration, in lever pushing experiments (Evarts, 1967, cited by Ghez, 1981 and by Eccles, 1977, pp. 111-112). Electromyographic traces for two pyramidal cells in the monkey's motor cortex were strikingly correlated as synergists for this task, but not for a different shoulder-moving action; they exhibited functional linkage. The idea of a monkey keeping transition duration constant between 400 and 700 ms seems at first sight to have bearings on speech production, but one might argue that the kinematic facts of Evarts (1967) study simply reflect, and do not explain, the control tactics. The monkey was rewarded with grape juice; what would be the rewards for particular kinds of transition control in speech production? They may be perceptual as well as or instead of productive; in speech research the requirements of production cannot be considered in isolation.

In tasks where subjects aimed a spot at a target on an oscilloscope screen, approximately linear relationships were found by Schmidt et al. (1979) for within-subject variability expressed as standard deviation in force \( \sigma_F \) as a function of the level of force \( F \). In tasks where subjects operated a lever to make a spot move between two target zones on an oscilloscope screen, within-subject variability in applied impulse duration \( \sigma_T \) increased linearly with movement duration, approximately \( 2 \times T \). For small forces of less than 1 kg and for durations between 200 and 500 ms, the slopes were:

\[
\frac{\sigma_F}{F} \approx 0.1 \quad \text{(II.4)}
\]

\[
\frac{\sigma_T}{T} \approx 0.05 \text{ or } 0.1 \quad \text{(II.5)}
\]
The uncertainty in Equation (II.5) arises because it is not clear in the published graph and text whether Figure 4 (Schmidt et al., 1979, p.422) shows impulse duration or twice impulse duration as abscissa. The minimum value of $\sigma_F$ is about 16 gm. A minimum value for $\sigma_T$ cannot be estimated from the data: linear extrapolation gives a negative $\sigma_T$ intercept of about -5 ms and it must be supposed that the slope of $\sigma_T$ against $T$ would be less at small values of $T$. At 200 ms for $T$ or $2T$, $\sigma_T$ is about 7 ms.

Schmidt et al. (1979) derived from their data and Newton's laws of motion predictions for the accuracy in distance traversed when the force-time pattern is altered.

Under the assumption of increased force levels but not different timing as a means of increasing distance $D$ traversed without altering the time taken, an increase in $D$ is achieved by larger $F$ and therefore larger $\sigma_F$ and a greater variability in the total nett impulse $\int F \, dt$ for the transition. Variability for distance $\sigma_D$ must be proportional to $D$ in this case.

If, on the other hand, there is a requirement to traverse the same distance $D$ in a reduced time $T$, the force levels need to be increased and in addition the pattern of forces needs to be applied over a shorter time interval. Two opposing variability factors are at work in this case: increased force level implies that total impulse variability increases proportionately, but Schmidt et al. (1979) argued that impulse variability is inversely proportional to the square of the movement duration, so that the nett effect is to reduce the distance variability. Combining force level and movement duration effects, it was predicted that $\sigma_D$ would be inversely proportional to $T$ in this case and that the combined effect of force
and timing changes could be expressed as:

\[ W \propto D/T \quad \text{(II.6)} \]

where \( W \) is the target width. Whereas Fitt's index of difficulty, Equation (II.3) above, based on hits of a target of different widths, depends on \( W/2D \), Schmidt et al. (1979) called \( D/T \) the movement difficulty. They used \( D \) and \( T \) as independent variables for movement of a stylus to a target position and measured \( W \) as variability, in two directions. They found that for durations of 140, 170 and 200 ms and distances of 10, 20 and 30 cm, target width in the direction of movement was proportional to \( D/T \). Linear extrapolation beyond the data points (Schmidt et al., 1979, Figure 9, p.427) suggests a minimum distance error, or target width, of about 2 mm. It is difficult to judge whether accuracy in speech production is likely to be similar to this: the durations are similar, but the distances are much larger and the velocities correspondingly smaller than in speech.

Actual values for parameters found in these and other studies of motor skills may be unreliable guides to speech production. But some of the principles suggested seem to offer the possibility at least of explanations for the forms of articulatory transitions and the ways they can be modified to increase speaking rate. For example, one factor in the optimisation of the articulatory variables may be the need to keep total muscle force as low as possible so as to minimise force variability.

II.1.4 Concepts applied to speech

There have been many recent analyses of articulation focussed on the concept of coordinative structures or functional synergies: groups of muscles and joints organised for specific speech tasks. This approach is sometimes called Action theory (see, for example, Fowler
et al., 1980). Transition durations are considered to be determined by a change in the values for stiffness and rest length of the structures. It is argued that stable inter-articulator coordination can be described as a phase angle in a distance-velocity plane without recourse to explicit timing; the phase angle may perhaps remain constant across stress or speaking rate changes (Kelso, 1986). There seems to be no clear evidence for a controlling clock in the central nervous system but some evidence points against invariant phase angle. The issues concerning absolute versus relative control of timing in the central nervous system have been reviewed by Keller (1989).

Parameters of force, mass, stiffness and viscous damping cannot be reliably measured for the structures and the muscles; when testing the mathematical models of the dynamics, kinematic data have to be used: position, velocity and acceleration as functions of time for individual articulators or for two or more articulators that are believed to be functionally linked. For example relationships between peak velocity $v_{\text{max}}$, distance $D$ and duration $T$ for tongue dorsum transitions have been interpreted in terms of a mass-spring system without damping and fitted to the expression:

$$v_{\text{max}}/D = c_s/T$$

(II.7)

where $c_s$ is a constant whose value is 1.57 in a mass-spring system Ostry and Munhall (1985). The $c_s$ values for their data, 1.8 to 1.9, are closer to the value of 2 predicted for a mass accelerated by a constant force as shown in Figure 3. Presumably goodness of fit must be balanced against plausibility of the model; perhaps the mass-spring model is more realistic than an assumption of a step function for the accelerating force with no stiffness or viscous loss.

One way of approaching the question of functional linkage is to apply an unexpected perturbation to one articulator and observe whether the
other one is involved in any compensation that occurs. The jaw and lips, for example, appear to be functionally linked for the action of closing the lips needed for a [p]. If the jaw was unexpectedly loaded during a closing phase the lip actions compensated (Folkins and Abbs, 1975). Relative timings of peak velocity between lower lip, upper lip and jaw were found to be the same whether the lower lip was unexpectedly loaded or not during the closing action for [p] (Gracco and Abbs, 1988).

An experiment described by Shipp supports the view that the lips and the larynx are genuinely independent and do not have a closed loop relationship. Oral air pressure was mechanically released at random moments during bilabial plosive production. Although the normal time course of oral pressure rise was thus interrupted, the timing of the laryngeal muscle activity for abduction or adduction of the vocal folds remained constant (Shipp, 1982).

Where two or more structures controlled by separate sets of muscles move cooperatively so as to achieve jointly the shaping of some region of the vocal tract, this will be called functional synergy. Articulators which shape or otherwise control more or less distinct, separate portions or properties of the respiratory tract are skilfully coordinated in speech production. Where the relative timings of their actions are controlled for the production of a specific sequence of allophones this will be referred to as inter-articulator coordination.

II.1.5 Feedback and feedforward

The role of feedback has not been discussed in detail so far, but is clearly of immense importance for the acquisition and performance of a motor skill such as speech. Indeed, speech production is often
described as a sensori-motor skill. Many feedback paths and modalities exist in the speech producing systems (see, for example, Hardcastle, 1976; Gentil, 1989). Visual feedback should be included, to take account of a listener's reactions in face-to-face communication (Fairbanks, 1954). Auditory feedback must play a central role in both acquisition and skilled performance since speech production is directed towards auditory goals. This feedback acts after the event and is part of the trial and error with monitoring needed to build up the kinds of sensori-motor schemata needed.

The times associated with all the neural pathways and processes involved in feedback are important for speech considered as a control system. Estimates of feedback loop or dead times vary greatly in the literature, but are believed to be as short as 10 to 20 ms from the cerebellum to the motor associative area of the cerebral cortex (Eccles, 1977, pp.133-137). Eccles conjectured that skilled, learned movements are highly preprogrammed with much interaction between these two portions of the central nervous system; exploratory movements, by contrast, are slower and make more use of on-line correction.

One value for the time delay between a stimulus and the electromyographic response signal is 8 ms for the tendon reflex of a jaw muscle and 14 ms for the muscles's stretch reflex. It is suggested that the time differences may indicate different path lengths: whereas the tendon reflex is spinal, the stretch reflex, muscle spindle, feedback loop may go back via a fast pathway to the cortex (Morton, 1982). It is apparent that most of the muscle motor units used to control the vocal tract are at the shortest possible distance from the motor cortex and cerebellum; feedback loops must surely be quicker for these speech actions than for any other sensori-motor skill.
Feedback times have been tracked along the auditory pathways for click stimuli. The time delays are longer for weaker signals. Lutman (1983) for example gives cochlear response latencies ranging from 2 to 5 ms as clicks decrease in acoustic intensity. Response delays increase along the brainstem up to about 10 ms; a much later cortical response can have a delay of between 50 and 300 ms. Similarly long delay values, 100 to 200 ms, are cited for motor reactions to auditory and proprioceptive stimuli in a review of speech production control models (Kent, 1976).

A dichotomy has sometimes been proposed for the relationship between a sequence of linguistic units such as phonemes as represented in the central nervous system and the actions associated with their production. In a preprogramming strategy, sometimes called a comb model, the actions succeed each other without feedback; in a feedback strategy, sometimes called a chain model, sensory feedback from the execution of one unit must be received before the action for the next unit is initiated. Kent (1976) has reviewed these models and the issues connected with serial ordering in speech. It seems to the present writer that motor control for speech production is likely to be too complex to be explained by either the chain or the comb model.

Schmidt (1982) suggested that a selected motor program determines the overall pattern of movement while reflex mechanisms ensure that this pattern is carried out faithfully. He considered human limitations on the use of feedback to correct actions already prepared. By reference to experiments in which subjects tried to reverse actions already prepared to go, he distinguished between short and long duration transitions. Visual signals were involved, so the results may not have direct relevance to speech, but it was suggested that transitions of less than about 150 to 200 ms duration are
irrevocable, once initiated within the central nervous system. If this is the case, then each articulatory transition in speech production is likely to be irrevocable.

An articulatory transition can be mediated by on-line feedback however, and this process has been demonstrated in loading experiments such as those cited in Section II.1.4 (Folkins and Abbs, 1975; Gracco and Abbs, 1988). The perturbed articulator itself performs compensatory movements tending to achieve the goal, for example, lip closure for [p]; this autogenic compensation is thought to be under closed-loop feedback control. Response delays are too long to be explained by peripheral reflex actions; lower brainstem reflexes have latencies of 12 to 18 ms for the facial muscles, while these latencies were between about 35 and 75 ms. Responses appear to take place in remote portions of the neuromuscular system also; with delay times as short as 35 to 75 ms. They seem likely to be achieved by open-loop feedforward control mechanisms (Abbs et al., 1984).

Signals in the central nervous system associated with the innervation of a muscle for one articulator are input to regions of the central nervous system controlling another muscle that moves a different articulator, but is able to act as a synergist to the first muscle for a particular goal. Potential errors in the innervation of the first muscle are detected in the experience-based representation in the central nervous system of the relationships between the two independent muscles. Error correction can thus take place through both muscles. This process of feedforward is quicker than reliance on closed loop feedback. Since it relies upon learned relationships, it seems likely to be found in skilled speech actions, for example in tongue-jaw interactions (Gentil, 1989) as well as in jaw-lip interactions as suggested by Abbs et al. (1984).
II.2. Mechanical properties of the structures

The organic materials of which speech production systems are composed are very complex in structure. In many cases their properties vary with temperature, with lung volume and with frequency. Temperature alters along the respiratory tract, lung volume changes during the course of an utterance and the pattern of energy distribution across frequency is the very fabric of speech signals; therefore all these three variables are relevant to speech production.

A distinction will be made between a low frequency range, from d.c. up to about 50 Hz, which is applicable to the aerodynamic system of speech production, and higher frequencies, the a.c. range encompassed by speech signals, from about 50 Hz upwards, applicable in the modelling of acoustic sources and filtering. It will be demonstrated that some components of mechanical impedance may be considered to be negligible at the low frequencies, even though they may need to be taken into account in the acoustic range.

Speech production uses many different kinds of materials, as described in standard anatomical textbooks (Hardcastle, 1976; Williams and Warwick, 1980). Control of the mechanical properties and the applied forces is exercised by the central and peripheral nervous systems, as discussed by Müller et al. (1981). However, neurological tissues will be considered here only to the extent that they contribute to the limitations and constraints expected or observed in the moving structures. Attention will be focussed here on two states of matter: solid objects, which form channels and cavities of complicated and changing shapes, and the air enclosed within these channels and cavities. Mass or inertance, resistance or viscous loss and compliance or its inverse, stiffness, are the three mechanical properties to be considered. Their electrical
counterparts are often used in modelling; the correspondence between
the two domains is given by Rothenberg (1968, pp.7-8) (see also
Sections II.3.3 and II.5.4). Although they share some underlying
mathematics, the electrical circuit theory approach is not ideally
suited to handling the flow-dependent resistances which are of
central importance in speech aerodynamics. Electrical resistances
are generally independent of current, obeying Ohm's law V=IR where V
is voltage drop across a resistance R and I is the current flowing
through it. The most important flow resistances in speech production
are flow-dependent; they are best expressed through the orifice
equation, with minimum cross-section area in the place of conductance
(see Section II.3.3). The electrical model presupposes a massless
conducting medium, whereas the transporting medium to be considered
here is air, which has a mass (Liljencrants, 1985, Section 1.12.).

The structures are non-uniform. In the vocal folds, for example,
layers having different mechanical properties have been identified.
The tissues are anisotropic, with different properties along and
across the direction of the muscle fibres. Hooke's law is not
obeyed: in each direction the stress-strain curves are nonlinear;
as the tissues elongate they become stiffer, so that there are
practical limits on their elongation. This has profound effects on
the vibration of the vocal folds (Titze, 1981). Below ultrasonic
frequencies the tissues are essentially incompressible: deformation
in one direction is accompanied by the opposite deformation in
another, leaving volume and density constant. The tissues can
stiffen or slacken as a result of the actions of muscles both outside
and within them, but they are neither purely elastic nor purely
viscous, being composed of elastic fibres surrounded by liquid.
Viscoelastic behaviour includes creep and relaxation to a steady
state, so that the tissue properties depend on their previous strain-
stress history. These effects are frequency dependent. There are
significant changes of structure and mechanical properties with the speaker's age (Titze, 1981; Hirano, 1983).

There have been very few experimental studies to determine the mechanical properties of the solid structures of the respiratory system as a function of frequency. Ishizaka et al. (1975) have investigated the mechanical properties of the walls of the cheeks and the throat at acoustic frequencies. Fant et al. (1976) showed through external vibration amplitude contours that wall impedance has two dominant regions: near the lips and near the larynx. The vocal tract walls act as a mass element which sets lower limits to the low frequency resonance of the vocal tract when closed, and appears also to be a significant factor contributing to the tuning of $F_1$ in [a:]-type vowels where there is a very small back cavity volume. If $F_{1i}$ is the lowest resonant frequency of the vocal tract with hard walls and $F_w$ is the resonant frequency for the wall mass, then the lowest resonant frequency with the wall mass included is $F_1$ where

$$F_1 = (F_{1i}^2 + F_w^2)^{0.5}$$  \hspace{1cm} \text{II.8}$$

The value of $F_w$ is about 220 Hz for women and about 190 Hz for men.

The properties of air relevant to acoustic wave propagation are known, if it is assumed as a simplification that the whole respiratory tract is at blood temperature, 37°C with relative humidity 100%. But yielding walls give an effective velocity of sound for the air inside the tube which is higher than it would be with rigid walls (Flanagan, 1972a, p.67). This would be expected to raise all the formant frequencies, as in Equation (II.8) above. Values for the electrical elements of inductance, capacitance and resistance, corresponding to acoustic inertance or mechanical mass, to compliance and to flow resistance respectively, have been estimated by Flanagan (1972a). Values appropriate for a reflected
acoustic pressure wave model have been given by Maeda (1982) and by Liljencrants (1985).

Tissue and air properties are considered in more detail in Section II.3.3.
II.3 Stages of speech production: theory and data in relation to models

Introduction

The stages of speech production as understood in this work were identified in Section I.4. All except the first of these, the neurolinguistic stage, will be considered here. Stages 2) and 3) will be conflated and discussed only briefly. The relationships between the stages as understood in this work, excluding sensory feedback paths, are shown in Figure 5. The model to be described in Chapter III is based on theory and data for the articulatory stage onwards.

II.3.1 Neural and muscular processes

This stage of speech production is the least accessible to measurement. Some of the observed characteristics of speech and other motor skills have been described in Section II.1, but the principles governing neural and neuromuscular processes in speech are not yet well understood. The technique of electromyography (emg) allows electrical changes in muscles to be observed. Magnitudes of emg signals must be interpreted with caution, but some studies, notably for the larynx, strongly suggest a cause and effect relationship between the emg signal for the muscle forces and the resulting articulatory movement or change of mechanical properties. For example, patterns of posterior cricoarytenoid (PCA) muscle activity and glottal width changes match well allowing for latency, the time interval between the muscle's application of forces and the actual movement (Hirose and Ushijima, 1974). The posterior cricoarytenoid muscle is the sole abductor of the vocal folds, but fundamental frequency $F_0$ seems to be controlled by more than one
larynx muscle and by aerodynamic forces also, as discussed in Section II.3.4.2. Patterns of cricothyroid (CTH) muscle activity seem to match patterns for $F_0$ rising and falling at middle to high frequencies; a larynx depressor such as the sternohyoid (STH) muscle appears to be associated with low falls in $F_0$, for example, at the end of utterances (Collier, 1975). The body and cover of the vocal folds can be independently controlled it seems: stiffening of the body alone by the action of the vocalis muscle with a slack cover produces a different kind of vocal fold vibration from stiffening of both body and cover by the combined action of the vocalis and the cricothyroid muscles (Hirano, 1977).

Many different muscles make up almost the entire volume of the tongue and these can combine to provide the forces required for changes of tongue shape. Some of them are difficult to investigate since they are interleaved to form the tongue. Examination of only a part of this system of forces may be misleading; emg studies seem to be of most benefit when several muscles are monitored at once and when they are combined with observation of the associated movement paths.

In the larynx there appears to be reciprocal activity between an adductor of the vocal folds, the interarytenoid muscle, and the abductor, the posterior cricoarytenoid muscle (Hirose and Gay, 1972). Velum height also seems to be controlled in speech by a pair of agonist-antagonist muscles, the levator palatini and the palatoglossus, acting reciprocally (Fritzell, 1969). Complex context-dependent effects are seen both in the emg patterns of these muscles' activities and in the associated changes of velum height. Data for American English suggested that speakers aim for a number of different velum heights for different vowel and consonant contexts. Increases in the height of the velum appear to be correlated with
increases in levator muscle activity as indicated by EMG traces (Bell-Berti and Hirose, 1975).

It may be supposed that speech operates efficiently, but it is not clear what constitutes efficiency in the nervous system. The apparently hierarchical nature of linguistic structures and the large number of linguistic and phonetic options available to a speaker is presumably reflected in many branching and interconnected neural structures, with many feedback loops and feedforward links used for error correction, as suggested in Laver's (1980a) model.

Any feedback control model of speech production would need to operate with realistic values of feedback loop times. For example, the question might be posed whether a speaker might prolong the articulatory state which produces frication noise in order to ensure that an English sequence /stressed vowel-voiceless fricative/ was heard rather than /stressed vowel-voiced fricative/ on the basis of his or her own on-line auditory feedback. The relative durations of vocoid and fricative contoid are important for this phonological distinction in English, as shown by Denes (1955); auditory correlates of these durational relationships could perhaps be assessed by a speaker while producing the fricative sound. Double dips in aerodynamically derived traces for vocal tract constriction cross-section area are often seen for [s] (Appendix 1, Scully, 1984a). They suggest the possibility of an additional noise-sustaining tongue action, but might equally well be a manifestation of overshoot, for a less than critically damped approach to an endpoint state.

It has been emphasised by Müller et al. (1981) that the mechanical behaviour of the structures needs to be related to the capabilities of the nervous system in altering mechanical properties such as
stiffness; the articulators and the forces acting on them are not simply second-order mechanical systems with fixed properties. It is important that current models of speech production should avoid violating established facts in this domain, but their incorporation into detailed quantitative signal models seems likely to be a goal for the future rather than an immediately attainable one.

Coordination across groups of muscles underpins the timing and coordination of articulatory actions. Some reciprocal organisation of muscle forces appears to be catered for specifically through the arrangement of pyramidal cells in the motor cortex of the central nervous system, for agonist, synergist and antagonist muscles (Eccles, 1977, p.113). But in addition muscles cooperate in functional synergy as discussed in Section II.1.

Section II.3.2 reviews movement paths during speech production for a number of quasi-independent articulators. A framework has been proposed for modelling some aspects of temporal control of these actions. This is one kind of representation of a presumed planning stage for speech production, based upon learned patterns of behaviour stored in the central nervous system of a speaker (Appendix 3, Scully, 1987). Variability is clearly an inescapable feature of neural control which is manifest as variability at all stages of speech production. It is beyond the scope of the present work to consider how variability in huge numbers of interacting pathways nevertheless produces articulatory results with manageable variability. Some scholars have considered problems of stability and this aspect of speech production is briefly reviewed in Section II.4.
II.3.2 Articulation

Introduction

The concept of a supraglottal articulator causing an acoustically significant reduction in tube size as it approaches the opposite wall of the vocal tract at a specific place along the tube provides the basis for the powerful though simplified insights of the IPA chart (IPA, 1979). By considering states and movements of whole structures, or quasi-independent portions of them, the number of components in analysis and simulation of speech production actions is reduced to a manageable number, up to about sixteen, listed in Section II.1 (Ladefoged, 1979).

The emphasis in the traditional IPA analysis is more upon static end point configurations than upon the gestures linking them. Many quantitative analyses of vocal tract shape also have focused on static shapes, especially for vowels.

In the case of supraglottal articulation it is necessary to consider both the solid structures and the ways in which they control the shape of the vocal tract, often through the combined actions of two or more articulators. The area function of the vocal tract, that is, the cross-section area of the complex tube at each point along its length, is the acoustically important aspect of the geometry, as discussed in Section II.3.5. Therefore movement paths need to be understood in terms of changing cross-section areas of regions of the vocal tract. The data from natural speech are sparse at present and they generally give information on distance moved in the medial sagittal plane. Distance moved across the vocal tract tube will be the more important of these two dimensions, but the mapping from the cross-dimension to the cross-section area of the air-filled tube
cannot be described simply and remains obscure. The vocal tract is rarely modelled in three-dimensional terms (but see Fujimura, 1977, discussed in Section II.5.2).

Several regions of the vocal tract need to be identified: the larynx tube, pharynx, oral cavity, uvula, soft palate or velum, hard palate, alveolar ridge, teeth and lips, as well as different portions of the tongue: the root, dorsum, blade and tip, with the tongue body comprising the main volume of the tongue, as shown in Figure 1(c). These names will be used as labels for moveable or fixed structures, and also for portions of the area function. The palatal region of the vocal tract, for example, is under the hard and soft palate. As in traditional phonetic descriptions of supraglottal articulators, it will not be possible to define exactly where the different portions of the tongue or vocal tract begin and end.

Actions of the larynx and of the subglottal respiratory tract are just as necessary for speech production as those of the vocal tract. All relevant moving structures will be subsumed under the term articulator in this work. In most cases an articulatory component is associated with a change of position of a structure. But in one case at least it is a change of mechanical state that is described, as in the laryngeal component of fundamental frequency control, called Q in the modelling. This articulatory component is related to the length and stiffness of the vocal folds, controlled mainly by the cricothyroid and vocalis muscles, and probably by larynx depressor muscles also, as discussed in Section II.3.1.

The dynamics of articulatory movements have been considered in a few studies, as discussed in Sections II.1.2 and II.1.4, with an individual articulator treated as a second order mechanical system moving under the influence of specified forces. Most of the studies
cited below are kinematic descriptions of movement paths. The latter approach is directly relevant for the modelling to be described here. No mathematical model of the movement dynamics is assumed in this study; the aim is to simulate some of the path pattern characteristics of natural speech.

The techniques used to investigate movement paths in natural speech include cinefluorography, cine-photography, palatography, strain gauge transducers, optical and magnetometer methods. Transillumination and fiberscope observations have been of particular importance for the larynx. The methods will not be described in detail here, but will be mentioned for some of the studies cited. Articulatory data are always partial descriptions of the whole system. Most of the relevant structures are inaccessible to direct and non-invasive observation; there is therefore much interest in inferring vocal tract shapes from acoustic information.

Two mainly acoustic studies have strongly influenced the basic concepts of speech production adopted by subsequent researchers. Lindblom's (1963) analysis of formant transitions has provided an interpretation of acoustic data in terms of articulatory targets and undershoot. In this model there is considered to be an invariant underlying target for each set of vowel formants; articulators move towards but do not always reach the target, thus exhibiting undershoot in both acoustic and articulatory domains. Öhman's (1966) study was mainly an acoustic analysis of formant transitions but included some cineradiographic data on vocal tract shapes during [VCV] sequences. Patterns of formant transitions associated with the plosive consonant were found to vary systematically with vowel context. An articulatory interpretation was provided: the actions needed to achieve the vocal tract closure for the plosive are considered to be superimposed on a diphthong-like transition of the
tongue body from the preceding to the following vowel. Different sets of muscle force commands are attributed to vowels and consonants. From acoustic data alone it appears difficult if not impossible to distinguish this model of articulatory organisation from another in which all parts of the tongue are made to move into a range of configurations appropriate for the consonant production and thence towards the states needed for the second vowel.

An attempt will be made here to assess present knowledge about articulatory end points reached at specified points in the acoustic structures of particular vowel and consonant sequences and the movement paths between them.

The terms end point and transition for each articulator will be used here; they are intended to contain the concepts of varying amounts of precision and variability in different portions of movement paths, as discussed in Section II.3.2.6. It seems to be the case that at least one articulator is moving during most of speech production, so that essential articulatory transitions probably impose major durational constraints on the acoustic structures. Whether end points are context-dependent or invariant, they can be located fairly easily on traces as time points of zero velocity as the articulator changes direction. The data will not be interpreted in terms of any particular model such as that of Lindblom (1963) or Öhman (1966) and the terms target and undershoot will be avoided.

Consistency of description across published studies would be desirable but is not completely attainable: definitions of end points vary and the diagrams of the movement paths analysed are not generally included in the published papers. The language analysed, and whether the speaker is a woman or a man will be stated when possible, but in several published works this information is not
explicitly stated. There are insufficient data for conclusions to be
drawn about general characteristics of women's and men's vocal tract
shapes.

II.3.2.1 Configurations of the vocal tract for vowels and
diphthongs

Introduction

X-ray studies for speakers of several different languages, whether
with stills or cinefilm and whether for vowels in isolation or in
some consonantal context, show the overall configuration of the vocal
tract. These are generally side views approximating to a medial
sagittal section. Although the phonetic qualities must vary, there
is agreement on the general vocal tract configuration for some
broadly defined vowel types relevant to the simulation of an RP
accent of British English. The data reviewed here do not include
phonologically contrastive nasalised vowels so that the vocal tract
is an unbranched tube, approximately closed at the glottis end and
open at the lips. The phonetic notation used where the language
investigated is British English will be as set out as on page 17.
The same symbols will be used for American English, although the
phonetic quality for the vowels in the key words would differ in some
cases. Where other languages are considered, symbols will conform to
IPA (1979) usage.

The descriptions which follow are based on the research of Chiba and
Kajiyama (1941) for Japanese and German, Fant (1960, 1970) for
Russian, Öhman (1966) for Swedish, the author's own speech (Delattre,
1968) for British English, Perkell (1969), MacNeilage and DeClerk
(1969), Carney and Moll (1971), Kent and Moll (1972a), for American

Some cinefluorographic studies have yielded important information about movement paths for selected portions of the tongue, often through the use of radio-opaque markers. In addition, they do give some indications of the tongue body shape for a limited set of vowels in particular consonantal contexts (for example Houde, 1968; Carney and Moll, 1971; Kent, 1972; Kent and Moll, 1972a, 1972b; Gay, 1974, 1977a, 1977b; Gay et al., 1974; Borden and Gay, 1979).

Most of the studies cited above have analysed one or two repetitions only, because of limitations on the allowed radiation dose. Recently the computer controlled X-ray microbeam technique has permitted analysis of tongue marker paths in multiple repetitions, so that variable and invariant aspects of trajectories and configurations may be investigated now more reliably than before. Reference is made to microbeam data in Section II.3.2.6 where variability in timing and path shape is discussed.

Vowels

For [i]-type vowels the bulk of the tongue body is forward and the upper surface approaches close to the palate, presumably with contact at the sides, as seen on direct palatography records (unpublished laboratory data). The anterior-posterior dimension of the pharynx is very large and rather uniformly so. The jaw is fairly high or very high. In some cases there appears to be a long, narrow channel for the constriction; in others it is more sharply localised, but there
always appears to be a gradually flaring tube increasing in cross-section area from the constriction to the mouth outlet, unobstructed by the lips.

[ɪ]-type, [ɛ]-type and [ã]-type vowels have a pattern similar to [i] but with progressively less extreme narrowing near the hard palate and with smaller pharynx dimensions in the anterior-posterior direction. This progression is seen in Wood's (1979) area functions, both for a Southern British speaker and for an Egyptian Arabic speaker. In Perkell's traces for English, [ɪ], [ɛ], [ʊ] and [æ] do not have extreme narrowing of the vocal tract. The shapes of the tongue body for [ɪ] and [ɛ] are close to those it assumes for some alveolar stops and fricatives (Perkell, 1969, pp.72,78,80,82,84,90).

[a]-type and [a]-type vowels show extreme narrowing in the pharynx and a large oral cavity. The jaw is lowered and there is a large mouth opening. The position of the narrowest constriction is not always clear here. There seems to be a progression for the series [a] to [a] to [æ], going from more to less extreme pharynx narrowing and correspondingly from larger to smaller front oral cavity volume, seen in Wood's (1979) area functions for the Southern British and the Egyptian Arabic speakers.

[u]-type and [o]-type vowels generally have the appearance of the tongue dorsum strongly bunched towards the uvula. The location of the narrowest constriction may perhaps be generally further forward for [u]-type than for [o]-type vowels. It seems particularly important to recall that British English [u:] is not necessarily near to Cardinal Vowel [u] but may be auditorily more like Cardinal Vowel [u]. One Southern British speaker (Wood, 1979) shows extreme lip outlet narrowing and lengthening for [u:], less extremely so for [o] (sic) and [ɔ:]. Another Southern British English speaker (data for
the author's speech, Delattre, 1968) shows the tongue bunched towards the hard palate for [u:]; the anterior-posterior dimensions for the pharynx are almost as large as for the same speaker's [i:]; the lips are not narrowed nor protruded.

[u] alone has protruded lips with a narrow lip outlet in Perkell's (1969) study of an American English speaker. Other examples of [u]-type vowels (Chiba and Kajiyama, 1941, for German; Fant, 1960, 1970, for Russian; Öhman, 1966 for Swedish) show the tongue bunched towards the soft palate or uvula and extreme lip narrowing and protrusion. Abry and Bœ (1983) found that, for French vowels, lip width decreased while lip protrusion increased in the following order:

\[
\begin{align*}
\text{[a e i a w/ε w i w/ε w o/œ o]/u/y]}
\end{align*}
\]

where the subscript _w_ denotes allophones of the vowels in [i] or [j] context. For American English vowels the corresponding order was:

\[
\begin{align*}
\text{[æ/æ i: e/ε i: ø/ø o (sic)]}
\end{align*}
\]

[u:] had the same amount of protrusion as [o:] but much less width (Fromkin, 1964).

Wood found four constriction locations for vowels in his own data for a Southern British English speaker: along the hard palate for [i:], [i] and [e], along the soft palate for [u:] and [u:], in the upper pharynx for [o] and [ɔ:], in the lower pharynx for [æ], [ʌ], [a] and [a:]. The same four places for similar vowels were found for an Egyptian Arabic speaker recorded and analysed by him. There were four repetitions of each vowel, with four additional repetitions at a different speaking rate for the English speaker. He reported confirmation of these same four places from his analysis of 38 sets of published X-ray traces for twelve languages (Wood, 1979).
Diphthongs

Vocal tract shapes for diphthongs of English seem to have been rarely studied with X-ray techniques. Kent and Moll (1972a) analysed vowel-to-vowel tongue body gestures, but these were across syllable or word boundaries. Diphthong data have been obtained for one woman speaker of Southern Icelandic (Pétursson, 1974). The tongue body movements apparent here seem very like what would be guessed at for British English, on the basis of the [a] and [I] end points in the case of [aɪ], at least. The kind of transition apparent in Pétursson's (1974) Icelandic data produced an auditorily acceptable [aɪ] with the model, as described in Section IV.1.2.

Problems of interpretation of the data

These data can be only a rough guide to vocal tract shapes, for two reasons. First, it is the area function that determines the frequencies of the first three or four formants, not the dimensions in a medial sagittal section as shown in X-ray traces; few data are available for the complex transformation from one to another, as discussed in Section II.3.2.3. Secondly, it is probable that no one portion of the vocal tract acoustic tube dominates in determining resonant frequencies for vowels. The case is different for at least some consonants, where the configuration behind the main articulatory constriction is not crucial.

Vocal tract models

The electrical analogue models of Fant (1960, 1970) and Stevens and House (1955, 1956) (see Section II.5.2) provide extremely valuable area function data for vowels produced by a vocal tract of a specific length, appropriate for some men speakers. But a given vowel quality
cannot be reproduced in a modelled vocal tract of a different length simply by compressing or expanding the distance uniformly along the vocal tract. Formant frequencies from natural speech are not uniformly scaled for different vocal tract dimensions either (Fant, 1966).

Three parameters of these modelled vocal tract shapes determined the formant patterns for simulated static vowels:

1) position of the constriction;
2) minimum cross-section area of the constriction;
3) lip outlet shape.

Support for the finding that the size and location of the constricted portion is more crucial to vowel quality than other portions of the tube is found in X-ray traces and formant frequency analyses under bite block conditions. With the jaw constrained to stay lower than normal for an [iː]-type vowel the tongue-palate constriction region conformed more closely to its normal shape without a bite block than did other parts of the vocal tract (Gay et al., 1981).

The nomograms published by Fant (1960, 1970) form the acoustic basis for any subsequent vocal tract modelling. Figure 2 in Section 1.3 reproduces some of Fant's findings for horn-shaped tube models. In natural speech the extreme mobility of the tongue may allow additional degrees of freedom. For example, Lindau's (1979) analysis of X-ray traces for Akan demonstrates the possibility of control of the advancement of the tongue root, independent of the rest of the area function.
Tongue shape factors

Nevertheless, several analyses of tongue shapes during vowel production in natural speech agree in finding only two or three dominant shape factors (for example: Kiritani and Fujimura, 1975; Kiritani, 1977; Nakajima, 1977; Shirai and Honda, 1977; Harshman et al., 1977; Maeda, 1979).

Where the lips and the jaw have been included these contribute additional factors. It seems that two parameters are sufficient to describe the part of the area function controlled by the tongue body and jaw in static vowels; one additional parameter is needed for tube outlet shape and another additional jaw parameter if the analysis is of solid structures rather than of tube area function directly.

One rather striking feature of X-ray side views of the vocal tract for vowels is the pivot-like effect seen near the soft palate. That portion of the tongue surface moves very little compared with the regions on either side: behind, controlling the area function of the pharynx and in front, controlling the tube shape in the hard palate region. The effect may be seen in, for example, vowels of Southern British English (Wood, 1979), American English subjects (MacNeilage and DeClerk, 1969; Perkell, 1969, and Icelandic (Pétursson, 1974).

In the factor analysis study by Ladefoged and his colleagues, two pivot points emerged, corresponding to the two main factors found for tongue body shape in ten vowels produced by five speakers of American English. The same basic set of gestures was found to be used by all the speakers but they differed in the relative amounts of each set. One tongue point in the oral cavity near the soft palate is the pivot for their first factor, front raising or lowering; a second tongue
point in the upper pharynx, slightly below the level of the uvula, is the pivot for their second factor, back raising or lowering. They analysed vocal tract wall and tongue body outline shapes measured at sixteen points from the epiglottis to the alveolar ridge. Jaw and lip positions were excluded. Vowels \([i: e \theta \varepsilon]\) had front raising, \([\alpha: o \text{ (the published symbol)}\upsilon u:]\) had front lowering. \([u:]\) had more back raising while \([\varnothing:]\) and \([\alpha:]\) had more back lowering than other vowels (Harshman et al., 1977; Ladefoged et al., 1978).

Other analyses agree with this one. Pivots are similarly placed for the first and second components of an analysis of one speaker of Japanese. A somewhat similar progression was found from front raised to front lowered extremes of their first component across the vowels \([i e u o a]\) (I E U O A in the publication) (Kiritani and Fujimura, 1975).

In a subsequent analysis of pellet positions from cineradiography, the five vowels of Japanese were associated with four parameters. Two jaw-independent tongue components \(T_1\) and \(T_2\) seem to correspond to front raising and lowering and to back raising and lowering respectively. A jaw parameter \(J\) was extracted and this accounted for a large fraction of the observed tongue movement. There is also a lower lip parameter \(L_R\), after removal of the jaw-influenced lip movement (Kiritani, 1977).

It might be suggested that these two studies are in broad agreement, but that Kiritani's analysis puts more emphasis on the movement components ascribable to distinct structures - the lower lip, jaw and tongue - while that of Ladefoged et al. emphasises the total shaping of the vocal tract tube by the tongue surface, however this may be achieved in terms of jaw and tongue muscles combined.
Other analyses indicate also that only two or three components are needed to describe tongue shape for vowels. For example, Shirai and Honda (1977) derived two principal components for tongue shape in ten men speakers, relative to a best mean shape. A linear transformation of this with two speaker-specific parameters gave quite good separation of five Japanese vowels in two-dimensional space.

Goldstein (1980) reviewed and analysed X-ray data for different speaker types and modelled the area functions for children, women and men speakers. It does not appear possible at present to give a general formula for the mapping of closely similar vowel qualities onto different vocal tract lengths.

The IPA vowel chart applied to English (Gimson, 1980) provided a starting point for the simulation of vowels and diphthongs in this study. Although this chart displays auditory rather than articulatory space, the approximate correspondence between regions of the chart and configurations of the vocal tract and the general agreement across X-ray studies discussed above made it possible to produce the right kinds of phonetic quality for some of the vowels and diphthongs of English without difficulty, as described in Sections IV.1.1 and IV.1.2. However, the task of generating specific vowel and diphthong qualities that are acceptable as an RP accent of British English is very much more difficult; many trials with auditory monitoring are needed to find an appropriate area function for a given simulated vocal tract length. Area functions for Russian vowels in Fant's (1960, 1970) study were used also as a starting point for modelling the vowels of English.
II.3.2.2 Configurations of the vocal tract for consonants

The IPA chart (IPA, 1979) categorises consonants according to place of articulation, manner of articulation and voicing. Vocal tract configurations for some plosives, fricatives and approximants have been included in cineradiographic studies of a number of languages. The evidence on the extent to which the whole tongue body is constrained to adopt a particular shape is slight, unclear and in some cases conflicting.

Most consonants of English require severe constriction at one portion of the vocal tract. The main articulator for the consonant is thus tightly constrained. The shape of the cavity in front of the main articulatory constriction is important also and this is likely to vary with vowel context. Öhman (1966) makes a distinction between languages such as Swedish and English, in which the tongue body seems to be free during the production of a consonant to move towards the shape needed for a following vowel, and languages such as Russian, in which plosives must be either palatalised or velarised, with correspondingly constrained tongue body shapes.

In the production of velar plosives [k] and [g] by a speaker of English the main articulator is the dorsal portion of the tongue body. Vocal tract shapes for one speaker's [g] indicate a closely similar high jaw position for four vowel contexts. There are similar tongue body configurations with slight shifts appearing consistent with the vowel contexts, as follows: the tongue is raised more towards the palate in front of the closure region and the contact point seems to be further forward for [i:] than for [u:], [æ] or [ə:] (MacNeilage and DeClerk, 1969). Another X-ray study showed the tongue dorsum moving up for a [g] closure, forward
during the closed phase and back as well as down during the release phase (Kent and Moll, 1972b).

For a French speaker's [g], the jaw was lower and the contact further back in [a-a] context than in [i-i] context; For [u-u] context the jaw height was intermediate and the contact furthest back of all three. The data were explained in terms of a four component model of tongue shapes (Maeda, 1979). A Swedish speaker, similarly, showed a more fronted contact region for [g] in [y-y] context than in [a-a] or [u-u] contexts. For [d] in the 3 different vowel contexts most of the tongue body shape differed across the 3 contexts, apparently towards the vowel shape (Öhman, 1969). In X-ray traces for an English speaker, alveolar consonants [n, z, s, d, t] in [hæ-e] contexts all showed similar tongue shapes, with the tongue tip in contact with the alveolar ridge and the rest of the tongue assuming a shape similar to that for [ɪ] or [e]. Whichever vowel followed, the tongue tip moved along a very similar path during the first 30 to 50 ms after a plosive release for [t] (Perkell, 1969, pp. 72-84). This release action could perhaps provide a relatively context-independent acoustic pattern feature for identification of an alveolar place of articulation. It is important that the rest of the tongue should move much more slowly than the main articulator so that this place information should be preserved (Maeda, personal communication).

Lip shape is not specified on the IPA chart for most consonants of the types used in English, apart from [w] which is shown with narrowing at the lips as well as in the velar region. It has sometimes been assumed, in studies of the spread in the time domain of actions such as lip rounding and protrusion needed for some vowels, that lip shape is not specified in an English speaker's stored plans for plosives and fricatives. However, in French at least this assumption may not always be justified. For some French
speakers the consonants [ʃ] and [ʒ] have noticeably more lip protrusion than other consonants (Benguerel and Cowan, 1974). Some French speakers more than others separate [s] and [z] from [ʃ] and [ʒ] in a three-dimensional lip space of vertical lip separation, width of lip outlet space and protrusion (Graillot et al., 1980). Some speakers of British English tend to use lip rounding in dark [l], especially when it is a syllabic [l] (Gimson, 1980, p.202). Many RP speakers have slight lip rounding for [r] whichever vowel follows; for others the amount of lip rounding depends on the following vowel (Gimson, 1980, p.207; Roach, 1983).

It is even more clear for consonants than it is for vowels that speech production consists of gestures, not of a succession of fixed postures. Therefore, no one set of frames extracted from cinefluorography can be said to characterise a given phonetic element. Some of the available evidence on vocal tract configurations during the production of some consonants and vowels of English and observed patterns of inter-articulator coordination, for different portions of the tongue and for the jaw or part of the tongue body relative to the main vocal tract articulator for consonants, is presented in Appendix 3 Sections II.15 to II.18 (Scully, 1987).

The question of whether the tongue body moves to a specific configuration or to a broader region during the extreme narrowing or closure of one part of the vocal tract for a consonant is unlikely to be answered in a simple way. It seems likely that phonetic context effects, such as whether adjacent vowels are stressed or not, together with communicative requirements will impose greater or lesser constraints on the speaker's freedom to vary tongue shapes.
II.3.2.3 Cross-dimension to cross-section area mapping

X-ray views of the vocal tract usually show dimensions in a medial sagittal plane, the lateral view; sometimes the mid-line of the tongue surface is enhanced by radio-opaque marking. But the acoustically relevant parameter at each point along the vocal tract is not distance across the tube in the mid-sagittal plane on its own; it is cross-section area at each point along the tube, so that lateral dimensions must be estimated. The information needed for mapping from X-ray view to area is very difficult to obtain, especially for the pharynx. Shapes across the tube vary at different points and they must be expected to change with vowel type and during speech-like sequences. Individual vocal tracts seem likely to vary in their 3-dimensional shapes; static palatography confirms this for the palatal-alveolar region (unpublished laboratory data). A single cross-dimension to area mapping must be a compromise.

Vocal tract area functions have been inferred from X-ray stills or from cineradiography frames showing the mid-sagittal plane (for example Fant, 1960, 1970, pp.93-98; Heinz and Stevens, 1964, 1965; Perkell, 1969, used by Maeda, 1971, 1972, and by Mermelstein, 1973; Lindblom and Sundberg, 1971; Heike, 1979); sometimes with the addition of the sides of the tongue (for example Chiba and Kajiyama, 1941, pp.108-114; Ladefoged et al., 1971). Plaster dental casts have been made of the roof of the mouth and sometimes of the floor of the mouth also (Chiba and Kajiyama, 1941); dental impression cast material has even been poured into an inverted speaker's pharynx, down to the level of the arytenoid cartilages (Ladefoged et al., 1971)!

The pharynx is particularly difficult to observe directly. Apart from the drastic method just mentioned, X-ray tomography (Fant,
1965), a laryngoscope mirror, and an endoscope or fiberscope (Chiba and Kajiyama, 1941; Ladefoged et al., 1971; Gauffin and Sundberg, 1978) have been used to view the pharynx cavity shapes at different levels. Measurements have been made on cadavers (Anthony, 1964). Ultrasonic transducers externally placed have been used to track lateral movements of the pharynx walls (Minifie et al., 1970).

The complex shapes of the vocal tract walls and tongue surface have been approximated by simplified geometric forms; for example the pharynx cross-section shape has been represented as an ellipse (Heinz and Stevens, 1964; Maeda, 1971, 1972). The lip outlet shape has been expressed by a function which relates its cross-section area to the width and height of the space between the lips (Fromkin, 1964; Lindblom and Sundberg, 1971; Bøe et al., 1980) or has been approximated by an ellipse (for example Mermelstein, 1973). Different mappings are needed for rounded and unrounded vowels, but lip protrusion can be approximately predicted from the mouth width (for example, Fromkin, 1964).

The relationship derived between cross-dimensions of the vocal tract in the mid-sagittal plane $d$ and cross-section area $A$ generally has the form

$$A = k d^b$$

where $k$ and $b$ are constants (for example, Heinz and Stevens, 1965).

The vocal tract is usually divided into at least the pharynx and the oral cavity, with different parameter values for each portion. The pharynx itself is separated into upper and lower portions, above and below the level of the upper tip of the epiglottis, by Lindblom and Sundberg (1971). They caution that their expression had restricted applicability for $d$ less than 2 cm; this limitation would exclude back vowels such as $[\alpha:]$. Heinz and Stevens (1964) increased the
lateral dimension axis of their pharynx ellipse between the larynx and the uvula. The palate region was separated into front and back by, for example, Heike (1979); here the back portion included an extra constant, with a mapping of the form

\[ A = k_1 + k_2 d^b \]  

(II.10)

where \( k_1 \), \( k_2 \) and \( b \) are constants. Maeda (1971, 1972) derived a shape factor which took seven different values, at different points along the vocal tract between the larynx and the teeth. An eighth shape factor represented the effects of the non-rigid walls. The mapping is expected to vary for different vowels, but from the detailed analysis of Chiba and Kajiyama (1941) a line giving an approximate fit with all five Japanese vowels along most of the vocal tract could be found. This is used in the present study, as described in Section III.1.2.

Apart from the difficulties of observation and the large changes in shape across different vowel and consonant allophones, other general problems have been identified. One is the problem of defining the midline of the tube, especially where it bends through approximately 90°, discussed by Goldstein (1980, pp.120-162). Another problem concerns the estimation of an acoustically appropriate termination plane at the mouth opening of the vocal tract. Heinz and Stevens (1965) located the termination just in front of the corner of the mouth; Lindblom and Sundberg (1971) located it further in front of or even behind the lower incisor teeth as the jaw was raised or lowered, respectively. A similar method was followed by Mermelstein (1973).

The relationship between glottal area and glottal width as seen with a fiberscope needs to be investigated further. Since the shape of the glottis as seen on published frames from cinefilms is approximately triangular, its width measured between the arytenoid
cartilages, as measured by for example Kagaya and Hirose (1975), can probably be considered to be proportional to glottal area, as a first approximation.

II.3.2.4 Shapes and general properties of movement paths

Articulatory transitions are defined for the purposes of this work as beginning and ending with zero velocity. They may be divided into three segments: acceleration, movement at maximum velocity, and deceleration. S-shaped curves of distance versus time have been obtained in many analyses of X-ray views of the vocal tract.

A movement path may be derived from the distance between two points identifiable on the film. One of the points is often fixed (for example Perkell, 1969), but both may be moving, as when tongue body movements are measured relative to the jaw (Kent and Moll, 1972a). Alternatively the vertical and horizontal components of a radio-opaque marker attached to a moving structure may be tracked, with a fixed reference marker, usually on the nose (for example Fujimura, 1986).

Two approaches to the kinematic description of vocal tract movement paths are possible. Either the movements of each identifiable solid structure may be considered individually and then combined to give vocal tract shape, or the overall effect on the shaping of the vocal tract tube may be observed directly. The choice of analysis relates to the question - debated but not resolved - of whether the jaw should be included as a quasi-independent articulator. For example, the jaw cooperates with the tongue to change the shape and position of the surface of the tongue; the total displacement results from a combination of jaw movement and tongue movement relative to the jaw. These two components may combine in different proportions in
different speakers for the same context (Kent and Moll, 1972a, Figure 2). A given speaker may use different amounts of jaw movement for phonetically similar sequences. For example one speaker analysed by Kent and Moll (1972a, Figure 11) gave a smaller displacement of the tongue dorsum with less jaw movement for [aɪ] in "I hold" than for [aɪ] in "hyoid". In this case reduced jaw movement was not compensated for by greater tongue movement.

Compensation between lips and jaw muscles is found under an abnormal loading as discussed in Section II.1.4. In unperturbed speech also lip closing actions may use different combinations of lip muscle and jaw raising muscle forces to achieve lip closure for bilabial plosives. Hughes and Abbs (1976) found evidence for cooperation between lower lip and jaw in the production of [p]. Vertical separation of the upper and lower lips at the end points of their transitions had less repetition to repetition scatter than that of the vertical positions of the lower lip and the jaw considered separately. The lower lip and jaw made varying and compensatory contributions to transition velocity also, in the reduction of the lip outlet size for [p], [f] and [v] (Hasegawa et al., 1976, cited in Abbs et al., 1984). Differences in the amount of jaw lowering for vowels related to stress context found by Macchi (1985) seem to be consistent with a flexible organisation of lip and jaw actions so as to achieve both lip closure where needed for [p] and the required size of lip outlet area where needed for vowels.

In one type of articulatory model the jaw, tongue and lips would be represented separately, with the advantage that the model is close to the actual physical system and shares its interactions and constraints. For example, jaw lowering in adults carries the tongue body backwards as well as downwards, thus narrowing the pharynx unless tongue muscle forces are employed to counteract the effect
(Lindblom and Sundberg, 1971). If it is assumed that speech uses the available muscle forces in simple ways, such a model has the advantage of behaving similarly. Solid structure oriented models might be expected to have the advantage of simpler shapes for the component movement paths, especially if two articulators with noticeably different time constants combine to shape the tube. Where regularities are found they are likely to shed light on the organisation of speech production. For example, if jaw raising is found to be used by a particular speaker to assist the lip closure for a bilabial plosive following an open, low jaw vowel, but not following a close vowel, then the time taken to achieve lip closure relates to the slower jaw action in the former case but to the faster lip action in the latter.

An advantage of the alternative approach to modelling, advocated by Ladefoged (1979), is that the shape of the vocal tract tube directly controls important aspects of the acoustic patterning of the output speech. It may be reasonable to suppose therefore that vocal tract shaping is more closely related to a speaker's goals than are the movements of individual solid structures. Closely similar total changes to vocal tract tube shape can be achieved by different combinations of component movements. The resulting vocal tract shape change may be a more invariant and so a simpler, more economical description. The flexibility, which appears to be characteristic of the human nervous system as discussed in Sections II.1 and II.3.1, might be taken for granted as a more microscopic scale of analysis than is required.

The emphasis in this study is on the shaping of the vocal tract. Unless otherwise stated the observations cited below for natural speech use the palate or another fixed point as a reference for movement paths.
MacNeilage and DeClerk (1969) and Perkell (1969) derived paths for many points along the vocal tract. Changes in shape and position of the tongue body have been analysed by, for example, Houde (1968), Kent and Moll (1969), Kent and Netsell (1971), Kent and Moll (1972a, 1972b). Other articulatory regions studied include the tongue tip (Kent and Moll, 1969), the tongue blade, jaw, hyoid bone and upper lip (Kent and Netsell, 1971), tongue tip and jaw (Kent and Moll, 1972b). Because their computer-controlled X-ray microbeam system permits multiple repetitions, Fujimura and his associates are able to analyse the variability in different portions of the transition for one articulatory region such as the tongue blade (for example Fujimura, 1986).

It is important to bear in mind that the cinefluorographic studies of the 1960s and 1970s nearly always included only one example of each utterance. Sometimes several similar phonetic contexts were clustered to give more than one data point but, in general, it is not possible to assert with confidence that observed differences for contrasted contexts or different speakers are caused solely by the independent variable; they may reflect partly, or even mainly, variability from one repetition to another. Thus the patterning discussed below must be taken as hints and possible guides to articulatory regularities rather than as irrefutable evidence.

Maximum velocity is reached only momentarily, at a time approximately half way through a transition. Kent (1972, Figure 8) and Kent and Moll (1972a, Figure 3) for example show fairly symmetrical acceleration and deceleration portions with peak velocity slightly after the mid point in time, as shown in Figure 6. Tongue tip paths are sometimes sharply cut off however when contact with the palate is made for a consonant, without the velocity falling to zero; indeed in a similar effect, the lips are sometimes moving at their maximum
velocity at the moment of closure and release for a bilabial plosive, as discussed below (Kiritani et al., 1978). A similar effect is shown for alveolar consonants in Figure 7 which is taken from Kent and Moll (1969, Figure 3; 1972b, Figure 12). Mechanical pressure recorded for [di:] by McGlone et al. (1967) shows a rounded peak during the tongue-palate contact phase; this suggests that the path of the tongue tip is towards a point above the palate surface, but that the movement is interrupted by contact. This seems a plausible explanation of the data and would give firm contact for a plosive. Although this [di:] seems to have been produced at a very slow speech rate, a composite sketch based on the two sets of data seems justified, as shown in Figure 7. This kind of approach, with a virtual end point lying above the palate, was used by Mermelstein (1973). For simulation of the [z] fricative paths included in Figure 7, a virtual end point above the palate would need to be for the sides of the tongue blade region only, leaving a small opening in the middle.

For one speaker whose tongue dorsum moved through different distances for [i:-æ], [i:-œ] and [i:-ɔ], peak velocity occurred approximately half way through each transition in time and a little after half way in distance. As the distance traversed increased from about 0.8 to 2.0 cm, peak velocity increased from about 10 to 40 cm/s. For this male speaker, and for the female speaker studied also, quite a strong positive linear correlation between distance traversed and peak velocity was shown. A similar correlation for distance and mean velocity in these two speakers showed that as a first approximation it would be reasonable to assume a constant duration for all transitions of this single tongue point for vowel-to-vowel actions of both speakers (Kent and Moll, 1972a). Many studies show that when a greater distance is traversed by a given articulator the time taken does not increase in proportion; the data
indicate instead that there is an almost constant duration for articulatory gestures, regardless of distance moved.

In the next section an appropriate range of durations is suggested for each type of articulatory transition. The focus is upon a slow or moderate speaking rate, but it may be noted that transition duration does not necessarily decrease at a faster speaking rate. Peak velocity is sometimes higher but the distance moved larger at a slower speaking rate, so that transition duration is sometimes approximately the same for slow and fast speaking rates. This has been shown, for example, for the lower lip (Chistovich et al., 1965, p.178) and for tongue movements (Kent and Moll, 1972b).

Recent evidence, obtained by different transducer systems as well as by cinefluorography, has provided additional support for the view that peak velocity increases in direct proportion to distance traversed during an articulatory transition. This has been shown for, for example, the tongue dorsum and tongue tip for each of five speakers (Kuehn and Moll, 1976), the lower lip with the jaw component subtracted (Hirose and Kiritani, 1979, the jaw (Kiritani et al., 1982), and the tongue dorsum (Ostry and Munhall, 1985).

II.3.2.5 The articulators and their movement path durations

Introduction

In this section the articulators will be listed and reasonable values of transition durations for each one, approximately independent of distance traversed, will be suggested. Some of the evidence on movement paths for the individual quasi-independent articulators will be cited. The various kinds of tongue body transitions will be discussed in some detail.
About nine parameters are needed to describe the control of the positions, states and movements of the articulators when simulating normal speech. The sixteen parameters which Ladefoged (1979) proposed as necessary and sufficient (see Section II.1.1 for the list) can be reduced to a smaller number if the various parameters listed by Ladefoged which shape the vocal tract are grouped together into solid structures, each of which may need more than one parameter to describe its movement paths in speech, as follows:

1) the tongue body with the pharynx walls;
2) the tongue tip or tip and blade;
3) the jaw;
4) the lips;
5) the velum;
6) abduction and adduction of the vocal folds;
7) stiffness and effective vibrating mass of the vocal folds;
8) vertical position of the vocal folds;
9) subglottal, respiratory control of the lung walls.

These nine parameters are taken to be independently controlled by a speaker to a first order of approximation. Their roles in speech production are, briefly, as follows:

1) the tongue body and pharynx wall shapes control, together with the lip outlet shape, the whole vocal tract acoustic filter area function; tongue body shape and lip protrusion change slowly, relative to other actions; raising and lowering the tongue dorsum, for example for velar consonants, may need to be considered as a distinct articulatory gesture with its own transition time;

2) the tip or tip and blade of the tongue forms the main constriction for many consonants of English; the tip is probably the fastest moving articulator, especially for the production of taps [央行]; in English the actions
are likely to be less rapid;

3) the jaw is included since, if its position needs to change for the production of a required allophone, the jaw transition acts as a time constraint;

4) two aspects of lip shape give different transition durations: lip protrusion and retraction are relatively slow; these actions combine with others as stated in 1) above. The lip closing and opening actions needed for bilabial consonants are more rapid, probably very rapid where jaw assistance is not invoked;

5) raising and lowering the velum, together with closing and opening actions of the upper pharynx wall structures, control the size of the velopharyngeal port;

6) abduction and adduction of the vocal folds are the main articulatory action which control the area of the glottis;

7) changes in the stiffness and effective vibrating mass of the vocal folds contribute the laryngeal, myoelastic, component of control of the fundamental frequency of voicing F₀;

8) vertical movements of the larynx as a whole, and thus of the vocal folds, contribute to the control of vocal tract length and cavity volumes, probably also constituting a factor in F₀ control;

9) Controlled movement of the walls of the lungs is responsible for the flow of air out of the lungs which provides the basis for most acoustic sources of speech; it contributes to the control of subglottal air pressure.

Two articulators not generally used in normal speech production will be mentioned also. First, the nostrils outlet size is sometimes controlled by abnormally nasal speakers, as a substitute for separation of the nasal and oral cavities by control of the area of the velopharyngeal port. Secondly, a segment of the wall near the
top of the oesophagus can be made to vibrate as a substitute for voicing in the production of oesophageal speech; some degree of voluntary control over the abduction and adduction of the opposing portions of the segment wall appears to exist in such speakers.

The boundaries between different portions of the vocal tract cannot be rigidly defined even for a single speaker. The various speakers analysed in the publications reviewed here have their own individually shaped structures and individual neuromuscular control characteristics. The discussion here focusses on points or small regions of the respiratory tract which are judged to contribute significantly to the acoustic sources and/or filters for particular classes of speech sounds.

It is not suggested that the movement paths of natural speech are inevitably of the forms described here; they are simply those which do occur. However, to the extent that different studies point to similar conclusions, the articulatory gestures found may be supposed to be the framework within which speakers normally operate for speech. As stated in Section II.3.2.4, the values which follow are for a slow or medium speaking rate.

**Tongue body, dorsum, and root**

Transition durations for the tongue body and tongue dorsum are about 150 to 250 ms or more, with vertical movements of the dorsum for velar consonants possibly at the faster end of the range (Houde, 1968; Perkell, 1969; MacNeilage and Declerk, 1969; Kent and Moll, 1969, 1972a, 1972b; Kent, 1972; Kiritani et al., 1975; Sonada and Kiritani, 1976; Kiritani et al., 1978; Ostry and Munhall, 1985), with indirect, acoustic evidence (Öhman, 1966; Gay, 1968).
Transition durations for lateral movements of the mid pharynx walls are about 100 to 180 ms (Minifie et al., 1970; Niimi, 1979).

Houde's (1968) analysis of movement paths of four marked points on the tongue, derived from cinefluorography for one speaker, suggested two possible principles concerning tongue movements. First, where the dorsum was the main articulator for the consonants in the speech-like sequence tongue movements were faster than with [b] as the consonant not involving the tongue in the closing and releasing actions. In a sequence such as [aːg'ɑːɡɑː] transition durations were about 160 to 170 ms along the nearly vertical line of the closing and releasing action of the dorsum for [ɡ] and about 150 to 200 ms in a backward to forward direction nearly parallel to the jaw line (op.cit. Figure 54(a), p.98). For [ɑːb'iːbɑː] the transition times were longer in duration, about 250 ms or more (op.cit. p.55, Figure 13(A), p.39 and Figure 29, p.59). Secondly, in this latter case of slower tongue body movement the four marked tongue points began and ended their movement together, although they traversed different distances. Spectrograms in Houde's analysis (op.cit., p.55 ff.) suggest that the speaking rate may have been very slow with very little shortening of vowels in unstressed positions, so that some of these articulatory transition durations may be too long for a normal speaking rate, but the observation that different points along the surface of the tongue body apparently move in synchrony for a diphthong-like gesture such as that apparent in the sequence [ɑːb'iːbɑː] seems important for simulating articulatory processes and is used in our modelling.

Shorter durations for diphthong-like gestures of the tongue body were found by Kent and Moll (1972a), about 120 ms or a little more (op.cit. Figure 3). Transitions of the dorsum for sequences [ɑːkɑː] and [iːkɪː] took about 130 to 140 ms (Kent and Moll, 1969) and were
180 and 220 ms for each of two speakers' production of [aːɡaː] (Kent and Moll, 1972b).

In an X-ray microbeam study of a speaker of Japanese, all the tongue points take the same time, about 200 to 225 ms, for vowel-to-vowel sequences, in agreement with Houde's (1968) findings. Dorsum transitions into and out of the closure for [k] in a repeated [aka] sequence and an associated vertical movement of the tongue root take about 140 ms each. The anterior-posterior dimension of the pharynx at the tongue root point hardly alters at all, so that the small amount of vertical movement visible near the root seems likely to be a concomitant of the dorsum raising and not part of a tongue body transition towards an [a] end point (Kiritani et al., 1975).

Perkell's (1969) study of articulation in [həC'V] sequences produced by a speaker of American English is of particular importance. The published traces track position changes for seventeen points and thus show how the whole vocal tract from the glottis to the lips alters in shape throughout each utterance. As with other cinefluorographic studies, only one example of each sequence is given, but some tentative conclusions may be drawn, since many of the sequences contain [t] as the consonant.

The forward movement of the tongue root for [həC...] (E path) has a clear S-shaped path and takes from 100 to 180 ms for the seven examples with [t] and for six examples containing other alveolar consonants or [p] or [k]. The mean value is about 150 ms. the transition of the tongue root from its consonant state to its end point for the following stressed vowel is more variable in shape. It has a similar but sometimes greater duration, ranging from 180 to 260 ms or more. In some cases the longer transitions seem to be made of two shorter portions; the action for [...]t'u:] for example might
perhaps be interpreted as containing two transitions of 120 ms
duration each (op.cit., p.96).

Perkell's data for [hət'V] show more complexity than simple
synchronisation of tongue body transitions. Nevertheless, in at
least two cases approximate synchronisation is apparent between
tongue root paths, points D and E, and the tongue dorsum path B. The
root moves forward while the dorsum moves upwards or vice versa.
This tongue body transition into the [t] position takes about 150 to
170 ms; the tongue body transition from [t] to the vowel end point
takes about 200 ms. There is an almost perfectly static portion of
the tongue body at the back part of the dorsum near the soft palate,
path C. The only clear departure from this pivot-like effect when
the consonant is [t] is for [hət'u:], where the C point rises up and
back to form a narrow constriction near the middle of the soft palate
for [u:]. Among sequences containing other consonants, only [hək'e]
shows a similar effect: this portion of the tongue forms the main
vocal tract constriction for [k] and there is closure near the soft palate-hard palate boundary. The pivot-like effects in Perkell's
(1969) traces appear to be entirely consistent with the tongue
factors found for vowel production, discussed in Section II.3.2.1.
It seems that components for quasi-static tongue body shapes are
likely to apply to tongue body gestures also. This extension of the
principal components approach to articulatory modelling has been
investigated by Maeda (1979).

Sonada and Kiritani (1976) showed movement paths for a point 3 cm
back from the tongue tip with a magnetometer technique. Their vowel-
to-vowel transitions indicate synchronisation of vertical and
horizontal components of movement, with a transition time of about
160 ms at both slow and medium speaking rates. A static end point
for [a] is seen at the slow rate but not at the medium rate.
Taken together, the various publications cited here suggest that the following rules for tongue body movement should be included in articulatory models:

1) In diphthong-like gestures all parts of the tongue move together. They coincide for the transition's start, end and middle region. Peak velocity in this middle region increases with distance traversed. Transition durations for a moderate speaking rate fall within a range of about 150 to 250 ms or even more.

2) The movement of the tongue dorsum into and out of a closure for a velar plosive may possibly occupy a shorter time, but the evidence is not strong. Here also a range of about 150 to 250 or even 300 ms seems plausible for dorsum transitions executed at a moderate speaking rate.

3) At a fast speaking rate transition durations for all portions of the tongue body may be reduced. A reasonable estimate for the range of durations in this case is about 120 to 160 ms.

4) The shorter durations may be appropriate also for vowel-to-vowel diphthong-like gestures if one of the vowels is unstressed.

5) Different speakers can use somewhat different ranges of durations. Speaker type might perhaps be characterised in part by a perceptually judged speaker rate. It is not clear whether speaker rate and speaking rate would be independent of each other as perceptual judgements.

6) The duration ranges are not to be interpreted in terms of increased duration for a larger distance traversed. As a first approximation, a constant transition time should be assumed for both larger and smaller distances, but with a large scatter about any mean value used.
Tongue tip and blade

The mean duration for the transition durations for the tongue tip or tip and blade are about 70 to 95 ms for actions similar to those used in RP English, with trills and taps excluded (Perkell, 1969; Kent and Moll, 1969, 1972b; Kiritani and Fujimura, 1975; Kuehn and Moll, 1976). As discussed in Section II.3.2.4 above, jaw raising assists closure or severe narrowing of the oral cavity for consonants, so that transition durations for the tongue tip region may be controlled by jaw movements rather than by small, probably very rapid tongue shape changes. This may perhaps explain, for example, some findings of Laferriere (1983): if a stressed vowel preceded a [t] or [d], the tongue blade moved up with an almost constant transition duration of about 100 to 170 ms, increasing only slightly with distance moved; but if an unstressed vowel preceded the consonant, tongue blade transition duration increased noticeably with distance moved, from about 40 ms for a displacement of about 2 mm up to about 110 ms for a displacement of 14 to 16 mm. Tongue blade lowering transitions for [...] analysed by Fujimura and Spencer (1983) have durations as long as 200 ms, which seem likely to reflect jaw movement constraints. In Perkell's (1969) study, the raising action of the tongue tip for an alveolar consonant of American English and the first 30 to 50 ms of the downward movement for an alveolar plosive release have higher velocities than the later portion of the release phase. The downward movements for the first 30 to 50 ms of the alveolar plosive releases appear to be approximately invariant across different following vowel contexts and approximately co-terminus with frication noise, suggesting that these are actions needed to provide specific acoustic cues for plosive judgements. It is important in modelling that the rest of the vocal tract articulators which are not the main releasing articulator for a plosive should be moving much more slowly than the plosive release path itself, so that the
acoustic cues for listeners' judgement of place of articulation should be clear (Maeda, personal communication). It seems necessary in modelling these complex paths to impose a more rapid tongue tip movement upon a slower jaw and tongue body based movement; this is done in our model as described in Appendix 3 (Scully, 1987) Section I.9. Values for the rate of the approximately linear increase of constriction area for the main vocal tract articulator during the first portion of a tip-blade release for [t'3:] were estimated for two women speakers of RP English as 9.1 and 12.5 cm²/s (Scully, 1989). These values fall within the range 5 to 20 cm²/s given for plosive releases by Fant (1960, 1970, p.199). The same two RP speakers had slower, parabolic increases of constriction area in the case of the affricates [tʃ]. For [t'3:] an area of 0.5 cm² was reached in 55 or 50 ms; for [tʃ'3:] the same area of 0.5 cm² was reached after 80 or 70 ms (Scully, 1989).

Transition durations for the jaw are likely to be the most reliable values of all, since the jaw is a rigid structure and its movements, mainly rotation about the temporo-mandibular joint, can be directly recorded by means of transducers attached to the lower teeth. Several studies show peak velocity increasing approximately in proportion to distance moved, for example Ohala (1970), Sussman and Smith (1971). Clear positive correlations between peak velocity of jaw movement and distance moved during a transition are seen in the magnetometer transducer data of Kiritani et al. (1982). Jaw lowering and jaw raising by the same speaker are similar at the slower speaking rate elicited. Typical values for peak velocity are 4 to 8 cm/s for a distance moved of 6 mm. Without a mathematical model of the relationship between the two, total transition duration cannot be reliably deduced from these data. From other studies a duration of about 140 to 170 ms is indicated (Kozhevnikov and Chistovich, 1965,
p.180; Kiritani and Fujimura, 1975; Nelson et al., 1984) or perhaps from about 160 ms to as high as 260 ms (Kent and Moll, 1972b).

Lips

More than one kind of transition needs to be considered for the lips. One of these is lip rounding and protrusion or the opposite action of spreading and retraction. A constant transition time across different speaking rates was found by Benguerel and Cowan (1974) but a value cannot be deduced from the published traces. It is not always clear whether a particular trace shows one or more transition gestures and the range of durations seems wide. Judging from peak to peak time for orbicularis oris muscle activity, as indicated by electromyography, changes of lip rounding seem to take about 200 ms (Hirose and Gay, 1972); however, this lip action is for bilabial plosives and not for vowels with different lip outlet shapes. The most reliable data seems to be that of McAllister (1978) and Lubker and Gay (1982) for Swedish, with durations between about 250 and 500 ms; and for three American English speakers (Lubker and Gay, 1982), with similar durations for the three speakers: about 220 to 280 ms.

Another lip action, carried in part by the jaw, is that for closing and opening the lips for bilabial plosives. The resultant movement path can be complex. For example, the cinefilm analysis of Fujimura (1961) showed that at the release of a [p] the lower lip moved down extremely rapidly at first: rates of increase of constriction area here are about 20 to 100 cm²/s (Maeda, 1987, referring to data in Fujimura, 1961). The traces published by Fujimura (1961) indicate that, in particular examples, the lip outlet area changed from zero to 0.3 cm² in only 5 ms, then from 0.3 to 1.0 cm² in 5 ms; a second phase was much slower, from 1.0 to 2.0 cm² in 30 ms. It seems likely that the initial rapid movement was due to lip muscle action, with
the jaw carrying the lips in the later, slower phase. There were other complications: the inner edges of the lips could be analysed into two components: a smooth path with a transition duration of about 80 to 120 ms, related to jaw lowering, and a highly damped oscillation of one or both lips with a period of 25 to 30 ms. The oscillation was seen for [p] and [b] but not for [m]; it seems likely to be an aerodynamic effect associated with the high airflow at a plosive release. On this basis, an oscillatory component would not be included under articulation. The initial rapid increase in vertical lip separation could be given a lip transition duration of about 20 ms. As far as the control of lip outlet area is concerned perhaps it does not matter whether this time interval is the first cycle of a highly damped oscillation or whether it is due to lip lowering muscles acting over a very short time, before being swamped by jaw actions.

A different kind of movement path is obtained from pellet movement in X-ray microbeam traces. Here the pellet shows lip movement after the outlet area has been reduced to zero. The traces obtained by Kiritani et al. (1978) show, for one American English speaker, the release of a [p] closure very close in time to the point of maximum velocity for a downward and backward movement of the lower lip pellet. Similarly, closure occurs near the maximum velocity in an upward and forward direction. Each of these lower lip gestures takes about 150 ms. As in the case of a suggested end point for the tongue tip above the palate for alveolar plosives, as sketched in Figure 7, Section II.3.2.4, here it may be reasonable to give the lips a cross-section area less than zero as the end point value for bilabial plosives. In modelling, the negative lip outlet areas would be reset to zero and the release, where area increases from zero, would have a high velocity consistent with the findings of Fujimura (1961) and Kiritani et al. (1978). Brooke and Summerfield (1983) found that for
British English [VCV] sequences the upper and lower lips appeared to move in synchrony. Their videotape data showed that downward movements of the upper lip for [b] closure took about 100 to 190 ms; for release the upper lip moved with transition durations between about 190 and 250 ms. The lower lip moved in a slightly shorter time.

Studies for American English vowels (Fromkin, 1964) and for French vowels (Abry and Boë, 1983) show that the width and height of lip separations need to be considered as separate parameters in a detailed model of lip actions for speech. Lip protrusion seems to be related to decreasing lip space width in most but not all cases (see also Section II.3.2.1).

Lip actions are further complicated by the lack of synchrony between lip actions and the jaw actions which help to bring the lips together. This aspect of inter-articulator coordination is discussed, with others, in Appendix 3, Section II.17 (Scully, 1987).

**Velum**

Durations for changes in velum height range from about 100 to 160 ms (Björk, 1961; Kozhevnikov and Chistovich, 1965; Fritzell, 1969; Scully, 1970; Moll and Daniloff, 1971; Vaissière, 1983). The velum does not simply move between two states: down for nasal consonants and nasalised vowels, up for non-nasal allophones. Its height appears to be adjusted according to the aerodynamic and acoustic requirements for the production of, for example, close and open vowels (Lubker, 1968), or plosive, fricative and affricate consonants in which oral air pressure must be raised high enough to generate sufficient transient and frication noise acoustic sources. Velum height is not all that is involved in achieving a sufficient closure
of the velopharyngeal port. Constriction of the upper pharynx accompanies velum raising in normal speakers. Lateral movements of the upper pharynx wall are shown by Niimi (1979) for Japanese sequences containing nasal consonants; their transitions seem to be of about the same duration as the velum movements and synchronised with them in most but not all cases shown.

**Adduction and abduction of the vocal folds**

Durations for changes in the degree of vocal fold adduction range from about 90 to 150 ms (Rothenberg, 1968; Frøkjær Jensen et al., 1971; Sawashima and Miyazaki, 1973; Kagaya and Hirose, 1975; Iwata and Hirose, 1976; Lofqvist and Yoshioka, 1981). The methods used include direct fiberscope observation, with measurement of relative glottal width and optical glottography. Electromyography for the posterior cricoarytenoid muscle, the sole abductor of the vocal folds, also indicates the duration of the abducting and adducting actions of the vocal folds (Hirose and Gay, 1972). There is no clear correlation between the amount of increase in glottal width as observed by fiberscope methods and the time taken for the vocal folds to become abducted in different repetitions of voiceless consonants (Sawashima, personal communication). It might nevertheless be the case that more time is required to change the vocal fold state to a large glottis state, as needed for voiceless fricatives and voiceless aspirated plosives, as compared with transitions to a breathy voiced state, such as may be needed for other phonetic classes of consonants. Rothenberg's (1972) study of glottal adjustments for tight or glottalised and loose or breathy voicing indicates transition durations within the range cited above, or a little longer. The shapes of optical glottography traces and, as suggested in Section II.3.2.3, of glottal width paths seem likely to be quite good indicators of the path shape for glottal area. On this basis,
the transition appears to end more suddenly at the large glottis change of direction of movement of the vocal folds than at its phonation state start and end. There is very rarely, if ever, a static segment during which the vocal folds remain adducted. There may be difficulties of interpretation of optical glottography traces at the phonation state end, however, since light transmission through closed vocal folds tissues must be considered possible.

**Vertical movement of the larynx**

Durations for changes in larynx height range from about 100 to 200 ms (MacNeilage and DeClerk, 1969; Riordan, 1979).

**Stiffness and effective mass of the vocal folds**

Durations for changes in the stiffness and effective vibrating mass of the vocal folds range from about 80 to 150 ms, but with values at the higher end of this range during speech. This articulatory parameter of larynx behaviour, to be called Q in the modelling described here, is not observable as the displacement of points on solid structures. Since Q is the major controlling factor for changing fundamental frequency $F_0$, it is on acoustic $F_0$ data that these estimates are based, mainly on $F_0$ changes produced in singing. Estimation of total S-shaped transition duration is made difficult here, as for other articulators also, by the fact that investigators often measure a time constant or response time which constitutes a part only of a complete transition as defined by zero velocity at start and end.

Several studies have shown that $F_0$ rises take more time than $F_0$ falls, when singers move across different musical intervals (Ohala and Ewan, 1973; Sundberg, 1979; Fujisaki, 1981). Some data from $F_0$
traces of speech appear to contradict this trend (Scully, 1973b; Anderson et al., 1984). The different indications for Q may perhaps be associated with different degrees of aerodynamic components of \( F_0 \) in singing and in speech; the singing data might be expected to be the better guide to Q. Transitions of \( F_0 \) can have different durations for one singer, from very short values for appoggiatura falls to very long values for portamento rises. From traces for one highly trained singer, these extremes of the range can be estimated at about 80 and 800 ms. The differences in duration for rising and falling \( F_0 \) are more noticeable for the fast transitions of appoggiatura and non-legato note changes (Fujisaki, 1981).

It is perhaps appropriate to consider vibrato frequency as a guide to the fastest possible Q transition. The mechanism of vibrato is not well understood, but it appears to be associated with periodically alternating levels of cricothyroid, lateral cricoarytenoid and vocalis muscle activity, in some cases at least (Hirano, 1981). For vibrato between 5 and 7 Hz transition durations between 100 and 72 ms respectively are needed. This range of durations lies at the fast end of Q transition durations indicated by several studies; singing with vibrato was successfully modelled in this way (Scully and Allwood, 1985). All except portamento transitions in Fujisaki's (1981) study seem to agree with the range of transition durations seen for other singers analysed.

\( F_0 \) changes appear to be a little faster for women than for men and a little faster for trained than for untrained singers (Sundberg, 1979). If his measures, for 75% of the transition distance, are doubled as an estimate for the total duration of an S-shaped transition, then the values range between about 120 and 190 ms for rising \( F_0 \) and between about 120 and 150 ms for falling \( F_0 \). When singers raise \( F_0 \), the time taken increases slightly as the frequency
change increases, but for falls in $F_0$ the time taken seems to be approximately independent of the musical interval involved (Ohala and Ewan, 1983; Sundberg, 1979). Some $F_0$ versus time traces for speech suggest that the peak velocity of $F_0$ change increases as the $F_0$ distance fallen increases, but that there is a constant peak velocity for the larger $F_0$ transitions (Anderson, 1984). In addition, $F_0$ transition durations were found by Fujisaki (1981) to be approximately independent of the level of effort. Faster transitions tended to have more overshoot and oscillation about the end point. Where a change in $F_0$ was made extremely fast, for a fall at higher effort, this was accompanied by a great deal of overshoot.

Subglottal articulators

Finally, estimates of durations for articulatory transitions associated with respiratory control in speech, the controlled reduction of lung volume, need to be considered. A few studies show measurements of lung volume changes during speech. The decrease of lung volume seems to be linear and generally slower for speech than for the expiratory phase of quiet respiration (Draper et al., 1960; Bouhuys et al., 1966).

The articulatory transitions that need to be considered are changes in the rate of lung volume change, that is, the second differential of lung volume with respect to time. The distribution of volumes between the chest and the abdomen is shifted before speech: there is, surprisingly, a rib cage horizontal expansion, but this is accompanied by a reduction in the circumference of the abdomen; the nett effect seems to be that lung volume is reduced (Konno and Mead, 1967; Hixon et al., 1973). This preparatory manoeuvre appears to take about 90 to 110 ms (Baken et al., 1979; Wilder, 1983).
The rate of lung volume decrement increases according to the acoustic intensity requirements: in one study, from about 120 cm³/s for reading at a normal loudness level to about 250 cm³/s or more in very loud reading (Bouhuys et al., 1966). It is difficult to make any estimate for the time needed to change this volume reduction rate during speech, if this does normally happen at all. Some, perhaps all, changes in the rate of lung volume reduction can be considered to be passive reactions to changing loads on the aerodynamic system (Ohala et al., 1980; Ohala, 1989).

The time taken for subglottal or lung air pressure to rise from about zero to its operating value in speech may perhaps be a guide to the durations of respiratory actions for speech, although it must be remembered that these pressures increase through the combined effects of subglottal, larynx and vocal tract actions, not solely because of lung wall movements. Durations range from about 100 or 150 ms up to 300 ms or more (Ladefoged, 1960, 1963; Rothenberg, 1968, p.26 ff.; Netsell, 1969; Löfqvist, 1975). The time needed to reach a high level of respiratory muscle force is about 200 ms (Mognoni et al, 1968) or 250 to 300 ms, independent of the amount of airways resistance (Mead and Agostini, 1964).

Some of the uncertainties here relate to the choice of an appropriate representation of respiratory articulation, whether by lung volume reduction rate, air pressure in the air sacs of the lungs or nett expiratory force applied. The last of these is considered by Ohala to be the most appropriate model, giving results under different amounts of loading in good agreement with natural speech (Ohala, 1989).
II.3.2.6 Timing, coordination and variability

Appropriate timing of actions and coordination between the moving structures are essential for the skilled sensori-motor activity of speech production. The nature of this skill and the kinds of ways structures may act together have been discussed in Section II.1. Inter-articulator coordination and the time span of nearly static configurations at transition end points for the various articulators in some different phonetic contexts are discussed in Appendix 3, Sections II.12 to II.18 (Scully, 1987).

Human actions, even ones as well practised and skilful as those of speech production, are subject to variability. This is discussed in Appendix 3 (Scully, 1987, Sections II.9 to II.12).
II.3.3 Aerodynamics

Introduction

Where slowly changing air flows and air pressures are considered, in the aerodynamic stage of speech production, the much more extensive literature on breathing and lung function is valuable as a source of data. Much of this section relies on respiratory textbooks, especially Bouhuys (1977), Cotes (1979) and West (1974). The physiological findings must be taken over with caution, however, since the use of the system differs notably in speech and in respiration, as discussed below in this section.

To a first approximation, the actions of any one articulator may be considered in isolation, independently of what all the other articulators are doing at the same time. The case is different for aerodynamic processes. Here the respiratory tract must be considered as a unified whole; the aerodynamic system, like an electrical circuit, is not reducible (Ashby, 1965, p.262).

The elements of this aerodynamic circuit are air-filled tubes, cavities and orifices, with their walls. Aerodynamic parameters are low frequency components of airflow and air pressure. There are parallels with electrical circuits, as mentioned in Section II.2. Volume flow rate of air corresponds to electrical current I and air pressure to electrical voltage V. The elements have mechanical and aerodynamic properties of inertance corresponding to electrical inductance, also stiffness, the inverse of which is compliance, corresponding to electrical capacitance, and flow resistance corresponding to electrical resistance R. Flow resistance is defined as the air pressure drop across an element divided by the volume flow rate of air through it. Unlike electrical resistance, most of the
important flow resistances in speech increase with flow rate, so that the aerodynamic equivalent of Ohm's electrical law \( R = \frac{V}{I} \) does not apply to them.

II.3.3.1 Lung volumes

The nomenclature for different lung states is shown in Figure 8. Vital capacity (VC) is the usable range of lung volume. It is much easier to measure than total lung capacity (TLC) and most respiratory studies use \%VC as an indicator of lung states. Residual volume (RV) is the volume that cannot be expelled from the lungs; it remains at the end of a maximal expiratory breath, the second part of the vital capacity manoeuvre. Functional residual capacity (FRC) is the volume assumed by the lungs at rest, as discussed below.

When applying basic physical principles to the aerodynamics of speech production it is important to use actual volumes, so estimates of TLC must be made. In healthy people TLC varies with size: its range is 3.0 to 7.3 L in adult females, 3.6 to 9.4 L in adult males, about 1.5 to 4.8 L for girls and about 1.6 to 6.4 L for boys, both these last two groups with height ranges 1.1 to 1.8 m. Other lung volume values for adults are: for women, VC 1.4 to 5.6 L (mean 3.14 L), RV 0.4 to 3.0 L (mean 1.10 L), FRC 0.7 to 4.9 L (mean 1.82 L); for men, VC 2.0 to 6.6 L (mean 4.78 L), RV 0.5 to 3.5 L (mean 1.19 L), FRC 0.8 to 6.5 L (mean 2.18 L). Using the mean values cited, FRC/TLC is about 0.35 to 0.4 and RV/TLC is about 0.2 to 0.25. The ranges of values are taken from Cotes (1979, pp.68,72-74,335); the mean values from Geigy (1973, p.550). VC decreases, while RV increases, with increasing age in adults (Cotes, op. cit., pp.371-381). The Geigy Scientific Tables (op. cit.) contain comprehensive data for adults and children with 95% confidence limits included.
The main forces available to a speaker for the control of lung volume comprise muscle forces (inspiratory and expiratory), elastic recoils and air pressure. These are shown schematically in Figure 9, as pressures, which is customary in the respiratory literature. Some other forces, such as gravitation and surface tension are omitted here, as are regional variations in the pressures although these are by no means insignificant. West (1974, p.98) shows pleural pressure ranging from -10 cmH₂O at the top of a lung to -2.5 cmH₂O at its base. The difference is ascribed to the weight of the lung. At FRC, the resting level reached at the end of the expiratory phase of quiet respiration, the elastic recoils of the lung walls and rib cage are equal and opposite; no muscle forces are required to maintain this lung volume. The value of FRC depends on posture, being about 25% smaller when the subject is supine. The figures given above are for FRC in healthy subjects in an upright posture. FRC increases with height and decreases with weight for women and men; it is shown as increasing with ageing in men; no ageing figures are given for women (Cotes, 1979, pp.373,381).

II.3.3.2 Use of the system in breathing and speech

From shortly after birth an accommodation between respiratory requirements and other functions of the airways is achieved (Selley, 1980). Starting with irregular patterns at birth, after about three days babies settle into a regular cycle of breathing alternating with sucking, quickly developing sensori-motor skills, as discussed in Section II.1.1.

Airflow requirements for speech are different from those for gas exchange. Breathing and speech both use the same airways and the same forces, but the forces are employed differently in each. The inspiratory phase is generally shorter for pauses in speech and has
higher flow rates, as compared with quiet breathing (unpublished laboratory data). Speech is produced on an expiratory phase and relies mainly on air flowing out of the lungs for the aerodynamic component in the generation of acoustic sources. The durations of the inspiratory and expiratory phases of quiet breathing are approximately equal, but speech is characterised by short inspiratory phases and long expiratory phases. Speech is not cyclical; the duration of the expiratory phase must depend upon the linguistic structures. It may be supposed, however, that homeostasis demands, notably for concentration of CO₂ in the blood, interact with linguistic requirements. In one study of airflow during spontaneous conversation, the overall average number of speech phases for an adult male English speaker was found to lie close to the subject's quiet breathing rate (Daniel, 1984).

The lower flow rates used in speech under normal conditions could mean that not enough gas was exhaled, even with the longer expiratory phase in speech. Quite long stretches of expiratory flow are observed for many speakers after speech has ended (unpublished laboratory data). This may be part of a speaker's tactics for meeting gas exchange requirements. The effect is apparently even more noticeable for oboe players, where flow rates are even lower than for speech, and also in speech when breathing 3% CO₂ in air. Under this condition or when exercising, for which ventilation needs to be increased, air is released after the speech phase; in addition, flowrate for speech is increased not withstanding the deleterious effect on the speech (Bouhuys, 1977, p.285).

Additional mechanisms include the control of supraglottal air pressure changes through a variety of actions. Speech differs from the expiratory phase of quiet breathing in two respects: control and obstruction.
In quiet respiration the expiratory phase is mainly or entirely passive: the inspiratory muscles relax and the lungs collapse under the influence of the nett recoil force for the lung walls and rib cage combined. This is an expiratory force for lung volumes above FRC. Inspiratory muscles must be used to expand the lung volume against the nett recoil, for the inspiratory phase of breathing. The resulting volume flowrate of air into and out of the lungs in the breathing cycle is approximately a sinusoidal function of time (West, 1974, p.105). Breathing rates decrease from about 50 cycles per minute in the newborn to about 15 per minute for girls and boys aged 16 years, to about 12 per minute for men and women at rest (Geigy, 1973, p.550).

In speech, the escape of air from the lungs is controlled so as to give volume flowrates lower than the highest values of about 500 cm$^3$/s occurring midway through the expiratory phase of quiet breathing. Overall, the flowrate in speech is about 100 to 200 cm$^3$/s (unpublished laboratory data) and lung volume decreases as an approximately linear function of time (Draper et al., 1960; Bouhuys et al., 1966), although Ohala et al. (1980) found that lung volume decreased more slowly during plosive closures and more rapidly during [h] and just after the release of aspirated plosives. These rate variations are explained by Ohala as short-term passive reactions to lung air pressure variations which are themselves passive reactions to changing downstream resistance, under an approximately constant applied respiratory force driving the lung volume decrement (Ohala, 1989).

This constant nett expiratory force is achieved by a shifting balance between inspiratory muscles, especially the external intercostal muscles, and expiratory ones, especially the internal intercostal muscles (Draper et al., 1959). Often, at the start of an expiratory
phase for speech the lung volume is above FRC and inspiratory muscle forces need to dominate in order to oppose the nett elastic recoil of the lung walls and rib cage combined. In ordinary speech, diaphragm activity ceases very soon after expiration begins, but in highly trained singing, where the requirements for respiratory control go well beyond that of ordinary speech, diaphragm muscle activity has been found (Sundberg et al., 1985).

In normal healthy breathing, the whole respiratory tract is relatively unconstricted with low total flow resistance, while during speech there must nearly always be at least one severe obstruction to airflow, since it is at such constrictions of the respiratory tract that acoustic sources arise. Increase in air pressure is associated with significant obstruction to airflow. For quiet breathing by healthy subjects the air pressure developed in the lungs is only about 1 cmH₂O, but it can be much higher for patients with airway obstruction (West, 1974, p.105). During speech, differences of air pressure develop across constrictions and the lung air pressure is likely to be about 5 to 10 cmH₂O; it can rise much higher than this for trained singing.

For both quiet breathing and conversational speech, the mid region of lung volume is generally used by healthy subjects. Quiet respiration occupies the tidal volume, from FRC to about 0.5 L above that; in an example of reading at different loudness levels and for spontaneous speech also the lung volume was centred approximately at FRC, but it was above FRC for the loudest reading (Bouhuys et al., 1966).

II.3.3.3 Distribution and aerodynamic properties for the air in the respiratory tract

In this section are considered the distributions and magnitudes for
the geometry and dimensions of the airways and the distribution of volume and flow resistance. It will be shown that volume and flow resistance are well separated in the subglottal portion of the respiratory tract.

**Geometry, volumes and flow resistances in the supraglottal airways**

The supraglottal volume of the pharynx and nose in quiet breathing is estimated by Bouhuys (1977, p.53) to be about 60 cm³ as discussed in this section below. The shape of the pharynx and oral cavity changes continuously during speech. Articulatory movements such as jaw lowering and raising can alter the volume of air in the vocal tract, but it is difficult to estimate the magnitudes of such volume changes in natural speech. The supraglottal volume enclosed behind a bilabial closure has been estimated from X-ray data as about 70 cm³ for a man (Fant, 1960, 1970, p.279). Flow resistances in the vocal tract vary greatly in both location and magnitude during speech. Between the supraglottal and subglottal airways, the flow resistance of the glottis is of major importance in speech, and is significant also during quiet respiration.

**Geometry, volumes and flow resistances in the subglottal airways**

The subglottal airways contain very large numbers of branching tubes such that a cast of the airways has the appearance of an inverted, double-limbed tree (see Figure 1(a)). This aesthetically pleasing organic complexity must be abandoned for the purposes of aerodynamic modelling and reduced to a highly simplified description consistent with the overall, observed effects. The total number of airways is fairly constant from birth and is independent of height, but the cross-section areas and the number of alveoli increase with body size (Cotes, 1979, p.57).
The airways are given generation numbers: from generation 0 for the trachea, 1 for each bronchial tube, up to numbers 19-23 for the alveolar ducts and 24 for the alveoli of the lungs. These are the terminating alveolar sacs across the walls of which gas diffusion takes place into and out of the blood stream. 'Large' airways, from generations 0 up to about 11, are supported by cartilage as well as by smooth muscle, while the 'small' airways of generation above about 11 have walls of smooth muscle. A few alveoli arise from respiratory bronchioles, but most arise from alveolar ducts, about $20 \times 10^6$ in number, with between 10 and 30 alveoli per duct (Cotes, 1979, p.59). Above about generation 12 air transport occurs mainly by diffusion (West, 1974, p.7). The total surface area of the alveoli has been estimated as about 50 to 100 $m^2$ (West, op. cit., p.10).

Dimensions of all airways of diameter 0.06 cm or more were measured in one human lung by Horsfield and Cumming (1967, cited by Pedley et al., 1970). The number of branches between the trachea and bronchioles of diameter 0.06 cm varied between 8 and 24. In the simplified dichotomous model of Weibel (1963), all branchings are assumed to be two-way. Table I, based on Bouhuys (1977, p.31), conflates estimates of the dimensions and numbers of the various generations of airways for the lungs of an adult at 75% TLC, derived by Weibel.

Theoretical predictions of flow resistance in the various portions of the subglottal airways were made by Pedley et al. (1970) and by Pedley (1977), using the geometry of Weibel with branching angles based on those found by Horsfield and Cumming (1967). Weibel's data were for a lung with a TLC of 6.4 L which could be for a woman or a man. It may be seen in Table I that total cross-section area remains approximately constant across generations 0 to 4, increases gradually through generations 5 to about 14, and then increases dramatically.
Table I  Geometry of the subglottal airways for lungs at 75% TLC, based on Buhuys (1977, p.31).

The airways are categorised as large or small. The columns show, from left to right: Generation number, name; data from Weibel's model: diameter, length, number per generation; counts of small airways in human lungs; values derived from diameter and length: total cross-section area $A$ and total volume $V$.

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Name</th>
<th>Diam. (cm)</th>
<th>Length (cm)</th>
<th>Number per airway</th>
<th>Small airway counts</th>
<th>$A$ (cm$^2$)</th>
<th>$V$ (cm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>trachea</td>
<td>1.80</td>
<td>12.00</td>
<td>1</td>
<td>2.54</td>
<td>30.5</td>
<td></td>
</tr>
<tr>
<td>L 1</td>
<td>main bronchi</td>
<td>1.22</td>
<td>4.76</td>
<td>2</td>
<td>2.34</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>A 2</td>
<td>lobar bronchi</td>
<td>0.83</td>
<td>1.90</td>
<td>4</td>
<td>2.16</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>R 3</td>
<td>segmental bronchi</td>
<td>0.56</td>
<td>0.76</td>
<td>8</td>
<td>1.97</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>G 4</td>
<td>subsegmental bronchi</td>
<td>0.45</td>
<td>1.27</td>
<td>16</td>
<td>2.55</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>small bronchi</td>
<td>0.35</td>
<td>1.07</td>
<td>32</td>
<td>3.08</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.13</td>
<td>0.46</td>
<td>1020</td>
<td>13.54</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>0.109</td>
<td>0.39</td>
<td>2050</td>
<td>19.13</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>bronchioles</td>
<td>0.082</td>
<td>0.27</td>
<td>8190</td>
<td>7600</td>
<td>43.3</td>
<td>11.7</td>
</tr>
<tr>
<td>14</td>
<td>terminal bronchioles</td>
<td>0.074</td>
<td>0.23</td>
<td>16380</td>
<td>13400</td>
<td>70.5</td>
<td>16.2</td>
</tr>
<tr>
<td>15</td>
<td>bronchioles</td>
<td>0.066</td>
<td>0.20</td>
<td>32770</td>
<td>24800</td>
<td>112.1</td>
<td>22.4</td>
</tr>
<tr>
<td>16</td>
<td>respiratory bronchioles</td>
<td>0.060</td>
<td>0.165</td>
<td>65540</td>
<td>51600</td>
<td>185.3</td>
<td>30.6</td>
</tr>
<tr>
<td>S 18</td>
<td></td>
<td>0.050</td>
<td>0.117</td>
<td>262140</td>
<td>224000</td>
<td>514.8</td>
<td>60.2</td>
</tr>
<tr>
<td>A 19</td>
<td>alveolar ducts</td>
<td>0.047</td>
<td>0.099</td>
<td>524290</td>
<td>909.7</td>
<td>90.1</td>
<td></td>
</tr>
<tr>
<td>L 23</td>
<td>alveoli</td>
<td>0.041</td>
<td>0.050</td>
<td>$8.39 \times 10^6$</td>
<td>11078</td>
<td>553.9</td>
<td></td>
</tr>
</tbody>
</table>

24 alveoli 244 $\mu$m  238 $\mu$m  $300 \times 10^6$  43 to 80 m$^2$
through generations 15 upwards. According to Cotes (1979, p.57),
Weibel's model overestimates the number of respiratory segments,
generations 16 and above. He states that total cross-section area
has its maximum value somewhere between the alveoli and the main
bronchi. But according to Bouhuys (1977, pp.30-33), and as shown in
Table I, it agrees well with actual counts of small airways.

Figure 10(a) is derived from the data of Table I; Figure 10(b) is
based on Pedley et al. (1970, Figure 2, p.392), (see also West, 1974,
Figure 5, p.7 and Figure 68, p.107). From Table I the total volume
contained in generations 0 to 15 can be estimated: it is about
150 cm³ only, in agreement with the values from Bouhuys (1977) cited
at the end of this section. The volumes in Table I for the high
number generations do not give a good indication of the total volume
contained in them, which must be about 4 to 5 L, since the lungs in
Bouhuys (1977, p.31) are inflated to 75% of TLC.

Together Figures 10(a) and 10(b) and the figures cited at the end of
this section demonstrate a strikingly clear separation of subglottal
volume and subglottal flow resistance. Nearly all the lung volume is
contained in generations above 15. The viscous flow resistance here
is negligible, with most flow resistance predicted by the theoretical
modelling of Pedley et al. (1970) to be located almost entirely in
generations 0 to 12. They assumed inspiratory flow with a lung at
75% TLC. Good agreement with these predictions was obtained from
measurements on dog and excised human lungs, which showed that
airways with internal diameter less than 2 mm, or generations above
about 10 in humans, contributed very little to flow resistance
(Macklem and Mead, 1967).
The cavities as lumped elements

It is reasonable to treat a volume of enclosed air as a lumped element as long as the wavelength for the highest frequency of interest is much longer than the largest dimension of the cavity. If the whole vocal tract from the glottis to the lip outlet is considered, the most rapid aerodynamic component of oral pressure rise is likely to be for a voiceless aspirated plosive and the most rapid fall for a voiced bilabial plosive, if clicks are excluded. For one woman speaker's British English [p] and [b] these changes took about 20 ms and 50 ms respectively. Taking a half-cycle time of 20 ms, the maximum frequency of oral pressure change, excluding acoustic frequencies of about 50 Hz and higher, is thus about 25 Hz. The rise and fall times could be less for velar plosives, since a smaller volume of air is enclosed in that case, but the value derived from bilabial plosives seems likely to be a reasonable indication. This is not likely to be a spurious value limited by the frequency response of the oral pressure transducer system. Although a low-pass filter with 50 Hz cut-off frequency is included, this only reduces in amplitude and does not eliminate voicing ripples at about 150 Hz or higher frequency. The minimum wavelength is thus about 1400 cm, assuming the velocity of sound under adiabatic conditions. This is about 90 times longer than the woman's cavity length of about 15 cm, so the assumption of uniform pressure throughout the vocal tract cavity is well justified. The conditions are even more securely met for smaller cavities formed by part of the vocal tract.

If the total lengths of the subglottal airways are estimated from Table I, the cavity containing nearly all the volume, generations 12 and above, appears to be only 2 cm in length. The resistance zone, generations 0 to 11, seem to have a total length of 26 cm. Both these dimensions are much less than 1400 cm and it seems reasonable
therefore to treat the whole subglottal system as only two lumped elements.

In fact, even for a tall man, the maximum dimension for the whole respiratory tract is much less than 1400 cm. It seems reasonable to assume therefore that changes in aerodynamic conditions take place simultaneously throughout the whole system.

Partition of volumes at different lung volumes

The conducting airways comprise the supraglottal airways and the subglottal airways generations 0 to 16; the total volume is known as the anatomical dead space. The subglottal airways generations 17 to 23 constitute the respiratory zone.

In a simplified representation (Ganong, 1981, p.511), a respiratory dead space is shown, having constant dimensions as the lungs expand from residual volume RV up to total lung capacity TLC. In reality, the diameters of the airways increase as lung volume rises and dead space is a function of lung volume. Estimates for the anatomical dead space at lung volumes 2 L and 5.5 L have been given as about 93 cm$^3$ not varying with lung volume, to which must be added 57 cm$^3$ at lung volume 2 L or 157 cm$^3$ at lung volume 5.5 L. The non-varying portion comprises the supraglottal airways and the trachea. Subtracting the tracheal volume of about 30 cm$^3$, as shown in Table I, gives an estimated volume of 63 cm$^3$ for the supraglottal airways (Bouhuys, 1977, p.53). Adding this tracheal volume to the varying component of dead space gives a total subglottal conducting zone volume of about 87 cm$^3$ at a low lung volume of 2 L, and about 187 cm$^3$ at a fairly high lung volume of 5.5 L. On this basis, the subglottal conducting zone constitutes about 4% of the total subglottal volume at low lung volumes and only 3% at fairly high lung volumes. It is
justifiable in modelling, therefore, to lump all the subglottal volume together in one element and all the subglottal flow resistance together in a separate element, over a wide range of lung volumes.

II.3.3.4 Resistance

Flow mechanisms

The dominant type of flow resistance in a particular region of the respiratory tract depends on the type of flow there. In general, flow resistance varies with volume flowrate.

There is bulk flow in the larger subglottal airways and diffusion in the smallest ones of the respiratory zone (Bouhuys, 1977, p.48). Fluid flow in branched tubes with the geometry of the subglottal system has been discussed by Pedley (1977). The geometry itself, rather than turbulence, is responsible for most of the pressure drop. The modelling of Pedley et al. (1970) was for inspiratory, divergent flow; the same branched tube geometry is expected to have a different effect on expiratory, convergent flow.

Expiratory, but not inspiratory, flow has an upper limit which decreases as lung volume decreases but is mainly independent of effort. The limiting mechanism is dynamic compression of the upper subglottal respiratory airways, called a Starling resistor mechanism by West (1974, pp.108-112). It occurs when the pressure drop between the alveoli of the lungs and a point in the airway downstream but still within the thoracic cavity is sufficient to make the air pressure there less than the pleural pressure around the airway tube. The nett inward force collapses the airway tube and greatly increases its flow resistance. Pleural pressure is always lower than the air pressure in the alveoli of the lungs, but the difference decreases as
lung volume, and with it the elastic recoil of the lung walls, decreases. Thus dynamic compression occurs at a lower flow rate if lung volume is low. The effect is enhanced also if subglottal airways resistance is higher than normal due to disease.

The limiting flowrate is nearly 2000 cm$^3$/s at lung volumes as low as 2 L and increases with lung volume, as explained above (Bouhuys, 1977, p.185). Since volume flowrate during speech very rarely exceeds 1800 cm$^3$/s, even during [h], it does not appear necessary to model flow limitation in simulations of normal speakers operating at the usual mid lung volumes. Flow limitation might need to be introduced at very low lung volumes or when simulating a speaker with bronchitis or asthma.

Laminar flow occurs within long tubes at flowrates low enough for the Reynold's number not to be exceeded: a parabolic velocity profile develops from a flat front at the tube entry point. Such conditions seem unlikely to exist in general during speech within the vocal tract, with its sharp change of direction near the uvula, its continuously changing tube shape and its constrictions of only a few cm in length at most. If there is laminar pipe flow, however, it can remain so up to a Reynold's number Re of about 2000.

\[
Re = \frac{vd}{\nu}
\]  

where \( \nu \) is a characteristic velocity, the linear velocity in the pipe; \( d \) is a characteristic dimension, the minimum cross-dimension of the pipe; \( \nu \) is the kinematic viscosity of air in the pipe. (See also Equation (II.20) in Section II.3.4.12.)

When a jet emerges from a circular nozzle, an unstable region, between laminar and fully turbulent flow, can occur for Reynolds
numbers as low as 160 to 1200 (Goldstein, 1976, cited by Shadle, 1989). This mechanism seems likely to be important in the generation of turbulence noise downstream from a constriction, as discussed in Section II.3.4.12.

Constrictions

The major sources of pressure drop in the respiratory tract during speech are the severe constrictions or orifices at which acoustic sources are generated. It is not appropriate to cite a resistance value here since the resistance is generally flow-dependent. Instead the orifice equation is applicable if turbulence at the outlet is the dominant mechanism for energy dissipation. The relationship between volume flowrate through the constriction $U$, the pressure drop across it $\Delta P$ and its minimum cross-section area $A$ is:

$$\Delta P = \frac{\rho_a K U^2}{2A^2}$$

This is known as the orifice (or hydrokinetic) equation. $\rho_a$ is the density of air at the constriction. $K$ is an empirical constant. The equation was first applied to speech in a model of the geometry and aerodynamic conditions at the glottis, by van den Berg et al. (1957); it was adapted from research on blood flow and applied to the geometry and aerodynamics of the velopharyngeal port by Warren and DuBois (1964). Its value varies, but not very greatly, with geometry and flow rates. Model experiments (for example Heinz, 1956; van den Berg et al., 1957; Warren and DuBois, 1964; Smith and Weinberg, 1980) showed that the value of $K$ varied with the sharpness of entry and exit to the constriction, the length of the constriction and volume flowrate of air through it. The larynx modelling of Gauffin et al. (1983) suggested a range of 0.9 to 1.5 for $K$; it has been
suggested that, for simplicity and consistency across different studies, a value K=1 should be used (Scully, 1986, Appendix 2).

The choice of a suitable single value for K has been complicated by the fact that different investigators used different values for \( \rho_a \), the density of air in the respiratory tract, from 0.001 gm/cm\(^3\) (Warren and DuBois, 1964) to 0.0013 gm/cm\(^3\) (van den Berg et al., 1957). In this study, including the associated publications, the values given by van den Berg et al. (1957) for both \( \rho_a \) and K were used. This gave an expression for the minimum cross-section area of a constriction of the respiratory tract, the working orifice equation, as follows:

\[
A = \frac{0.00076 \, U}{\Delta P^{0.5}}
\]

with area A in cm\(^2\), volume flowrate \( U \) in cm\(^3\)/s and pressure drop \( \Delta P \) in cmH\(_2\)O. The value of \( \rho_a \) should be near 0.00114 gm/cm\(^3\) for moist air at body temperature 37°C (Flanagan, 1972, p.35). Although a wrong value was used for \( \rho_a \), since the K value used was the one derived by van den Berg et al. (1957) using that wrong value for \( \rho_a \), the constant in the working orifice equation (II.13) is the same as if a correct value for \( \rho_a \) had been used, combined with K=1, as recommended above. The density of air appears in our modelling only in the derivation of the orifice equation, so the use of a wrong value is not deleterious to the modelling.

Since the pioneering work of Warren and DuBois (1964) and other studies since 1964 in which the equation has been applied to the velopharyngeal port, this orifice equation has been applied to vocal tract constrictions and to the mean value of glottal area during voicing in a few analyses of speech production (for example Hixon,
1966; Stevens, 1971; Scully, 1984a, Appendix 1; Fritzell et al., 1985, 1986). It is of central importance in the modelling described in this work, since it forms an internally consistent link between articulatory and aerodynamic processes.

**Expressions for resistance of the whole respiratory tract**

In an account of the dynamics of breathing, Mead and Agostini (1964) expressed each component of resistance by a pressure-flow equation of the form:

\[ P_{res} = K_1 \dot{V} + K_2 \dot{V}^2 \]  

(II.14)

where \( P_{res} \) is the pressure drop across the resistance, \( \dot{V} \) is the rate of change of lung volume and \( K_1 \) and \( K_2 \) are empirical constants, for a laminar and a turbulent term respectively. Their Figures 3 and 4 (op. cit. p.415) show that, at middle lung volumes, only the gaseous portions are non-ohmic with \( K_2 \) above zero. Glottal and supraglottal regions are included here. The pressure-flow expression above seems to imply that resistances of two types can coexist in one flow resistance element. This approach is adopted in our modelling, even though a more realistic representation of the processes would specify under what conditions different types of flow occurred and use different pressure-flow relationships under the different conditions (Malcolm Bloor, personal communication).

**Resistance of the tissues**

For the lung tissues and the 'chest wall', under which Mead and Agostini (1964) subsume all the thorax outside the lungs including the rib cage, with the diaphragm, all the abdominal contents and the abdominal wall, \( K_2 \) in Equation (II.14) is zero. These structures
offer ohmic resistance, independent of volume flowrate, up to at least 1500 cm³/s. At mid lung volumes for a man K₁ is 0.2 and 1.0 cmH₂O/LS⁻¹ for the lung tissues and the 'chest wall' respectively.

Bachofen and Scherrer (1970) found that lung tissue resistance increased when tidal volume and mean volume flowrate in the breathing cycle increased, and at very high lung volumes near TLC also. Hysteresis effects were observed.

When Macklem and Mead (1967) measured pressure differences between the pleural space and points in the airways they were including lung tissue resistance in their measures. Since the pressure drops from the pleural space to points in the respiratory zone were negligible, it may be concluded that, in addition to these small airways having negligible resistance, the resistance of the lung tissues is negligible also.

The ohmic flow resistance of the subglottal airways

Most values cited for total flow resistance in the respiratory literature are misleading if taken as good indicators of subglottal flow resistance. This is because the glottis and supraglottal airways have been included, and they contribute most of the pressure drop.

When the volume flow rate through the glottis is high, the pressure drop from the alveoli of the lungs to just below the glottis increases in proportion to the volume flow rate increase as long as the flow resistance of the subglottal airways R_{sg} is ohmic. If air pressure in the alveoli is assumed to remain constant while flow increases, then subglottal pressure falls from P_{sg1} to P_{sg2} when U increases from U₁ to U₂ and:
From the fall in subglottal pressure, estimated from oesophageal pressure, during the release phase of a voiceless aspirated plosive, Ladefoged (1963) deduced an effective total subglottal resistance of 1.0 to 1.5 cmH\textsubscript{2}O/Ls\textsuperscript{-1}. Rothenberg adopted a value of 4 cmH\textsubscript{2}O/Ls\textsuperscript{-1} for this, based on previous work including references in the respiration literature. This gave much larger percentage falls of subglottal pressure in his model than are apparent in natural speech (Rothenberg, 1968, pp.59-61). The interpretation of these data is complicated by the fact that a fall in subglottal pressure is predicted when the total glottal and supraglottal obstruction load is reduced, as is the case here, under the assumption of a constant applied nett expiratory force (Ohala, 1975, 1989, see Section II.3.3.2).

Too high a value for the subglottal airways resistance had a catastrophic effect upon our early attempts to model [s]. The more the vocal folds were abducted the higher the airflow rate and therefore the more the fall in subglottal pressure. All attempts to vary larynx and vocal tract coordination resulted in very weak frication noise.

A few studies do provide the necessary data. The theoretical predictions of Pedley et al. (1970, Figure 4) agreed with two earlier studies in showing low values for the pressure drop between the pleural surface and the lower part of the trachea, with an increase in pressure drop proportional to an increase in volume flowrate. These data gave a subglottal resistance of 0.5 cmH\textsubscript{2}O/Ls\textsuperscript{-1} at 75\% TLC, that is, for rather high lung volumes. They predicted non-linearity,
but over the range of flow rates used in speech the effect was insignificant.

Subglottal airways flow resistance as a function of lung volume

Two studies agree rather well about airflow resistance as a function of lung volume. Both show a linear relationship between subglottal air conductance and lung volume, as shown in Figure 11. Blide et al. (1964) measured the pressure difference between the alveoli in the lungs and the top of the trachea at a flow rate of 0.5 L/s. Their data are for the airways only. Vincent et al. (1970) included lung tissue resistance by measuring the pressure difference between the pleural cavity and the top of the trachea. The slight differences in conductance in the two studies is consistent with the different measures made. The larger discrepancies at very high lung volumes near TLC are consistent with the findings of Bachofen and Scherrer (1970) discussed above, that lung tissue resistance rises near TLC. It might be more appropriate to base the modelling of this conductance on the results of Blide et al. (op. cit.), excluding lung tissues. But a practical advantage of using linear approximations to those of Vincent et al. (op. cit.) is that subglottal conductance is thus given an upper limit. Discrepancies are not likely to be apparent in modelling ordinary speech since this operates at lung volumes well below TLC.

For the mid range of lung volumes, both studies give a subglottal airways flow resistance of about 0.5 cmH₂O/Ls⁻¹, in good agreement with the predictions of Pedley et al. (1970) and in quite good agreement with the estimates from natural speech of Ladefoged (1963).

Blide et al. (1964) found that day to day variations in total airways conductance, including the glottal and supraglottal regions, for an
individual person were quite large; perhaps this might be true for
the subglottal airways on their own, although changes in the mucosa
of the nasal cavities seem likely to dominate.

**Supraglottal and glottal flow resistance**

The severe constrictions at and above the glottis at which the
acoustic sources of speech arise are the dominant obstructions to
flow. Turbulence downstream of these constrictions is essential for
noise sources; turbulence near the upper surface of the vocal folds
is probably part of the voicing mechanism also (Broad, 1979). For
the mid portions of some vowels there may perhaps be only one
significant obstruction, the glottis; for most of speech, however,
at least two and probably more constrictions need to be considered.
The glottis and regions of the vocal tract that form the main
articulator for vowels and consonants are likely to be simultaneously
significant as the configuration changes from one to the other
constriction dominating. Lip rounding adds another constriction, the
lip outlet. The velopharyngeal port constitutes another
constriction, through which aerodynamically defined nasalisation is
introduced. The patterns of orifice flow obstruction are, of course,
continuously changing during speech.

**II.3.3.5 Compliance**

Compliance is the volume change per unit pressure change. As far as
the air in the lungs is concerned there are two components to be
considered. Air is compressible; if ideal gas laws and isothermal
conditions are assumed, then increased air pressure results in a
reduction of volume for a given mass of gas. Boyle's law PV=constant
will apply. But because the walls of the lungs are compliant,
increased air pressure within the lungs will cause their volume to
increase. Energy is stored in the walls as an elastic recoil, equivalent to a pressure, as shown in Figure 9.

The lungs have walls which yield easily; the walls are very compliant except at high volumes near to TLC, and have a value of about 200 cm$^3$/cm H$_2$O near mid lung volumes (West, 1974, p. 93). The resultant elastic recoil pressure always acts so as to reduce lung volume. The rib cage has a recoil acting in the opposite direction for all except high lung volumes near TLC. The nett effect of the lung walls and rib cage combined is an elastic recoil which acts so as to expand the lungs at operating levels below FRC and reduce their volume above FRC. Relaxation pressure is a measure of this nett recoil. It is obtained by relaxation of muscle forces against closed airways. When performed at different lung volumes this manoeuvre allows the static recoil line to be drawn. Static recoil lines were obtained for ten subjects by Rahn et al. (1946). Near FRC and above, up to about 80% VC, the combined lung-rib cage compliance of each subject was approximately constant, at about 100 to 150 cm$^3$/cmH$_2$O.

**II.3.3.6 Inertance**

Force is needed to accelerate or decelerate a mass. The equivalent process to be considered here is the inertance pressure associated with acceleration or deceleration of the air and tissues. Fant (1960, 1970) and, following him, Rothenberg (1968) state that inertance of air may be neglected at the low frequencies of interest in aerodynamics of speech production. For the low flow rates and volume accelerations associated with phonation, inertance of the lung walls and rib cage may be neglected also, according to Bouhuys et al. (1966).
Sharp et al. (1964) applied pressure sine waves to a subject's chest to obtain the resonant frequency of the whole respiratory system. Compliance C was measured independently with a body plethysmograph and inertances of the air and tissues were separated by using different gas mixtures for inhalation. Inertance I values, derived from the expression for resonance:

\[ f_{\text{resonance}} = \frac{1}{2\pi \sqrt{IC}} \]  

(II.16)

were estimated for the subglottal airways and the chest wall tissues in normal and obese subjects in units of cmH\textsubscript{2}O/Ls\textsuperscript{-1} as:

- \( I_{\text{air}} = 0.008 \) (normal); \( I_{\text{tissues}} = 0.002 \) (normal)
- \( I_{\text{air}} = 0.01 \) (obese); \( I_{\text{tissues}} = 0.02 \) (obese)

Bouhuys (1977, p.176) states, with reference to respiration and citing this study (Sharp et al., 1964), that: "Energy losses due to acceleration of the chest wall are negligible in healthy persons and are increased in obesity, where they conceivably may limit maximum flow and cough velocities."

An estimate of the upper limit for volume acceleration during speech has been made from unpublished data of a man saying "hay" loudly. Total volume flowrate, with oral and nasal airflows added, increased from zero to 500 cm\textsuperscript{3}/s in just under 39 ms, giving a volume acceleration \( \frac{dU}{dt} \) of about 13 L/s\textsuperscript{2}. Using this value in the expression for inertance, viz.

\[ p = I \frac{dU}{dt} \]  

(II.17)
the pressures associated with the total lung air and chest wall
tissues inertance for the normal subjects for whom inertance figures
are cited above are 0.13 cm\textsuperscript{2}O at most. This pressure is small
enough to be neglected in speech production where air pressure in the
alveoli of the lungs is likely to be above 5 cm\textsuperscript{2}O most of the time.
Thus, it is confirmed that inertance does not need to be included in
aerodynamic models of speech production.

II.3.3.7 Aerodynamic data for natural speech

There have been many studies in which aerodynamic parameters of
volume flowrate and air pressure have been measured for natural
speech. A detailed review will not be attempted here. As regards
the construction of a composite model of speech production, what has
to be judged is the set of articulatory actions needed to generate a
reasonably close approximation to data for natural speech.
Aerodynamic traces are generated by the model described in this work
and are compared with traces of oral and nasal volume flowrate and
oral air pressure for a few English speakers. Vocal tract
constriction cross-section area paths in the model are based upon
corresponding traces for natural speech, derived from air pressure
and airflow by means of the working orifice equation, Equation
(II.13) in Section II.3.3.4, as described in Scully (1979) and in
Appendix 1 (Scully, 1984a) and Appendix 2 (Scully, 1986). Volume
flowrate of air through the mouth and nose can be estimated fairly
accurately by means of a mask and an orally inserted pressure tube,
although there are certainly problems of interpretation with these
kinds of measurements (Scully, 1969). Subglottal pressure presents
the greatest difficulties. There have been understandably few
studies of this in natural speech, but those data which have been
reported are very important for the modelling; some of the findings
will be mentioned.
The average level of subglottal pressure for speech at a normal effort varies across speakers and studies, from under 5 cm$^2$O (Löfqvist, 1975) to about 11 or 12 cm$^2$O (Netsell, 1969; Murry and Brown, 1971). The values for one man were 3 and 7 cm$^2$O for low and high intensity tones respectively (Kitzing and Löfqvist, 1975).

Subglottal pressure is maintained fairly constant during speech (for example Collier, 1975; Löfqvist, 1975). During the central portion of each expiratory breath, two speakers analysed by Murry and Brown (1971) showed variations in subglottal pressure of no more than 10 to 15%. Subglottal pressure does not clearly rise for stressed syllables (Ohala, 1989); it is about the same for aspirated and unaspirated plosives in Hindi (Ohala and Ohala, 1972), for [t] and [d] in English (Netsell, 1969), for pairs [p] and [b], [t] and [d] in Swedish (Löfqvist, 1975) and for pairs [f] and [v], [t] and [d] in Dutch (Collier et al., 1979). The wide variety of plosive types represented here, together with one pair of homorganic voiced and voiceless fricatives, supports the view that subglottal pressure is generally kept approximately constant during speech, apart from the passive reactions to downstream resistance discussed and modelled by by Ohala (1975, 1989). In agreement with Ohala's model of respiratory control, subglottal pressure rises during some plosive closed phases and falls by up to about 1.5 cm$^2$O, about 25%, just after a plosive release (Löfqvist, 1975).

Subglottal pressure varies in the short term also, as the vocal folds vibrate during voicing, by between less than 5% (van den Berg et al., 1957) or by 31 to 37% (Kitzing and Löfqvist, 1975). As the glottal area becomes larger during voicing subglottal pressure falls; as glottal area becomes smaller subglottal pressure rises (Koike and Hirano, 1973). Acoustic components of pressure, showing resonances of the subglottal airways (see Section II.3.5), can be seen on some traces (for example Perkins and Koike, 1969).
II.3.4 Acoustic sources

Introduction

The acoustic theory of speech production, developed in the 1950s, states that speech sound waves may be uniquely specified in terms of acoustic sources and filters. To a first order of approximation the sources and filters are independent of each other. Acoustic sources in speech are changes of air pressure and airflow rate superimposed upon an inaudible air stream, generally flowing in an expiratory direction. The relevant disturbances comprise the range of frequencies perceptible to normally hearing human listeners as sound, that is from about 20 Hz up to about 20 kHz.

The basic physical principle responsible for the modification of a sound source through filtering is resonance, the property of a body, in this case the air enclosed in the respiratory tract, to vibrate when excited by vibrations at or close to its own natural resonant frequencies. The most important resonator in speech production is the air in the vocal tract, the tube of complex and changing shape between the glottis and the outlets at the lips and nostrils. Sound is radiated mainly from the mouth and nose, but also from the throat walls and, to a lesser extent, from other surfaces of the body (Kirikae et al., 1964).

The emphasis in this study is on simulation of the interactions between articulatory and aerodynamic conditions at sound-generating constrictions and the resulting patterning of source characteristics. It is necessary to filter the sources if the modelling is to produce speech-like sounds that can be assessed as acceptable or unacceptable versions of words of English. Source-filter theory has been very fully developed and expounded by Fant (1960, 1970), Flanagan (1972a)
and others. Here some limitations on the validity of particular approaches to the acoustic modelling will be discussed. Some portions of the large body of published research will be highlighted.

There are three types of acoustic sources in speech: voice, turbulence noise and transient. Silence, the absence of all of these, is important also for listeners, for example when judgements are made of the presence or absence of a plosive in word pairs "split" and "slit" (Summerfield et al. 1981). Different mechanisms are responsible for the three types of source. Voice is associated with the rapid release or, more importantly, closing off of an airstream; turbulence noise arises when eddies are formed as a high-speed jet of air emerges from a constriction; a transient arises from pressure differences between adjacent regions of the respiratory tract. Source characteristics have been discussed by Fant (1960, 1970, Appendix A.2). It is important to note that, for each source type, both the local geometry and the local airstream must be appropriate and that these two controlling factors interact. Each source is defined by a combination of articulation and aerodynamics.

II.3.4.1 Voice: introduction

Voice originates in the larynx when the vocal folds vibrate quasi-periodically under the influence of muscle and tissue forces combined with aerodynamic forces. The structure and function of the larynx is anatomically complex both because of the number of muscles which control it and because of the non-uniform composition and the non-linear stress-strain curve of the tissues of the vocal folds, discussed in Section II.2. In descriptions of vocal fold structure a distinction is made between the passive ligamental cover and the body, composed mainly of the vocalis muscle. The importance of
voicing mechanisms and the difficulties of representing them in models may be judged from the number of symposia and publications devoted to this topic alone: see, for example Fant and Scully (1977), Boë et al. (1979), Cohen and Broecke (1983, pp.147-213), Broecke and Cohen (1984, pp.157-170), the continuing series of Vocal Fold Physiology conferences (Stevens and Hirano, 1981; Bless and Abbs, 1983; Titze and Scherer, 1983; Baer et al., 1987; Fujimura, 1988).

One of the limitations of current speech synthesisers is the fixed pulse generally employed to represent the excitation of the vocal tract resonances by the voice source. In natural speech the acoustically relevant properties of voice change with the underlying physiological controlling parameter values.

The oscillations of the vocal folds are neither articulatory nor acoustic, but they give rise to the acoustic source called voice which is the acoustic component of the quasi-periodic waveform of volume flowrate of air through the glottis, called $U_g$ here. Sketches of the range of voice source waveshapes are shown in Figure 12. The data on which the sketches are based are discussed in Sections II.3.4.8 and III.2.4.1. The natural speech data are interpreted in terms of a parametric representation of each cycle of the voice waveform, shown in Figure 13. The ways in which the parameter values vary under different controlling conditions in natural speech are suggested in Figure 14.

The frequency spectrum of the source, and so the extent to which resonances at different frequencies are excited, depend on the shape of this wave. The area under the curve is an indication of the total energy and the amount of excitation at low frequencies; higher frequencies are more strongly excited if the time derivative of the
flowrate increases (Fant, 1980a). Since the steepest portion of the curve is usually at or just before closure, this is where the main excitation of the vocal tract resonances takes place. The steeper this falling part of the flow curve, the more the excitation corresponds to a pulse, with strong excitation across the frequency spectrum.

In addition to the rapidly changing acoustically significant flow there may be a continuous leak of air as shown in Figure 13. A continuous posterior glottal chink is commonly observed in normal speakers (Hirano, personal communication, 1985). A continuous component of airflow does not contribute to the voice source properties but it is likely to affect acoustic energy losses through the glottis, as discussed under filtering in Section II.3.5. The portion labelled TC in Figure 13 will be referred to as the closed phase, even if there is a continuous air leak through a chink; the rest of the cycle, of duration $T_{q} - TC$, is the open phase, during which the vocal folds are apart. The total volume flowrate of air $U_{t}$ equals the sum of the continuous flow and $U_{g}$. The aerodynamic parameter $U_{g}$ is the mean value of the volume flowrate of air through the glottis; it is related to other aerodynamic parameters as discussed in Sections II.3.3, III.1.3 and III.2.3. The relation between the aerodynamic parameter $U_{g}$ and the whole glottal flow depends on the wave shape, but if the open phase is approximated by a triangle then the maximum amplitude of the acoustic component $U_{g}$ called VOIA, the continuous d.c. flow and the aerodynamic parameter $U_{g}$ are related by:

$$U_{g} = (\text{d.c. flow}) + 0.5 \times (1 - TCR) \times \text{VOIA}$$

where $TCR$ is the ratio of the closed phase to the total cycle duration (see Figure 13). A distinction needs to be made between a
phonation state for the vocal folds and phonation itself, the
vibration of the vocal folds which creates the flow waveform of
voice. A combination of suitable conditions is required in the
larynx for phonation to occur: the vocal folds must be brought
together (adducted) to the right extent, they must be suitably
stiffened, and there must be appropriate air pressure forces applied
to the vocal folds. Different positions and states of the vocal
folds are likely to require different aerodynamic forces to sustain
vibration, but, in general, it seems that there is a range of degree
of vocal fold adduction and corresponding ranges of glottal width and
glottal area over which voicing can take place, as discussed in
Section II.3.4.8.

II.3.4.2 Theory of voicing mechanisms

The generally favoured account is the myoelastic-aerodynamic theory
of voicing, first advanced by Johannes Müller in 1843 according to
Zemlin (1968, p.213). Strong supporting evidence was provided by
measurements on excised larynges (van den Berg and Tan, 1959) and on
plaster casts of larynges (van den Berg et al., 1957). In its
simplest form the mechanism is described as follows: air pressure
applied to the lower surfaces of the closed vocal folds forces them
apart; because of continuity of mass flow air flows through the
glottis at higher velocity than in the trachea; because of
conservation of energy at the entry to the glottis higher velocity
means lower air pressure inside the glottis than below it and this
Bernoulli effect sucks the vocal folds together; air pressure
builds up below the closed vocal folds and the cycle repeats.

Increased understanding of how the vibration of the vocal folds can
be predicted from laws of mechanics came from the research of
Ishizaka and Matsudaira (1972). In his review of the theory, Broad
(1979) emphasised the importance of differences in flow patterns at the inlet and outlet of the glottis for vocal fold vibration. The flow entering the glottis is believed to remain relatively attached to the tracheal wall and vocal fold surfaces. The case is different at the outlet: here the high velocity airstream is believed to emerge as a jet separated from the boundaries. Evidence for this has been obtained with flow visualisation methods (Shadle et al., 1989). Energy is lost through turbulent mixing, so that the air pressure lost at entry is not all regained as the air stream leaves the glottis. Stevens (1977) has explained the mechanism whereby energy from the airstream is fed into the vocal fold tissues so as to maintain oscillation. The domain of the Bernoulli force associated with high velocity flow inside the glottis was analysed by Titze (1980): he showed that it is effective only when the glottis is very narrow. Under appropriate conditions the force is asymmetrical during the cycle and this is what is required to sustain oscillation. The complex relationship between average values of the aerodynamic parameters and the forces on the tissues throughout a voicing cycle has been analysed also. The lower portions of the vocal folds begin to open before the upper portions and for this converging glottis the driving pressure within the glottis is close to the pressure drop across the glottis; later in the cycle, as the lower portions of the vocal folds begin to close before the upper portions and the glottis is diverging, the driving pressure inside the glottis is close to the supraglottal pressure just beyond the outlet of the glottis (Titze, 1986).

In the 2-mass model of the vocal folds, the upper and lower portion of each vocal fold is represented by a mass. The two masses for each vocal fold are linked by a spring and each mass is attached by a spring and a viscous loss to fixed walls. Ishizaka and Matsudaira (1972) derived an equivalent aerodynamic stiffness associated with a
force acting on the lower portion of the vocal fold only and proportional to the differences in the excursions of the upper and lower portions of the vocal folds. Oscillation could occur only for low enough stiffness of the coupling between the two masses of each vocal fold; the equivalent aerodynamic stiffness needed to be larger than this coupling stiffness but smaller than the total stiffness for the lower portion of the vocal fold.

Broad (1979, p.235) notes that flow separation at the glottal outlet permits oscillation in other vocal fold representations besides the 2-mass model. Oscillation could be obtained with a one-mass model for each vocal fold but only under a limited range of vocal tract shapes; the 2-mass model is more robust in this respect. Computer simulation of the 2-mass model (Ishizaka and Flanagan, 1972) extended the conditions beyond the small signal analysis of Ishizaka and Matsudaira (1972) and showed that vocal fold vibration is limited in amplitude by the effects of collision at closure, non-linear tissue properties and non-linear relationships between the aerodynamic forces and vocal fold displacement. This and subsequent 2-mass model simulations (Broad, 1979, p.236 referring to Isshiki et al., 1977; Guérin and Boë, 1980) generated a variety of waveforms of glottal area as a function of time which were in good agreement with observations on excised larynges.

The vocal folds have been represented in other ways also, for example with each vocal fold represented as sixteen coupled masses (Titze, 1973, 1974), as a continuous medium (Titze and Talkin, 1979) or as a beam (Perrier, 1982).

The vocal folds are made up of a number of layers with different mechanical properties (Hirano, 1983). In particular the cover of a vocal fold is a ligament, passive tissue which can be tensed by the
action of the cricothyroid muscle. The body of a vocal fold can be
tensed when the vocalis muscle which is contained within it becomes
active. The effects of changes in vocal fold stiffness have been
analysed. It was proposed by Stevens (1977) that a voluntary
increase in stiffness might be used by speakers to assist in the
cessation of voicing for some voiceless consonants. Titze (1979) set
out a table of different modes of vibration of the vocal folds
arising from different combinations of body and cover stiffness with
different combinations of cricothyroid and vocalis muscle activity,
interacting with different degrees of vocal fold adduction. He
suggested that efficient voicing having a large $F_0$ range occurs when
the stiffness of the body and cover are matched; when they differ
saturation effects occur more easily and voicing becomes inefficient
or aperiodic or ceases altogether. Figure 12 includes only an
overall stiffness-mass factor $Q$, and does not attempt to represent
the kind of body-cover stiffness match parameter suggested by Titze
(1979).

The left and right vocal folds need to be well matched in their
mechanical properties for regular oscillation. This matching needs
to be learned: babies can exhibit a vocal fold vibration mode
similar to a disordered state of voicing associated with unilateral
vocal fold palsy (Fourcin, 1978). The 2-mass model can be made to
exhibit similar effects (Ishizaka and Isshiki, 1976). Direct
observation by Honda et al. (1980) showed the left and right vocal
folds having the same length as each other at minimum and maximum
glottal area but differing for intermediate openings; perhaps they
behave like coupled systems, more closely coupled at the ends of
their vibrating amplitudes.
II.3.4.3 Phonation types and registers

In view of the complexity of the larynx and the many subtle and varied uses to which it is put in speech and singing, it is not surprising that there are many descriptive frameworks in use, often conflicting both in basic concepts and in terminology. The terms voice quality, register and phonation type are used in many analyses. Categorisation is partly based on auditory judgements and it is difficult to relate these kinds of descriptions to voice waveforms and the physiological controlling conditions.

There is some considerable measure of agreement among researchers that the vocal folds of normal speakers and singers can be made to vibrate in distinctly different modes or registers. Sudden jumps from one vibration mode to another were observed in excised larynges by van den Berg and Tan (1959). Hollien (1974) presented evidence in support of three major registers called modal, falsetto and vocal fry (alternatively called glottal fry or creak). He suggested that in falsetto aerodynamic forces dominate, in vocal fry muscle and tissue elasticity forces dominate, and in the modal register neither factor dominates; both are important.

Laver characterises phonatory settings, defined as long-term muscular adjustments which give a background auditory colouring persisting over more than one segment. He reviews the laryngeal characteristics of different modes of vocal fold vibration. His neutral phonation mode is characterised by, among other features, the absence of audible noise. His modes of laryngeal vibration include some that can occur alone and also in combination, others that can occur only in combination, as modificatory settings. The components are called modal voice, falsetto, whisper, creak, harshness and breathiness. Possible combinations include, for example, harsh creaky voice and
whispery falsetto. The label ventricular implies involvement of the false vocal folds; ventricular voice is equivalent to extremely harsh voice (Laver, 1980b, pp.99-140).

The present study operates in a somewhat different conceptual framework. Here the terms 'voice' and 'voicing' are taken to refer to the quasi-periodic component of volume flowrate of air through the glottis $U_g$ or time-derivatives of this function. Associated with particular patterns of glottal airflow there may be random components also; this is the aspiration noise source originating at or a little above the glottis. The term 'phonation type' is used for voice waveshapes associated with different degrees of adduction of the vocal folds. Following the terminology used by Sundberg and Gauffin (1979), it is understood here that the normal setting of the vocal folds for speech gives normal phonation; pressed phonation implies a greater amount of medial compression and vocal fold adduction than for normal phonation; breathy phonation implies less vocal fold adduction. Normal voice here corresponds to modal and mid voice in Titze's (1979) descriptive framework, with the stiffness of the ligamental cover and the muscular body of the vocal folds varying in step. Voice waveshapes change, not only with changes of register and phonation type, but also with changes of voice effort, transglottal pressure drop, larynx height, fundamental frequency control setting and vocal tract shape, as discussed in Sections II.3.4.8, II.3.4.9 and II.3.4.10. Only some of these parameters are included in Figure 12.

II.3.4.4 Glottal area, glottal volume flowrate and vocal fold contact area during voicing

The Bell Telephone Laboratories high speed film of the vocal folds (Farnsworth, 1940) showed that during voicing the closing phase was
generally more abrupt for glottal airflow, obtained by inverse filtering, than for glottal area (Miller, cited by Timcke et al., 1958). The phase of decreasing area was generally longer than that of increasing area and this asymmetry or skewing increased at higher acoustic intensity; other effects were a noticeable increase in the maximum glottal area during the cycle and a larger proportion of the cycle occupied by the closed phase (Timcke et al., 1958). Rothenberg (1981) suggested a maximum glottal area of 0.16 mm$^2$ as typical of adult male voicing, the value shown by Flanagan (1958). This value seems very small, considering that many kinds of data point to a mean glottal area of about 0.05 cm$^2$ for voicing in adults. Titze's (1980) value of 0.3 cm for maximum glottal width, implying a maximum glottal area of about 0.15 cm$^2$, seems more plausible; this agrees with the glottal areas modelled by Van den Berg et al. (1957) and Fant's (1960, 1970, p.269) estimate of 0.1 cm$^2$ for maximum glottal area, or more, taking into account the glottal widths actually observed on the Bell Telephone Laboratories film (Farnsworth, 1940). Direct measurement of vocal fold lengths and glottal width between the vocal processes of the left and right arytenoid cartilages by means of a stereo-fiberscope suggests a maximum width during voicing of about 0.3 cm and a maximum glottal area of about 0.6 cm$^2$ (Honda et al., 1980). A maximum glottal width of 0.3 cm during voicing seems to be close to the maximum width observed during voiceless fricatives, observed in the same direct way (Sawashima and Miyuzaki, 1974).

Several studies have shown rather symmetrical rise-fall functions for glottal area or glottal width (Gill, 1962; Hiki et al., 1970; Childers et al., 1986; Zemlin, 1968, pp.177-210; Hirano et al., 1981). Van den Berg et al. (1957) showed with plaster casts that volume flowrate of air through the glottis is proportional to glottal area $A_g$ when the glottis is large but proportional to $A_g^3$ when it is small, so that the leading and trailing edges of the airflow waveform
are steeper than those of the area waveform. Flanagan (1958) estimated magnitudes and shapes of flow waveshapes corresponding to observed glottal area functions and the corresponding frequency spectra.

Optical glottograph traces have been shown to match glottal width although they can be misleading for some placements of the light source and sensor; the traces are either symmetrically rising and falling or skewed with a shorter opening and a longer closing time (Baer et al., 1983). A laryngograph trace shows a sharp rise just as the glottal width falls to zero. This signal is believed to indicate vocal fold contact area, reducing gradually as the vocal folds part but increasing suddenly as they come together (Childers et al., 1986).

The amount of skewing to the right of volume flowrate of air through the glottis may differ systematically across different vowels. Boves and Cranen (1982) found skewing for most vowels but more rising-falling symmetry for [u]. Rothenberg (1981) attributed the skewing of the flow waveform relative to that of glottal area to inertance of air in the glottis, air displaced by the moving vocal folds and vocal tract inertance loading. This last is expected to be more important where the first formant frequency $F_1$ is well above the fundamental frequency $F_0$. The inverse filtered waveforms he obtained did show more skewing to the right for [aː]-type than for other vowels, in agreement with the theoretical prediction.

Apparent discrepancies between different indicators of vocal fold adduction and abduction can be reconciled by considering vertical and longitudinal phase differences in the displacements of the vocal folds (Childers et al., 1986).
II.3.4.5 Source-filter interaction

The skewing of the voice source waveform, the volume flowrate through the glottis, arises partly through voice source-vocal tract filter acoustic interaction, as discussed above. An additional effect of such interaction may be seen as a ripple superimposed on the smooth curves sketched in Figures 12 and 13. This is believed to be due to air pressure variations just above the glottis, primarily at the frequency of the first formant.

In considering how best to represent the voice source in acoustic models, it is important to take into account the changing conditions over the small time span of each voicing cycle. The major excitation of the vocal tract and subglottal resonances takes place where the time derivative has its largest magnitude, generally at or just before the vocal folds become closed. While the vocal folds remain closed, and especially if any continuous glottal chink is insignificant, there is essentially complete separation of the supraglottal and subglottal regions of the respiratory tract and least acoustic energy is lost from the vocal tract; the resonances are lightly damped. As the vocal folds part, acoustic coupling between the vocal tract, the glottis and the subglottal system increases. It is here that the loading of F1 on the glottal flow shows as a ripple, in the rising portion of the flow curve. Bandwidths of vocal tract resonances increase here as energy is lost through the glottis and is partially absorbed in the subglottal airways. Vocal tract formant frequencies rise because the acoustic inerterance associated with the plug of air in the glottis decreases. Probably most of the energy from one major excitation is dissipated before the next one, although there can be truncation of formant time oscillations decay here, especially for F1 in back vowels where
acoustic interaction between source and vocal tract filter is greatest (Fant, 1986).

In Guérin's modelling acoustic coupling between the voice source and the vocal tract filter is included as a circuit with subglottal pressure providing the driving power. There is a specified glottal area time function and loading by F1 and F2. The actual glottal volume flowrate obtained with this circuit contains a ripple in the rising portion of the curve (Guérin et al., 1976). The naturalness of syntheses with different amounts of voice source-filter interaction specified as different amounts of ripple in this glottal flow model was assessed by Childers et al. (1985). They found that it was possible to introduce too much interaction for acceptability as well as too little.

Besides acoustic interactions, a change of filter shape can affect the mode of vibration of the vocal folds through aerodynamic interactions. If oral air pressure rises because the vocal tract is severely constricted or closed at some point, the driving force within the glottis decreases since it is close to the pressure drop across the glottis for part of the voicing cycle (Titze, 1986), as discussed in Section II.3.4.2 above. It is sometimes suggested that aerodynamic forces associated with changes of oral air pressure of this kind could be a significant factor in the control of average glottal area, which is related to the degree of adduction of the vocal folds. At a plosive release, for example, the fall in oral pressure might be responsible for a rapid adduction of the vocal folds; raised oral pressure could assist the abduction of the vocal folds. However, in one example of relevant traces maximum glottal area does not coincide with where oral pressure would be highest, just before the release of a Danish [b] (Frøkjæer-Jensen et al., 1971); and it is assumed in the present study that the slowly
changing component of glottal area is a purely articulatory parameter, controlled entirely by voluntary muscle forces, to a first order of approximation.

II.3.4.6 Flow in the closed phase

Besides the d.c. flow arising from a continuous glottal chink, there may be acoustically significant air movements during the closed phase of the voicing cycle. Effects seen on high speed cinefilms of the larynx include vertical displacement of the vocal folds and a wave motion travelling out across their upper surface. Ripples in the closed portion of glottal flow waveforms obtained by inverse filtering gave significant changes to the frequency spectrum above 500 Hz (Holmes, 1976). When lateral and vertical displacements of the vocal folds were included in 2-mass model simulations small changes in the glottal flow waveshape, including ripples in the rising flow portion, were produced. The effects were judged to be of negligible importance perceptually, but it was considered that the additional computational complexity involved might be justified when modelling pathological states of the larynx (Flanagan and Ishizaka, 1978).

II.3.4.7 Choice of a voice source model

1) The closest approach to the real production system is achieved when a simulation of the structures of the larynx is made to self-oscillate under an applied subglottal or lung air pressure (Ishizaka and Flanagan, 1972); there is no separately identifiable acoustic source in this case.

2) Another approach is to specify the quasi-periodic glottal area waveform $A_g$ as the voice source (for example Guérin et al., 1976).
Disadvantages include a complicated nonlinear time-variable filter function and difficulty in obtaining realistic area waveforms.

3) The true glottal flow may be specified as the voice source combined with a simple filter function which does not change during the voicing cycle. Difficulties here include the calculation of the flow waveform, which depends on the configuration of the vocal tract and subglottal acoustic filters.

4) An approximate model proposed by Fant (1979a, 1979b, 1980a 1981) specifies the glottal volume flowrate waveform without the ripples of F1-source interaction but suitably skewed in shape. Acoustic losses through the glottis to the subglottal region need to be incorporated as time-varying within the voicing cycle. In this way the bandwidth effects of the ripple in a true glottal flow waveform rather than the ripple itself are modelled. Fant explains that this is the acoustic equivalent of a "constant current source to the vocal tract supraglottal system in parallel with the time-varying glottal-plus-subglottal impedance"; when the glottis is large, during the open phase, airflow is diverted to the subglottal filter: "...the shunt impedance draws current which is the negative of the ripple component in the vocal tract input flow." (Fant, 1985, p.19). The glottal flow waveshape must be made to vary appropriately with phonetic context, so as to include some of the effects of source-filter interaction (Fant, 1980a; Fujisaki and Ljungqvist, 1986).

5) In a purely acoustic domain approach, a smooth voice source waveform, of constant shape and without interaction ripples such as that commonly used in terminal analog formant synthesisers, could be the starting point. Here representation of the physical processes of speech production could be improved by varying the skew of the wave and the formant bandwidths appropriately as phonetic context changed.
The varying acoustic loss through the glottis over a voicing cycle could be approximated by two values of glottal loss coefficients, defined for the closed phase and the open phase (Brookes and Naylor, 1988). For the present study the fourth approach was adopted. Data from natural speech will be reviewed below related to a suitable skewed waveform of volume flowrate of air through the glottis, without ripples either in the rising portion or in the closed phase.

II.3.4.8 The range of voice waveshapes

The glottal volume flowrate waveforms of natural speech exhibit a range of pattern features which depend upon the controlling conditions. Most of the data are provided by inverse filtering, mainly from the speaker's total output volume flowrate at the mouth and nose combined. There have been rather few investigations in this field, but from the pioneering research of Lindqvist (1970) and that of others after him a fairly clear picture emerges of how the glottal flow waveshape changes as controlling conditions change. Other important kinds of evidence come from measurements on excised larynges (van den Berg and Tan, 1959), simulation of the physical processes with the 2-mass model of the vocal folds (Ishizaka and Flanagan, 1972; Monsen et al., 1978), and perception experiments (Rosenberg, 1971).

Waveshape trends are discussed in more detail, with reference to the particular parametric model shown in Figure 13, in Sections III.1.4.1 and III.2.4.1. The sketches of Figure 12 are intended to summarise the general trends indicated by the literature. Normal, or modal and mid voice (see Section II.3.4.3) is given a central position. At the ends of the various dimensions portrayed in Figure 12 voicing becomes acoustically or auditorily less successful; voicing may weaken and cease for a combination of reasons. Very low pressure drop across
the glottis may be associated with low respiratory force or with raised oral pressure or both. The domain of voicing may be limited by physiological constraints such as the speaker's maximum possible respiratory force; there may be a threshold such as the minimum pressure drop across the glottis or minimum volume flowrate through the glottis below which voicing ceases. Trained singers may be able to maintain the quasi-periodic component of flow acoustically and auditorily dominant even at very high flowrates. If the vocal folds become too much abducted or, on the contrary, too tightly pressed together, the voicing changes its waveshape characteristics and eventually ceases.

It seems that there is kind of plateau range of vocal fold adduction and glottal area within which acoustically strong voicing with the voice wave amplitude almost independent of glottal area is possible, given appropriate aerodynamic conditions at the glottis. At the edges of this good region the vocal fold vibrations gradually alter and finally cease altogether. At one end of the range, as the vocal folds are parted (abducted) so that glottal area increases, voicing becomes breathy in type, with less excitation of high frequencies, and it is accompanied by aspiration noise; towards the other extreme of the good region, as the vocal folds are strongly approximated (adducted), voicing becomes glottalised in type, with more excitation of higher frequencies (Stevens, 1977). With extreme adduction, as for a glottal stop, the fundamental frequency of voicing falls extremely low and then voice ceases.

The voice waveshape alters for voice effort changes (Lindqvist, 1970; Fant, 1979a; Sundberg and Gauffin, 1979; Karlsson, 1986; Holmberg et al., 1988). It alters with changes of fundamental frequency $F_0$, register and larynx height in singers and speakers (Sundberg and Gauffin, 1979), with $F_0$ and subglottal pressure (Lindqvist, 1970;
Rothenberg, 1973), for different degrees of vocal tract constriction (Rothenberg and Mahshie, 1986; Bickley and Stevens, 1986), for different vowel types (Rothenberg, 1981; Boves and Cranen, 1982), for different phonation types or degrees of vocal fold adduction (Rothenberg, 1972, 1973; Sundberg and Gauffin, 1978, 1979; Gauffin and Sundberg, 1980; Fant, 1980a; Bickley, 1982), and for different speaker types (Monsen and Engebretson, 1977; Karlsson, 1988). Voice waveshapes generated by the 2-mass model of the vocal folds alter in shape when the values of the control parameters are changed (Flanagan et al., 1975; Guerin and Degryse, 1977; Monsen et al., 1978). It seems likely that the naturalness of a terminal analog synthesiser will be improved if its voice waveshapes are made to vary as in natural speech (Brookes et al., 1988).

II.3.4.9 Control of sound pressure level

Van den Berg and Tan (1959) found that each larynx "sang after its own fashion" but with common trends. When the applied lateral forces on the vocal folds were small and conditions probably close to conditions in a speaker's larynx, with a pressure drop of only about 1 cmH$_2$O needed for voicing to occur, sound pressure $p_s$ increased with pressure drop across the glottis $\text{PDIFF}$ to the power 1.5. Several other studies are in good agreement with a relationship of the form

$$p_s \propto \text{PDIFF}^{1.5} \quad (\text{or} \quad \text{SPL} \propto \text{PDIFF}^{3.0} ) \quad (\text{II.19})$$

with exponents in the expression for sound pressure of 1.6 (Ladefoged and McKinney, 1963), 1.7 for low and mid pitch ranges (Isshiki, 1964), 1.5 (Bouhuys et al., 1968), between 0.7 and 1.8 for six men and between 1.3 and 2.0 for one of them depending on $F_0$ (Cavagna and Margaria, 1968). Exponent values are doubled for sound pressure level (SPL) or intensity level (IL).
A minimum subglottal pressure is needed to make the vocal folds vibrate. This threshold seems to be higher for more adducted vocal folds in excised larynges, perhaps as high as 10 cmH₂O with high lateral forces, but about 2 cmH₂O with low lateral forces (van den Berg and Tan, 1959). Lindqvist (1972) found that the vocal folds continued to vibrate down to about 1 cmH₂O at low and normal F₀, but voicing stopped at a higher air pressure when F₀ was high. It may be necessary to estimate minimum flowrates for voicing also. In van den Berg and Tan's (1959) data a minimum flow of 30 to 50 cm³/s seems to be suggested; for the six men analysed by Cavagna and Margaria (1968) the minimum flowrate for voicing at F₀ of 130 or 260 Hz, near the speech range, seems to lie between about 60 and 70 cm³/s.

The maximum amplitude VOIA of the voice waveshape is the main determiner of the amplitude of the voice source spectrum fundamental or first harmonic and this, together with the maximum negative slope of glottal flow, \( \frac{dU_g}{dt} \), is a good predictor of the SPL of the radiated sound pressure wave for sung vowels (Sundberg and Gauffin, 1979). Raised subglottal air pressure probably increases the suddenness of vocal fold closure and so of the cut-off of glottal airflow during voicing, by a localised impact when the glottis is very narrow. At the same time, higher air pressure drives the vocal folds into larger amplitude of vibration and so raises F₀ as well as increasing peak flowrate (Titze, 1980), as discussed in Section II.3.4.10. Thus SPL and F₀ often increase or decrease together during speech (Hiki, 1967).

When a singer wishes to increase loudness on a constant note, cricothyroid muscle activity may need to be reduced in order to counteract the rise in F₀ due to the increased air pressure needed for greater acoustic intensity. This effect is seen in a study by Yanagihara and von Leden (1966), but only at high F₀. Their data at
low $F_0$ showed cricothyroid activity actually increasing with increased subglottal pressure and SPL. It seems that there may be two distinct mechanisms for raising subglottal air pressure: firstly by increasing respiratory muscle forces, in which case glottal flow and glottal area increase; or secondly by adducting the vocal folds more strongly, in this case reducing glottal area. If an increase in glottal area gives a higher natural frequency of vibration to the vocal folds (Titze, 1980), then the first mechanism increases vocal fold stiffness while the second decreases it; cricothyroid $F_0$ control would need to be modified in opposite ways for a crescendo in the two cases, as seen in the data of Yanagihara and von Leden (1966).

Tests with the model used in the present study showed that it gave the appropriate relationship between subglottal pressure or pressure drop $PDIFF$ and radiated sound pressure level. Loudness in simulated singing was controlled using each of the two mechanisms and this gave a realistic auditory crescendo effect, as long as subglottal pressure was raised at the same time as glottal area was decreased when simulating the second mechanism (Scully and Allwood, 1985b).

II.3.4.10 Control of fundamental frequency

When interpreting the changes of voice waveform at different fundamental frequencies or pitch levels, it is important to note that fundamental frequency is a dependent variable. Two components of $F_0$ control may be identified. The major effect is associated with laryngeal muscle activity, but there is an aerodynamic effect also: $F_0$ is sensitive to changes in subglottal pressure or to pressure drop across the glottis. These two expressions of air pressure will be used interchangeably in this section. Published works usually refer to subglottal pressure, but if the vocal tract is in a vowel-like
state so that supraglottal pressure is near atmospheric then the pressure drop across the glottis is very close to subglottal pressure.

Hollien and his co-workers found that thicker vocal folds have a lower overall F₀ range. Within one speaker, as F₀ is raised the vocal folds lengthen and become thinner and less massive; vocal fold thickness decreases as F₀ rises across a range of singing voice types and registers, although changes in F₀ in the falsetto register do not seem to be accompanied by thickness changes and aerodynamic forces seem to dominate here (Hollien, 1974).

Rises and falls of F₀ in the medium and high frequency range appear visually similar to patterns of cricothyroid muscle activity; low falling F₀ may perhaps correlate with activity of the laryngeal depressor muscles (Collier, 1975). Van den Berg and Tan (1959) were able to raise F₀ in excised human larynges by stretching the vocal folds, simulating cricothyroid but not vocalis muscle activity. Increased length is apparently more than compensated for by increased stiffness and a reduced effective vibrating mass. F₀ rose also with increased air pressure below the glottis and with increased volume flowrate of air through it.

It is not easy to explain the dependency of F₀ upon the aerodynamic parameters of pressure and flow, although the reality of a pressure-dependent effect can hardly be doubted. The evidence cited in Section II.3.3.7 showed subglottal pressure remaining relatively constant during [VCV] sequences while oral pressure rose. Highly significant positive correlations were found between reductions in F₀ and rises in oral pressure for [VCV] sequences produced by one speaker (Di Cristo and Teston, 1979); a falling F₀ is commonly seen
directly proportional to the square root of the vibrating mass and to the small distance the vocal folds had to move apart before the mean pressure in the glottis became negative. On this theoretical basis a lower pressure drop across the glottis, as during a voiced fricative for example, reduces the aerodynamic restoring force, the Bernoulli effect, so that the vocal folds move further apart, thus increasing the cycle time.

When the elasticity of the vocal folds is taken into account there is a different explanation of the effect of pressure on \( F_0 \), consistent with the observations of van den Berg and Tan (1959) that as subglottal pressure, flow and \( F_0 \) rise the amplitude of vocal fold vibration increases. The argument developed by Titze (1980) is based on the nonlinear stress-strain curve of the tissues of the vocal folds and the narrow range of glottal widths over which the aerodynamic forces on the vocal folds are significant. He argues that, at a subglottal pressure of 10 cmH\(_2\)O for example, the impulse-like Bernoulli force operates for glottal widths of 0.01 to 0.08 cm whereas the maximum glottal width is about 0.3 cm. Between the application of these impulse forces the vocal folds oscillate at their natural frequency.

A higher pressure drop across the glottis increases the impulse forces so that the vocal folds vibrate with a larger amplitude. As the vocal folds vibrate they are bowed and lengthen and become stiffer; this is the nature of the nonlinear stress-strain curve for organic tissues in general and the tissues of the vocal folds in particular. The greater the deformation of the vocal folds during the free oscillating portion of the voicing cycle the greater is the effective stiffness and the higher is the fundamental frequency of the oscillations. From his continued theoretical analysis and measurements Titze (1989) predicts values of 0.5 to 6 Hz/cmH\(_2\)O for
on laboratory traces as oral pressure rises. A reasonable inference is that \( F_0 \) falls as pressure drop across the glottis decreases.

The effect has been quantified: Lieberman's (1967) study indicated as much as 20 Hz rise for 1 cm\( \text{H}_2\text{O} \) increase in subglottal pressure or transglottal pressure drop, but most investigators have found values under 10 Hz/cm\( \text{H}_2\text{O} \) (for example, Öhman and Lindqvist, 1965; Ladegoged, 1963; Netsell, 1969; Lieberman et al., 1969; Ohala and Ladefoged, 1970; Hixon et al., 1971). The data of Öhman and Lindqvist (1965) suggested a constant decrease in period of voicing; the value was 0.16 ms for 1 cm\( \text{H}_2\text{O} \) increase in pressure drop across the glottis. The values vary across speakers and may vary with \( F_0 \), intensity level and register but probably not with vowel type (Ohala and Ladefoged, 1970; Hixon et al., 1971). Reviewing the findings from 1957 onwards, Titze (1989) finds that the sensitivity of \( F_0 \) to subglottal pressure seems to be least in the high-pitched chest register (1 to 3 Hz/cm\( \text{H}_2\text{O} \)), greater in the low-pitched chest register (2 to 6 Hz/cm\( \text{H}_2\text{O} \)) and greatest in the falsetto register (5 to 10 Hz/cm\( \text{H}_2\text{O} \)). A typical value for men speakers might be 4 to 5 Hz/cm\( \text{H}_2\text{O} \). Using a value of 5 Hz/cm\( \text{H}_2\text{O} \) and an expression of the form

\[
F_0 = Q + 5(\Delta P_{gl} - \Delta P_{gl,\text{ref}})
\]

where \( \Delta P_{gl} \) is the pressure drop across the glottis and \( \Delta P_{gl,\text{ref}} \) is a reference value of this pressure drop, Ladefoged (1963) obtained a laryngeal control function \( Q \) which had a smoother contour than that of \( F_0 \) and did not always run in the same direction as \( F_0 \).

From the van den Berg et al. (1957) measurements of air pressure within the glottis in static models, Fant (1960, 1970, p.266-267) concluded that, if vocal fold elasticity were to be neglected, the time for one cycle of vocal fold oscillation would be inversely proportional to the square root of the subglottal pressure, and
the dependence of $F_0$ on subglottal pressure changes, consistent with the data from natural speech; the largest values are likely to be associated with short, slack, low stiffness vocal folds. In natural speech this condition corresponds to the low pitch end of the chest or modal register. Behaviour in the falsetto register, where a different mathematical treatment seems to be required, remains to be studied and explained.

The 2-mass model of the vocal folds incorporates nonlinear stress-strain curves for the tissues and exhibits increase of $F_0$ with increased subglottal pressure, in contradiction to the small-amplitude theory of Ishizaka and Matsudaira (1972). Simulations with the 2-mass model suggest that the sensitivity of $F_0$ to transglottal pressure drop may increase noticeably for pressure values below about 5 cmH$_2$O (Ishizaka and Flanagan, 1972). This could be particularly relevant for voiced fricatives of RP English for example, in which the pressure drop across the glottis does appear to fall below this value and where falls in $F_0$ are strikingly clear on laboratory traces.

Whether there are specific $F_0$-raising mechanisms which apply to the release phase of some voiceless consonants is unresolved. Here high $F_0$ seems to be associated with voicing onset occurring at very high airflow rates. The phenomena observed in speech and possible mechanisms and explanations have been discussed by Silverman (1987). For the larynx described by van den Berg and Tan (1959) $F_0$ rose as volume flowrate increased, but the effect was always associated with increased subglottal pressure. At a constant subglottal pressure, of 10 cmH$_2$O for example, moving from larger to smaller lateral forces on the vocal folds, $F_0$ fell while airflow rate increased greatly due to increased mean glottal area. It seems reasonable to suppose that high flowrate does not of itself cause an increase in the frequency
of oscillation of the vocal folds. It may be the case that voicing at large values of mean glottal area implies increased mean stiffness of the vocal folds over the free-oscillating portion of the cycle because the bowed, stretched vocal folds are stiffer overall than when the vocal folds are more adducted. The argument advanced by Titze (1980, 1989) might explain the effects observed. It has been suggested that voluntary increasing of vocal fold stiffness might be used as a way of assisting the cessation of voicing (Stevens, 1977), an additional but not necessarily a conflicting explanation.

It is apparent that different mechanisms may need to be invoked in different portions of the voicing cycle and for different settings of the vocal folds to explain and quantify the aerodynamic components of fundamental frequency control. This is still an area of disagreement and controversy (Stevens and Hirano, 1981, pp.265-270).

**II.3.4.11 Jitter and shimmer**

Besides the slowly changing laryngeally controlled component of $F_0$ and the relatively localised perturbations to the intonation contour associated with the aerodynamic effects discussed above, there is a certain amount of cycle-to-cycle variation or jitter in $F_0$, even in normal speakers without pathological conditions of the larynx. Distributions of $F_0$ obtained from laryngograph signals, displayed as histograms for individual cycles or pairs or trios of cycles or as scattergrams taking successive cycles in pairs, show a variety of patterns. These may be interpreted in terms of $F_0$ range and regularity of vocal fold vibration. The displays and their analysis are used in the assessment of treatment as well as for the characterisation of normal and disordered speech (Fourcin, 1986).
Some degree of jitter, period-to-period $F_0$ variation and shimmer, amplitude variation, contributes to naturalness in synthetic speech it seems. Sustained vowels were synthesised with approximately normal distribution for period and amplitude of the voice pulse source in a cascade-connected formant synthesiser. Jitter and shimmer were measured by standard deviation expressed as a percentage of the mean voicing period or amplitude. 1% jitter and no shimmer gave the best ratings of acceptability for [i:]; 2% jitter with no shimmer or with about 5% shimmer was most acceptable for [æ:] and [u:] respectively (Rozsypal and Millar, 1979).

Jitter has been measured in natural speech, both normal and pathological, by, for example Lieberman (1963). Ten parameters to characterise $F_0$ range and jitter and shimmer perturbations have been assessed for their ability to discriminate between normal states and some pathological states of the larynx. The best indicator seems to be a directional perturbation factor for shimmer with a 3% threshold (Laver et al., 1986).

**II.3.4.12 Turbulence noise**

Only a few theoretical or experimental studies have considered the physical processes responsible for the generation of turbulence noise in speech. It is probable that relevant research has been done in other fields such as aeronautics, but this literature was too extensive to be investigated for the present study. It is, in any case, questionable whether flow results can be mapped onto systems of different size.

Turbulence noise arises where a moving air stream has a high enough particle velocity for eddies to form, creating irregular air movement and air pressure changes at acoustic frequencies. In speech a high
velocity jet is formed as air emerges from a very narrow constriction of the respiratory tract; the jet gradually mixes with the air around it. The assumption that a turbulence noise sound source is separable from the filter needs to be justified, since the formation of the sound involves the geometry of the main resonator, the cavity downstream from the source constriction.

Shadle has demonstrated by means of mechanical models and theoretical analysis, including measures of coherence for spectra of signals at probes inside the vocal tract models and at a microphone some distance from the model's outlet, that a specific source location can be found. She states: "A pressure source located within the front cavity will excite all poles of the entire tract, and in addition a real zero near 0 Hz and a complex conjugate pair of zeros at a frequency related to the distance between the pressure source and the anterior end of the constriction. Since the distance is so short, a change in it of the order of a millimeter will change the frequency of the free zero by a few hundred Hertz." (Shadle, 1989).

Shadle identified more than one sound source mechanism: either the jet hits an abrupt obstacle, the teeth, or it hits a relatively rigid wall at an oblique angle. In the former case, which applies in fricatives [s] and [ʃ], the sound source appears to be localised at the teeth; in the latter case, applicable to other fricatives with their constrictions further back in the oral cavity, the source appears to be distributed and formed just downstream of the constriction. The acoustically weak front fricatives [ɸ, f, θ] may perhaps be cases of no obstacles for the jet to strike. At flow velocities below the speed of sound the efficiency of conversion of flow to turbulence noise sound is less efficient in a free jet and more efficient at rigid boundaries (Shadle, 1986a, 1986b, 1989).
The degree of turbulence increases with the Reynolds number $Re$ which is defined as

$$Re = \frac{\rho v l}{\mu}$$  \hspace{1cm} (II.21)

where $\rho$ is the density of air 0.00113 gm/cm$^3$, $v$ is the velocity in cm/s, here taken to be the value at the centre of the constriction outlet, $\mu$ is the dynamic coefficient of viscosity for air 0.000184 gm/cm.s and $l$ is a length in cm related to the smallest dimension of the constriction. The critical value of $Re$ depends on the geometry, including the roughness of any walls, but for a jet emerging from a circular cross-section hole, laminar flow is likely to become unstable and become fully turbulent for $Re$ between 160 and 1200 (Goldstein, 1976, cited by Shadle, 1989). (See also Equation (II.11) in Section II.3.3.4.)

Processes responsible for the turbulence noise sources of speech were analysed by Fant (1960, 1970, pp.272-280). Stevens included in his analysis both theory and empirical data. Relationships between aspiration noise generated at the glottis and frication noise generated at a supraglottal constriction were derived. Appropriate regions of glottal area and vocal tract constriction area were established and an expression for the radiated sound pressure of turbulence noise $p_T$ was obtained as:

$$p_T \propto \Delta P^{b/A^{0.5}}$$  \hspace{1cm} (II.22)

where $\Delta P$ is the pressure drop across the constriction at which the noise source develops, $A$ is its minimum cross-section area, and the exponent $b$ is between 1.0 and 1.5. The modulation of noise sources when voicing gives a periodically interrupted airstream was considered also (Stevens, 1971).
The spectral shape of a turbulence noise source is expected to be related to conditions for the jet. There is a broad spectral peak at a frequency $f_n$ Hz, given by:

$$f_n = \frac{Sv}{d}$$  \hspace{1cm} (II.23)

where $v$ cm/s is the linear velocity in the centre of the jet as it leaves the constriction, $d$ cm is the jet diameter and $S$ is the Strouhal number, equal to 0.15 (Shadle, 1989). These values are bound to be constantly changing during the production of a fricative, plosive or affricate. The centre frequency $f_n$ was predicted by Stevens (1971) to range from about 500 Hz to 3 kHz for noise sources of speech. These predictions were supported by experiments on natural speech in which a reflectionless tube method was used to derive estimates for aspiration noise sources. The spectra were the same for the women and the men speakers analysed (Hillman et al., 1983).

It is difficult to separate aspiration noise from frication noise on the one hand and from voicing on the other in natural speech, since the sources overlap in many cases, for example after the release of a voiceless aspirated plosive and within the frication noise segment of a voiceless fricative, as sketched in Figure 3. Rothenberg used high-pass filtering at a cut-off of 5 Hz to remove most voicing energy and so indicate the time domain of co-occurring aspiration noise. Combined with inverse filtering to give the voice source waveform, the existence of a modulated noise component was demonstrated. It was shown to be strongest for glottal areas corresponding to volume flowrates of 200 to 300 cm$^3$/s (Rothenberg, 1974). These values are considerably lower than the range from about 500 to 1500 cm$^3$/s derived by Stevens for aspiration noise although
they agree quite well with Stevens' values between 200 and 500 cm³/s for frication noise (Stevens, 1971).

II.3.4.13 Transient

This source arises from sudden changes of air pressure. It is important in clicks, where air pressure in an enclosed cavity is reduced by volume enlargement and then connected to a region of higher pressure. It contributes to plosive bursts also. Here air pressure inside the vocal tract is raised during the plosive closure and falls when the closure is released with pressure equalisation achieved by airflow through the opening constriction. The discharge of the air from the enclosed volume has been analysed by Stevens (1956), as discussed in Section II.5.3.

What appears to be a transient as well as noise may often be seen on spectrograms of natural speech as a plosive closure phase begins, as well as at the release. Oral pressure can build up rapidly and there seems no reason to arbitrarily confine transients to plosives and their release phases.

Maeda's analysis of acoustic conditions in and very close to a constriction rapidly increasing in cross-section area showed that a coherent acoustic source or transient could arise from airflow due to localised conditions very close to the constriction and not simply from the overall discharge of pressure from the vocal tract cavity. A rapid release would give a strong transient and abrupt frication noise of short duration; a slower release would give a weak transient and gradually increasing frication noise extending over a longer time. Glottal area at the release influenced frication noise but not the release transient (Maeda, 1987).
II.3.5 Filtering

Different ways of handling the concept of source and filter are discussed in Sections II.3.4.5 and II.3.4.7, with particular reference to the voice source. Here, the various ways of representing the filters or resonators will be considered briefly, especially the main filter, the vocal tract: all the air-filled tubes and cavities between the glottis and the mouth and nose outlets. The transfer function comprises the poles and zeros of the air, but the wall properties contribute to losses and tuning of resonances as discussed in Section II.2; some sound is radiated from the walls also. There are several conflicting filter models, each of which is applicable over a limited range of geometries and frequencies. During speech production the vocal tract is continuously changing shape, so that it cannot be expected that any one model will be valid over all.

For the aerodynamic processes of speech production Boyle's law $PV=\text{constant}$ can be applied to a given mass of gas, since it can be argued that at low frequencies an approximation to isothermal conditions might obtain, but at higher frequencies this is certainly not the case. In the acoustic range of frequencies conditions are assumed to be adiabatic and the appropriate gas law is $PV^\gamma=\text{constant}$, where $\gamma$ is the ratio of specific heats at constant pressure and at constant volume $c_p/c_v$, with a value of 1.4 (Flanagan, 1972a, p.30). Disturbances travel with the velocity of sound.

Lumped elements must be small enough for their dimensions to be much less than the wavelength; this sets an upper limit for frequencies. A Helmholtz resonator applies if the vocal tract can be approximated by one or two cavities with narrow outlets; generally the wavelength limitation means that this model is valid only for the first
resonance, or F₁ in the speech signal. Otherwise, the vocal tract needs to be modelled as a tube along which acoustic waves travel. The wave travel may be considered to be one-dimensional along the curve of the vocal tract tube as long as the largest dimension across this tube is well under half a wavelength.

There are considerable uncertainties about the value to be given to the velocity of sound inside the vocal tract, where temperature and humidity are changing along the length and where the walls yield under pressure. According to Regnault, 1868, cited by Richardson, 1940, p.27), the effective value of the velocity of sound inside a pipe falls as the pipe diameter decreases. On the other hand, yielding walls can be looked upon as increasing the effective velocity of sound inside the pipe (Flanagan, 1972a, p.67) or as tuning the resonances to higher frequencies (Fant et al., 1976), see Section II.2 and Equation (II.8). A value of 350 m/s, little affected by flow profile or cross-section shape, but dependent upon gas composition, temperature and mean flow speed, has been suggested by Rice (1980). Wavelengths are 350 cm at 100 Hz, 7 cm at 5 kHz and 3.5 cm at 10 kHz. The vocal tract, about 14 to 18 cm in length in adults, lies within the wavelengths of interest for speech. If the maximum cross-dimension is taken to be 4 cm, then half-wave cross-modes occur above 4 kHz and the assumption of one-dimensional wave propagation is valid up to about 4 kHz.

For one-dimensional wave propagation the vocal tract tube may be treated by the Webster horn equation for loss-free conditions, as long as changes in cross-section area are so gradual that internal reflections can be ignored (see Flanagan, 1972a, p.85). Where there are sudden changes of area, reflections must be included. The vocal tract acoustic tube can be represented as an electrical transmission line made up of a number of sections. The maximum dimension of each
lumped element, or its electrical equivalent, must be less than one eighth of a wavelength. For sections 0.5 cm long the errors will be 5% at 6 kHz and less at lower frequencies (Stevens et al., 1953). Thus, with this length of section the upper frequency of the model of filtering is limited by the cross-dimensions of the vocal tract and not by the quantisation along its length.

Losses of various kinds - viscosity, heat conduction, losses at severe constrictions, transmission of sound through the walls and radiation - need to be included in models. There have been many recent developments in the treatment of vocal tract acoustics; it is beyond the scope of this work to consider these in detail. Some of the problems of representation and implementation of losses and other geometric complexities in the reflected pressure wave model have been tackled by Liljencrants (1985).

The subglottal airways participate in the processes of filtering and it is possible to include these, but with the geometry very much simplified, in models. Traces of subglottal pressure can show significant oscillations, suggesting a lowest subglottal resonance near about 500 Hz (for example Koike, 1981).

II.3.6 Radiation

The radiation termination of the vocal tract filter and the transform from an acoustic component of volume flowrate of air out of the mouth and nose to sound pressure as received at a microphone some distance away from the speaker is treated in, for example Fant (1960, 1970) and Flanagan (1972a). An important development in the simulation of the acoustic processes of speech production is the use of the $z$-transform to enable frequency-dependent effects to be included in
time-domain models. Lip radiation impedance effects have been treated in this way by Laine (1982).
II.4 Mapping across stages: stability and sensitivity

Differences of acoustic structure such as are necessary for the transmission of contrasts of meaning under specific communication conditions must be preserved in spite of the kinds of articulatory variability discussed in Appendix 3 (Scully, 1987). It has been suggested that speech preferentially exploits those articulatory patterns and regions of acoustic space for which acoustic pattern features are particularly insensitive to articulatory perturbations (Stevens, 1972; Stevens and others, 1989). This quantal theory must take into account the fact that articulatory control is multidimensional. An important acoustic pattern feature might be stable with respect to some articulatory perturbations but sensitive to others. It will be argued in Section IV.1.3 that sensitivity to one or two articulatory parameters is just as important as stability with respect to others.

Stability considerations apply across all stages of speech production. Non-linearities between muscle forces and tongue shapes were demonstrated for [iː]-type vowels by means of a 3-dimensional finite-element model of the tongue: constrained by rigid boundaries as in the vocal tract, this tongue model exhibited saturation effects, suggesting that for this vowel at least the vocal tract area function may be stable with respect to muscle force perturbations (Fujimura and Kakita, 1979). Similar saturation effects might be expected to apply to some other vowels also.

Stevens (1972) proposed that [iː], [ɑː] and [uː] should be considered to be examples of highly stable, quantal vowels because their formant frequencies are insensitive to small perturbations of constriction position, located as they are at a minimum or maximum of graphs of
formant frequency versus distance from the glottis of the
constriction in Fant's (1960, 1970) nomograms, as may be seen in
Figure 2.

It has been argued that the arrangement of the extrinsic muscles of
the tongue and pharyngeal constrictor muscles is particularly
suitable for the formation of a constriction at the four places along
the vocal tract which give formant stability. These places seem to
fit area function data for natural vowels, discussed in Section
II.3.2.1 (Wood, 1979). The X-ray microbeam data of Perkell and
Nelson (1985) for [i:] and [a:] vowels of American English lend
support to the view that position scatter for points on the tongue
surface near the major vocal tract constriction is lower in the
direction across the vocal tract tube than in the direction along its
length. This would be consistent with the need to control closely
the constriction size for acoustic reasons, while allowing some
degree of variability in its place, to which the formant frequencies
are relatively insensitive. The data might instead be interpreted as
reflecting the muscle force - position saturation effects described
above. As is common in theoretical considerations of costs and
benefits in speech production, an observed mode of behaviour in
natural speech can be hypothesised to be chosen for muscular or
acoustic reasons, and sometimes for aerodynamic reasons also; it is
difficult to judge what weight should be given to one or another
explanation.

Many different vocal tract area functions can result in the same set
of formant frequencies (Atal et al., 1978). Changes of formant
frequencies in response to realistic changes in the parameters of an
articulatory model have been investigated by, for example, Lindblom
and Sundberg (1971), Ladefoged and Harshman (1979) and Abry et al.
(1989). Plateau regions of formant frequency stability with respect
to small and large changes in vocal tract shape have been investigated, for example with Maeda's (1979) articulatory model. It was found that $F_1 - F_2$ stability needed cooperation between more than one of the command parameters: lips, jaw, tongue body, dorsum and tip. Macro-sensitivity functions for [iː]-type vowels, for example, showed large formant frequency changes if only one parameter such as tongue body or jaw was varied. Abry et al. (1989) suggest that this kind of multi-articulator gesture enables trade-offs and compensations that can account for the constancy of shape achieved for the crucial region of the vocal tract during experiments in which a speaker is prevented from raising the jaw in the normal way by a bite-block between the upper and lower teeth.

The speakers investigated did not use completely different articulatory configurations to achieve acoustically equivalent outputs under bite-block and normal conditions. With the bite-block they achieved almost the normal area function, especially in the region of the vocal tract constriction. Large shifts in tongue and lip position relative to the jaw were used to achieve this area function constancy; most dramatically for [iː] and [uː] (Gay et al., 1981).

An additional kind of stability is proposed by Stevens (1989): for the mapping from acoustic pattern to perceptual quality. Supporting evidence is provided by Abry et al. (1989) for [iː]-type vowels with vocal tract shapes near where cavity affiliation switches (discussed in Section I.3). They found that, in the vicinity of this exchange point or focal point, the two close $F_2$ and $F_3$ peaks in the 2 to 3 kHz region altered in their relative heights, or formed a single spectral peak. They invoke the perceptual large-scale integration concept of Chistovich and colleagues (for example Chistovich and Sheikin, 1979)
to suggest that the three different acoustic patterns are perceptually equivalent.
II.5 Specific models of speech production

Introduction

Petit Robert defines a model as a simplified representation of a process. The complexities, "les grandeurs réelles", are to be represented by appropriate signals (Rey and Rey-Debove, 1982). The models with which this study is concerned are ones which capture, through mathematical expressions, some of the properties of natural speech production. As Flanagan states in his account of the history of speech synthesis: "Much treachery lies in trying to model nature verbatim, rather than in terms of underlying physical principles." (Flanagan, 1972b, p.1380). Throughout this chapter, reference has been made to many studies which include both analysis of natural speech processes and the representation of these through quantitative expressions. Thus modelling of speech production is linked with data and theory. This section will do no more than list some publications which are representative of different aspects of the modelling of speech production, under the headings of the various stages identified in Section I.4 and Figure 5. Many of the works cited in this section have been discussed in more detail in other sections. Those cited here are given as examples only, to indicate the range of approaches; this is not a comprehensive review of the field.

II.5.1 Neural and muscular models

A model of the flow of information in the central nervous system, as required for the linguistic planning stages of speech production, has been proposed by Laver (1980a), with speech errors providing some of the data against which to test the model.
There have been a few models of the muscle forces: for the larynx with particular reference to the control of $F_0$ (Kakita and Hiki, 1975), and for the tongue (Hiki and Oizumi, 1975; Perkell, 1974; Fujimura and Kakita, 1979). Physiological interactions between tongue configurations and $F_0$ control have been modelled by Honda (1983).

II.5.2 Articulatory models

Most models of the articulatory stage focus on shapes and movement paths for the structures above the larynx, those which control the area function of the vocal tract; changes in lung configuration are usually considered as a component of aerodynamic models and are included in Section II.5.3. The complete area function of the vocal tract is derived and mapped onto formant frequencies or an acoustic waveform. This mapping is discussed under acoustic models, Section II.5.4.

The larynx is included, as well as supraglottal structures, in the articulatory model of Lindblom and Sundberg (1971); here the emphasis is upon the interactions between the solid structures. Other models have a similar focus (for example Mermelstein, 1973; Coker, 1976). About nine parameters seem to be needed to shape the vocal tract if consonants as well as vowels are simulated. For example, Coker's (1976) model contains three parameters to control the tongue body, two to define lip actions, two for raising and curling back the tongue tip, and one for velum actions. The ninth parameter controls the mapping between the side view of the vocal tract and the cross-section area of the tube. Models which focus on the solid structures need to perform this mapping, in the ways discussed in Section II.3.2.3. The lips have been modelled in detail by Abry and Boe (1983), with the emphasis on the mapping between
gestures, the resulting positions of the solid structures and the associated total lip shapes.

Some models treat tongue shapes as the sum of linear components (for example Kiritani, 1977; Nakajima, 1977; Shirai and Honda, 1977; Harshman et al., 1977; Maeda, 1979). As discussed in Section II.3.2.1 there is considerable agreement among them in the articulatory interpretation of the main factors. Maeda (1979) for example has used his statistical analysis as a model of articulatory control, with jaw, tongue-body, dorsal and apical components from which realistic gestures of the whole tongue can be derived. In the modelling of Lindblom and Sundberg (1971) vertical displacement of the larynx is an additional control parameter.

Semi-polar coordinates are often used to describe tongue surface outlines and vocal tract shapes in the midsagittal plane (for example Lindblom and Sundberg, 1971; Coker, 1976; Shirai and Honda, 1977; Maeda, 1979). The tongue body outline is often approximated by a circle with a moving centre (for example Coker, 1976), with deformation from a circle included as a tongue body parameter (Lindblom and Sundberg, 1971). The timing constraints imposed by articulatory transitions are an important feature of some models (for example Coker, 1976; Flanagan et al., 1975).

Other models work directly with the vocal tract area function. Perhaps the simplest representation of vocal tract shape as an area function was Dunn's (1950) transmission line model. In this all the sections except one represented the same cross-section area; one section represented a narrowing or constriction of the tube. The vocal tract tube is represented in a parametric form in the models of Stevens and House (1955, 1956) and Fant (1960, 1970). The three parameters are the size of the minimum cross-section area of the
vocal tract, its position along the tube, and the area/length ratio for the lip outlet. Parabolic or horn expressions specify the total area function. The vocal tract has been represented instead as four tubes (Fant, 1960, 1970; Badin and Boë, 1987) or divided into regions (Mrayati et al., 1988). Here, the emphasis is on the investigation of the mapping from area function to formant pattern; these models need to be classified as acoustic as well as articulatory since they simulate acoustic filtering processes. The problem of mapping from formant frequencies to vocal tract shape has been approached through analysis by synthesis, for example with a six parameter model of vocal tract area function (Charpentier, 1984).

Area function models have been used to simulate the production of fricative sounds (Rosen, 1958; Shirai and Masaki, 1983) and plosives (Maeda, 1987b).

II.5.3 Aerodynamic models

The complex geometry of the subglottal airways was represented as many generations of two-way branchings in Weibel's (1963) dichotomous model, and other more complex models have been proposed (see Bouhuys, 1977, pp.30-34). Changes in cross-section area of the airways during speech production are considered to be a part of articulation as defined in this work. But modelling of the branching tubes of the subglottal airways is an important aspect of the aerodynamics also. Arguments for the separation of subglottal flow resistance and lung volume, discussed in Section II.3.3.3, depend on predictions from the modelling of Pedley et al. (1970) and Pedley (1977), which made use of Weibel's (1963) model.

Stevens derived, from Boyle's law and the principle of conservation of mass, patterns of low frequency components of air pressure and volume flowrate during the release of a vocal tract closure with an
orifice of increasing cross-section area. The importance of the size of the glottis for the pressure decay and turbulent flow time, and thus for the acoustic properties of different plosive types, was emphasised (Stevens, 1956; see also Fant, 1960, 1970, pp.276-280). Stevens' (1971) analysis and theoretical modelling of aerodynamic and acoustic processes predicts the range of aerodynamic conditions at the glottis and at a supraglottal constriction for which turbulence noise sources are expected to arise. This study demonstrated the unifying role of aerodynamics and provides a theoretical basis for the generation of turbulence noise sources in our model.

Rothenberg's detailed and pioneering study established the basis for subsequent modelling of aerodynamic processes in speech production. In his hardware electrical circuit model mechanical elements such as muscle force, stiffness and viscous resistance of tissues, and aerodynamic elements such as pressure, volume flowrate, flow resistance and compressibility of a volume of air were represented by the equivalent electrical circuit elements. His approach was based on that of van den Berg (1960).

Rothenberg modelled the many factors which contribute to the observed aerodynamic and acoustic properties of plosives and demonstrated the importance of aerodynamic effects, such as increase or reduction of supraglottal air pressure associated with changes of cavity volume, and changes in subglottal air pressure resulting from changing muscle forces and articulatory actions. The vocal tract was modelled as a cavity with compliant walls, between the glottis and the point of closure and release for the plosive. The subglottal system included three pairs of capacitances and resistances to represent the subglottal airways. It is argued in Section II.3.3.3 that a simpler representation, with a lower value for flow resistance, is appropriate for this portion of an aerodynamic model. The importance
for the aerodynamics and acoustics of the time required for actions such as abducting and then adducting the vocal folds was made clear (Rothenberg, 1968).

Further modelling of this kind has been done, to investigate how patterns of oral air pressure for stop consonants relate to place and manner of the consonant and to vowel context; the model was implemented as computer programs in this case and the important flow resistances of severe constrictions of the respiratory tract were more nearly correctly defined in this case by the use of the orifice equation, Equations (II.12) and (II.13) in Section II.3.3.4 (Müller and Brown, 1980). Westbury (1979) included aerodynamic as well as articulatory considerations in his analysis and modelling of the production of consonant clusters in English speech.

Ohala's modelling of aerodynamic processes takes as respiratory control the application of a pulmonic force. Rate of decrease of lung volume and air pressure and volume flowrate parameters are derived. Changes of subglottal pressure and lung volume decrement are shown to depend on the total obstruction presented to the airflow; patterns generated by the model matched data for natural speech (Ohala, 1975, 1989; Ohala et al., 1980).

II.5.4 Acoustic models and electrical analogues

Modelling of the voice source is a research area of major importance and is the subject of many conferences and publications. The modelling of this and other types of acoustic sources has been discussed in some detail in Section II.3.4 and the references cited there will not be repeated here. It may be mentioned that although most models are computer implementations of mathematical representations of the physical processes, hardware mechanical models
on which actual acoustic measurements can be made are of great value also. This approach is used by, for example, Shadle (1986a, 1986b) to investigate mechanisms for fricatives. A human larynx has been combined with a hardware model vocal tract to investigate source-filter interaction in voicing (Laine and Vilkman, 1987). The perceptual effects of cross-modes of the standing waves in the vocal tract have been investigated with a mechanical model by Holmes (1981). Rothenberg et al. (1975) have described a model of the voice source, driven by three controlling parameters, implemented as electrical circuits.

The acoustic behaviour of the vocal tract has been modelled by several different methods, as discussed in Section II.3.5. The Webster horn equation, which applies to one dimensional loss-free wave propagation along a tube with gradual changes of cross-section area, was solved numerically by Chiba and Kajiyama (1941). More often, reflections at sudden area changes do need to be taken into account when modelling the filtering processes of speech production. The vocal tract is generally represented for this purpose as a set of abutting cylinders of different cross-section areas. The magnitudes of incident, reflected and transmitted pressure waves at the cylinder boundaries were obtained by Kelly and Lochbaum (1962). In Mermelstein's (1972) model reflections are expressed in $z$-transform terms and the computed transfer function is expressed as a recursive filter. The reflected pressure wave method, which is lossless in principle, but to which small losses can be added, is included in many models of speech production, including the one described in this work. A nasal branch was added to the reflected pressure wave method by Husband et al (1977); sinus cavities were represented in Maeda's (1982a) model, which represents the vocal tract by lumped elements of an acoustic transmission line; multiple branching tubes have been introduced into the reflected pressure wave method for the better
representation of the nasal cavities in French nasalised vowels by Castelli et al. (1989). A Helmholtz resonator lumped element is an appropriate model for low frequencies if a cavity can be identified; the whole skull is proposed as an explanation of low frequency spectra for nasalisation by Castelli et al. (1989).

In the modelling of Flanagan et al. (1975) the continuous acoustic tube is represented by a multisection electrical transmission line in which each section has component values appropriate for an acoustic tube of a particular cross-section area, as follows:

<table>
<thead>
<tr>
<th>Acoustic tube</th>
<th>Electrical transmission line</th>
</tr>
</thead>
<tbody>
<tr>
<td>cross-section area $A$</td>
<td>inductance $L = \rho \frac{l}{A}$ (II.24)</td>
</tr>
<tr>
<td>length $l$</td>
<td>capacitance $C = lA/\rho c^2$</td>
</tr>
</tbody>
</table>

The representation of losses and wall effects in this kind of electrical analogue is derived by Fant (1960, 1970, p.27 ff.; see also Flanagan, 1972a, p.25 ff.).

It is difficult to represent frequency-dependent effects in a time-domain simulation, but the z-transform provides the techniques to do so. This approach to the representation of losses in the reflected pressure wave method has been investigated in the modelling of Liljencrants (1985).

There has been much development of acoustic models for the mapping from vocal tract area function to acoustic patterns (for example Degryse, 1981; Maeda, 1982a; Wu et al., 1987; Laine, 1989). At least one implementation using a signal processing chip has been proposed (Cross et al., 1986).
II.5.5 Composite models and the need to include an aerodynamic stage

Probably the two best known complete models of speech production are those of von Kempelen (1791) and of Flanagan and his co-workers at Bell Laboratories (Flanagan et al., 1975). Both of these synthesise speech by simulation of all the main processes involved. They both model as one system low frequency aerodynamics and higher frequency acoustics, though by very different means. Von Kempelen's was a mechanical model, with bellows for the lungs to provide an airstream, a vibrating reed as the voice source and narrow pipes for generating turbulence noise, and a mobile india-rubber mouth for the filters (see Linggard, 1985, Chapter 1). The model developed at Bell Laboratories is a digital simulation of the physical processes, implemented on a high-speed computer. It includes an applied subglottal air pressure, vocal folds set into oscillation when conditions are suitable, and a mobile vocal tract which filters and interacts with the voice source. Turbulence noise sources are inserted at different places along the filter, wherever and whenever conditions are appropriate (Flanagan and Ishizaka, 1976). In recent developments of this modelling, the model has been matched to a specific speech signal, by means of computationally heavy analysis-by-synthesis procedures (Flanagan et al., 1980). As computing power increases the computational times decrease. The modelling of Sondhi and Schroeter (1987), with a model developed from that of Flanagan et al. (1975) takes only about twice real time.

Models which incorporate articulation and subsequent stages so as to generate synthetic speech-like sounds have been used for perception research. In a model implemented at Haskins Laboratories (Abramson et al. 1979; Rubin et al., 1981) a series of stimuli along an articulatory continuum, the size of the velopharyngeal port, was generated. Here the acoustic sources of voice and frication noise
were specified by the experimenters, and not derived from the articulatory and aerodynamic conditions at the glottis and at a vocal tract constriction.

In Haggard's (1979) modelling, on the other hand, also directed towards perceptual research, the processes which give rise to the acoustic sources were included: voice and frication noise sources were made to depend on both the articulation and the aerodynamic conditions. Frication noise varied with vocal tract constriction size and oral pressure; voice was sustained or ceased during plosive closures according to changes in aerodynamic conditions, which were influenced by vertical larynx movements and other actions. An important advantage of this kind of approach, as opposed to an ad-hoc introduction of acoustic sources, is that the whole set of acoustic effects generated by a step along a single articulatory continuum should have a complexity and internal consistency paralleling that of natural speech.

Much better understanding of the patterning and processes is needed for all stages of speech production. All natural speech data and all models of the processes contribute to this. Better mathematical models are needed for every stage. The implications of possibilities and patterning at one stage for the succeeding one need to be explored further.

It is an advantage of simulations which include more than one or two stages of speech production that reasons for the observed patterning at any one earlier stage may become clearer when the results of that patterning at all the later stages, including the output speech signal, are made manifest. For example, coarticulation is a term widely used to mean the ways in which articulatory patterns for one linguistic unit are dependent upon phonetic context. But the
gestures of the vocal tract, larynx and respiratory system are as they are in order to yield a particular required sequence of sounds, broadly defined, with some modifications permitted due to changes of phonetic context. If the organisation of articulation is to be fully understood its consequences need to be carried through to the production of acoustic sources, their modification by filtering, and the radiation of a soundwave.

The particular focus of the model to be described in this study is the interaction of articulation and the resulting aerodynamic patterns to generate acoustic sources. The final acoustic processes of filtering and radiation are included so that the achievement or otherwise of the specified auditory goals may be assessed.

It seems important to try to incorporate into signal models of speech explicit descriptions of the aerodynamics, the first stage at which quasi-independence of the components no longer holds. Because the aerodynamic system of speech production is unified, linking the different acoustic sources and the sources with the filters, its effects are seen in the multiplicity of acoustic pattern features which change together when a single articulatory change is made. These covarying bundles of acoustic pattern features seem to be used by human listeners and need to be related to the physical processes of speech production which give rise to them. The composite model described in the next chapter was developed in response to the need as the present author saw it to explain the interactions between acoustic sources and filters on the basis of the unified aerodynamic system of speech production which links the actions to the sounds.
Chapter III  A COMPOSITE MODEL OF SPEECH PRODUCTION

Introduction

The speech production processes inside the central nervous system, together with neuromuscular activity, articulation, aerodynamics, acoustic sources generation and filtering, form a link between the representation of linguistic structures and rules in the nervous system of a particular speaker, the application of these rules to a particular linguistic message and the transmitted speech signal. The overall view of speech production adopted here is shown in Figure 5.

Ideally, a model of speech production should include all these processes; as discussed in Section II.5.5 a composite model is needed. The model described here takes as a starting point the voluntary actions, the articulatory stage. A time scheme planning and organisation stage is included. This precedes the generation of actual movement paths and might perhaps be considered to represent part of the timing organisation function of the central nervous system. A neuromuscular stage, with its large number of degrees of freedom, is not included in our model. Articulation, the positions and movements of only about nine to sixteen quasi-independent structures, is amenable to experimentation through variation of just a few parameters. Another advantage of starting with articulatory processes is that there are enough data from natural speech on which to base the modelling, as discussed in Section II.3.2. and Appendix 3 (Scully, 1987).

Subsequent blocks of the model represent automatic physical processes; these are under the speaker's control only via articulation. In reality, all these processes of speech production occur nearly simultaneously and almost instantaneously. In the model
described here they occur as successive stages: articulation, aerodynamics, acoustic sources, acoustic filtering and radiation of a sound pressure wave. The same structures can play a part in several stages. The processes of the stages are not spatially separated; they are only conceptually distinct. In the model they must be taken in the correct order.

In natural speech the various processes are bound to be consistent with each other. If a model is to achieve some of the complexities and covariations of acoustic patterning found in natural speech it needs to be internally consistent. To some extent this is realised in our model. For example, vocal tract articulation is described in terms of cross-section areas of the vocal tract air-filled tube. One articulatory parameter point defines the most severe constriction of the vocal tract and this one is used as an orifice providing an obstruction to the flow of air, in the aerodynamic block of the model. The active component of the volume of the cavity behind this constriction, another component of the aerodynamics block, is determined by the changing vocal tract area function, which is derived in the previous, articulatory, block of the model. The total area function of the vocal tract serves also as the filter which modifies the acoustic sources. Thus, the vocal tract shape has a consistent effect in the domains of articulation, aerodynamics and filtering.

Consistency between the aerodynamic and the acoustic voice components of glottal airflow has not been achieved, but the acoustic sources are made to depend on the combinations of articulatory and aerodynamic conditions defined by earlier stages which give rise to them. The vocal tract shape has a consistent effect for articulation, aerodynamics and filtering.
This chapter is divided into three main sections. In Section 1 the general principles of the model are set out and related to the data and theory discussed in Chapter II. Section 2 of the chapter includes descriptions of the actual computer programs and the commands within them that are available to an experimenter. Some details are given about the organisation and structure of the programs. There is some overlap in the material included in the different sub-sections. In particular, the Introduction and Section III.2.1 present an outline of the whole model; in subsequent sections, some of the commands and operations of the modelling are described in more detail. Section III.3 gives a brief account of the ways in which the patterns generated in the modelling are assessed, through auditory judgements and other sorts of comparisons with real speech.
III.1 Principles

III.1.1 Inputs

The speaker's actions are taken to be goal-directed; therefore one input to a model of the processes must be the auditory target aimed for: a real word or a phrase or a phonologically allowable sequence of allophones of an RP accent of British English. A second set of inputs characterises the particular speaker type to be simulated. A third set specifies the conditions of the systems at the start of the simulation.

Within the articulatory block of the model, to be described below, a complete set of movement paths is to be established. The first set of choices made is very unlikely to generate the specified auditory goal; suitable paths are to be determined by trial and error with auditory and other types of monitoring. Whereas a baby must learn the mapping from actions to sounds from scratch, phonetic knowledge available to the experimenter can direct the initial choice of actions. This knowledge is, in effect, another kind of input to the model. Few data were available during the course of the study for speech produced by women and children, so in order to make use of available knowledge, the modelling described in this work is almost entirely confined to that of a man's speech.

III.1.2 Articulation

The articulatory block of the model is described in detail in Appendix 3 (Scully, 1987); only an outline will be given here.
The basic approach is to treat the moving structures not in terms of the forty or so muscles used in speech but as about nine to sixteen articulators (see Sections II.1.1 and II.3.2.5).

As discussed in Section II.3.2, for an internally consistent descriptive framework it is advantageous to include as articulators all the structures whose configurations, mechanical states and movements take part in the production of speech. The number of quasi-independent articulators varies a little in different articulatory models. The articulators controlled in our model comprise the following, as indicated in Figure 15:

1) the lung walls, coupled to the chest wall, controlling lung volume VLU;
2) the vocal folds, adducted and abducted so as to control the area of the glottis AG;
3) the effective vibrating mass and stiffness of the vocal folds as the laryngeal control of \( F_0 \), called Q;
4) the tongue body and pharynx walls, called TB, controlling most of the vocal tract tube shape;
5) the tongue tip or tip and blade portion of the tongue, controlling the cross-section area in the alveolar ridge or hard palate region of the vocal tract and defining AF or AF and AC in the model;
6) the jaw, controlling the vocal tract tube cross-section area AJ, in the region of the jaw;
7) the lips, controlling the lip outlet area AL;
8) the velum (or soft palate) which, together with the upper pharynx walls, controls the area of the velopharyngeal port AV;

Eight of the nine articulators listed in Section II.3.2.5 are represented in the model; only vertical movements of the larynx are
omitted. This third aspect of laryngeal articulation is not included because it would introduce changes in the number of vocal tract filter sections and this cannot conveniently be managed in the filtering method implemented in the model.

For vocal tract actions, some models emphasise the solid structures themselves, others the local shaping of the vocal tract tube, as discussed in Section II.5.2. The latter approach is adopted here. Although each supraglottal articulator is here associated with one or more solid structures as a convenient aid to visualisation, the actual vocal tract articulatory parameter is, in each case, the cross-section area at a point in the vocal tract air-filled tube.

The joint actions of more than one set of muscles and structures moved by the muscle forces to control a particular portion of the vocal tract is assumed to underlie the movement paths, but these processes of functional synergy (see Section II.1.4) are not explicitly modelled. Instead, each of the supraglottal articulators is individually responsible for the shaping of a particular portion of the vocal tract. Supraglottal articulation is expressed as changing values of cross-section area $AX$ defined at the position $DX$ along the vocal tract tube of parameter point $X$.

Parameter points $S, E, P, N, B, C, F, J, T$ and $L$ each define the cross-section area of a point along the curved air-filled tube, as shown in Figure 15. Interpolation is applied between the defined values $AE, AP, AN$ and so on, so as to link the parameter points together and define the total area function of the vocal tract.

Each vocal tract parameter point has an associated $D$ parameter $DE, DP, DN$ and so on, defining its distance from the glottis along the length of the vocal tract tube; thus the lengths and proportions of
the vocal tract in the model can be set to different values. In addition, transitions can be given to one or more D parameters so as to change the position of the parameter point during the course of a simulation.

The articulators are considered to be quasi-independent: each one is given instructions which do not automatically affect the path of another articulator. The articulatory model does not have physiological constraints built into it; it is underconstrained. One advantage of this is that the modelling is not confined to normal adult speech nor to a limited range of speaker types. But for simulating normal male adult speech reasonable estimates of constraints on the shapes and movements of different portions of the tongue need to be added.

Articulatory constraints and the interactions observed in natural speech, discussed in Section II.3.2, are introduced into the modelling by the choice of instructions. For example, based on the assumption of an incompressible tongue volume and the tongue shape factor analyses reviewed in Section II.3.2.1, diphthong-like actions of the tongue body are made to conform approximately to two factors, as shown in Figure 16. For example, an [aɪ]-type diphthong can be generated very easily, simply by using tongue factor 1: the pharynx size is made to increase by increasing the cross-section area at the P point AP and at the E point AE; meanwhile, the tongue dorsum is raised towards the hard palate by decreasing AB and AJ; AN at the hinge point for factor 1 remains constant. The effect, corresponding to the front-raising factor of Harshman et al. (1977), is sketched in Figure 16(a). The representation of back raising, the second factor of their study, is sketched in Figure 16(b).
It is allowable for the area of, for example, the F point to go to a value less than zero, so as to simulate the movement paths into and out of a tight closure of the vocal tract for some consonants, as shown in Figure 7 and discussed in Section II.3.2.4. The cross-section area, in this example AF, is reset to zero later.

It seems that during diphthong-like actions different portions of the tongue body have synchronised articulatory transitions (Houde, 1968, see Section II.3.2.5). In our model, the tongue body TB is defined by about three or four parameter points and their associated D values and cross-section areas, for example AE, AP, AN and AB, and usually AJ also. They are all made to move in synchrony for diphthong-like gestures, acting together as a single articulator. The implementation of this approach is described in Section IV.1.2, using [ai] as an example.

Although the tongue tip, or tip-blade, region is treated as an individual articulator distinct from the tongue body, the extension of the tongue tip, in simulating the closure for an alveolar fricative, for example, is not permitted to be unnaturally large. The whole tongue body is made to move to a suitable configuration for the fricative, by a change in the cross-section area at all the tongue body parameter points. To the extent that the modelling matches vocal tract gestures found in natural speech, the constraints of natural speech are observed also.

The dorsum of the tongue is included as part of the tongue body. For diphthong-like movements of the tongue as a whole AE, AP, AN and AB are given the same transition, but for raising and lowering the dorsum in actions associated with palatal or velar consonants one or more of these points, probably AN and AB, can be given a different transition type by means of which the vocal tract shape is modified
locally. In the modelling described here this approach is sometimes used for alveolar consonants, as shown in Appendix 3, Figure 6 (Scully, 1987, p.93). First the overall vocal tract shape is defined by the tongue body points, the jaw AJ and the lips AL; then the shape is modified locally by transitions of the tip-blade AF. AF may be but need not be included among the tongue body parameter points. The resulting complex transitions of the F point or of the cross-section area AF can be made to match data from natural speech.

Different articulators are skilfully coordinated for particular actions. The elements at the heart of articulation are taken to be the movement paths for all the articulators taken separately and inter-articulator coordination describing how two independent articulatory events are timed relative to each other. A framework to handle timing in a quantitative but flexible way has been developed; it is described in Appendix 3 (Scully, 1987). The construction of a time plan, expressed as simple duration offsets D between pairs of events E, is the first operation performed, after a particular type of vocal tract has been defined. The values in this time scheme are incorporated into the instructions given to each articulator, from which its actual path is calculated, as described in Appendix 3 (see, especially, Appendix 3, Chap.IV, Scully, 1987, p.125 ff.).

An articulatory transition is defined in the following way: the articulator starts at rest, accelerates, reaches a maximum velocity, decelerates, and comes to rest again at the end point of the transition. This end point state forms the starting state for the next transition of that articulator. It is possible in the model to allow too little time for the transition to be completed; in that case a moment of zero velocity will arise, but not at the end point of the originally defined transition. This could be considered to
represent the phenomenon of undershoot, discussed in Section II.3.2. The situation is generally avoided in the modelling. Evidence for the existence of undershoot in natural speech is not conclusive, and the view adopted in this study is that speakers execute complete articulatory transitions as intended, perhaps modifying them, for example at a faster speech rate, but not interrupting one transition by a subsequent one.

Articulation is treated kinematically, not dynamically: there is no attempt to analyse movement in terms of forces, mass, stiffness and viscous resistance. Movement paths are simply described. They are intended to match those observed in real speech, but in a very simplified way.

As an approximation to the kinds of data reviewed in Section II.3.2 each articulator moves along an S-shaped distance-versus-time path as shown in Figure 16 (c) and in Appendix 3, Figure 5 (Scully, 1987, p.92). The time taken to accelerate from rest, reach maximum velocity and decelerate to rest is made independent of the distance traversed. A slow (S) transition time of 190 ms is appropriate for diphthong-like movements of the tongue body, especially in the front-back direction, and for lip protrusion and retraction. A medium (M) transition of 120 ms is appropriate for the vocal folds, the soft palate, the jaw, and raising or lowering the tongue dorsum. A fast (F) transition of 40 ms seems appropriate for the rapid tongue tip movement in a tap [t], and perhaps for [l] also. Intermediate durations may be a better match to real speech data when the tongue tip and blade move together towards and away from the roof of the mouth for [s] and [z]. A transition time of 80 ms has been used for this in modelling "a hiss it said" and "a his it said" (Appendix 7, Scully and Allwood, 1985a).
Articulatory transitions for the vocal tract expressed as changes in tube cross-section area at selected points are assumed to have the same general distance-versus-time shape as changes in distance of selected mid-sagittal points of the tongue, jaw, lips or velum from the hard palate or other fixed reference lines. This neglects the cross-dimension to cross-section area transform which was discussed in Section II.3.2.3. The simplification was felt to be justified in the absence of a clearly defined and universally applicable expression for this transform. However, this procedure is a weakness in the modelling, since most of the natural speech data on which the model's articulatory description was based were mid-sagittal section X-ray views of the vocal tract.

Areas are always used for the modelling of articulation, but a side view of the vocal tract, derived from the area function, is available as a graphical display; this can help to indicate whether the total vocal tract shape seems plausible or not. Based on the research reviewed in Section II.3.2.3 the following expression for mapping is used:

\[ X = \left( \frac{A}{3.5} \right)^{0.7} \quad \text{or} \quad A = 3.5 \cdot X^{1.43} \]  \hspace{1cm} (III.1)

where \( A \) is the cross-section area and \( X \) is the distance across the vocal tract tube in the mid-sagittal plane. The data of Chiba and Kajiyama were used to derive this mapping; it makes cross-dimension \( X \) grow rather less rapidly with increasing cross-section area \( A \) than does the expression used by Flanagan et al. (1980) for example. It seems probable that an equally realistic medial sagittal section view could have been obtained by using the cross-section area values directly.
Nasalisation can be included in the modelling: AV in Figure 15 represents the area of the velopharyngeal port, which links the oral and nasal cavities.
III.1.3 Aerodynamics

Introduction

The aerodynamic system must be considered as a unified, irreducible whole. The aerodynamic model is shown in Figure 17. It contains two cavities, the lungs and the portion of the vocal tract between the glottis and a constriction of the vocal tract, which represents the main vocal tract articulator. Generally this is the point at which the vocal tract is most severely constricted; for example, if alveolar consonants are included in the allophone sequence to be modelled this constriction would be the one formed by the tongue tip or tip and blade with the alveolar ridge.

The three orifices of the system are the main vocal tract constriction, the glottis and the velopharyngeal port. The nasal portion of the vocal tract and the portion in front of the main vocal tract constriction are assumed to be at atmospheric pressure.

It was argued in Section II.3.3.3 that only two lumped elements are needed for the subglottal portion of the aerodynamic system - a single cavity, containing nearly all the volume of air but essentially no flow resistance, and a single rigid-walled tube containing an ohmic-type flow resistance, but essentially no volume. These two elements are shown in Figure 17: the lung cavity represents the subglottal airways of generations 12 and above; the rigid tube connecting this cavity to the glottal orifice represents generations 0 to 11. The flow conductance of this tube is taken to be proportional to lung volume but with a maximum value which applies for lung volumes of about 0.45xTLC and above. This function, shown in Figure 18, is based on Vincent et al. (1970), as discussed in Section II.3.3.4.
The aerodynamic block of the model is closest to a true physical model, albeit in a very simplified way; its basis will be described in more detail than other blocks of the model. A set of simultaneous differential equations is needed to describe conditions and rates of change of the aerodynamic parameters at each time point. Solution of the equations gives volume flowrate of air and air pressure at selected places in the respiratory tract.

Initial conditions need to be stated. Numerical methods are used to work forward from there, so as to specify values at all subsequent sampled time points.

The main physical principle invoked is continuity of mass flow. A number of simplifying assumptions are made, based on the theory and data reviewed in Section II.3.3.

The components are lumped elements:

1) volumes of air, the cavities, throughout each of which airflow and air pressure are taken to be constant;
2) orifices through which air flows and across which a pressure difference develops, with flow-dependent resistances;
3) an ohmic, flow-independent resistance located in the larger subglottal airways, the conducting zone for respiration, across which a small pressure difference develops;
4) the yielding walls of the vocal tract and of the lungs.

Expressions and values for the compliances and flow resistances of each of the components must be stated. Inertances can be neglected, as discussed in Section II.3.3.6. In most cases reasonable estimates have to be made for the speaker type to be simulated, based on the evidence from respiration and speech reviewed in Section II.3.3.
III.1.3.1 Physical principles applied and assumptions made

From continuity of mass flow, the rate at which air molecules flow into a region added to the rate at which the number of molecules in that region increases is equal to the rate at which molecules flow out of the region. The principle applies to each cavity and is part of the derivation of the orifice equation also. Conservation of energy is taken to apply at the entry side of an orifice but not at the outlet where most of the turbulence and dissipation of energy takes place. The following assumptions are made:

1) the air in the respiratory tract is a perfect gas under isothermal conditions;
2) the air throughout the respiratory tract moves at velocities well below Mach 1, the velocity of sound in the air of the respiratory tract;
3) changes in air pressure are small compared with atmospheric pressure;
4) continuity of mass flow applies for each cavity and through each orifice;
5) conservation of energy applies at the inlet to each orifice but not at its outlet;
6) inertance of the air and the tissues may be neglected;
7) the subglottal airways may be represented by a single lung cavity and a single ohmic flow resistance;
8) the compliance of the lung cavity walls is effectively zero;
9) the compliance of the vocal tract cavity walls needs to be included;
10) changes in vocal tract cavity volume due to wall compliance are not fed back into the articulatory block of the model.
In Sections III.1.3.2, III.1.3.3 and III.1.3.4 which follow, the correct symbols for the physical variables will be used, written by hand for clarity. For example:

\[
\begin{align*}
P & = \text{absolute air pressure} \\
P_a & = \text{atmospheric air pressure} \\
P_a + P_l & = \text{pressure of air in the lungs} \\
\rho_l & = \text{density of air in the lungs} \\
\frac{dV_l}{dt} & = \text{rate of increase of lung volume with time.}
\end{align*}
\]

and so on. Each of the physical variables which is needed as a parameter in the model is given a more convenient working symbol, for example:

\[
\begin{align*}
P_l & \quad \text{becomes PLU} \\
\frac{dV_l}{dt} & \quad \text{becomes DVLU}
\end{align*}
\]

and so on. The two sets of equivalent symbols are shown on pages 14 to 16, together with the default values for the lung parameters and values of the physical constants.

**III.1.3.2 Respiratory control**

It has been assumed at different times in the modelling that the rate of lung volume increase \(\frac{dV_l}{dt}\) or air pressure in the lungs \(P_l\) may be defined as an input to the aerodynamic model, assumed to be an aspect of articulation under the speaker's control. From either of these input parameters, the other must be obtained. \(P_l\) builds up to a suitable positive value for speech whereas \(\frac{dV_l}{dt}\) has a suitable negative value.

For both approaches, lung wall compliance, or the effective compliance due to the lung wall and rib cage combined, was included
originally and is included in the continuity equation for the lung cavity set out in Section III.1.3.3. But the speaker knows his or her own mechanical system: the recoil from this combined compliance must have been taken into account when the speaker discovered suitable respiratory muscle control functions at different lung volumes that would give a constant nett expiratory force, as discussed in Section II.3.3.2. This means that, as long as conditions remain as predicted, the lung cavity walls appear to be infinitely stiff, that is, the cavity has rigid walls.

When lung volume increase \( \frac{dV_L}{dt} \) was first tried as the respiratory control parameter with lung cavity wall compliance of 150 cm\(^3\)/cmH\(_2\)O, the simulation was unsuccessful. A smaller value such as 50 or 100 cm\(^3\)/cmH\(_2\)O was tried also, combined with a realistic value for an active component of lung volume decrease, but as air pressure began to rise in the lung cavity the lung walls expanded greatly and air pressure took far too long to rise to speech operating values; the resultant rate of lung volume decrease was far too low. This was a simulation of an unskilled speaker. For the modelling described in Chapter IV and the Appendices, lung cavity wall compliance was set to an extremely low value, generally 0.001 cm\(^3\)/cmH\(_2\)O, and this gave realistic results.

A third type of respiratory control was devised in which nett expiratory pressure \( P_{\text{exp}} \) is specified, but this was not used for any simulations described in this study.

The method with \( P_L \) controlled is easier to work with since it gives subglottal pressure essentially equal to this defined articulatory parameter. When \( \frac{dV_L}{dt} \) is used instead suitable values to generate appropriate subglottal pressure must be found, partly by trial and error. This seems likely to be a more realistic approach however,
with subglottal air pressure determined not only by lung actions but also by the degree of obstruction offered at and above the glottis.

III.1.3.3 The simultaneous differential equations

The parameters and their symbols are shown in Figure 17.

The equation of state for a perfect gas is:

\[ PV = \frac{kM}{m} T \]  

(III.2)

with \[ \rho = \frac{M}{V} \]  

(III.3)

where \( P \) is pressure, \( V \) is volume, \( M \) is the mass of gas occupying the volume, \( \rho \) is density, \( T \) is absolute temperature, \( k \) is Boltzmann's constant and \( m \) is the mass of a gas molecule.

From equations (III.2) and (III.3):

\[ \rho = \frac{mP}{kT} \]  

(III.4)

Isothermal conditions are assumed, with a temperature of 37°C throughout the respiratory tract, so that Boyle's law applies, viz. for a given mass of gas:

\[ PV = \text{constant} \]  

(III.5)

From Equation (III.4):

\[ \rho \propto P \]  

(III.6)
Supraglottal cavity

From Equation (III.6) the air pressure in the supraglottal cavity \((P_a + P_c)\) and the density of air in the supraglottal cavity \(\rho_c\) are related by:

\[
\frac{\rho_c}{\rho_a} = \frac{(P_a + P_c)}{P_a}
\]

(III.7)

where \(P_a\) is atmospheric pressure, 1030 cmH\(_2\)O and \(\rho_a\) is the density of air at atmospheric pressure.

Since \(P_c\) typically rises to about 10 cmH\(_2\)O in speech, much less than \(P_a\), it may be assumed that:

\[
\frac{(P_a + P_c)}{P_a} \approx 1
\]

(III.8)

The nett rate of mass flow of air into the supraglottal cavity is:

\[
\rho_c(u_g - u_c - u_v)
\]

The rate of mass increase for air in the supraglottal cavity is:

\[
\frac{d}{dt}(\rho_c V_c)
\]

Continuity of mass flow for this cavity gives:

\[
\rho_c(u_g - u_c - u_v) = \frac{d}{dt}(\rho_c V_c)
\]

(III.9)

Or, substituting for \(\rho_c\) from Equation (III.7):

\[
\rho_a \frac{(P_a + P_c)}{P_a} (u_g - u_c - u_v) = \rho_a \frac{(P_a + P_c)}{P_a} \frac{dV_c}{dt} + V_c \frac{d}{dt}\left(\frac{\rho_a (P_a + P_c)}{P_a}\right)
\]
From Equation (III.8) this gives:

$$U_g - U_c - U_v = \frac{dV_c}{dt} + \frac{V_c}{P_a} \frac{dP_c}{dt}$$  \hspace{1cm} (III.10)

$V_c$, the vocal tract cavity volume, changes as a result of articulatory movements and as a result of movements of its compliant walls when air pressure in the cavity departs from atmospheric pressure. These are considered to be an active and a passive component of cavity volume change respectively.

The cavity volume at atmospheric pressure, obtained from the articulatory block of the model, is $V_c^{\text{act}}$ from which an active component of volume change $(dV_c/dt)_{\text{act}}$ is derived.

The compliance of the vocal tract wall is $c_{cw}$ and this gives a passive component of vocal tract cavity volume change $(dV_c/dt)_{\text{pas}}$ as air pressure inside the cavity alters which is given by:

$$\left(\frac{dV_c}{dt}\right)_{\text{pas}} = c_{cw} \frac{dP_c}{dt}$$  \hspace{1cm} (III.11)

so that the total rate of change of cavity volume is given by:

$$\frac{dV_c}{dt} = (dV_c/dt)_{\text{act}} + c_{cw} \frac{dP_c}{dt}$$  \hspace{1cm} (III.12)

Substituting for $dV_c/dt$ in Equation (III.10) gives:

$$U_g - (U_c + U_v) = (dV_c/dt)_{\text{act}} + (c_{cw} + \frac{V_c}{P_a}) \frac{dP_c}{dt}$$ \hspace{1cm} (III.13)
Lung cavity

Corresponding to Equation (III.7) for the supraglottal cavity, air pressure in the lung cavity \((P_a + P_l)\) and density of air in the lung cavity \(\rho_l\) are related by:

\[
\frac{\rho_l}{\rho_a} = \frac{(P_a + P_l)}{P_a} \tag{III.14}
\]

Since \(P_l\) is typically about 10 cmH\(_2\)O in speech, much less than \(P_a\), it may be assumed that:

\[
\frac{(P_a + P_l)}{P_a} \approx 1 \tag{III.15}
\]

The nett rate of mass flow of air into the lung cavity is:

\[-\rho_l U_l\]

The rate of mass increase for air in the lung cavity is:

\[
\frac{d}{dt}(\rho_l V_l)
\]

Continuity of mass flow applied to the lung cavity gives:

\[-\rho_l U_l = \frac{d}{dt}(\rho_l V_l) \tag{III.16}\]

Or, substituting for \(\rho_l\) from Equation (III.14):

\[-\rho_a \frac{(P_a + P_l)}{P_a} U_l = \rho_a \frac{(P_a + P_l)}{P_a} \frac{dV_l}{dt} + V_l \frac{d}{dt} \left(\rho_a \frac{(P_a + P_l)}{P_a}\right)\]

From Equation (III.15) this gives:

\[-U_l = \frac{dV_l}{dt} + \frac{V_l}{P_a} \frac{dP_l}{dt} \tag{III.17}\]
Although the effective wall compliance of the lung cavity needs to be made extremely small, as discussed in Section III.1.3.2, it is included here, for completeness.

The compliance of the lung cavity wall is $c_{l\omega}$ and this gives a passive component of lung cavity volume change $(dV_L/dt)_\text{pas}$ as air pressure inside the cavity alters, which is given by:

$$\left(\frac{dV_L}{dt}\right)_\text{pas} = c_{l\omega} \frac{dP_L}{dt} \quad (\text{III.18})$$

The active component $(dV_L/dt)_\text{act}$ is an input to the aerodynamic block of the model if this form of respiratory control is used. If, on the other hand, $P_L$ is the input control parameter, then $(dV_L/dt)_\text{act}$ is derived.

As in the case of the vocal tract cavity, the $dP_a/dt$ term is zero. The total rate of change of cavity volume is given by:

$$\frac{dV_L}{dt} = (dV_L/dt)_\text{act} + c_{l\omega} \frac{dP_L}{dt} \quad (\text{III.19})$$

Substituting for $dV_L/dt$ in Equation (III.17) gives:

$$-u_L = (dV_L/dt)_\text{act} + c_{l\omega} \frac{dP_L}{dt} + \frac{V_L}{P_a} \frac{dP_a}{dt} \quad (\text{III.20})$$

**Orifices**

Across each of the three orifices - the glottis, the main vocal tract constriction and the velopharyngeal port - the air pressure drop is taken to be the sum of two components: a viscous loss term and a turbulence loss term.
The viscous loss, a flow-independent term, is given by the Poiseuille formula for slow steady flow through a pipe of circular cross-section, as follows (van den Berg et al., 1957):

$$\Delta P = \frac{8\pi\mu l u}{gA^2} \quad \text{(III.21)}$$

where $\mu$ is the dynamic coefficient of viscosity for air, $g$ is the acceleration due to gravity, $\Delta P$ is the pressure drop across the pipe in cm H$_2$O, $l$ is its length in cm, $U$ is the volume flow rate through it in cm$^3$/s, $A$ is the cross-section area in cm$^2$.

The turbulence loss, flow-dependent resistance term is given by the orifice equation (Equation II.12):

$$\Delta P = \frac{\rho K U^2}{2A^2}$$

or by the working orifice equation (Equation II.13):

$$A = \frac{0.00076 U}{\Delta P^{0.5}} = \frac{K_e U}{\Delta P^{0.5}}$$

where $A$ is the minimum cross-section area in cm$^2$, $U$ is the volume flow rate through the orifice in cm$^3$/s, $\Delta P$ is the pressure drop across it in cmH$_2$O and $K_e$ is an empirical constant. (See page 16 and Section II.3.3.4 for explanations of the choice of $K_e$ value.)

The pressure drop across a constriction is given by the sum of the viscous loss and turbulence loss components, viz.

$$\Delta P = \frac{8\pi\mu l u}{gA^2} + \frac{(0.00076 U)^2}{A^2} \quad \text{(III.22)}$$
This expression becomes, for the vocal tract constriction:

$$P_c = \frac{N_c 8\pi \mu l_c |u_c|}{g A_c^2} + \frac{(0.00076 U_c)^2}{A_c^2}$$  \hspace{1cm} (III.23)

for the velopharyngeal port:

$$P_c = \frac{N_v 8\pi \mu l_v |u_v|}{g A_v^2} + \frac{(0.00076 U_v)^2}{A_v^2}$$  \hspace{1cm} (III.24)

and for the glottis:

$$P_{sg} - P_c = \frac{N_g 8\pi \mu l_g |u_g|}{g A_g^2} + \frac{(0.00076 U_g)^2}{A_g^2}$$  \hspace{1cm} (III.25)

Empirical constants $N_c$, $N_v$ and $N_g$ are included in the viscous loss term, in case research establishes an appropriate range of values for conditions inside the constrictions during speech production. For all the modelling described in this study $N_c$, $N_v$ and $N_g$ were set to 1.

Absolute values $|u_c|$, $|u_v|$ and $|u_g|$ are used here to ensure that the two components of pressure drop act in the same direction, whether airflow is egressive or ingressive.

**Subglottal resistance**

Based on the findings of Blide et al. (1964) and Vincent et al. (1970) discussed in Section II.3.3.4., the flow conductance of the subglottal airways $G_{sg}$ is made to increase linearly with lung volume up to a maximum value $G_{sg,\text{max}}$. Conductance is zero and resistance infinite when lung volume decreases to the residual volume RV.
As shown in Figure 18:

For \((V_t - RV) < 0.45 \, VC\)

\[
\frac{1}{R_{sg}} = G_{sg} = \frac{G_{sg\text{max}}(V_t - RV)}{0.45 \, VC}
\]  
(III.26)

For \((V_t - RV) \geq 0.45 \, VC\)

\[
\frac{1}{R_{sg}} = G_{sg} = G_{sg\text{max}}
\]  
(III.27)

Residual volume \(RV\) and vital capacity \(VC\) are given values appropriate to the type of speaker to be simulated.

A flow-independent ohmic resistance is assumed so that:

\[
P_{\lambda} - P_{sg} = \frac{U_{g}}{G_{sg}}
\]  
(III.28)

The larger subglottal conducting airways, generations 0 to about 12, which contain nearly all the subglottal ohmic resistance are assumed to have rigid walls, with flow rates well below those at which dynamic compression begins, so that:

\[
U_{g} = U_{\lambda}
\]  
(III.29)

### III.1.3.4. Composite constrictions

Where two adjacent parameter points are made to move together so as to define a single long constriction, one of the two points is arbitrarily chosen as the vocal tract constriction for use in the aerodynamic model. No allowance was made in the simulations described in this study for different lengths of constrictions; the value of \(K_e\) was not altered in this case.
For two distinct constrictions of the vocal tract, for consonants with two different places of articulation, on the other hand, their distinctness needs to be maintained in the articulatory and acoustic blocks of the model, but they have to be represented by a single equivalent orifice in the aerodynamics. This is an appropriate procedure if they are close enough together for the volume enclosed between them to be small compared with the volume of the whole enclosed cavity of the vocal tract. The turbulence term only is used here, as follows and as shown in Figure 19:

\[
\begin{align*}
P_x - P_y &= \frac{(0.00076 U_c)^2}{A_x^2} \\
\text{or:} \\
P_x &= \frac{(0.00076 U_c)^2}{A_c^2}
\end{align*}
\]

Where

\[
\frac{1}{A_c^2} = \frac{1}{A_x^2} + \frac{1}{A_y^2}
\]

\(A_c\) is the equivalent composite constriction. It is obtained as shown here, in the articulatory block and is used as the single orifice in the aerodynamic block.
III.1.4 Acoustic sources

Introduction

Three types of acoustic source need to be modelled: voice, turbulence noise and transient. At least two, or sometimes three, turbulence noise sources are considered to be significant: aspiration noise at the glottis, frication noise at the main constriction or constrictions of the vocal tract, and frication noise at the velopharyngeal port. The intention is to describe the sources by means of empirical formulae based on observations of natural speech. The basic physics of the sound generating mechanisms is not modelled directly in contrast to, for example, the methods of Flanagan et al. (1975) and others, reviewed in Sections II.3.4 and II.5.5. Results from these true physical models provide valuable data for the phenomenological approach used in the present study, as discussed in Section II.3.4.

It is an important principle in the modelling that the acoustic sources should depend upon the prevailing physical conditions in the respiratory tract, specifically conditions at and near severe obstructions to the flow of air. It is here that a slowly changing airstream is partially converted into rapid changes of flowrate and pressure with frequency components in the acoustic range. Appropriate constriction geometry combined with appropriate aerodynamic conditions are needed. For voice, in addition, the mechanical state of the vocal fold tissues is relevant. The sources need to occur across a range of conditions but, within that range, their waveshapes should vary. Suitable threshold conditions need to be specified for the onset and offset of each source.
In natural speech acoustic frequencies may arise at many places simultaneously, but in the model acoustic sources are confined to those constrictions which are included as orifices in the aerodynamic modelling, as shown in Figure 17. The time variations of these constrictions and of the volumes of the enclosed cavities directly reflect changes in articulatory configuration as the articulators move. The resultant computed time variations in aerodynamic conditions at each constriction combine with the time varying articulation at that constriction to generate acoustic sources with time varying properties. In this way the acoustic sources are made to vary with articulation in a realistic way, as determined by an internally consistent representation of the processes linking them.

Voice (VOICE in the model) is derived from conditions at the glottal orifice; turbulence noise, called aspiration noise (ASP), is generated at the glottis also. Both these sources are assumed to be localised at a point just downstream from, that is just vertically above, the glottis. The vocal tract constriction included in the aerodynamic model is generally that associated with the consonants to be modelled; the turbulence noise generated there is called frication noise (FRIC). This source is considered to be located at a point somewhere downstream from the constriction: its position needs to vary depending upon the place of constriction for the consonant, as discussed in Section II.3.4.12. When the main vocal tract articulator is the tip-blade region of the tongue, frication noise is considered to be in the teeth section of the vocal tract tube. In nasal modelling, turbulence noise generated at the velopharyngeal port is located at the constriction AV and is called nasal frication (FRICV). A transient source (PULSE) is associated with the main vocal tract constriction and is taken to be at the same place as the frication noise source (FRIC).
Because of the aerodynamic contribution, the acoustic sources are not, in general, independent of each other nor of the vocal tract filter shape. For example, when the vocal tract is severely constricted for a fricative consonant oral air pressure rises and conditions for frication noise improve. But at the same time, since subglottal pressure remains nearly constant in the model, the pressure drop across the glottis decreases. Under reduced pressure drop the voice waveshape in the model is made to change, becoming weaker acoustically and lower in fundamental frequency. Aspiration noise is affected by the change in transglottal pressure drop also. In this way the various sources in the model covary in, it is hoped, a natural sort of way.

III.1.4.1 Voice

Most of the data referred to in Section II.3.4 are for a man with a normal and not dysphonic voice. This is the type of speaker to be simulated by the model. The voicing model is intended to be flexible enough to represent different speaker types within the normal range.

The voice source in the model is a parametric representation of $U_g$, the acoustic component of volume flowrate of air through the glottis, as shown in Figures 12 and 13. The intention is to give the wave parameters realistic values which change with some of the controlling parameters in a way consistent with observations of natural speech, as suggested in Figure 14. For some simulations it is convenient to bypass the aerodynamic modelling and obtain a source of excitation for the vocal tract by calculation from specified time functions of the amplitude and fundamental frequency of a voice waveshape. This approach was used for example in the diphthong study described in Section IV.1.2, where the vocal tract was vowel-like throughout. The emphasis there was on suitable vocal tract articulations to yield
appropriate formant patterns rather than on acoustic sources. For transfer function studies flat frequency spectra can be generated.

Generally, the voice waveshape description is to be derived from a combination of articulatory and aerodynamic parameters whose values are obtained in the articulatory and aerodynamic blocks of the model. For the modelling of the processes which give rise to voice two parametric models are used, based on Rosenberg (1971) or on Fant (1979a, 1979b, 1980a). These are approximate models of the source defined as independent of the vocal tract transfer function and therefore without ripples associated with acoustic source-filter interaction. The waveform is skewed appropriately to take into account some of the aerodynamic and acoustic interactions between the larynx and the vocal tract and other controlling parameters discussed in Section II.3.4.8. This is a source of type (4) in Section II.3.4.7. The waveforms used in the modelling and their parameters are shown in Figure 13.

Intonation is not a prime concern of this study but a laryngeal component for the control of \( F_0 \) called \( Q \) is included; the actual \( F_0 \) value in the modelling is made to depend on both \( Q \) and the pressure drop across the larynx \( PDIFF \), following the approach proposed by Ladefoged (1963) described in Section II.3.4.10. As \( Q \) varies the voice waveshape is made to change. The register represented is intended to be appropriate to ordinary speech; double periodicity cannot be simulated in the model and there is no jitter and shimmer. Phonation type is taken to be a continuum, corresponding to different amounts of vocal fold adduction, represented as different values for the articulatory component of glottal area \( AG \) in the modelling. For convenience labels are attached to different parts of the range: breathy phonation has a large value of \( AG \) about 0.15 cm\(^2\), normal
phonation has an intermediate value near 0.05 cm², pressed phonation has a very small value of AG, less than 0.05 cm².

It is an important principle in the modelling that conditions for voicing are best over a range of glottal area AG values; within this range VOIA, the amplitude of the acoustic component of glottal flow \( U_g \), does not vary greatly with glottal area. At either edge of this plateau for voicing the amplitude of the voice wave falls off, whether glottal area becomes too big or too small. The waveshape changes to different forms at the two edges as sketched in Figure 12 and there is an associated fall in \( F_0 \) for the pressed phonation end of the range. Other \( U_g \) waveshape parameters besides the amplitude are made to vary appropriately with glottal area AG.

As the pressure drop across the glottis increases the voice waveshape is made to change in keeping with data from natural speech. The effect should to increase the acoustic intensity and alter the spectral balance towards more excitation of higher frequencies.

The general principle in the modelling of the voice source is that its waveshape should depend on three controlling parameters: the air pressure drop across the glottis \( P_{DIFF} \), the effective vibrating mass and stiffness of the vocal folds \( Q \), and the articulatory component of glottal area AG. Generally, for most of this study, the three \( U_g \) parameters varying with the three controlling parameters are: VOIA the amplitude, TCR the ratio of the closed phase duration to the total period, and \( K \) the asymmetry or skewing factor in the Fant (1979a, 1979b, 1980a) wave formulation or FBA the corresponding skewing factor in the Rosenberg (1971) wave formulation (see Figures 13 and 20). Changes in overall acoustic intensity and in spectral balance are associated with changes in waveshape as discussed in Sections II.3.4.8 and II.3.4.9, so that the acoustic structure of the
synthetic speech signal is made to vary under different production conditions. A description of the trends observed in natural speech is given in Section III.2.4.1, where waveshape parameters in the Fant model shown in Figures 13 and 20(c) are related to PDIFF, Q and AG. The general trends, so far as they can be judged, are suggested in Figure 14.

The voice waveshape is not made to depend directly on vocal tract shape, so that different amounts of skewing for different vowel types, discussed in Section II.3.4.5, are not introduced. Differing amounts of acoustic loss across the glottis for different vowel types can occur in the model, because the fraction of acoustic pressure reflected back to the glottis depends on the pattern of cross-section area changes in the vocal tract (Fant, 1979c). In a study of vocal tract transfer functions, the first formant bandwidth was greater for [aː]-type than for [iː]-type vocal tract shapes as expected (Appendix 4, Scully and Allwood, 1982).

There are different ways of representing the dependence of voice waveshape parameters on the three controlling parameters, with different degrees of closeness to the complexities of natural speech. These are described in Section III.2.4.1. The simulation of thresholds for the onset and offset of voicing are described there also.

The modelling as described in this study does not attempt to be internally consistent as regards the mapping from the low frequency aerodynamic component of volume flowrate of air through the glottis UG and the corresponding acoustic component U_g, the voice source. The relationship between UG and U_g is not a simple one, depending as it does upon the magnitude of any continuous leak of air through the glottis and the shape of the voice wave. Airflow well above zero
during what is referred to, in an acoustic context, as the closed phase may have several possible interpretations. A first attempt has been made at using inverse filtering to establish an aerodynamic-to-acoustic mapping for two speakers (Scully, 1989).

III.1.4.2 Turbulence noise

A theoretical study by Stevens (1971) provides the basis for noise sources. The same empirical formula is used for aspiration noise at the glottis (ASP), frication noise at the main vocal tract constriction (FRIC) and nasal frication noise at the velopharyngeal port (FRICV).

Radiated sound pressure is given approximately by

\[ p_r = \Delta p^{1.5} A^{0.5} \]  

(III.31)

where \( A \) is the minimum cross-section area of a constriction and \( \Delta p \) is the pressure drop across it. A linear relationship is shown between relative sound pressure level in dB and the expression on the right hand side of Equation (III.31) plotted on a logarithmic scale (Stevens, 1971, Equation 8, p.1186 and Figure 8, p.1187). The noise source amplitude dependence seems to be neither linear nor logarithmic, but instead near to

\[ \frac{(p_{\text{noise}})^{20}}{(p_{\text{reference}})^{20}} = (\Delta p^{1.5} A^{0.5})^{0.5} \]  

(III.32)

(Stevens, op. cit., Figure 8). But for larger values of the expression on the right hand side of Equation (III.31) above, a linear relationship seems reasonably accurate and this is used in the modelling. The amplitude of each turbulence noise acoustic pressure
source is derived from the right hand side of Equation (III.31) above. The logarithmic expression is available also, but it gives a frication noise output which seems to be too sharply peaked and insufficiently spread in time as compared with frication noise in natural speech.

It may be appropriate to vary the exponents of $\Delta P$ and $A$ in Equation (III.31) for different fricatives (Badin, 1989), but there was no basis for doing so at the time of this study.

Reynold's number is important for the onset and offset of turbulence noise (see Equation (II.11) in Section II.3.3.4 and Equation (II.21) in Section II.3.4.12). It is not used in the model because the expression of Equation (III.31) above gave a clearly defined onset and offset to the noise sources, probably because of the 1.5 power pressure dependency; no threshold appeared to be needed.

If there are two important vocal tract constrictions in one simulation, for example at the lips for [p] and at the alveolar ridge for [1] and [t] in "plight" or "polite", so that a composite constriction is used in the aerodynamic modelling, then the frication noise source derived is split at a time point in between the portions with differing places of origin for frication noise. Each of these split frication noise sources is saved separately for use in the filtering block of the model.

**III.1.4.3 Transient**

This source arises from sudden changes of air pressure as discussed in Section II.3.4.13. Rate of change of pressure drop across the vocal tract constriction was used to define a transient source. Strong theoretical grounds for its location could not be found during
this study and it was assumed to be at the same place as a frication noise source associated with that constriction. The analysis by Maeda (1987) was not incorporated in the modelling.

III.1.5 Acoustic filtering

The acoustic pressure or volume flowrate waveform of each acoustic source is to be modified by resonance. The filter is the air-filled acoustic tube of the vocal tract, with its complex and time varying shape. The total area function is the relevant aspect and this comes out of the articulatory block of the model.

Different approaches to the concept of source, especially voice, and filter are possible as discussed in Sections II.3.4.5 and II.3.4.7. The choice of filter description is related to the choice of voice source representation, which is discussed in Section III.1.4.1.

The acoustic filter for all the sources is taken to be the vocal tract only, with the subglottal airways excluded. One-dimensional wave propagation along the vocal tract is assumed, valid for frequencies up to about 4 kHz, as discussed in Section II.3.5. It is advantageous to extend the acoustic modelling to higher frequencies than this in order to include, for example, frication noise segments for alveolar consonants.

The bandwidth for the acoustic signal output from the model is 5.9 kHz. The errors introduced into the modelling do not seem likely to be serious: the simulations focus on alveolar fricatives and plosives in which the noise source is effectively filtered by a small front cavity only. All its dimensions are likely to be below about 3 or 4 cm, permitting one-dimensional flow to be assumed up to frequencies somewhat above 4 kHz without unacceptably large error.
High frequency errors in the vowel-like portions of simulations are not likely to be very significant acoustically, since voice sources have little energy above 4 kHz. Significant errors at frequencies above about 3 or 4 kHz are to be expected in the response to aspiration noise since this source has a significant amount of energy at these high frequencies and is filtered by the whole vocal tract. A similar effect may be expected for frication noise sources also in those cases where there is a large amount of acoustic coupling between the front and back cavities. Perceptually, lack of modelling of cross-tube modes of vibration of the air in the vocal tract may be not very serious, because of the wide critical bandwidths for human hearing above about 3 kHz (Holmes, 1981).

The open ends of the branched acoustic tube are represented very simply in the filtering: reflection of the acoustic pressure wave occurs here but there is no attempt to provide the mouth and nose terminations of the tube with appropriate frequency-dependent impedances, discussed in Section II.3.5.

It is assumed that acoustic losses are small enough for acoustic pressure and acoustic volume flowrate to be in phase. Filtering is modelled as pressure waves reflected and transmitted at each change in the cross-section area of the air-filled vocal tract tube (Kelly and Lochbaum, 1962). Some acoustic losses must be introduced when simulating filtering since, without losses, it does not seem possible to obtain even qualitative agreement with natural speech acoustics.

Acoustic energy is lost to the subglottal system by an amount which increases with glottal area. In most of the simulations described here transglottal losses change only with the slowly changing articulatory component of glottal area AG, representing increased loss when the vocal folds are abducted. This is a very important
aspect of the modelling: before the introduction of such loss variations, too much low frequency energy was trapped in the vocal tract tube when it was severely constricted for a voiceless fricative following a vowel. More realistic transglottal losses not only cured that problem but also widened the bandwidths of the formants seen through frication and aspiration noise, making them closer to natural speech. Acoustic losses through the glottis vary also with the oscillatory changes of glottal area which give rise to the voice source. In one study transglottal losses were made to vary as glottal area increased and decreased within each voicing cycle as well as with the articulatory component AG (Appendix 4, Scully and Allwood, 1982, see also Section III.2.5).

Some loss is introduced at the lip and nostril outlets also, the amount of loss being controlled by the choice of the fixed values for the reflection coefficients here. Vocal tract wall effects are included as a time spreading or smoothing of some of the energy in the acoustic pressure wave. No acoustic losses are introduced at severe constrictions although loss of energy there is expressed in the aerodynamic model through the orifice equation, Equation (II.13), discussed in Section II.3.3.4). The aerodynamic and acoustic blocks of the model are not consistent with each other in this respect. It is considered reasonable to assume basically lossless conditions for the travelling pressure waves but to add some losses; this is a theoretical approach used in other complicated systems (Titze, personal communication).

Turbulence noise and transient sources are expressed as acoustic pressures and are injected into the appropriate section of the filter.
The voice source is a volume flowrate $U_g$. The magnitude of acoustic pressure $P_g$ is derived from the magnitude of volume flowrate $U_g$ through the relation

$$P_g = Z (U_g - c U_g)$$

where $Z$ is acoustic impedance (see Fant, 1960, 1970, p.27) and $A_1$ is the cross-section area of the first section of the filter, just above the glottis, the place where the voice and aspiration noise sources are injected. The filtered waveforms appear at the lip and nostril outlets of the vocal tract; here the response to the voice source is converted back to volume flowrate through an expression like that of Equation (III.33), but with the area of the final outlet section of the vocal tract instead of $A_1$.

Each source is filtered in turn. If two vocal tract constrictions have been combined to form one composite constriction for the aerodynamics, the two halves of the split frication noise source are filtered separately, with the two frication sources injected at different places in the vocal tract. For example, in "plight" and "polite" a frication noise source for the portion containing [p] could be injected at the lips, while that for the portion containing [t] could be injected at the teeth; frication noise associated with the [l] would not be correctly filtered. This crude method of partially overcoming the limitations of the aerodynamic system as modelled, with only one orifice in the oral vocal tract, must introduce errors and prevent the full complexity of natural speech, with shifts and overlaps of noise source location, from being realised in the synthetic speech from the model. For the "plight" - "polite" series modelled, complexity in the modelling was avoided by generating a single frication noise source. Realism for the [p] was
sacrificed by injecting this source at the teeth, an appropriate location for the constrictions associated with [l] and [t].

Spectrograms of the synthetic speech signals are shown in Appendix 3 (Scully, 1987, Figure 17, pp.136-138).

### III.1.6 Radiation

As an approximation to the transform from acoustic volume flowrate at the lips or nostrils or both to acoustic pressure at a microphone some distance away from a speaker, the far-field acoustic pressure, mentioned briefly in Section II.3.6, high-pass filtering of +6dB/octave is applied to the output signals of volume flowrate for the filtered voice waveform. Although the outputs for the other filtered sources of aspiration noise, frication noise and transient are derived as pressures, not flows, they are high-pass filtered in the same way. This theoretically inappropriate treatment is not likely to give serious problems, for non-nasal sequences at least, since all these pressures would be transformed into flows by the same factor of lip outlet area in Equation (III.33). The actual magnitudes of these radiated acoustic pressures are not important since arbitrary units are used for all the acoustic sources and responses except for the voice source and its filtered output. Suitable relative weightings for the three types of output, for voice, turbulence noise and transient, have to be found empirically.

### III.1.7 Outputs

In the modelling described here the velum is raised, the area of the velopharyngeal port is zero and there is no acoustic output from the nose.
There is no representation in the model of sound radiated from the walls of the vocal tract, nor of transmission across any solid barriers; the continuation of voicing as an acoustic output when the vocal tract is completely closed, as in a fully voiced plosive, cannot be simulated.

There are generally four oral output waveforms: the responses to the voice, aspiration noise, frication noise and transient sources. When a split frication noise source is introduced there is a fifth output component. The four or five signals are added together with a weight given to each. It is considered important to give the same weights to both frication signals, and preferable to give the same weights to the aspiration noise and frication noise signals, since both types of source are derived from the same expression for similar physical processes. If aspiration and frication weights do need to differ for the most natural-sounding effect, this can be justified on the grounds that the turbulence-creating geometries differ, as discussed in Section III.1.4.2, and that the approximations and errors introduced by the filtering methods used are likely to differ in the two cases, as discussed in Section III.1.5.

Weights are chosen by trial and error with auditory assessment and spectrographic comparisons with natural speech. Once a set of weights has been selected it is kept constant throughout a related series of simulations. The aim must be to find weights which will be suitable for simulation of any speech-like sequence, at least while one speaker is being modelled, but this will require a better theoretical basis for the generation of all the sources in the modelling and, in particular, a better representation of the complex physical processes involved in the generation and filtering of turbulence from a jet emerging from constrictions at different places along the vocal tract (Shadle, 1989).
The final output from the model is a single waveform. Like all previous stages of the modelling, this is digitally described. The signal is passed through a 12-bit digital-to-analog converter and then, as an analog signal, through an anti-aliasing hardware filter. This gives an analog waveform with a bandwidth of 5.9 kHz. It is heard over headphones and recorded onto audio tape.

Many graphical displays are available in the model, in all its stages. These outputs can be assessed and compared in detail with corresponding traces from natural speech. The methods used are described in Section III.3, but only briefly there, since articulatory, aerodynamic and acoustic sources graphs from the modelling are included and assessed throughout the experiments outlined in Section IV and reported in the Appendices.
III.2 Implementation

Introduction

The block structure of the model is shown in Figure 21. The blocks are organised into three main computer programs, which will be referred to in this study as follows:

- **ARTIC** articulation;
- **AERO** aerodynamics and construction of the acoustic sources;
- **FILT** filtering of the acoustic sources;

An auxiliary program called **SIGNAL** in this work provides spectral shaping and frequency analysis of waveforms. It is used to shape noise sources with the command POLE and to examine the formant patterns of the synthetic speech generated by the model with FFT. It includes **ZERO** for the radiation transform. Separate programs called **LOPASS**, **SEND** and **XFER**, all subsumed under the general label **OUT** in this work handle the final processing and outputting of the synthetic speech signal.

Each program contains parameters which can be reset from their default values; these define physical constants, speaker types and initial conditions.

An articulatory scheme is defined in **ARTIC** by the inputs; actual movement paths are calculated and viewed within **ARTIC** and are saved as output files. The files saved from **ARTIC** form inputs for both **AERO** and **FILT**. In general, files saved from an earlier stage are read in as inputs to later stages.
Each program has a menu from which options need to be chosen in the right order when performing a simulation of natural speech. The selections can be made from the keyboard or, more conveniently and more reliably, from a previously prepared command file, which may include calls for file names to be entered at the keyboard. When END is reached in the command file the program menu is displayed and control returns to the experimenter at the keyboard.

The programs are interactive, with many options for graphical displays; these facilities allow an experimenter to explore the movement paths resulting from different articulatory schemes and the aerodynamic consequences also. At each stage, the effects of changed parameter values can be assessed visually and, at the filtering and output stages, auditorily. Generally, a first attempt is defined by an articulatory file. This is run and the results are viewed. The instructions for the scheme are modified via the keyboard; most of the modelling time is given to these explorations of variations on a basic scheme.

In the first place the aim is to obtain a set of movement paths in ARTIC that displays appropriate inter-articulator coordination and transition end point values, avoiding unintended undershoot where insufficient time has been allowed for all transitions to be completed, and avoiding unintended steady state portions of time. The aerodynamic consequences are determined with AERO and assessed, usually by comparison with corresponding traces from some examples of natural speech. For an improved match at this stage it is necessary to return to the articulatory stage, except in cases where the values of aerodynamic parameters themselves are being investigated. It is often helpful to construct in ARTIC, or in an associated program called VOCAL, a number of different component path. For example, within a given time scheme the glottis might be increased to
different maximum values for a fricative, or there might be different rates of lung volume decrement to define different respiratory gestures, or a change in glottal area might be differently timed. Each one of the set of related movement paths is given a different name when it is saved as a file from ARTIC or VOCAL; different combinations of larynx, respiratory and supraglottal articulations can be tried out in AERO. A small repertoire of actions for two or three different articulators can be made to generate a large number of different aerodynamic traces.

Once a set of appropriate-looking aerodynamic traces has been obtained, the acoustic sources are derived, within the same AERO program. If the sources appear to be appropriate they are saved as output files from AERO and combined with the vocal tract shape saved from ARTIC, in the filtering program FILT.

The four responses of the vocal tract, to each source separately, or five if there is a split frication source, are saved from FILT as separate files; each one is high-pass filtered for radiation and then is down-sampled by a factor of 3 with low-pass filtering. The four waveforms are added together in the output program SEND. The combined waveform is the final digital signal which is sent from the VAX11/780 to a PDP11/03 for output via a 12-bit digital-to-analog converter. The small capacity of the PDP11/03 microprocessor limits each simulation to 3 sec maximum duration. The final processing before recording is low-pass filtering of the analog signal by a hardware anti-aliasing filter.

All the modelling is in software. There are different sample times at different stages of the model: 5 ms for articulation, 5 ms for aerodynamic traces but much smaller internally during the solution of the differential equations, 28 μs for acoustic sources, 14 μs
internally during filtering, 28 $\mu$s for the signals output from the filtering stage, and 84 $\mu$s after down-sampling with low-pass filtering, giving an acoustic signal with a Nyquist frequency of 11.9 kHz. The anti-aliasing filter gives a final output signal with a bandwidth of about 5.9 kHz (60 dB down by half the Nyquist frequency 5.95 kHz).

When a series of routine runs through the model is to be done, for example if some aspect of articulation is being altered in steps, overnight batch processing is often used. A command file is set up with calls to run the various programs: for inputting, saving and naming of files, including graph plot files, deletion of intermediate files, resetting of switches as appropriate, and for saving and keeping on disc the acoustic output digital waveforms ready to be mixed and output.

The model is implemented mainly in FORTRAN. Some non-standard VAX FORTRAN features are used. Some routines are written in MACRO, the VAX assembler code, for efficiency. The output program for the PDP11/03 microprocessor, where maximum computation speed is required, is written in PDP11 assembler code.

The VAX11/780 computer system runs under a VMS operating system, with 2.5 Mbyte of main memory at the time of this study and a 32-bit word. The system included two RM05 exchangeable discs each of 256 Mbyte and one RP07 fixed disc of 512 Mbyte. Half-inch magnetic tape is used for back-up storage of programs and for output files.

Each sample of all the files produced in the modelling, whether intermediate or output, is represented as a real number requiring 4 bytes of storage. These are data files, with names such as VOICE123.DAT, but .DAT need not be specified in the modelling. For
the acoustic output signal sampled at 84 $\mu$s, one second of synthetic speech occupies approximately 100 blocks of storage.

Each program in the modelling, ARTIC, AERO, FILT and the auxiliary programs also, contains a HELP file. The computation methods for the modelling are described in detail by Allwood (1984).

III.2.1 Inputs

There are default parameter values in all blocks of the model. To change a parameter its label followed by a new value is entered, either from the keyboard or by calling up a file which includes the command. Signals generated in each block of the model can be viewed as graphs, as outlined below, and can be listed numerically with the command SHOW.

(1) A file for ARTIC usually contains the following commands in the order given below:

TPOS defines and lists on the screen the D values for each parameter point. DL defines vocal tract length. Fixed values such as AV=0 and fixed relationships such as AT = AJ=0.5 can be included here.

MARK shows the timing and coordination expressions and allows changes to be made to any of them and to values for the durations in the expressions. MARK Event and Duration time values appear automatically in the set of commands for an articulator if the TMOD instruction lines include MARK labels to define the timing of the beginning or end of some of the transitions. See Appendix 3 (especially Chapter IV, Scully, 1987, p.125 ff.) for a fuller
description of the organisation of the coordination expressions which are listed under MARK and of the way the labels are incorporated into the instructions for the articulators.

TMOD contains instructions for each articulator separately. The order is not important. The LIST (or L) command puts an updated set of instructions for the current articulator onto the screen.

ACAL computes and displays the individual articulatory paths. A list of articulator labels, DVLU, AG, AE, AP and so on, is specified. Here also the order is not important.

LINK connects the vocal tract parameter points together by interpolation of area values between adjacent points along the tract and constructs the total vocal tract area function at each sample time. The points to be linked are listed as AE, AP, AN and so on. Not all the vocal tract points specified in ACAL are necessarily included, but those that are must be listed in the correct sequence, going from the glottis towards the lip outlet.

FOUR TYPES OF VISUAL DISPLAY are available if wanted, viz.:

PLOT displays graphs of the vocal tract area function, i.e. cross-section area versus distance from the glottis along the vocal tract D, or of cross-dimension versus D, at specified sample times. Alternatively cross-section area or cross-dimension is plotted as a function of time for a specified articulator label, or for a specified D value along the vocal tract.

LOTS is used to display inter-articulator coordination, with multiple graphs showing parameter values against time.

OVER is similar to LOTS but here the graphs are overlaid for detailed comparison.

AXES can be specified for LOTS and OVER, to give the same scaling or a specified scaling factor; otherwise scales are adjusted automatically to fill the screen.
TRAC is a side view of the vocal tract overlaid at specified time samples, derived from the area function by the area to cross-dimension mapping set out in Section III.1.2, Equation (III.1).

TMOD can be used to modify instructions and ACAL can be repeated, for any of the articulators. If the overall vocal tract shape is to be recalculated it is first reset with the command ZERO. If on the other hand local modification is required so as to represent departure of the main vocal tract articulator for a consonant from the overall tongue body configuration, as described in Section III.1.2 and Appendix 3, Figure 6 (Scully, 1987, p.93), ZERO is not used. Instead TMOD and ACAL are performed, for example for AF if the consonant is an alveolar one; LINK with AB, AF, AJ connects the F point to points either side of it, thus modifying the shape between the B and J points.

COMP is needed only if two different vocal tract parameter points need to be combined to form a composite aerodynamically equivalent single constriction AC in AERO, as described in Section III.1.3.4, Equation (III.30).

Files needed for the next block of the model AERO must be saved in ARTIC. The following menu items are chosen; the order is not important:

ASAV saves the articulatory paths needed in AERO: the main vocal tract constriction area AC, the glottal constriction area AG, the velopharyngeal port area AV, the respiratory articulator DVLU or PLU, and the vocal fold mass and stiffness factor Q. Each file has TR added to its label. For an alveolar consonant the path for AF would be saved as a file labelled AFTR, to be used as AC in AERO.

VOLS is used to compute the cavity volume enclosed between two points, generally from the glottis to the main vocal tract.
constriction, AC in AERO, AF in this example. The volume as a function of time is saved as VSFTR in this example, to be called VC in AERO. If AC in AERO is to be a composite constriction from ARTIC, say AF and AL combined, then one of these parameter points is used for the cavity volume, for example this might be VSFTR between the S point just above the glottis and the F point. Significant errors would be introduced into the aerodynamics if the two points were far apart.

ALTS saves the whole vocal tract area function specified at each 0.5 cm section along the vocal tract and at each 5 ms sample time, for use in FILT.

TSAV can be used if the articulatory instructions of TMOD are to be saved. This is not needed for later stages of the modelling.

TGET, AGET and ALTG can be used in ARTIC if articulation is to be defined from previously saved instructions, articulatory paths or an area function respectively.

An auxiliary program called VOCAL allows individual articulatory paths to be constructed from TMOD and ACAL, without the time-consuming computation of the vocal tract area function.

(2) A file for AERO usually contains the following commands in the order given below:

SGET reads in the files needed, viz.:
DVLU or PLU (or PP, not used) filename, e.g. DVLUTR123;
Q, filename, e.g. QTR123;
AG, filename, e.g. AGTR123;
AC, filename, e.g. AFTR123;
VC, filename, e.g. VSFTR123;
AV, filename, e.g. AVTR123.
The filenames are either specified in the file or a filename can be requested from the experimenter, to be typed in at the keyboard. The second of these options is more flexible, allowing one AERO file to be used for a series of simulations.

The file can include a call for another file which specifies values for all the parameters needed to simulate a particular larynx type in the voice source generating portion of AERO.

FLOW calls up the aerodynamic calculations to solve the simultaneous differential equations set out in Section III.1.3. A method for the numerical integration is selected, then a threshold value for PC the pressure drop across the vocal tract constriction, then a tolerance value for the numerical integration.

WAVE computes all the voice waveshape parameters and also the amplitude envelopes for aspiration noise ASPA, frication noise FRICA and frication noise at the velopharyngeal port FRICAV in nasal simulations.

SORV constructs the voice source waveshape VOICE. This must be done before the noise sources are constructed because their waveforms are modulated by VOICE.

The following commands construct the noise sources:

SCAL allows signals to be defined analytically.

RANDOM is chosen, with a sample time specified as 0.000028 sec (28 µs). Minimum and maximum amplitudes, generally -1 and 1, and two intial values between 0 and 100 are specified. A pseudo-random number sequence is generated.

ASP is the command to construct the aspiration noise source waveform ASPS, multiplying ASPA by the pseudo-random number sequence at each time sample.

FRIC similarly, constructs the frication noise source FRICS.

FRICV similarly, constructs FRICSV.
SAVE is used to save each acoustic source in turn onto disc. A filename is specified for each one, either from the keyboard or, in batch processing, from a file.

LOTS, OVER, AXES are used as in ARTIC for graphical displays. The down-sampling for plotted points is specified. This is important for the acoustic waveform files which have sample times of 28 \( \mu s \) and 35,714 samples in one second of synthetic speech. If ALL is specified after the graphics command a suitable number of samples between plotted points is chosen automatically.

AERO contains a large number of parameters which can be listed on the screen as follows:

- FLIP lists the FLOW parameters needed for the aerodynamics;
- SLIP lists the acoustic sources parameters;
- WAIL lists the wave table and other voice source parameters;
- TIME lists, for all the signals created, the sample time, the number of samples and the total duration.

(3) A file for FILT usually contains the following commands in the order given below:

- SGET reads in the files needed for filtering of the voice source:
  - INPUT, voice source filename, e.g. VOICE123;
  - AG, filename, e.g. AGTR123;
  - AV, filename, e.g. AVTR123;
- ALTG reads in the vocal tract area function file. The filename is specified, e.g. ALT123.

The number of samples is determined by the INPUT file and must be extended with added zeros to allow the filter response to die away.
NSAMP is specified as about 1000 more than the number of samples in the voice source file.

INJECT is a switch set to zero for sources at the glottis.

RESL is the command to perform the reflected pressure wave method of filtering with losses included.

SAVE, RESPONSE are the commands to save the response signal. It is given a filename, e.g. RONE.TMP.

The following commands are used to obtain and save the response to an aspiration noise source, injected at the glottis:

SGET, INPUT, filename, e.g. ASPS123, NSAMP reset to give 1000 extra samples, RESL, SAVE, RESPONSE, filename, e.g. RTHREE.TMP.

Nearly the same commands are needed for the frication noise source, as follows:

DP is specified as the injection point for the frication noise source, for example DP=16.5 cm generally corresponds to the distance from the glottis of the teeth; this is used for alveolar consonants.

SGET, INPUT, S000 to read in a dummy source of zeros at the glottis.

SGET, FRICATION, filename, e.g. FRICS123 to read in the frication noise source.

NSAMP reset to give 1000 extra samples, INJECT, switched to 1 for a sound source not at the glottis, RESL, SAVE, RESPONSE, filename, e.g. RTHREE.TMP.

(4) For the radiation transform, each of the response files saved from FILT is given a +6 dB/octave frequency shaping from input commands to the signal processing program SIGNAL as follows:

INPG to get a response file, its filename, for example RONE.TMP;
ZERO to apply high-pass filtering +6 dB/octave;
 RSAV to save the high-pass filtered signal. The output of ZERO is given a filename, RZONE.TMP.

This file is low-pass filtered and down-sampled in a program called LOPASS. Default parameter values are always used and so they do not need to be specified by inputs. The output file is saved as RZONEF.DAT. The prefixes and the affix to the filename show that it is a Response, R, has gone through ZERO, Z, and has been low-pass filtered with down-sampling, F. The intermediate files RONE.TMP and RZONE.TMP are deleted.

(5) The three or more output signals RZONEF, RZTWOF and so on are added together, with specified weightings, by means of the command MIX in the output programs SEND and XFER.

III.2.2 Articulation

Articulation is implemented as an interactive graphics program ARTIC. Each of the articulators listed in Section III.1.2 is given a set of commands or instructions to define a sequence of transitions. Figure 15 shows the complete set of articulatory parameters, including the parameter points for the vocal tract. Time is expressed in 5 ms units throughout ARTIC. The word "I" [ai] will be used as an example of the organisation of the instructions, in Section IV.1.2. Appendix 3 (Scully, 1987) describes the full capabilities of the articulatory block including the use of MARK with its set of D and E parameters for the construction of a quantitative but flexible time scheme for articulation. Some of the studies described in Chapter IV made use of event labels like those in MARK in order to vary inter-articulator coordination. In this section time will be treated in a non-parametric way.
The sequence of options from the menu in ARTIC is as follows:

**TPOS** The D values of all the vocal tract parameter points are defined. DS is always zero. DL defines the total vocal tract length from the glottis to the lip outlet.

**TMOD** Instructions for articulatory paths are set out and modified if necessary. They can contain up to 20 lines, each defining one transition. The articulator may be any of the ones listed in Section III.1.2 and also D values for the vocal tract parameter points.

Each line contains several columns, viz.:

1) a line number, starting with 0;

2) an articulatory end point value which may be preceded optionally by a type code. The default code is TO for an absolute end point value; BY defines a relative one. TO allows local modification of an already defined tongue shape. The experimenter does not need to calculate how far the tongue tip, for example, should be moved so that it may reach a specified small cross-section area AF for [s], starting from the value it has as part of the overall tongue body shape. BY 0 means an already calculated current end point. It is used to make the AF value, for example, return from a small value for [s] back to its previously defined value as part of the overall tongue body shape. AT 0 means that the articulator's path is defined by links between the points adjacent to it whose paths are defined as parts of the tongue body articulation. Thus, the AF value can be calculated as part of the overall tongue body contour until its path needs to be specified precisely for [s]. In this way the F point can execute a smooth path, changing its role from an undefined part of the tongue body to the main vocal tract articulator and back again, as shown in Appendix 3, Figure 6 (Scully, 1987, p.93);
3) the time allowed for the transition in 5 ms units;
4) the transition type, which has one of two forms. In one there is a code showing the time needed to complete a transition: S for Slow, 38 time units or 190 ms; M for Medium, 24 time units or 120 ms; or F for Fast, 8 time units or 40 ms. In the other, there is a duration in 5 ms time units followed by a slope factor which defines the relative gradient at the mid and end portions of the transition; 1 means a linear transition while a large number, say 10000, means a gradual acceleration and deceleration with high velocity mid way between the end points. The S, M and F transition types correspond to slope factors of about 100 (see Appendix 3, Figure 5, Scully, 1987, p.92).

The use of the commands ACAL, LINK, TMOD to construct an overall vocal tract shape and to modify it, with COMP to construct a composite constriction and VOLS to compute the volume of air between the glottis and a constriction of the vocal tract, has been described already in Section III.2.1. Examples of articulatory instructions and the resultant articulatory paths are shown in Chapter IV and the Appendices; in particular, the derivation of an [ar] diphthong is shown in Section IV.1.2.

The sigmoidal function used to construct movement paths for the articulators imposes a path of the form shown in Figure 16 with a specified constant duration regardless of the magnitude of the change in cross-section area between the start of the transition and its end: 38 time units, 190 ms for S, the slow transition; 24 time units, 120 ms for M, the medium transition, and 8 time units, 40 ms, for F, the fast transition; or a path of the form shown in Appendix 3 (Figure 5, Scully, 1987, p.92) with a specified duration in 5 ms units.
ACAL computes the cross-section value for the articulator at each 5 ms time sample and thus interpolates between the area values specified for the start and the end point of the transition.

As each of the vocal tract articulator's series of transitions is computed, its slot in the ALT table, of cross-section areas, distance from the glottis and time, is filled in. To obtain the total area function of the vocal tract interpolation is performed in distance, between some or all of the values already defined, in the filled slots. The interpolation, which is called by the LINK command, is derived from an exponential curve. There are four types of link available, but generally the default type is chosen. This gives zero gradient dA/dD at each end of the interpolation, that is, at the already defined parameter points, to ensure a smoothly changing cross-section area along the length of the vocal tract. One of the other options would permit a flared outlet at the lips.

III.2.3 Aerodynamics

The files needed for the aerodynamic block of the model, which is the first part of AERO, are as listed in Section III.2.1, viz.:  
  DVLU or PLU (or PP), the respiratory articulator;  
  Q, the effective mass and stiffness of the vocal folds;  
  AG, the articulatory component of glottal area;  
  AC, the cross-section area of the vocal tract constriction, which may be a composite of two constrictions in ARTIC;  
  VC, the cavity volume enclosed behind the vocal tract constriction;  
  AV, the cross-section area of the velopharyngeal port.

Default values of the parameters and the values of the physical constants needed in FLOW to solve the simultaneous differential
equations of the aerodynamics are shown on pages 15 and 16. They are listed to the screen by the command FLIP. They include physical properties of the air, wall compliances CLW and CCW for the lungs and vocal tract cavity respectively, parameters needed to calculate subglottal airways flow conductance \( G_{sg} \), and initial conditions.

The equations to be solved are set out in Section III.1.3 and are referred to here by the same equation numbers. For the mapping from the symbols used in the equations to the parameter symbols in the model, see page 14. The input articulatory paths for DVLU or PLU (or PP), Q, AG, AC, VC and AV, all defined at 5 ms intervals in ARTIC, are interpolated, generally to a 1 ms sample time, if the solution method requires it.

A time-marching method is used. At time \( t = 0 \) lung air pressure PLU, oral air pressure (and also pressure drop across the vocal tract constriction) PC and lung volume VLU are known; they have their initial values as defined in the FLIP parameter list (PLUO, PCO and VLUO respectively). VLU usually starts at a low to mid lung level, about 3000 cm\(^3\) or higher. RVLU the residual volume of the lungs is generally 2000 cm\(^3\) and VCLU the vital capacity is usually 4500 cm\(^3\), so that total lung capacity TLC is usually 6500 cm\(^3\), an appropriate value for a man. Generally PLU and PC are given initial values of zero and the simulation is intended to begin at the start of an expiratory phase of controlled breathing for speech. An inspiratory phase can be modelled also; this was done for the study of Appendix 6 (Scully and Allwood, 1984) for example. At subsequent sample times the values obtained by integration in the previous loop are used. The whole time span of the files is stepped through in this way.
At the current time, the value of $dV_C/dt_{act}$ (the active component of rate of change of vocal tract cavity volume from ARTIC) in Equation (III.13) is obtained from $VC$ by numerical differentiation.

If the respiratory control is lung air pressure $PLU$, then numerical differentiation gives $dPLU/dt$ at the current time.

$UG$ is obtained from $PLU$, $PC$, $VLU$ and $AG$ by combining Equations (III.25) and (III.28) to give:

$$PLU - PC = \frac{UG + NG*8T/\gamma*\mu*LG|UG|}{GSG} + \frac{(0.00076*UG)^2}{AG^2}$$  \hspace{1cm} (III.34)

If $AG$ is 0 then $UG$ is set to 0. The corresponding equations for $UC$ and $UV$ are Equations (III.23) and (III.24). These three equations are quadratics with the general form:

$$a*UX^2 + b*|UX| + c = 0$$  \hspace{1cm} (III.35)

where $a$, $b$ and $c$ are constants. Following Muller and Brown (1980), the magnitude and the direction of $UX$ are separated. It can be shown that $a$ is greater than zero and that $b$ is greater than zero also, and that of the two possible solutions:

$$|UX| = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$  \hspace{1cm} (III.36)

only the positive option yields $|UX|$ greater than or equal to zero, which must be the case; the negative option is inadmissible.

$|UX|$, the magnitude of $UX$, having been obtained in this way, its sign is chosen so as to make the direction of airflow consistent with the
direction of the pressure difference. This is (PSG-PC) called PDIFF for UG and is PC for UC and for UV.

dPLU/dt is obtained from Equations (III.20) and dPC/dt is obtained from Equation (III.13). These equations are derived by making the approximations of Equations (III.15) for the lungs and (III.8) for the vocal tract cavity. In the actual implementation of the aerodynamics, these approximations were not made. As a result, the following expressions were obtained for dPLU/dt and dPC/dt:

\[
\frac{dPLU}{dt} = \frac{((PAT+PLU) \times CLW + VLU)}{dt} = \frac{PAT \times (-UG)}{dt} - \frac{(PAT+FLU) \times dVLU}{dt} \quad (III.37)
\]

\[
\frac{dPC}{dt} = \frac{((PAT+PC) \times CCW + VC)}{dt} = \frac{PAT \times (UG-UC-UV)}{dt} - \frac{(PAT+PC) \times dVC}{dt} \quad (III.38)
\]

dVLU/\text{dt(act)} is obtained by interpolation or by calculation depending on the respiratory articulatory control chosen. If rate of lung volume decrease is the articulatory control DVLU, then dVLU/\text{dt(act)} is equal to DVLU interpolated to give a value at the current time. If, instead, the air pressure in the lungs PLU is taken to be under the speaker's control, then PLU is interpolated to give PLU at the current time and dPLU/dt is obtained by numerical differentiation.

The methods of solution depend on the respiratory control chosen: DVLU with LUNGSW= 1, PP with LUNGSW= 2 (Not used in this study) or PLU with LUNGSW= 3.

The solution methods are described more fully by Allwood (1984).

The main problem encountered was instability in the numerical integration which gave large oscillations in the aerodynamic signals
of volume flowrate and pressure. At first a smoothing routine was used to reduce the oscillations but for the work described here the problem was solved by the use of improved numerical integration methods which became available in the Numerical Algorithms Group (NAG) library and by the use of a mixed model for the aerodynamics. The advice of Dr. Malcolm Bloor is gratefully acknowledged. The problem was severe because of the disparity in the magnitudes of volume flowrates and oral pressure for vowel-like vocal tract shapes, where PC is essentially zero while UC is of the order of 50 to 500 cm$^3$/s.

When PC is zero and dPC/dt is zero also, Equation (III.13) reduces to:

$$\left( \frac{dV}{dt} \right)_{\text{act}} = UG - UC - UV \quad (\text{III.39})$$

If PC is less than a threshold value PCTOL, a composite model is established as follows:

$$UC = w*UC1 + (1-w)*UC2 \quad (\text{III.40})$$

where $w$ is the ratio PC/PCTOL. UC1 is obtained from the full quadratic equation, Equation (III.23), needed if PC is large and changing with time; this term dominates if PC is nearly as large as PCTOL. UC2 is UC in Equation (III.39) and this term dominates when PC is much less than PCTOL. For PC larger than PCTOL the full model with Equation (III.23) is used. PCTOL was always set to 0.1 cmH$_2$O for this study.
Integration is performed by one of five methods:

1) Euler;
2) Runge-Kutta;
3) Merson (NAG library);
4) Overall Merson (NAG library);
5) Overall Gear (NAG library).

Methods 4) and 5) have internal sample times and give outputs with the same sample time as the inputs, 5 ms. A tolerance is required for methods 3), 4) and 5). This should be 0.005 or less; generally 0.0000001 is used. Gear's method is particularly suitable for handling the stiff equations of the aerodynamics in which there are widely differing magnitudes for the variables.

The files generated are for PSG and PC, the subglottal and oral pressures respectively; UG, UC and UV, the volume flowrates of air through the glottis, the vocal tract constriction and the velopharyngeal port respectively; DVLU, the rate of change of lung volume VLU, and VLU itself. The graphical displays generally include both articulatory and aerodynamic traces so that articulatory cause and aerodynamic effect may be inferred. By iteration round an articulatory-aerodynamic loop, some aspects of the aerodynamic patterning in the model can be better matched to some data for natural speech, for example, the magnitude and timing of the oral airflow peak following the release in a plosive, or of the double peaks of airflow associated with a voiceless fricative.

The equations of Section III.1.3 show that, if UV is zero, UC is nevertheless different from UG if the vocal tract cavity volume VC or oral pressure PC or both are changing. From a first estimate of larynx actions the model is made to generate UC. This trace can be compared directly with oral airflow in natural speech as estimated
from airflow through a mesh in a face mask, as described in Appendices 1 and 2 (Scully, 1984a, 1986). As a result of the discrepancies observed between the two traces, the larynx actions in the model are modified and a new UC trace is generated. With this analysis-by-synthesis approach, better estimates for articulatory paths, especially for respiratory control, for glottal area AG, and for changes in the vocal tract cavity volume can be found. Quite close matches of aerodynamic traces between the model and each of two real speakers have been obtained by this method (Scully, 1989).

It did not seem practicable to analyse natural speech data with the full expressions for the relationships between pressure drop, volume flowrate and cross-section area as given in Equations (III.23), (III.24) and (III.25). Instead, the second term only, the working orifice equation Equation (II.13) was used. For consistency between the model and the analysis of natural speech, the same orifice equation was used in the model. This was easily done by setting NC, NV and NG (Nc, Nv, Ng) in Equations (III.23), (III.24) and (III.25) respectively to zero. The simplification was felt to be justified since it seems probable that the turbulence term dominates for the small cross-section areas of most interest (see Appendix 2, Scully, 1986). However, for the modelling studies described in Chapter IV and Appendix 6 (Scully and Allwood, 1984) and Appendix 7 (Scully and Allwood, 1985a), the full forms of the expressions were used.

Although lung wall compliance CLW is included in the parameters and in the aerodynamic equations, it needs to be set to a value near zero, as explained in Section III.1.3.2. A value of 0.001 cm$^3$/cmH$_2$O was used in the modelling.
When satisfactory aerodynamic graphs have been obtained, the sound sources are constructed within the second part of AERO, as described in Section III.2.4.

### III.2.4 Acoustic sources

#### III.2.4.1 Voice

Several different methods are available in the model for deriving the acoustic component of volume flowrate of air through the glottis, the VOICE waveform $U_g$ (see Figures 12 and 13). These voice sources approximate to different degrees the complexity of real speech. The waveforms used in the model are shown in Figure 20. The term 'closed phase' is used even if there is actually a continuous leak of air through the glottis during voicing. The various methods described below are all called by the command WAVE.

This section is concerned with implementation, based on the mechanisms believed to apply in natural speech as reviewed in Section II.3.4 and the general principles invoked in the modelling, set out in Section III.1.4. A brief account of the range of voice waveshapes found in natural speech under different conditions is included in this section, rather than in Section II.3.4, in order to relate these findings to the set of voice waveshape parameters used in the modelling; the general trends discussed in this section are suggested in Figure 14. In one of our implementations of the Fant model (see Figures 13 and 20(c)), a table of values (see Figure 23) shows the trends for three wave parameters as linear functions of three controlling parameters; in other implementations individual wave parameters are defined as functions of individual controlling parameters, but in both cases the values and functions used are intended to reflect some of the trends indicated in Figure 14.
1) The simplest method gives a train of pulses with specified amplitudes and repetition rates. This pulse amplitude can be considered to represent the absolute value of the maximum wave gradient $\frac{dU_g}{dt}$ at or before closure, indicated by USLOPE in the model. The magnitude of this is closely related to the amount of excitation of higher frequencies of the vocal tract resonator. This form of voice source is sketched in Figure 20(a). A single pulse can be used to obtain an approximation to the transfer function of the vocal tract, defined as its response to a single infinitely short impulse. Alternatively, a time function $\frac{\sin x}{x}$ is available for this purpose, with a band-limited flat frequency spectrum.

2) Another option, based on Rosenberg (1971) and shown in Figure 20(b), gives a $U_g$ waveform in which the open phase is built up from a half cycle of a sinewave of frequency $F_r$ for the rising flow portion and from a quarter cycle of a sinewave of frequency $F_f$ for the falling portion. The ratio $F_f/F_r$ is called FBA. There are three parameters besides $F_0$: FBA, which is half the ratio of the durations between rising and falling portions to control the skewing; the ratio of durations between the closed phase and the total period TCR; and the amplitude VOIA. These are similar to parameters found by Sundberg and Gauffin (1978) to be useful and acoustically relevant in the description of glottal flow waveshapes derived for real speakers by inverse filtering.

As a modelling short-cut, a natural-sounding voice source wave can be constructed by specifying contours of amplitude and $F_0$ for a Rosenberg wave, with fixed values of FBA and TCR. Cubic or other polynomial splines are used to make VOIA rise from zero to a maximum and then fall again, and to give $F_0$ a falling shape, for example. This is useful when the focus is on vocal tract shapes for vowel-like
sounds rather than on the voice source, as in the diphthong study described in Section IV.1.2.

3) The most versatile parametric description of voicing in the model and the one which can be best made to match natural speech is that due to Fant (1979a, 1979b, 1980a), sketched in Figure 20(c). There are three parameters besides $F_0$. Here, as in the Rosenberg wave, the balance between the rising and falling portions and between the closed phase and total period can be controlled, by parameters $K$ and TCR respectively, but breathy voicing is better represented, with the possibility of zero slope $dU_g/dt$ at closure. The wave amplitude is given by the parameter VOIA.

The Fant parametric representation is used for much of the modelling described in this study and is shown in detail in Figure 13. This includes the equations through which the parameters are related to each other. The complete set of Fant wave parameters, given the labels used in our modelling, comprises: amplitude VOIA, closed phase duration TC, period of the wave $T_0$, closed phase to period duration ratio TCR, asymmetry factor K, magnitude of negative gradient at closure USLOPE, closing duration TD derived from VOIA and USLOPE, fundamental frequency $F_0$ (VOIF), frequency of the rising portion $F_g$ (GLOF), the ratio $F_0/F_g$ (VGR).

K is an asymmetry factor: as K increases the wave becomes more skewed to the right and higher frequencies are strengthened while the amplitude of the first harmonic remains essentially unchanged. $K = 5$ with $T_h/TD = 6$ would represent extreme skewing. The minimum possible value of K is 0.5 with TD infinite; this gives an extremely prominent first harmonic relative to higher ones.
The rising portion is half a cycle of \( F_g \) (GLOF in the model) which appears in the frequency spectrum of the speech signal as a low frequency baseband source peak. \( F_0/F_g \) (VGR in the model) is an indication of the voice type. Its value is generally near 1 but for a weak voice the rising portion may occupy less than half the total cycle time and then \( F_g \) is significantly higher in frequency than \( F_0 \). \( F_g \) is sometimes visible below \( F_1 \) on spectrograms for open vowels with high frequency \( F_1 \), giving a split formant effect which can be confused with spectral complexity due to nasalisation.

USLOPE is the absolute magnitude of the negative flow gradient \( dU_g/dt \) at closure. As discussed in Sections II.3.4.1 and II.3.4.9, it is an indicator of the spectral balance of the voice source, especially in cases where \( K \) is greater than or equal to 1, since the gradient at closure is then the maximum value within the cycle. At frequencies above \( 2F_g \), the spectrum level is proportional to USLOPE. For \( K \) less than 1 the steepest gradient occurs before closure. If USLOPE is large, with sharp cut-off of airflow, there is strong excitation of higher frequencies.

The area under the curve of \( U_g \) versus time \( t \), the total volume of air in one cycle, omitting any steady d.c. flow, is an indicator of the acoustic intensity of the low frequency part of the voice source spectrum up to about 500 Hz.

If VOIA increases with TD constant, both USLOPE and the area under the curve increase so that the acoustic intensity is increased across the spectrum.

In order to construct a voice waveshape, values for \( F_0 \) and one of several possible combinations of three parameters chosen from VOIA,
TD, USLOPE, TC or TCR, K, GLOF, VGR needs to be chosen. The combinations available in the model are shown below in this section.

In AERO a switch WMODE is set to 1, 2 or 3 to select the pulse train, Rosenberg or Fant wave model respectively. If the switch SAME is set to 1 the wave parameters are kept constant at values set by the following parameters: PW pulse width, FBAC the skewing in the Rosenberg wave, TCRC the closed phase to period ratio in the pulse train, Rosenberg or Fant method, and KC the skewing in the Fant method. Default values are PW=2, FBAC=10, TCRC=0.3 and KC=1.

The acoustic properties of the voice source can be related to some controlling articulatory and aerodynamic conditions to a greater or lesser extent if the switch SAME is set to zero. Choices are available for each WMODE selected.

1) In one formulation the amplitude VOIA is made to increase as a function of the aerodynamic component of pressure drop across the glottis PDIFF. In addition, it decreases as the articulatory component of glottal area AG increases above a specified value AGMAX which defines one edge of the voicing plateau discussed in Sections II.3.4.8 and III.1.4.1. The following expression is used:

\[
VOIA = \frac{KV \times PDIFF^{1.5}}{\max(AG, AGMAX)}
\]  

(III.41)

KV is a constant and the denominator is the larger of AG and AGMAX. The exponent 1.5 is included to ensure that this aspect of the acoustic strength of voicing grows as fast as the amplitude envelope for aspiration noise as pressure drop across the glottis increases.
2) To give greater flexibility and thus allow for different larynx types, Equation (III.41) was modified to:

\[
\text{VOIA} = \frac{K \times \text{PDIF}^{1.5}}{\max(AG, AG_{\text{MAX}})^{A_{\text{EXP}}}}
\]  

(III.42)

The parameter \( A_{\text{EXP}} \) can make \( \text{VOIA} \) fall less or more rapidly as \( AG \) increases. Because \( AG \) is less than 1 cm\(^2\) in all the modelling, a larger exponent \( A_{\text{EXP}} \) makes the denominator smaller and \( \text{VOIA} \) larger; therefore if \( A_{\text{EXP}} \) is altered the weights applied when mixing the output responses to the different sources need to be changed. In this study \( A_{\text{EXP}} \) generally has the value 2 to give a rapid fall-off in amplitude at the high \( AG \) edge of the voicing plateau. The default value for \( AG_{\text{MAX}} \) is 0.1 cm\(^2\).

3) In another formulation two additional wave parameters are made to vary with \( AG \) to allow for the fact that amplitude \( \text{VOIA} \) is not the only acoustically significant aspect of the waveshape. These are the closed phase as a fraction of the whole cycle, called \( TCR \), and the skewing of the open phase: \( FBA \) in the Rosenberg wave option and \( K \) in the Fant wave option (see Figures 13 and 20).

When the vocal folds are adjusted articulatorily to be on average over the voicing cycle very close together, in pressed phonation with \( AG \) very small, there is a large amount of skewing and a sudden closure followed by a long closed phase. Acoustically the voice excitation of the vocal tract is rich in high frequency harmonics and aspiration noise is unlikely to be significant. When \( AG \) is very large, to simulate low vocal fold adduction and breathy phonation probably accompanied by significant aspiration noise, there is a gentle closure of the voice wave and a very short closed phase; the wave has little skewing and is nearly sinusoidal in shape. There is
low frequency emphasis in the vocal tract excitation. Between these two extremes, for an intermediate degree of adduction of the vocal folds with glottal area AG near 0.05 cm$^2$, the voice wave and its spectrum have intermediate properties, intended to be appropriate for normal phonation.

To simulate this continuum TCR in the model is related to AG by:

$$TCR = \frac{A}{(AG+A)}$$  \hspace{1cm} (III.43)

where A is a constant such that TCR has a specified value less than 1 when AG=0.05 cm$^2$, set by a parameter TCR0. As AG tends to zero TCR tends to 1, its maximum possible value, but an upper limit is set by a parameter TCRMAX which can be less than 1. As AG becomes very large TCR tends to zero.

Other inverse functions define the skewing, K in the Fant model or FBA in the Rosenberg model.

In the Rosenberg wave representation FBA is given a very high value of 100 at AG=0 and a value defined by the parameter FBA0 at AG=0.05 cm$^2$. The function used is:

$$FBA = \frac{A}{(AG+B)}$$  \hspace{1cm} (III.44)

where A and B are constants and FBA tends to zero as AG increases.

In the Fant wave representation K is given a very high value of 100 at AG=0; when AG becomes large K tends towards its theoretically allowable minimum value of 0.5. The inverse curve is further defined
by specifying parameter values $K_0$ at $AG=0.05 \text{ cm}^2$ for normal phonation and $K_{O1}$ at $AG=0.15 \text{ cm}^2$ for breathy phonation. $K_{O1}$ must be smaller than $K_0$. The expression used is:

$$K = \frac{A}{(AG+B)^R} + 0.5$$ (III.45)

where $A$ and $B$ are constants, dependent upon $K_0$ and $K_{O1}$. To avoid unrealistically large values when $AG$ is very small an upper limit is given to $K$ by a parameter $K_{MAX}$. In the study of vocal attack for example, Section IV.2.1, the graph of $K$ versus $AG$ was as shown in Figure 22.

The three methods just described of varying some wave parameters are selected by setting WAVSW to zero.

Default values for the parameters are:

- for $AG = 0.05 \text{ cm}^2$  
  $\text{F}_{BAO} = 10$  
  $\text{T}_{CR0} = 0.1$  
  $K_0 = 1.0$

- for $AG = 0.15 \text{ cm}^2$  
  $K_{O1} = 0.6$

and

- $\text{T}_{CRMAX} = 0.6$  
  $K_{MAX} = 5.0$

4) In a later development of the voice source model, available as another option in AERO by setting WAVSW to 1, all three parameters besides $F_0$ needed to construct the wave through the Fant formulation, selected with $W\text{MODE}=3$, are made to vary with three controlling parameters $PDIFF$, $Q$ and $AG$, through a linear mapping. Combinations of values needed to define the necessary number of points in the mapping for solution of the equations are expressed as values in a table as shown in Figure 23. The command WAIL displays the table and lists the other WAVE parameters.

Each line in the table shows a particular combination of controlling parameters and the corresponding voice wave parameters. Values of
PDIFF, Q and AG to be used are fixed except that Q1, the high value for Q, is a parameter, generally set to 200 Hz in this study. The fixed values are intended to be represent a reasonable range of values for speech: PDIFF normal as for a vowel and reduced as in a voiced fricative; Q for a mid and high pitch; AG for normal and breathy phonation types. Upper and lower bounds are included in the table as shown. Some are controlled by parameters: VOIAMAX, TCRMAX, KMAX, TDMIN; others are fixed: VOIAMIN = 0.01 cm$^3$/s, TCRMIN = 0, KMIN = 0.5, TDMAX = 1000000 ms.

Each of the three wave parameters selected is obtained by solution of four simultaneous equations. For example VOIA is defined by:

\[
\begin{align*}
\text{c11} \times \text{PDIFF} + \text{c12} \times Q + \text{c13} \times AG + \text{c14} \times \text{VOIA} + 1 &= 0 \\
\text{c11} \times \text{PDIFF} + \text{c12} \times Q + \text{c13} \times AG + \text{c14} \times \text{VOIA} + 1 &= 0 \\
\text{c11} \times \text{PDIFF} + \text{c12} \times Q + \text{c13} \times AG + \text{c14} \times \text{VOIA} + 1 &= 0 \\
\text{c11} \times \text{PDIFF} + \text{c12} \times Q + \text{c13} \times AG + \text{c14} \times \text{VOIA} + 1 &= 0
\end{align*}
\]

(III.46)

Two other sets of simultaneous equations are needed, for example for TCR and K, with different sets of constants: c21, c22, c23 and c24 in the equations for TCR; c31, c32, c33 and c34 in the equations for K. As long as the values in one line are not a multiple of values in any other line the equations can be solved. A routine from the NAG library called F04ATF is used to do this.

Besides F0 (VOIF), five wave parameters, VOIA, TCR, TD, K and USLOPE, are emphasised in the modelling because they are acoustically significant and can be measured on published voice waveforms for natural speech and singing. The linear mapping is a phenomenological approach to the simulation of voicing mechanisms: it is intended to reflect general trends and magnitudes found by analysis of natural speech and by true physical modelling of the processes of voice generation.
The research on which the values for the tables are based is reviewed in Section II.3.4, especially II.3.4.8. The data for natural speech come from inverse filtering of the output volume flowrate through mouth and nose combined, in speech and singing with different efforts, pitches and phonation types (Lindqvist, 1970; Rothenberg, 1973; Monsen and Engebretson, 1977; Sundberg and Gauffin, 1978, 1979; Gauffin and Sundberg, 1980; Sundberg, personal communication, 1981; Karlsson, personal communication, 1983; Karlsson, 1986); from inverse filtering of sound pressure signals (Holmes, 1962, 1976; Fant, 1979a; Bickley, 1982); from 2-mass models of the vocal folds (Ishizaka and Flanagan, 1972; Monsen et al., 1978; Boë et Guérin, 1979); from measurements on excised larynges (van den Berg and Tan, 1959); and from perception experiments to assess the range of voice wave parameter values over which synthetic speech is acceptable to listeners (Rosenberg, 1971).

The natural speech data include different speakers and different languages. Subglottal pressure and hence pressure drop across the glottis has to be estimated by a method described in Appendix 2 (Scully, 1986). Besides these limitations on the reliability of the values extracted from published papers there are other ways in which the model fails to take account of the complexity of natural speech processes. For example, in the modelling, reducing PDIF by raising supraglottal pressure with subglottal pressure constant and by reducing subglottal pressure are taken to be equivalent, but this is not the case. The air pressure within the glottis which provides one of the forces driving the vocal folds is not equal to the mean, aerodynamic component of pressure drop across the glottis, but varies throughout each voicing cycle as discussed in Section II.3.4.2: when the vocal folds are opening from below and the glottal flow is convergent, pressure in the glottis is close to subglottal pressure; when the lower part of the vocal folds are beginning to close and the
glottal flow is divergent, pressure in the glottis is close to supraglottal pressure (Titze, 1986). Data on raised oral pressure in consonantal shaped vocal tracts (for example Rothenberg and Mahshie, 1986; Bickley and Stevens, 1986) should really be considered separately from data on subglottal pressure varied during vowel production.

The assumption of linear mapping is for convenience only and lacks theoretical underpinning. The aim in this study is to capture some general trends. Although some figures are given below these are to be taken as plausible values only. The view taken here is that specific values for dependencies would characterise different speaker types, but the voicing models used for this study are not matched to particular speakers. In another series of experiments, two different voice wave tables have been obtained for two different women speakers with apparently contrasting larynx types, based on inverse filtering and aerodynamic data for their speech (Scully, 1989). It would be premature to suggest which wave parameter values, if any, might be considered to be universally applicable.

In general, for men speakers and singers, some trends are suggested by the research cited above. Data from different studies have been collated and AG values estimated from the working orifice equation

$$AG = 0.00076 \times UG / PDIF^{0.5}$$

(Equation (II.13), Sections II.3.3.4 and III.1.1.3, also Appendix 2, Scully, 1986). It should be emphasised that in most studies cited there were only one or two speakers.

As far as possible trends were related to only one controlling parameter changing at a time by selecting suitable points from constructed graphs, derived from published waveforms. However, it was not possible to apply this criterion rigorously in view of the limited range of conditions investigated in the literature. For
example, in the data from which the dependence of VOIA on AG was assessed PDIFF was not constant but decreased as AG increased; this had to be allowed for in a qualitative way.

VOIA has a minimum value of 200 to 300 cm$^3$/s found when AG is small or PDIFF is low; it increases approximately linearly with PDIFF or as PDIFF$^{0.5}$ or PDIFF$^{0.6}$ from 190 cm$^3$/s at PDIFF= 4 cmH$_2$O to 410 cm$^3$/s at PDIFF= 15 cmH$_2$O. A maximum value of VOIA might be 700 to 800 cm$^3$/s. VOIA increases as AG increases if AG is not too large but remains relatively constant as F1 varies throughout a breath group in connected speech: when AG is between 0.05 and 0.1 cm$^2$ VOIA is about 550 cm$^3$/s. It falls to a lower value of 200 cm$^3$/s as AG decreases to about 0.01 cm$^2$; it falls as AG increases above 0.1 cm$^2$, becoming 250 cm$^3$ at AG= 0.14 cm$^2$. VOIA decreases as Q rises.

TCR increases as PDIFF increases and decreases when AG or Q increases; its maximum value seems to be 0.5 or 0.6 and its minimum value near zero.

TD decreases as PDIFF increases, to a minimum value of 0.6 or 0.8 ms for speech. If TD is as low as 0.4 ms, synthetic speech sounds unnatural, but TD may go as low as about 0.3 ms for loud professional singing. TD increases as AG or Q increases.

VOIA/TD is an indicator of the acoustically important parameter, the maximum negative value of dU$^g$/dt, at or before closure. The ratio is given the label USLOPE. If VOIA and TD vary oppositely with a controlling variable the trend for USLOPE can be estimated; if VOIA and TD vary in the same direction the effect on USLOPE depends on the magnitudes of VOIA and TD. On this basis it is predicted that USLOPE increases as PDIFF increases and decreases as Q increases; the dependence of USLOPE on AG is not clear.
Since increases in Q and PDIFF seem to have opposite effects on VOIA, TCR and TD and since F0 depends on both Q and PDIFF, the variation of VOIA, TCR and TD with F0 is not clearly predictable: it depends on which mechanism, Q or PDIFF, is predominantly used to alter F0. The fact that F0 is not an independent but a dependent variable, as discussed in Section II.3.4.10, adds to the difficulties of interpreting inverse filtering data in terms of controlling and controlled parameters.

There are five combinations available in the modelling, selected with the switch TKSW, as follows:

\[
\begin{align*}
\text{TKSW}=1 & \quad \text{VOIA} \quad \text{TCR} \quad \text{TD} \\
\text{TKSW}=2 & \quad \text{VOIA} \quad \text{TCR} \quad \text{K} \\
\text{TKSW}=3 & \quad \text{VOIA} \quad \text{TD} \quad \text{K} \\
\text{TKSW}=4 & \quad \text{TCR} \quad \text{TD} \quad \text{K} \\
\text{TKSW}=5 & \quad \text{TCR} \quad \text{K} \quad \text{USLOPE}
\end{align*}
\]

TKSW=2 is the default option and is most often used, but in modelling singing the first option with TKSW=1 was chosen since there were data for VOIA, TCR and TD in professional singing (Sundberg and Gauffin, 1979; Gauffin and Sundberg, 1980) which could be matched in the modelling (Scully and Allwood, 1985b).

TKSW=5 is different in kind from the other four methods, since it includes the expression USLOPE = KU*PDIFF^UEXP, where KU and UEXP are constants set by parameters. This is based on the strong indication that USLOPE increases with PDIFF.

If TKSW is set to 1, 2, 3 or 4, three of the four wave parameters are calculated from the linear equations and the others are derived. It is necessary for internal consistency to apply modifications first. For example, in TKSW=2, VOIA reductions are made before calculating
TD which depends on VOIA. With TKSW=4 the equations cannot be completely solved. TKSW=1 and TKSW=2 will be outlined since these are the methods used in the modelling so far.

With TKSW=1 VOIA, TCR and TD are obtained from the equations. Various reduction factors are applied to VOIA as described below. After application of all the reduction rules VOIA is windowed by its upper and lower bounds as shown in the table, Figure 23. After this TD is obtained; then it is limited by its upper and lower bounds. TC is obtained from TCR and \( F_q \), the magnitude of USLOPE is obtained from VOIA and TD; USLOPE will be bounded because VOIA and TD are bounded. K must be obtained by an iterative procedure, with no guarantee of a unique solution; its upper and lower bounds are applied. Finally GLOF is derived from USLOPE, VOIA and K and then VGR from \( F_q \) and GLOF.

The method used with TKSW=2 is similar but here K is obtained directly and iteration is avoided.

Based on the approach of Ladefoged (1963) discussed in Section II.3.4.10, \( F_0 \) or VOIF as it is called in the modelling is given by:

\[
F_0 = Q + KF \times PDIFF
\]  

(III.47)

where Q represents the effective mass and tension of the vocal folds, an articulatory parameter. PDIFF is the pressure drop across the glottis, an aerodynamic parameter. KF is a constant set to 4 in this study, based on the findings for natural speech discussed in Section II.3.4.10.
To improve the naturalness of glottalised voicing, an extreme pressed phonation type where the vocal folds are strongly adducted, both the voice wave amplitude VOIA and its fundamental frequency $F_0$ (VOIF) are made to fall when AG is less than some threshold value AGMIN. The previously calculated value of VOIA or $F_0$ is multiplied by the factor $(1 - SMF\cdot AG/AGMIN)$, where SMF is the sigmoidal function used in ARTIC to define an S-shaped path through articulatory space (for Slow S, Medium M, or Fast F transitions) sketched in Figure 16(c) and in Appendix 3 (Figure 5, Scully, 1987, p.92). When AG is equal to AGMIN, SMF has the value zero; as AG approaches zero, SMF approaches the value 1.

Figure 24 indicates how VOIA varies with AG if PDIFF is constant and how $F_0$ varies with AG for constant values of Q and PDIFF. $F_0$ can fall to the extremely low frequencies seen as a realisation of a / gev / in intervocalic position in Arabic (Ahmed, 1979), during a glottalised voicing manoeuvre (Rothenberg, 1972) and when a glottalised plosive follows a vowel in English (unpublished laboratory traces). If $F_0$ falls below a threshold frequency VFT, voicing ceases. In the apparent absence of data for natural speech, VFT was set to a very low value of 2 Hz in this study.

Thresholds for voicing onset and offset need to be stated and these depend in complex ways on the mechanical state of the vocal folds (van den Berg and Tan, 1959). A minimum pressure drop PDIFF is specified in the model by the parameter DPMIN, generally set to 2 cmH$_2$O. By analogy with the onset and offset of turbulence it seems quite likely that a higher pressure difference may be needed to make vocal fold vibration start than to make it cease. As an alternative to a single threshold DPMIN, the modelling allows a higher threshold for voice onset DPMON, set to about 4 cmH$_2$O and a lower one for voice offset DPMOFF, set to about 2 cmH$_2$O.
If VOIA falls below a threshold value VAT it is set to zero so that voice ceases, but VAT is generally zero in this study and in that case this threshold is not used. It was found necessary to impose quite a high VAT threshold when trying to match the cycles at voice onset in the model with those of an individual real speaker (Scully, 1989).

Default values in the voice wave parameter model, which is called by WAVE, are: DPMIN = 2 cmH2O; VAT = 0 cm3/s; AEXP = 2; AGMAX = 0.1 cm2; AGMIN = 0.025 cm2; VFT = 2 Hz.

Whichever method is used, the voice wave parameter values are obtained at each 5 ms time sample. SORV constructs the actual voice wave, defined at 28 μs samples. After the rising and falling portions of the wave have been constructed from the expressions shown in Figures 13, the closed phase duration is derived from TCR and F0 (VOIF) already calculated, so that the duration of the whole voicing cycle is defined. When VOIA falls below its threshold minimum value VAT its value is reset to zero and the voice source is likely to be cut off in mid cycle; this must be avoided because the rapid change in VOICE constitutes a transient and gives spurious clicks in the output. At the start of voicing the problem is avoided by setting Wg in the expression for the rising portion of the wave (Figure 13) to zero at this point. Where voicing ceases as VOIA falls below its threshold value a partial cycle is removed by back-tracking.

The graphs of the voice source waveform generated in the model are compared with natural speech. Even though acoustic sources are not directly observable in natural speech, the goodness of fit between the modelling and the natural speech it attempts to simulate can be assessed by parameters such as voice onset time and vocoid durations on spectrograms (see Figure 3). This is with the proviso that an
acoustically weak source may be hardly apparent in the speech signal after filtering by, for example, a very constricted vocal tract.

The voice model parameters can be modified and the voice source recalculated from the same articulatory and aerodynamic files as before. With this kind of analysis-by-synthesis iteration, the match may be improved.

When an apparently reasonably good match has been obtained the voice source VOICE signal is saved as a file from AERO, to be used as an input source file in FILT.

### III.2.4.2 Turbulence noise

The linear relationship of the expression in Equation (III.31) in Section III.1.4.2 is used to define an acoustic pressure source amplitude envelope, with suffix A in the label. The articulatory and aerodynamic parameters needed are shown in Figures 15 and 17. The mapping from the symbols used in Section III.1.4 to those used for parameters of the model is shown on page 14.

For aspiration noise (ASP) arising just above the glottis

$$\text{ASPA} = K\times A G^{0.5} \times P D I F F^{1.5}$$  \hspace{1cm} (III.48)

For frication noise (FRIC) arising at or in front of a vocal tract constriction

$$\text{FRICA} = K\times A C^{0.5} \times P C^{1.5}$$  \hspace{1cm} (III.49)

For frication noise (FRICV) arising at the velopharyngeal port

$$\text{FRICAV} = K\times A V^{0.5} \times P C^{1.5}$$  \hspace{1cm} (III.50)

KN is a constant the value of which can be changed so as to control the balance between voice and the noise sources before filtering takes place. This would be useful if some sources were filtered.
simultaneously, but the value does not need to be varied in this study since each source was filtered separately.

Glottal area AG, vocal tract constriction cross-section area AC and velopharyngeal port area AV are available as functions of time; they are files saved from ARTIC and read into AERO. PDIFF is the pressure drop across the glottis, equal to (PSG-PC); PSG and PC are calculated as functions of time within the aerodynamic block as explained in Section III.2.3. Each noise source amplitude envelope ASPA or FRICA or FRICVA is multiplied by an unmodulated random number sequence to give a noise pressure waveform. A random number sequence is generated with SCAL, RANDOM by specifying a minimum, a maximum and two initial seed values. It is possible to vary the simulated white noise each time a turbulence noise source is made in this way, although in practice standard values are generally used. The amplitudes ASPA, FRICA and FRICAV have sample times of 5 ms, while the random number sequence and the sources derived are given a sample time of 28 μs.

Before constructing the noise source waveforms, by the commands ASP, FRIC and FRICV, the voice source waveform Ug, called VOICE in the model, needs to have been made, through the command SORV, as described in Section III.2.4.1, since this is needed for modulation of the noise sources in the presence of voicing.

It was thought that, without thresholds for the onset and offset of turbulence noise, the noise sources might extend unrealistically far in time. For this reason a switch is provided called AFSW, to permit selection of a logarithmic dependency for noise sources: when AFSW is set to 1 the amplitude envelope of frication noise is made equal to 10^{FRICA*AREF}, where AREF is a constant. However, it was found that the direct linear relationship of Equation (III.49) for
frication noise did not give the anticipated problem and the expressions of Equations (III.48), (III.49) and (III.50) with AFSW set to 2 were always used for the modelling described in this work.

Modulation and amplitude reduction for noise source waveforms in the presence of voicing are achieved by means of an absolute or a relative threshold applied to the voice wave amplitude VOIA; the threshold values are called ATOL and VTOL respectively. The relative threshold, which is the one used in this study, is expressed as a fraction of the largest voice wave amplitude VOIAMax occurring during that simulation. When VOIA is above the threshold value the expressions for noise source amplitude envelopes, Equations (III.48), (III.49) and (III.50), are multiplied by VFAC. VOIA is the current time value of the voice waveshape amplitude. VFAC is a constant for the run or series of simulations, generally a reduction factor, less than 1.

The sampled waveform thus defined for a noise source is, finally, divided by VOIAMax and multiplied by the current value of the voice waveshape Ug (VOICE), varying between zero for the closed phase and VOIA during the open phase. In this way the periodicity of voicing is imposed on the noise waveform, the noise source amplitude is reduced to allow for changing flow conditions which are likely to reduce the efficiency of turbulence generation, and the magnitude of the noise source envelope increases as voice amplitude increases. Thus for example in the word series "plight" - "polite" of Appendix 3, Section IV.7 (Scully, 1987, pp.135-139), VTOL was 0.3 with VFAC 0.75. When VOIA was above 0.3 times its maximum value in each one-word simulation, ASPA and FRICA were multiplied by 0.75•Ug/VOIAMax at each time sample.
Figure 25 exemplifies the effect. The increasing amplitude of an aspiration noise waveform arises from Equation (III.48). By the fourth cycle of voicing the threshold VFAC has been reached and the noise is modulated at the fundamental frequency of voicing. One problem with the method is that there is likely to be a discontinuity in the noise amplitude as modulation begins. A limitation of the method is that no account is taken of a possible continuous d.c. flow of air through the glottis which would give rise to a continuous, unmodulated component in the noise source.

The peak value for the noise source envelope occurs at the generally right-skewed maximum of glottal flow $U_g$. No allowance is made when modulating frication noise sources for the time delay when an acoustic disturbance travels from the glottis to a vocal tract constriction.

As in the rest of the modelling, trial and error with matching to natural speech is part of the approach used. Since there is no firm theoretical foundation for the noise reduction factor, the parameters VFAC and VTOL are varied and the relative weightings of the voice and noise filtered outputs are adjusted to give the auditorily best synthetic speech. It is not expected that the parameter values should be fixed and completely consistent from one series of simulations to another, although several of the series described in Chapter IV and the Appendices do share many parameter values.

The place of injection of a frication noise source within the vocal tract filter is important, as discussed in Sections II.3.4.12 and III.1.5.

In some cases the aspiration noise source is given a broad spectral peak to make it a little closer to the frequency spectrum associated
with an orifice, as discussed in Section II.3.4.12. No allowance is made for the changing orifice geometry. A complex conjugate pole pair with centre frequency 1 kHz and bandwidth 2 kHz is specified by means of one of the auxiliary programs in SIGNAL and applied to the aspiration noise source signal ASPS.

When a single frication noise source FRICS obtained in AERO is to be split, it is processed as follows in one of the auxiliary programs. A signal S is made with value 1 from time $t=0$ to time $t=T_1$, the time point where the frication noise source signal is to be cut, and with value 0 from $t=T_1+1$ to the last time sample needed. FRICS is multiplied by S to give FRICAS which equals FRICS up to time $t=T_1$ and is zero thereafter. Subtracting FRICAS from FRICS gives FRICBS which equals zero up to $t=T_1$ and equals FRICS thereafter. FRICAS and FRICBS are injected at different sections of the vocal tract and the responses to the two frication noise sources are saved as two files. FRICS can be split at more than one time point if necessary by constructing a suitable signal S.

III.2.4.3 Transient

This presents no particular problems of implementation. The rate of change of pressure drop across the vocal tract constriction $dP_C/dt$, called DPC in the modelling, is obtained by numerical differentiation, within the aerodynamic block AERO, where it is needed in any case for solution of the simultaneous differential equations, as described in Section III.2.3. Only interpolation is needed to change the sample time from 5 ms to 28 $\mu$s. The transient source DPC is injected into the same section of the vocal tract as is the frication noise source FRICS. The output signal response to DPC is mixed with other output signals in the final stage of the modelling.
As in the case of the noise sources, the lack of a threshold does not seem to result in an inappropriately wide-spread source. Although DPC is often different from zero, its acoustic significance seems to be confined to a few places of large magnitude, positive or negative. Transient pulses are likely to appear for fricatives as well as for plosives.

**III.2.5 Acoustic filtering**

It is assumed that cross-section area but not shape at every point along the tube determines the acoustic filter properties. A continuously changing area function is approximated by abutting cylinders of matching mean cross-section area and 0.5 cm in length. The reflected pressure wave method for filtering is used (Kelly and Lochbaum, 1962). The program was developed from software provided by Mr. Nicholas Husband of the Joint Speech Research Unit. Cylinder boundaries are taken in alternating pairs to calculate reflection coefficients for incident pressure waves. Transmitted and reflected waves travel a distance of 0.5 cm across a single cylindrical section with the speed of sound $c$ in air at the temperature and humidity found in the vocal tract. Temperature and humidity gradients between the larynx and the mouth and nose outlets are ignored. $c$ is given the value 35000 cm/s, for moist air at body temperature, 37°C (Flanagan, 1972, p.35). The time of traverse of one 0.5 cm section is very close to 14 μs, with a correspondingly very high sample frequency of 71.429 kHz. The final output sample frequency is 11.905 kHz so that the synthetic speech signal is defined up to 5.952 kHz.

The reflected pressure wave method is described in, for example Flanagan (1972a, p.273-276). The implementation in the model is
described by Allwood (1984). Some aspects of particular importance for the modelling will be discussed here.

The acoustic source waveforms and the output response waveforms need to be sampled at 28 μs, not at 14 μs, for reasons explained below in this section.

Area function quantisation errors are acceptably small as long as each section length is well below the shortest wavelength modelled (Flanagan, 1972, p.25). For good articulatory modelling of consonants it was decided to make the sections as short as 0.5 cm in spite of the associated heavy computational load. This section length introduces no additional frequency bandwidth limitations, since the critical frequency for a wavelength of 5 cm, ten times the section length, is 7 kHz, as discussed in Section II.3.5.

An entire vocal tract area function table, called ALT, is read in. This is defined at time intervals of 5 ms. Linear interpolation is used to derive the areas at 14 μs intervals. Cross-section areas in the ALT table are allowed to go to zero or even to negative values in the articulatory block ARTIC as discussed in Section II.3.2.4 and sketched in Figure 7. Negative areas are reset to zero before filtering is performed. To avoid problems with two adjacent zero areas in the reflected pressure method, a minimum area greater than zero is imposed in the filtering block FILT. The threshold AMIN is a parameter generally set at 0.001 cm².

For simulations of nasalisation a parameter NASW is switched from its usual non-nasal value of 0 to a value of 1. Additional files are read in to define the time path of the velopharyngeal port AV and a fixed nasal cavity area function. Linear interpolation converts AV from its 5 ms sample time in ARTIC to the 14 μs sample time needed.
for FILT. The position of the nasal coupling point is defined by a parameter NPOINT, the distance of the velopharyngeal port from the glottis in 0.5 cm units. Either one or two cylindrical sections of the nasal branch, selected by the parameter AVNUM, are given the value AV. AVNUM depends on the relative lengths of the oral and nasal tubes; its value is chosen to ensure that oral and nasal outputs are in phase. A minimum value for AV, AVMIN, is imposed, generally 0.001 cm².

The lips and nostrils terminations of the vocal tract are represented by reflection coefficients RCM and RCN respectively, with negative values appropriate for a sudden change of cross-section area to one about ten times greater or more. The air beyond each outlet of the vocal tract is represented by the larger cross-section area. A fixed value is used for each simulation, representing the magnitude at some frequency for each outlet lip or nostril area but not the frequency-dependent characteristics of a radiation impedance termination of the tube. RCM can be made to depend on the lip outlet area AL, in the same sort of way as described below for the reflection coefficient at the glottis, but this refinement was not used in the modelling of this study. A suitable value for the reflection coefficient at the lips RCM is found empirically, so as to give, in conjunction with the choice of reflection coefficient at the glottis, formant bandwidths comparable to those of natural speech.

For a pressure wave reflected at a boundary between two sections where the cross-section area changes along the direction of travel of the wave from $A_1$ to $A_2$, the fraction of the pressure reflected, the reflection coefficient $r$, is given by:

$$r_{A_1,A_2} = \frac{(A_1 - A_2)}{(A_1 + A_2)}$$  (III.51)
Typical values chosen for RCM, the reflection coefficient at the lips, are near -0.8. For a lip area AL of 5 cm² this implies that A₂, the notional cross-section area beyond the lip outlet of the vocal tract is 45 cm², since

\[-0.8 = \frac{5 - A_2}{5 + A_2}\]

At the glottal end of the acoustic tube also, a reflection coefficient is used to represent a sudden change of cross-section area from that of the first section of the vocal tract, the lowest part of the larynx tube, just above the glottis, to the very small area of the glottis. For a larynx tube cross-section area of 1 cm² and a glottal area of 0.05 cm², appropriate for vowels, the reflection coefficient at the glottis RCG would be:

\[
\frac{1 - 0.05}{1 + 0.05} = 0.95 \approx 0.9\]

and for a larynx tube of 2 cm², RCG would be 0.95. In practice, lower values of RCG are used for vowels, near 0.8, determined empirically in combination with RCM, the reflection coefficient at the lips. It is assumed that no energy is reflected back from the subglottal airways to the glottis.

The acoustic energy lost across the glottis is nearly always made to depend on the slowly changing articulatory component of glottal area AG. The file for AG is read into the filtering program and given linear interpolation to convert it from 5 ms to 14 µs. The reflection coefficient at the glottis RCG can be kept at a preset value but is usually made to degrease as AG increases. Typically, maximum glottal areas of 0.5 cm² are used in modelling voiceless fricatives.

It was found essential to increase transglottal losses in this case, as discussed in Section III.1.5. If RCG was left at a value of about
0.8 appropriate to the small glottis for a preceding vowel, voice energy trapped inside the strongly constricted vocal tract showed in output spectrograms as unnaturally strong dark formant-like bands at very low frequency. When abduction of the vocal folds was more realistically modelled in the acoustic filtering block FILT by reducing RCG, two improvements in acoustic matching resulted: the unrealistic low frequency energy disappeared from the spectrograms and formant bandwidths seen in frication noise for [s] became wider, giving the diffuse pattern characteristic of natural speech. This observation gave some confidence that the variation of glottal losses represented a real improvement in the physical model.

The reflection coefficient at the glottis RCG, which needs to increase with glottal area AG, is derived from an expression of the form:

\[
\text{RCG} = \frac{A}{(AG + B)} \tag{III.52}
\]

where A and B are constants. Their values are not specified by the experimenter. Instead two values for RCG are specified: RCG1 for a glottal area AG=0.05 cm², a usual value for normal phonation and RCG2 for AG=0.15 cm², appropriate for breathy phonation. The maximum theoretically permissible value for RCG is 1 but a lower limit at small AG values is set by specifying another parameter RCGMAX. This is the same sort of approach as is used in some of the methods of constructing a voice source waveshape, described in Section III.2.4.1, in which some of the parameters vary with AG. The resulting variation of RCG with AG is similar to that shown in Figure 22 for K as a function of AG.

Bandwidths for poles of the vocal tract transfer function were measured for two different vowel types, with two different values for
the mouth outlet loss and two different values for the loss through the
glottis. The bandwidths were of the right order of magnitude,
compared with available values, but the range of values published in
the literature is so great that precise guidelines for modelling are
not apparent (Appendix 4, Scully and Allwood, 1982).

Glottal losses have been made to vary within each voicing cycle also,
although this is not done for most of the modelling described in this
study. An acoustic or oscillatory component of glottal area $A_g(t)$
was obtained from the voice source $U_g$, neglecting differences in
their waveshapes, specifically the greater skewing in $U_g$ than in
$A_g(t)$ due to inertance of air in and above the glottis, as follows:

$$A_g(t) = \frac{A_G U_g}{|U_G|} \quad (III.53)$$

where $A_g$ is the articulatory component of glottal area and $U_G$ is the
mean aerodynamic component of volume flowrate of air through the
Glottis.

$A_G$ and $A_g(t)$ were combined in three different ways to give total
Glottal area $A_g$:

1) $A_g = A_g(t) + A_G \quad (III.54a)$
2) $A_g = w A_g + (1-w) A_G \quad (III.54b)$
3) $A_g = A_G (U_g - VOIA/2)/U_G + A_G \quad (III.54c)$

(The final $A_G$ term in Equation (III.54c) was omitted by mistake in
the published paper and has been added in Appendix 4 (Scully and
Allwood, 1982). $w = VOIA/VOIA_{max}$ and VOIA is the amplitude envelope
of $U_g$; $VOIA_{max}$ is its maximum value during a single simulation.
Besides losses due to reflection coefficients of magnitudes less than
1 at the lips, nostrils and glottis, additional spreading of spectral
peaks is obtained by a wall movement effect, treated as low-pass filtering. For each set of section boundaries, taken in alternating pairs, a small fraction of the reflected and transmitted pressure waves is delayed by one 14 μs time sample and contributes to an incident pressure wave one time sample later. The fractions delayed are controlled by three parameters: RPLC and RMLC for the pharynx and mouth cavity respectively, both generally set to 0.01, and RNLC for the nasal cavity, generally set to 0.15 to represent the effects of a larger, more lossy wall area in the nasal portion of the vocal tract.

At each 14 μs sample only half of the vocal tract section boundaries are included in the calculation of reflected and transmitted pressure waves. At one time sample boundaries 1, 3, 5 and so on are included; at the next sample 2, 4, 6 and so on are included. Acoustic source waveforms add to the pressure defined at the boundary at which they are injected when that boundary is included in the calculations, at 28 μs intervals.

The sources are injected, one at a time, at appropriate section boundaries. The voice source waveform is multiplied by the cross-section area of the first section, just above the glottis, to convert from flow to a value proportional to pressure (see Equation (III.33) in Section III.1.5). Voice and aspiration noise are injected as pressures at the glottal termination boundary; for those two runs of FILT a frication switch INJECT is set to zero. Frication noise and transient are injected as pressures in front of the vocal tract constriction included in the aerodynamic block. The switch INJECT is set to 1 and zeros are substituted for the glottal source. If the constriction is a composite one then the fricative noise is generally split, as described in Section III.1.5. The two portions of frication noise are used in turn, injected at different boundaries.
As at other boundaries, pressure wave calculations for the final mouth and nose outlet boundaries are performed at 28 \( \mu s \) intervals, so that acoustic outputs from the filtering block have 28 \( \mu s \) sample times. Generally, for non-nasal sequences, four or five output waveforms are obtained, in response to voice, aspiration noises, frication noise and transient. When the nasal cavity is included there can be up to ten outputs: eight are the oral and nasal outputs in response to voice, aspiration noise, frication noise and transient; the ninth and tenth are responses to a frication source called FRICVS generated at the velopharyngeal port constriction AV and injected for filtering just above the constriction.

About a thousand extra 28 \( \mu s \) samples with zero values are added at the end of the sources and the final shape of the vocal tract filter is maintained for that time. This allows the filter's responses to the sources to decay; without the extra samples the outputs are chopped off suddenly and there are extraneous clicks.

The oral output pressure is multiplied by the lip outlet area \( AL \) to convert it to a value proportional to volume flowrate (see Equation (III.33)). Similarly, nasal outputs are multiplied by the nostril outlet area.

Where filter shapes and the resulting formant patterns are explored it is sometimes convenient to bypass the aerodynamic and acoustic sources blocks of the model. Instead, a glottal source with an approximately flat frequency spectrum for the frequencies of interest is constructed; this is a single rectangular pulse or a train of such pulses or a \( \sin x/x \) function. Where the filter rather than the voice source is the focus of the modelling, but a natural-sounding vowel-like output is required, the glottal source for the filtering
is constructed from specified function for its amplitude envelope and fundamental frequency.

III.2.6 Radiation

The transform from acoustic volume flowrate at the lips to acoustic pressure at a microphone some distance away is performed by applying an auxiliary filter command called ZERO. A zero at zero frequency is applied to each output of the filtering block FILT in turn.

Before this stage of the modelling was added, unrealistically large values for K had to be used when constructing the voice wave parameters in order to obtain enough high frequency emphasis in synthesised vowels. With this radiation simulation added realistic K values became possible, giving some confidence that the main acoustic processes of natural speech production were now represented in the model.

Figure 26 shows the effect of ZERO on a sinx/x signal defined with a 28 μs sample time up to 10 kHz. A fast Fourier transform (FFT) with a Hanning window was used in SIGNAL for the frequency analysis. The shaping was close to the intended +6dB/octave over the frequency range of interest.

III.2.7 Outputs

Each of the signals output from ZERO is low-pass filtered and downsampled by a factor of three to convert it from a sample frequency of 35.714 kHz with a sample time of 28 μs to a sample frequency of 11.905 kHz with a sample time of 84 μs.
A non-recursive finite-impulse response low-pass filter is used, adapted from a program supplied by Dr. Steve Terepin. A Hanning window is shifted along the signal by the down-sampling factor; parameters define the window length and the filter length. The filter length used, 151 samples at 28 \( \mu s \), is a compromise to give adequate results without very long processing time. The \( \sin x/x \) signal of Figure 26(a) was passed through this process to give the frequency spectrum shown in Figure 26(c). 11.9 kHz is the Nyquist frequency for an analog signal defined up to 5.95 kHz.

The separate outputs from this process in the modelling, for responses to each acoustic source separately, are weighted and summed. The composite output signal, the result of the modelling the response of the vocal tract to all the sources combined, is down-line loaded from the VAX 11/780 to a PDP 11/03 microprocessor which has a 16-bit word and 56 kilobytes of main memory. The samples are represented as either 8-bit or 16-bit differences in order to make best use of the limited memory. Up to three seconds of synthetic speech can be stored. The digital signal thus defined passes through a 12-bit digital-to-analog converter. 12-bit samples (4096 discrete levels) are output under the control of a hardware clock at 84 \( \mu s \) intervals. The clock, a presettable interrupt driver, was designed and constructed by Mr. Stuart Allen and Mr. Stan Call.

It is possible to alter the clock rate so as to give synthetic speech speeded up a little or slowed down. The smallest sample time that can be handled in the microprocessor is 75 \( \mu s \). Two programs handle the output process: SEND in the VAX 11/780 and XFER in the microprocessor; the latter is written in PDP11 assembler code for high speed computation.
The digital signal thus defined passes through a 12-bit digital-to-analog converter. The analog waveform is 16 volts peak-to-peak and it is given 7 dB attenuation before it passes through the hardware anti-aliasing filter. This is a seventh-order elliptical filter with a flat frequency response within + or - 0.1 dB up to 5 kHz, and is more than 60 dB down by 5.95 kHz. The filter design was chosen on the advice of Dr. John Holmes for its flat frequency response and its optimum combination of sharp cut-off and low phase distortion. It was designed and constructed by Mr. Eric Brearley with advice from Dr. Gordon Lockhart and Dr. Muhammed Zaid.

The analog signal is played once or with a specified number of repetitions or until halted. The delay between successive repetitions can be varied. It is heard over headsets and recorded onto a Ferrograph reel-to-reel tape recorder. The spectrograms shown in this work were made with a 700 spectrograph (Voice Identification) with high frequency emphasis.

The final output synthetic speech signal has a bandwidth extending somewhat beyond the upper frequency for one-dimensional wave propagation in the vocal tract. As discussed in Section III.1.5, this is considered to be acceptable.

Graphical outputs make use of a locally developed device-independent graphics package. Graphs are displayed on a Tektronix 4010 storage screen, with keyboard and cursor control. Hard copy is obtained via a screen dump or with a Computer Instrumentation Economist plotter.

A LOG command is available in each of the model's programs. This allows the keyboard and screen dialogue to be saved. The information can be printed to give a permanent record of a run of the model.
III.2.8 Feedback and computation times

The processes of trial and error with monitoring clearly play a vital role in building up a real speaker's knowledge of the combinations of actions that will generate a sequence of speech sounds that is appropriate for the sequence of linguistic units specified, and that does not contain any additional, intrusive sounds. Human beings can, and do, try for the auditory goals many, many times; they can tell whether they are improving or not on each trial, unless their hearing is impaired; and this information comes back to the central nervous system at once.

Although phonetic knowledge provides additional inputs for an adult experimenter, similar processes of trial and error with monitoring are essential if a model of speech production processes is even to approximate to a specific accent of English for a limited set of allophone sequences. Probably the most severe constraint in our modelling has been the auditory feedback delay. The procedures in running the various computer programs must be gone through as outlined in this chapter, files must be processed and saved, response files must be weighted and added, and, finally, hardware devices must be attached to the computer before the output analog waveforms can be heard and recorded.

The actual computation time has been checked in some runs of the model. For a sample run simulating 600 ms of speech production, the total cpu time used was 687.3 s and so the computation rate is a little over 1000 times real time. This would cause few problems on its own. Much more severe are two other constraints, the interactive time and the amount of disc space needed for the acoustic output waveforms, defined at a sample frequency of 84 μs. The implications are considered further in Section III.3.
III.3 Methods of assessment

Introduction

There are many levels at which assessment can be discussed. Throughout the development of the computer model its functioning had to be checked as comprehensively as possible. Figure 26 is included in this report as an example of the many tests of that kind which were done. In this case the simulation of an approximation to the radiation transform from flow at the mouth or nose outlet to sound pressure at a microphone, and the final software process of low-pass filtering with down-sampling were being assessed.

Different kinds of assessment are of the essence in the modelling and attempted simulation of natural speech production: there are comparisons of many kinds between the synthetic speech patterns and some corresponding patterns taken from natural speech. These comparisons can be articulatory, aerodynamic, acoustic or auditory; comparisons were made in all these domains.

III.3.1 Comparisons with traces from natural speech

Speech patterns were obtained for a few speakers of British English with a near-RP accent. This is defined and the speakers' linguistic backgrounds are outlined in Section I.1.1. The speakers made studio recordings of multiple repetitions of the speech sequences to be modelled. They spoke into a mask with an orally inserted pressure tube, also, to give aerodynamic data. The working orifice equation, Equation (II.13) was used to infer estimated articulatory paths for tongue tip-blade constrictions and for glottal area in the natural speech, by procedures described in Appendices 1 and 2. Some examples
of the articulatory paths obtained are shown in Appendix 1 (Scully, 1984a, Figures 2 and 3).

For most of the modelling of the study the aim was to capture qualitatively and with the right order of magnitudes the patterns of natural speech, produced by the real speakers as a group. Convenient graphical procedures, with features such as automatic drawing and labelling of axes, a wide choice of functions to be displayed together, and, optionally, the setting of axes to fixed values for detailed examination of related traces, are an important aspect of the implementation of the model.

The graphs used frequently in ARTIC include articulator paths as functions of time, alone or in combination, vocal tract area functions or a side view of the vocal tract at different selected times, and composite constriction areas as functions of time. In the AERO block the traces shown usually include subglottal air pressure, oral air pressure, volume flowrates through the glottis and the vocal tract constriction, together with some articulatory traces, generally vocal tract and glottal area constriction traces, with the vocal tract cavity volume. Acoustic sources in the model are examined both as overall patterns for the whole simulation, or as detailed patterns, for example for the voice waveshape over a small stretch of time.

Assessment of the appropriateness or otherwise of aerodynamic traces generated in the modelling is through direct comparison with features in some of the natural speech traces. Not all the traces in the model are accessible in the natural speech data, notably subglottal pressure and glottal airflow. Subglottal pressure was estimated for controlled phonetic contexts and then iteration was used to get
closer to the natural speech data with the model, as described in Appendix 2 (Scully, 1986).

In a few cases the aerodynamic traces or acoustic outputs from the model were to be matched with specific speakers. In the experiment reported in Appendix 2 (Scully, 1986), an attempt was made to match the aerodynamic patterns of two real speakers who produced examples of [s] and [z] in controlled phonetic contexts. Appendix 7 (Scully and Allwood, 1985a) includes spectrogram comparisons between the model's synthetic speech and five real speakers. The measurements employed the spectrogram segmentation methods shown in Figure 3.

Detailed comparisons between the traces for the natural speech and the modelling were difficult to make because the graphs had different scales and were presented on different media: these were mingogram traces for the natural speech and a screen display with hard copy for the modelling. It was not practicable to log the real speech data onto the computer. It was necessary, instead, to focus the comparisons on pattern features such as the magnitudes of peaks of flow and pressure, values of, for example, nearly steady airflows for vowel-like articulations, and time intervals between pairs of features such as airflow release and the peak of airflow following it.

Thereafter, the aim was to match, similarly, traces for the acoustic sources in the model to their time domains as deduced from spectrograms and mingograph traces for the natural speech. For example, on spectrograms, the domain of frication noise can be estimated, as suggested in Figure 3; the relative strength of frication noise as a function of time can be estimated from traces of intensity level, high-pass filtered at 3.9 kHz, with a logarithmic, dB scale.
Different weights were tried for the, generally four, files added together at the final stage of the modelling. From informal auditory assessment a few combinations were chosen and recorded. Spectrograms were made and compared with those of natural speech, in order to find the weighting combination which matched best. Sometimes the model's output signals for voice, aspiration noise, frication noise and transient were given separate radiation transforms and were recorded separately onto audio tape. Their spectrograms were a form of analysis-by-synthesis for the interpretation of natural speech noise sounds as aspiration noise, or frication noise, or both.

**III.3.2 Auditory assessment**

Auditory assessment in the modelling, followed by further modified modelling with the aim of approaching closer to the auditory target, mimics the processes of trial and error with monitoring in the acquisition of natural speech. As discussed in Section III.2.8, the quality of the feedback loop presented problems for the experimenters. They could use their own, not noticeably impaired, hearing, but the outputs from the model often seemed like speech from a severely hearing-impaired person (Fourcin, personal communication). This is not a coincidence; the monitoring procedures were weak in the modelling because of the long time, generally one or two days, that elapsed between trying out an articulatory scheme in the model and listening to its results. Batch processing was often used in order to increase the number of runs of the model. It was not easy to decide how many variations to try in one overnight batch run. It was necessary to make extreme changes to parameter values to start with, in the hope of crossing important regions of nonlinear mapping from articulatory parameter values onto acoustic and auditory pattern features. Even so, all the variations in a large number of simulations might well sound essentially the same. The disc space
available for storing the outputs generated as a batch limited the number of runs that could be done before off-loading the files onto magnetic tape, in any case. Ideally, each attempt with the model would have been individually assessed, but in that case progress would have been too slow. A compromise number of runs, about five to ten, was usually chosen.

Detailed phonetic transcriptions were attempted and auditory peculiarities noted. For example, in the early modelling there were insufficient acoustic losses and an effect described as "rattling tea cups" was noted. It was often observed that speech-like allophones were generated beyond the number aimed for. This effect lent the strength of practical experience to the view of speech production held by the present author: the articulatory gestures are not seen as the links between a chain of allophones serially ordered in production as they usually are in phonetic transcriptions; instead, fragments of articulatory transitions are seen as the elements, the units of production themselves. A totality of articulatory gestures can generate one, two, or more allophones depending on the particular ways in which the actions are coordinated.

III.3.3 Perception experiments

In some cases the auditory judgements of the model's outputs were made by other listeners in a formal listening test. For example synthetic diphthongs were mixed in with diphthongs produced by the real speakers and presented to listeners for identification and for judgements of naturalness.

In the experiment described in Appendix 7 (Scully and Allwood, 1985a), stimuli were ordered along the articulatory continuum of maximum glottal area, the degree of abduction of the vocal folds, for
alveolar fricatives. The listeners were responding to the totality of acoustic changes caused by this single articulatory change. The results of the listening test are partly a formal assessment of the extent to which the model is capable, as programmed so far, of generating sounds which have enough of the acoustic patterning of natural speech to elicit phonologically relevant responses by listeners.
CHAPTER IV: INTRODUCTION TO THE APPENDICES AND SOME RESULTS OF THE MODELLING

CONCLUSIONS FROM THE STUDY

LIMITATIONS OF THE STUDY

SUGGESTIONS FOR FUTURE WORK

Introduction

In this chapter, some examples of results from experiments using the composite model are shown. The experiments are described in detail elsewhere (Scully and Allwood, 1978, 1979, 1981). The eight appendices describe some experiments in more detail, while the sections of this chapter and the figures included here are intended to serve as illustrations of some of the effects found with the model. The appendices are introduced here as appropriate.

The basis for the whole study is the hypothesis that much of the detailed acoustic patterning of speech can be explained with reference to the systems used in production. A particular set of coordinated actions is selected so as to generate a broadly defined sequence of sounds, for the specific linguistic message; details of acoustic patterns follow from this. In particular, when a given linguistic unit such as a phoneme of RP English is placed in different phonetic contexts, the requirements of the overall goals, the broadly defined sequence of sounds, impose changes of acoustic detail. These are likely to be bundles of covarying acoustic pattern features, because they arise from a unified aerodynamic system. In natural speech it may be supposed that the resultant signal has an internal consistency such that listeners can make use of the rule-governed changes of structure and regularities in the patterning. A good model of speech production processes should illustrate these
processes and thus contribute to an understanding of the rules that are likely to be useful for perception.

The overall aim of the experiments described here was to explore the detailed acoustic implications of the actions needed to achieve a broadly defined sequence of sounds, assessed auditorily. Acoustic comparisons with some patterns of natural speech were made, and aerodynamic comparisons also. It was hoped that, if the aerodynamic patterns matched, covarying acoustic pattern features would follow automatically.

The final sections of this chapter, and of the thesis, contain a statement of the general conclusions from the whole study, an assessment of its limitations, and some suggestions for future development and work in the modelling of speech production.

The references cited in the work follow and, after them, the appendices. Each appendix has its own individual page numbering.

IV.1 Modelling of some phonetic classes

IV.1.1 Vowels

Some vowels were generated in the model with its vocal tract area function set as specified by Fant (1960, 1970, p.115). Formant frequencies for [i], [a] and [u] were compared. The values agreed quite well (see Scully and Allwood, 1978) and gave some confidence that the acoustic filtering simulation was operating as it should. Good auditory vowel qualities were obtained for some selected vowels of English, approximating to an RP accent. These included [i:], [ɑ:],[ɜ:], [ɪ], [e] and [ʊ].
Figure 27 shows the effects of different phonation types illustrated with an [i:] vowel. As the vocal folds were made more abducted, with a larger value for the glottal area AG, the voice source spectrum and the strength of the aspiration noise source changed together.

Most of the modelling attempted to simulate a man's speech, but in some cases different vocal tract lengths and proportions were tried, combined with a voice source waveform, generated by the voice table method, Method 4) in Section III.2.4.1, using table values more appropriate for a woman (Karlsson, personal communication, 1983). It was necessary, in addition, to try out a number of different area functions, in order to obtain appropriate formant frequencies for a woman. Appropriate area functions for several different vowels for a few real speakers were obtained by trial and error with frequency spectrum comparisons which included a measure of goodness of fit of spectral peaks. Most of these vocal tract shapes were not combined with respiratory and larynx actions so as to compute acoustic sources, nor were they taken through the filtering, radiation and output stages of the model with a simply constructed voice source. In a few cases this complete run of the model was done; Figure 28 exemplifies the results; these sounds were extremely successful auditorily as a woman's [i:], with a close quality, near Cardinal Vowel 1. This serves to demonstrate that many of the problems in making the model sound like natural speech are connected with the complicated and time-consuming work that has to be done to find appropriate vocal tract shapes. Initial attempts at most speech sequences tried with the model were almost invariably very unnatural in sound, but they improved greatly later.
IV.1.2 Diphthongs

It was easy to discover instructions for the vocal tract parameters which would generate the right kind of formant movements for some diphthongs. The instructions for simulating an [ai] diphthong follow, as an example of a run of the model.

The vocal tract dimensions and proportions are defined by TPOS:

<table>
<thead>
<tr>
<th>DS</th>
<th>DE</th>
<th>DP</th>
<th>DN</th>
<th>DB</th>
<th>(DC)</th>
<th>(DF)</th>
<th>DJ</th>
<th>DT</th>
<th>DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>11</td>
<td>(12)</td>
<td>(13)</td>
<td>13.5</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

(cm from the glottis)

Instructions for the articulators in TMOD are:

<table>
<thead>
<tr>
<th>AE</th>
<th>2</th>
<th>AP</th>
<th>1</th>
<th></th>
<th></th>
<th></th>
<th>2, 19, S</th>
<th>10, 100, S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10, 100, S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10, 100, S</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AN</th>
<th>5</th>
<th>AB</th>
<th>10</th>
<th></th>
<th></th>
<th></th>
<th>5, 119, S</th>
<th>10, 19, S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5, 100, S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AJ</th>
<th>5</th>
<th>AL</th>
<th>10</th>
<th></th>
<th></th>
<th></th>
<th>5, 19, M</th>
<th>10, 19, S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2, 100, M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first line gives the initial configuration, the cross-section area for each parameter point needed, in cm$^2$. DC and DF are given values but the C and F points are not needed for the [ai].
The first column of the instructions is the next transition end point in cm\(^2\); the second column is the time allowed for the transition to the next end point in 5 ms units; the third column is the transition type: S is slow, taking 38 time units, M is medium, taking 24 time units.

Once the end point has been reached, in the time specified, the parameter value remains at that end point value. Thus the third line for AE, AP, AB and AL instructs the area values to change with a slow transition, taking 38 of the allowed 100 time units. For the remaining 62 time units the values remain constant.

In a complete run of the model, PLU or DVLU (air pressure in the lungs or rate of lung volume decrement, respectively), AG (glottal area), Q (the laryngeal component of \(F_0\) control), and AV (the area of the velopharyngeal port) would be given suitable instructions of this type. The procedures described in Section III would be followed.

A series of diphthongs was generated in the model, using a Rosenberg voice source constructed by method 2) in Section III.2.4.1, with the short-cut method described there. The experiment is described in detail in Scully and Allwood (1979). The voice source waveform amplitude VOIA was made to rise and then fall, while its fundamental frequency VOIF rose slightly from 120 to 140 Hz and then fell gradually to 100 Hz. The synthetic sounds were presented to listeners, mixed with diphthongs recorded in a studio by the real speakers. The listeners could identify the synthetic diphthongs as words of English without difficulty. The natural diphthongs were given better ratings for naturalness; this is not surprising, since there were many acoustic pattern features, such as noise, visible to a greater or lesser extent on the spectrograms of the natural speech,
not included in this partial modelling. The $F_0$ contours of the natural speech were less smooth than those from the model also.

One diphthong was generated by the full model. The [ei] diphthong was chosen because the location of the major constriction of the vocal tract remains approximately constant throughout, thus conforming to the requirements of the aerodynamic model. The auditory match did improve here, although it is not possible to draw many conclusions from this one example. The experiment showed that natural transitions of the vocal tract area function could easily be achieved in the model. It was very much more difficult to achieve appropriate diphthong qualities for simulations of men with different length vocal tracts. It was not sufficient to merely scale the dimensions up or down and use the same set of instructions to the vocal tract parameter points.

IV.1.3 Approximants

The modelling concentrated on [l] and [r]. It was surprisingly easy to model an auditorily successful [l], even though the central tube, closed at one end, which is formed by the tongue tip in [l] could not be simulated. Fant (1960, 1970, p.167) states that an acceptable [l] can be produced without a zero and that a sudden shift of $F_1$ from a lateral to adjacent vowels is important for perception. In the modelling for [l], the parameter point was given a fast, 40 ms (F) transition to simulate this effect, consistent with the findings of Lisker (1957) and O'Connor et al. (1957), and AF was reduced to 0.2 cm$^2$ at a distance from the glottis of 15 or 15.5 cm, appropriate for the alveolar ridge in a vocal tract 17.5 cm long. Oral pressure rose by an appropriate amount in the model.
It was very much more difficult to model an acceptable English [r]. An essential acoustic feature of [r] in the allophone sequence modelled, [əɹeɪ], seems to be a dip in F₃, to near F₂ (Lisker, 1957). For the real speakers analysed, F₂ and F₃ were close together in frequency and dipped, to different frequencies in different speakers. For example, where the formant frequencies were lowest on one speaker's spectrogram F₁ was at about 400 Hz, F₂ at 1100 Hz and F₃ at 1700 Hz. This kind of formant transition was not obtained until many articulatory parameters in the model had been varied.

It was found that the articulatory descriptions of [r] in the literature, although not actually misleading, do not emphasis the crucial features of vocal tract shape for mid-[r], which distinguish [r] from the other approximants of English, [w], [l] and [j]. Gimson, for example, describes the most common allophone of RP /r/ as a voiced post-alveolar frictionless continuant [ɹ]. The central part of the tongue is described as "lowered, with a general contraction of the tongue, so that the effect of the tongue position is one of hollowing and slight retroflexion of the tip.". The lips are indicated as rather open (Gimson, 1980, p.201). Delattre, on the other hand, in his X-ray study of American English /r/, showed that retroflexion of the tongue tip was not used by all speakers. In some cases the tongue dorsum formed the constriction. Tracings show the lips approximated and protruded, perhaps to different degrees by different speakers. There is marked narrowing in the mid-pharynx region. The other feature apparent in this study is the large volume of the cavity in front of the constriction in each case (Delattre, 1967, pp.200-202).

After a series of unsuccessful [r] sounds had been generated, as a result of simply moving the tongue constriction further back, in an attempt to simulate retroflexion and a post-alveolar place of
articulation, it was decided to vary systematically the following factors:

1) the place and shape of the constriction;
2) the volume of the front cavity;
3) the cross-section area of the lip outlet;
4) the minimum cross-section area of the constriction;
5) the cross-section area of the pharynx.

About seventy simulations were assessed auditorily and spectrographically. It was not possible to try every possible combination of these five factors. But from the selected combinations examined, some trends emerged quite clearly.

As long as the front cavity was kept small the sounds were unsuccessful. If the constriction had gradually tapering inlet and outlet shapes the sound was [g]-like. If the constriction was more sudden [w]-like or [l]-like sounds were heard, depending on the position of the constriction. Keeping the constriction cross-section area at 0.2 cm$^2$ and the lip outlet area at 0.3 cm$^2$, the constriction was moved forward and a clear patterning emerged, consistent with acoustic theory as shown in Figure 2(a) (see also Fant, 1960, 1970, pp.76,77,82). During the constricted portion, $F_1$ dipped for all positions of the constriction. For $F_2$ and higher formants the dip was transferred to progressively higher formants as the place of the constriction moved forward. With the constriction position DC at 11 cm from the glottis $F_2$ dipped and $F_3$ dipped slightly. At the other end of the range with DC at 15 cm from the glottis, $F_2$ and $F_3$ hardly moved, but $F_4$ and $F_5$ dipped sharply.

The constriction area needed to be between about 0.1 and 0.2 cm$^2$. If it was smaller, there was confusion with a plosive; if it was larger, even with a large enough front cavity, the formant dips were
not strong enough. Figures 29 and 30 show some of the [r] results.

Front cavity tuning seems to be the key to the control of [r]. For and English [\textipa{\textl}] the place of the constriction must be such that $F_2$ and $F_3$ are the formants most strongly dependent upon the front cavity and therefore the ones most reduced in frequency if that cavity volume is increased or its lip outlet area is reduced. This is the region on Figure 2(a) between about 11 and 13 cm from the glottis, where $F_2$ and $F_3$ both dip from curve 1 to curve 5, as lip rounding increases. But for a good [\textipa{\textl}] the front cavity volume needs to be bigger than it is in the vowel-like configurations of the vocal tract modelled by Fant (1960, 1970, p.82). The formant frequencies must dip strongly with only slight lip rounding; otherwise [\textipa{\textw}]-like sounds result. When a successful [\textipa{\textl}] had eventually been made, it seemed impossible to get rid of it! That was because, as long as the front cavity volume was right, the values of other articulatory parameters were not crucial. While the configuration was wrong, lip rounding gave slight formant dips, of short duration, which could be, wrongly, interpreted as associated with articulatory transitions of equally short duration. But there is a clear disparity between acoustic and articulatory transitions at regions of stability.

It is important that some aspect of the acoustic patterning, in this case a dip in specific formant frequencies, should be sensitive to a change in one or two articulatory parameters, in this case, the volume of the front cavity and lip rounding. If, at the same time, that acoustic pattern feature is stable with respect to other articulatory changes or unavoidable perturbations, then the problems of acoustic stability are likely to be soluble. Sensitivity, as well as stability, is important in quantal theory (Stevens, 1972; Stevens and others, 1989).
IV.1.4 Fricatives

The modelling focussed on alveolar fricatives, [s] and [z], for which aerodynamic data from natural speech could conveniently be obtained. It was necessary to inject the frication noise source in front of the alveolar constriction; otherwise, large internal reflections at the sudden changes of area removed too much acoustic energy from the output. [s] and [z] were investigated in some different contexts, as discussed in Section IV.2.2. Appendices 8 and 2 describe how the model was used to try to simulate particular speakers' productions of [s] and [z].

IV.1.5 Plosives

Only a few attempts were made at modelling plosives. An example is shown in Figure 31. The model's output showed some of the complexity of acoustic structure of the natural speech, but more trial and error would be needed to match the formant transitions.

IV.2 Implications of some minimal and non-minimal articulatory contrasts

Introduction

Most of the appendices explore some aspect of the changes in acoustic structure which result from sequences containing the same phoneme but in a different context. Appendix 1 describes the methods that were used to obtain the data from natural speech. These were used as a basis for constriction paths of changing cross-section area in the modelling and for aerodynamic comparisons.
IV.2.1 Coordination in vocal attack

In Appendix 6 it is shown that inter-articulator coordination is important even for a static vowel in isolation. The more realistic lung volume decrement parameter for respiratory articulation was used. Many kinds of [i:] vowels with extraneous sounds were made; moving into speech from breathing, and out of speech back to breathing require skilled actions, just as much as in the middle of a sequence of allophones. For this reason, the breathing state is given the status of an allophone in the time scheme framework that was developed. This is described, with some examples, in Appendix 3.

IV.2.2 Fricatives: voiced and voiceless; in different contexts

Fricatives [s] and [z] were placed in contrasting vowel contexts in one study (Allwood and Scully, 1982). Even though the articulatory transitions into and out of the fricatives had the same durations for close and open vowel environments, differences of segment duration arose, both for the fricative contoids and for the vocoids, consistent with effects noted for natural speech. Figure 32 shows the effects for sequences intended as /i:si:/ and /a:sa:/, but heard as /i:si:/ and /a:sa:/.

Figure 33 shows that several acoustic changes can result from a single articulatory change. Here the glottal area AG needed to be increased from its value for the vowels for both fricatives, but more for [s] than for [z]. The glottal area transition had the same duration in both cases. Both the fricative contoid segments and the adjacent vocoid segments altered in duration. The experiment is described in Appendix 7. Most aspects of the acoustic patterns generated by the modelling were in good qualitative agreement with measurements on natural speech.
IV.2.3 Devoiced [r]

Section IV.1.3 has described how appropriate values for the articulatory parameters were obtained for an approximant [u]. When the same vocal tract articulation was combined with abduction and adduction of the vocal folds during the time that the vocal tract was constricted, the result was a fricative version, a devoiced [ʃ], as shown in Figure 34. The shift from one manner category to another, from approximant to fricative was due not to a change of vocal tract configuration, but because of changing aerodynamic conditions at the vocal tract constriction as the larynx states changed. Even a vowel-like vocal tract configuration can change into a fricative under similar circumstances, as is illustrated in Appendix 3 (Scully, 1987, Figure 17(d), p.137).

IV.2.4 Sequences of the form [CCV] and [CeCV]

Some phonologically relevant contrasts of this kind were investigated. Some examples for [bl...], [pl...] and [pəl...] are given in Appendix 3 (Scully, 1987, Figure 17, pp.136-138). Another set of phonologically contrasting sequences was modelled also. Some of the results are shown in Figure 35, for simulations intended as /ətr'eɪ/ and /ətər'eɪ/. Extra, intrusive sounds were heard, so these are far from achieving the auditory goals set, but the spectrograms do nevertheless illustrate some apparently rule-governed effects. Because the glottis is large at and just after the [t] release, the vocal tract configuration that is appropriate for the formant patterns needed for /r/ judgements by listeners gives a devoiced fricative [r] for the /r/. Gimson states: "A completely devoiced fricative [ʃ] may be heard following accented /p,t,k/". It seems that this must be the case, if the actions for the /r/ are sufficiently linked to the release actions for the plosive to ensure
that no /CaC'V/ judgement is made because of an intrusive [ə]-type segment (Scully, 1973a).

IV.2.5 Articulatory perturbations

After some trial and error with auditory monitoring, sequences that was reasonably successful as intended /pɜː:s/ and /pɜː:z/ were generated. Many variations on a basic articulatory time plan were made, to simulate timing perturbations. When the formation of an alveolar constriction by means of a suitable AF path was timed to come earlier or later with respect to a glottal area increase for [s] or [z], there were changes in spectrogram features for both the fricative contoid segments and the vocoids. Timing perturbations did not disrupt the patterning in the case of [s] and /pɜː:s/ continued to be heard. The case was different with [z]. Here, the AF constriction needed to be formed later with respect to abduction of the vocal folds than was appropriate for [s]. Otherwise, an additional weak vocoid segment was seen on the spectrograms, after the [z] contoid; this would be interpreted by English listeners as a vowel after the fricative, which was not intended as the auditory goal. In avoiding the generation of an extra vocoid after the [z] contoid, the resultant fricative contoid for [z] was invariably voiceless. On the basis of a limited exploration such as this it is not possible to assert that /z/ must be devoiced in utterance-final position if a vowel following it is to be avoided, but this experiment suggests the possibility at least. Figure 36 shows the range of acoustic patterns generated and the informal auditory judgement made for each sound.

The time scheme framework that is described in Appendix 3 was developed in response to the need to apply timing changes which simulate chosen changes of inter-articulator coordination, as well as
for the simulation of timing variability within one basic plan. Appendix 5 states the problems of constructing a framework for mapping from a string of linguistic units onto the actions of speech production, organised along the time dimension. It is believed that the framework described in Appendix 3 is a possible solution to that problem, although its potential remains to be explored further.

**IV.3 Conclusions from the study**

The representation of the main processes of speech production, even in the highly simplified form of the composite model developed in this study, can generate pleasingly complex acoustic patterns. For the allophones and the phonetic contexts studied here, the modelling generated covarying bundles of acoustic pattern features which showed trends similar to those in natural speech. The experiments reported here show that even though articulation is a reducible system, to a first approximation, with quasi-independent actions for the different articulators, the acoustic effects are not reducible; through aerodynamic interactions, the effects of changes in larynx configuration, for example, are seen in the acoustic patterns associated with a constriction elsewhere. Even the class of manner of articulation for a consonant can change by this means. In a few examples, it has been demonstrated that the allophonic variations found in natural speech seem likely to be rule-governed and not arbitrarily chosen by speakers.

**IV.4 Limitations of the study**

It is clear from a comparison of the structures outlined in Chapter III and the theory and data discussed in Chapter II that the composite model developed in this study is far from capturing all the complexities of natural speech production. It is not the "Accurate,
agile" model of articulation with "exquisite, dynamic representation
of the details of laryngeal and tract functions and which is
controlled adaptively to 'mimic' an unknown input" that is needed,
according to Flanagan (1984), for solution of the problems of
synthesis, recognition and coding. Some specific shortcomings may be
mentioned, for each of the processes simulated.

Timing

The scheme for timing control, set out in Appendix 3, while perhaps
containing an appropriate level of complexity, does not yet capture
in an economical and manageable form the set patterns of inter-
articulator coordination that seem to be developed during speech
acquisition. The events in this scheme are probably more numerous
than they need to be. It seems likely that the mid portion of an
articulatory transition, near the maximum velocity, would be a better
candidate for an event, more closely related to what is known about
skilled sensori-motor activity. The scheme profoundly lacks data
from natural speech and probably it could benefit from being
expressed in some computer language such as LISP or PROLOG.

The analysis-by-synthesis process of representing timing data from
natural speech by parameter values, for example D and E expressions
in the timing framework, then generating traces with the model, then
comparing these with corresponding traces for natural speech, and
then altering values in the model to try again is laborious and not
automated; no optimisation methods have been used.

Articulation

The description of articulation is flexible and fairly convenient to
use, but it lacks any formal expression of basic physiological
constraints including what appear to be rules for constraints on tongue shapes actually occurring in natural speech. The flexibility in the model does have the advantage that it is not confined to a particular range of normal adult speakers, but each type of speaker to be simulated, whether disordered or normal in speech production capabilities, must have some built-in constraints, which ought to be included in a model.

It would probably be useful to extend the description of movement paths beyond the purely kinematic approach adopted here, so as to gain a better understanding of the ways in which the central nervous system handles muscle forces. But the problems of obtaining data on which to base even a kinematic description are very great; we are only at the beginning of the task. The first simplification made in order to get started on the modelling, of a constant transition time for larger or smaller distances traversed by a given articulator, is certainly an over-simplification; it needs to be modified in a theory-based way.

The need to obtain estimates of vocal tract area functions more reliably from acoustic speech signals and the problems of doing so are well recognised by researchers. The present imprecision in area function values in our model weakens the links between articulatory and aerodynamic processes.

**Aerodynamics**

The aerodynamic model is severely simplified. With only one constriction of the vocal tract included, it is bound to miss much of the interesting and important complexity of acoustic source structure of natural speech. Only a few, rather obvious, aerodynamically-determined interactions between acoustic pattern features for a given
phoneme and its phonetic context have been demonstrated with the model.

The respiratory control models used, which assume that either lung volume decrement or air pressure in the lungs is under the speaker's control, are not as realistic as the expiratory force model developed by Ohala (see Ohala, 1989). Some effects clearly visible in natural speech, such as rising oral air pressure during the closure of a voiceless aspirated plosive, cannot be generated in a realistic way by our model.

**Acoustic sources**

The principle that acoustic sources should be generated from prevailing physical conditions is a central feature of our modelling. But the forms of representation are weak as physical models. The voicing model is a purely phenomenological one, not embodying physical processes as does, for example, the 2-mass model of the vocal folds. The lack of a sound theoretical basis for the generation of turbulence noise and of data derived from natural speech or good models, both mathematical and mechanical, has imposed severe limitations on the naturalness of this stage of the model until the present time.

**Acoustic filtering**

The acoustic processes of filtering are very inadequately represented in the model. Out of the many shortcomings a few may be mentioned. The lips and nostrils terminations do not include frequency-dependent effects; in general, frequency-dependent effects of various kinds, which can now be handled in the time domain through the z-transform, have not been included in the model. Vocal tract wall effects are
included only in a very simplified way and there is no acoustic output via wall transmission and vibration. This is a serious limitation on the naturalness of the model's outputs, since voicing during a plosive closure, which is phonologically extremely important in many languages, cannot be simulated. The treatment of nasality in all its aspects, articulatory, aerodynamic, and, especially, acoustic, is oversimplified.

Some important links between articulation, aerodynamics and acoustic sources and filters have been ignored in our modelling because the filtering implementation does not allow for changes in vocal tract length during a simulation.

**Simulation time**

An overall limitation of the modelling described in this study is the long time needed to perform a complete simulation up to listening to the result. This was discussed in Section III.2.8.

**Range of investigation**

Finally, only one speaker type has been investigated for most of this study, and only a very few phonemes of English in a very limited set of linguistic contexts. The match to an RP accent is not very good so far. The synthetic speech generated has been confined to isolated words and short phrases. In principle the modelling can be expanded to longer sequences, but the problems outlined above mean that the time needed to achieve a reasonably good approximation to the natural speech would be prohibitive.
IV.5 Suggestions for future work

The limitations discussed above are pointers to what needs to be done in the future, if better models of speech production processes are to be developed. There is much research in progress and exciting developments in theory, data and modelling may be anticipated.

Great benefits would be reaped from any developments which speed up computations and simplify file handling for the experimenter. Clearly, much time would be saved by the inclusion of hardware real-time devices, especially for the filtering. But it will be essential to provide flexibility in any devices developed. They will need to be programmable, so that they may keep in step with research findings. The parameters must be resettable, and processes such as wall vibration must be available as options, so that experimenters have an amenable tool for progress in basic research. There needs to be "continuing development of a model of speech production and generative rules on all levels up to the linguistic framework and down to an advanced vocal tract model..." (Fant, 1984).

Just as important as representation of the processes of speech production in models is the requirement for more and better data from natural speech. The current rapid progress in transducer development and data acquisition, storage and labelling techniques, if they become available to a large number of researchers, should lead to a greatly improved understanding of the basic processes of normal and disordered speech production, with profound implications for speech recognition and synthesis, and for providing better help to speech-disabled people. Perhaps models of speech production will provide, eventually, really appropriate and life-like synthetic voices where they are needed.
Figure 1 Some of the anatomical structures used in speech production: (a) A cast of the subglottal airways from the trachea to the terminal bronchioles in the lungs (from West, 1974, p.5); (b) A posterior view of the larynx (from Williams and Warwick, 1980, p.1236).
Figure 1  Some of the anatomical structures used in speech production: (c) A side view of the head, identifying the structures and the regions of the vocal tract.
Figure 2 Nomograms relating formant frequencies $F_1$, $F_2$, $F_3$, $F_4$ and $F_5$ to parameters of the vocal tract area function. The distance of the constriction from the glottis $d_{fg}$, and also its distance from the lips, is shown as abscissa; lips at the left, glottis at the right.

(a) Effects of different degrees of lip rounding and protrusion: curve 1 with no narrow lip outlet; curves 2 to 5 with increasing lip rounding (from Fant, 1960, 1970, p. 82).
Figure 2 Nomograms relating formant frequencies $F_1$, $F_2$, $F_3$, $F_4$ and $F_5$ to parameters of the vocal tract area function. The distance of the constriction from the glottis $d_{fg}$, and also its distance from the lips, is shown as abscissa; lips at the left, glottis at the right. (b) Effects of different constriction sizes, from curve 1 with more severe narrowing, smaller $A_{min}$, to curve 3 with less severe narrowing, larger $A_{min}$ (from Fant, 1960, 1970, p.84)
Acoustic sources

V voice
A turbulence noise: aspiration
F turbulence noise: frication or mixed frication and aspiration noise (onset and offset defined at 5 kHz for alveolar fricatives)
T transient S silence

Spectrogram pattern features

VOT voice onset time
VPT voice persistence time
VCT vocoid onset time
voc vocoid
cont contoid
(vocoid and contoid boundaries are defined at the frequency of $F_2$)
cl acoustic 'closure'

Formant labels

$F_1$, $F_2$, $F_3$ etc. where the dark bands probably result from resonances of the whole vocal tract
$F_a$, $F_b$, $F_c$ etc. where the dark bands probably result from resonances affiliated with the front cavity

Figure 3 Terminology and segmentation criteria for spectrograms:
(a) For /VCV/ sequences containing a plosive consonant;
(b) For /VCV/ sequences containing a fricative consonant.
The abscissa is time; the ordinate is frequency, up to 8 kHz; sound pressure level is shown by the degree of darkening.
Figure 4 Simplified one-dimensional movement path characteristics in the absence of viscous loss and stiffness: (a) applied force $F$, a step function; (b) velocity; (c) distance moved; each plotted against time.

For the first half of the movement path:

$\frac{d}{dt} = \frac{1}{2} at^2$

$D = \frac{1}{2} t v_{\text{max}}$
Figure 5  Stages of speech production, excluding feedback paths.
Fig. 8. Velocity-versus-time curves for a point on the tongue body during the articulation of diphthong /ai/ by three speakers. The fleshpoint velocities were estimated by determining the incremental displacements for increments in time (20 msec).

Fig. 3. Displacement and velocity functions for TP1 of speaker 2 during the following vowel-to-vowel transitions: a /i/ to /æ/ in free allies; b /i/ to /æ/ in he honored, and c /i/ to /ɔr/ in he ordered. The functions are based on data recorded with a maxillary reference.

Figure 6 Movement paths for the tongue body during diphthong-like gestures, shown as displacement versus time and velocity versus time (from Kent, 1972, Figure 8; Kent and Moll, 1972a, Figure 3).
Fig. 3. Superimposed closure and release gestures for the apical consonants: /t/ in /atma/ (solid line), /d/ in /adma/ (dashed line), and /n/ in /anpa/ (dotted line). The measure plotted is LA, the distance between the tip of the tongue and the point of lingua-alveolar contact.

Figure 7 Movement paths for the tip-blade region of the tongue during closing and releasing actions for some alveolar consonants, showing near maximum velocity at closure and release (from Kent and Moll, 1969, Figure 3; 1972b, Figure 12); with a suggested path for tight closure in a model, based on these data and on McGlone et al. (1967).
Figure 8 Nomenclature for lung volumes and characteristic patterns of volume changes for quiet breathing, a vital capacity manoeuvre, and speech. The durations shown are indications of the order of magnitude only.
Key:

- $P_a$: Atmospheric pressure ($1030 \text{ cmH}_2\text{O}$)
- $P_t$: Pressure of air in the alveoli of the lungs with atmospheric pressure as reference zero
- $P_{pl}$: Pressure inside the pleural cavity with atmospheric pressure as reference zero
- $P_m$: Pressure associated with muscle forces
- $P_e$: Pressure associated with elastic recoil forces

Figure 9 Schematic representation of the main forces acting on a portion of the lung walls, rib cage, diaphragm and abdominal walls, at mid ranges of lung volumes.
Figure 10  Distribution in the subglottal airways of:
(a) Total cross-section area for each generation, also total volume for each generation, estimated from the values shown in Table I;
(b) Viscous flow resistance in each generation, based on Pedley et al., (1970, Figure 2, p.392).
Figure 11 Estimates for the resistance and conductance of the subglottal airways as a function of lung volume, based on Blide et al. (1964), solid lines and Vincent et al. (1970), dashed lines.
Figure 12: Sketches showing the range of voice source waveshapes found in natural speech, based on the literature cited in Sections II.3.4.8 and III.1.4.1. The abscissa is time; the ordinate is volume flowrate of air through the glottis. The abscissa is time; the ordinate is volume flowrate of air through the glottis. The continuous d.c. flow commonly observed in normal speakers is not shown, except where it is likely to be particularly large. The area under the curve and the steepest, negative, derivative determine low and high frequency spectrum levels respectively (Fant, 1980a, Figure II-A-4). The sketches are intended to give a rough indication of likely waveshapes and are speculative rather than quantitative.
Figure 13 A parametric representation of each cycle of the voice source waveshape (based on Fant, 1979a, 1979b, 1980a). Some of the relationships between the parameters are shown.

\[ T_0 = \frac{1}{F_0} = \frac{1}{\text{voi}f} \]

\[ TCR = \frac{TC}{T_0} \]

\[ \text{USLOPE} = \left| \frac{\text{voi}A}{TD} \right| \]

TIME (ms) \( \rightarrow t \)

\[ K = 0.5 + 0.125 \left( \frac{T_B}{TD} \right)^2 \]

\[ T = \frac{t}{L} = \frac{t}{\text{voi}f} \]

\[ \text{VRG} = \frac{F_0}{F_g} \}

\[ \omega_3 = 2\pi F_g \]

RISING FLOW PORTION:

\[ U_g = \frac{\text{voi}A}{2} \left[ 1 - \cos \omega_3(t - T_1) \right] \]

 FALLING FLOW PORTION:

\[ U_g = \text{voi}A \left[ K \cos \omega_3(t - T_2) - K + 1 \right] \]
Figure 14  Suggested ways in which some of the parameters of the voice source waveform, as shown in Figure 13, might vary with some controlling parameters in natural speech (see also Figure 12).
Figure 15 The composite model: the articulatory model with its parameters. Each vocal tract parameter point has an associated A value and an associated D value. DE, DP etc. means distance from the glottis along the vocal tract tube. A means the cross-section area of the tube.
Figure 16 The composite model: vocal tract shapes and tongue body transitions approximating to tongue shape factors found in natural speech. (a) Front raising; (b) Back raising (both based on Harshman et al., 1977, and other studies discussed in Section II.3.2.1). (c) The form of transitions in the articulatory model, based on the studies discussed in Section II.3.2.4, shown as change of cross-section area versus time.
Figure 17 The composite model: the aerodynamic model with its parameters.
Figure 18 The composite model: the function used for subglottal airways conductance, based on Section II.3.3.4 (Blide et al., 1964 and Vincent et al., 1970).

Figure 19 The composite model: formation of a single aerodynamically equivalent composite constriction from two constrictions of the vocal tract.
Figure 20 The composite model: three representations of the voice source.
(a) A train of pulses; (b) Rosenberg wave (Rosenberg, 1971);
(c) Fant wave (Fant, 1979a, 1979b, 1980a).
Figure 21 The composite model: block structure, processes simulated and sample times at each stage.
Figure 22 The composite model: the function to make the Fant wave parameter $K$, the skewing factor, vary with $AG$, the articulatory component of glottal area.

\[ K = \frac{A}{(AG + B)^R} + 0.5 \]
### WAVE SHAPE PARAMETERS

<table>
<thead>
<tr>
<th>PDIFF (cm H₂O)</th>
<th>Q (Hz)</th>
<th>AG (cm²)</th>
<th>VOIA (cm³/s)</th>
<th>TCR</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.00</td>
<td>Q1</td>
<td>0.05</td>
<td>VOA1 300.00</td>
<td>TCR1</td>
<td>K1</td>
</tr>
<tr>
<td>5.00</td>
<td>Q2</td>
<td>0.05</td>
<td>VOA2 350.00</td>
<td>TCR2</td>
<td>K2</td>
</tr>
<tr>
<td>10.00</td>
<td>Q3</td>
<td>0.05</td>
<td>VOA3 500.00</td>
<td>TCR3</td>
<td>K3</td>
</tr>
<tr>
<td>10.00</td>
<td>Q4</td>
<td>0.15</td>
<td>VOA4 750.00</td>
<td>TCR4</td>
<td>K4</td>
</tr>
<tr>
<td>MINIMUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>MAXIMUM</td>
<td>VOAMAX</td>
<td></td>
<td>TCRMAX 0.60</td>
<td>KMAX</td>
<td>5.00</td>
</tr>
</tbody>
</table>

Figure 23 The composite model: the voice table for linear mapping. The three controlling parameters are PDIFF, the pressure drop across the glottis, Q, the stiffness and effective mass of the vocal folds, and AG, the articulatory component of glottal area. The three dependent voice waveshape parameters are, in this example, the wave amplitude VOIA, the ratio of the closed phase duration to the total cycle duration TCR, and the skewing factor K.
Figure 24 The composite model: voice amplitude VOIA and fundamental frequency VOIF as functions of the articulatory component of glottal area AG. A plateau region is defined by AGMAX and AGMIN. Simulation of extreme pressed phonation or glottalisation as AG becomes very small, and of breathy voicing as AG becomes very large, are shown.
Figure 25 The composite model: modulation of a turbulence noise source waveform when voicing is present.
Figure 26 Assessment of part of the composite model: the radiation transform applied to a \( \sin x / x \) function. Frequency spectra are shown.
(a) \( \sin x / x \) before the radiation transform has been applied;
(b) \( \sin x / x \) after radiation, +6dB/octave;
(c) \( \sin x / x \) after low-pass filtering with down-sampling by a factor of three.
Figure 27 Spectrograms: modelling of different phonation types for an [i:] vowel.
Figure 28 Spectrograms: modelling of different speaker types. [i:] vowels to simulate (a) a man; (b) and (c) a woman.
Figure 29 Spectrograms for unsuccessful attempts at modelling [ər'ei].
Figure 30 Spectrograms for successful modelling of [ər'ei] and an example from natural speech.
Figure 31 Spectrogram for modelling of [at'ɛx] and an example from natural speech.
Figure 32 Acoustic sources as functions of time for [s] in close and open vowel contexts: (a) intended and heard as /iːsiː/;
(b) intended as /əːsaː/ and heard as /ɜːsaː/. 
Figure 33 Spectrograms of modelling of "A hiss it said" and "A his it said", with multiple acoustic contrasts arising from a single change of articulation, the maximum glottal area reached.
Figure 34 A spectrogram of a devoiced [ɒ], modelled with the vocal tract articulation appropriate for [ɒ], but with glottal area AG increasing to 0.5 cm².
Figure 35 Spectrograms of modelling [ətr'ei] and [ətrər'ei], showing the effects on acoustic structure of a change of phonetic context for /t/ and /r/.
Figure 36 Spectrograms of modelling the effects of timing perturbations for sequences intended as /p3:s/, above, and /p3:z/, below. Time perturbations about a central coordination are shown in ms. A positive value means that the AF transition for the vocal tract constriction is delayed relative to the abduction of the vocal folds. (Continued overleaf.)
Figure 36 (continued). Timing perturbations for /p3:z/ and /p3:z/.
Abbreviations:

Ann. Bull. RILP Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, Japan
Conf. Conference
Congr. Congress
Ed. Edition
ed(s). editor(s)
GALF GCP Groupement des Acousticiens de Langue Francaise, Groupe Communication Parlée
ICASSP Proc. of the IEEE Intl. Congr. on Acoustics, Speech and Signal Processing
ICPhS International Congress of Phonetic Sciences
Intl. International
J. Journal
JASA Journal of the Acoustical Society of America
Proc. Proceedings
STL-QPSR Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden.

REFERENCES

Papers in Phonetics, Dept. of Linguistics, 1, 10-14.


* Blide et al. (1964): see page 366.
References

Carlton, L.G. (1978) Retention characteristics of movement rate increase, J. of Motor Behaviour, 10, 105-112.
unpublished Final Year Laboratory Project, Department of Linguistics & Phonetics, University of Leeds.
Delattre, P. (1967) Final report: General Phonetic Characteristics of Languages, University of California at Santa Barbara, Research contract OEC-4-7-061990-0176 U.S. Office of Education.
Fant, G. (1980b) The relationships between area functions and the
acoustic signal, Phonetica, 37, 55-86.


Fitts, P.M. (1954) the information capacity of the human motor system in controlling the amplitude of movement, J. of Experimental Psychology, 47, 381-391.


speech: responses to resistive loading of the jaw, J. of Speech and Hearing Research, 18, 207-220.


Geigy (1973) Documenta Geigy Scientific Tables, 7th Ed.
References


Heike, G. (1979) Articulatory measurement and synthesis, Phonética, 36, 294-301.


Heinz, J.M. and Stevens, K.N. (1964) On the derivation of area functions and acoustic spectra from cineradiographic films of speech, JASA, 36, 1037.


Houde, R. (1968) A study of tongue body motions during selected speech sounds, SCRL Monograph No. 2, Santa Barbara: Speech Communications Research Laboratory.


References


Kempelen, von (1791) Le Mechanisme de la Parole, Suivi de la Description d'une Machine Parlante, J.V. Degen, Vienna.


Ladefoged, P. (1976) The phonetic specification of the languages of
the world, Working Papers in Phonetics, UCLA, Department of Linguistics, 3-21.


Maeda, S. (1987a) On the generation of sound in stop consonants, Speech Communication Group Working Papaers, Research Laboratory of
Mead, J. and Agostini, E. (1964) Dynamics of breathing, in Handbook of Physiology, Section 3 - Respiration I.
Murry, T. and Brown, W.S.Jr. (1971) Subglottal air pressure during
two types of vocal activity: vocal fry and modal phonation, Folia Phoniatrica, 23, 440-449.
Netsell, R. (1969) Subglottal and intraoral air pressure during the intervocalic contrast of /t/ and /d/, Phonetica, 20, 68-73.
Öhman, S.E.G. (1967) Peripheral motor commands in labial
Pedley, T.J. (1977) Pulmonary fluid dynamics, Annual Review of Fluid Mechanics, 9, 229-274.
presented at the 2nd. Conf. on Palato-Glossal Malfunction, Exeter.
Shadle, C.H. (1986b) Models of fricative consonants involving sound
generation along the walls of a tube, Proc. of the Intl. Congr. on
Acoustics, Toronto, Canada, paper A3-4.
consonants, in Hardcastle, W.J. and Marchal, A., eds. Proc. NATO
ASI on Speech Production and Speech Modelling, July, 1989, Bonas,
into the acoustics and aerodynamics of the larynx, Preprint of
paper to appear in the Proc. of the Vocal Fold Physiology Conf.,
Sharp, J.T., Henry, J.P., Sweany, S.K., Meadows, W.R. and
Pietras, R.J. (1964) Total respiratory inertance and its gas and
tissue components in normal and obese men, J. of Clinical
Investigation, 4, 503-509.
Speech Motor Control, Pergamon, Oxford, 105-112.
Shirai, K. and Honda, M. (1977) Estimation of articulatory motion,
in Sawashima, M. and Cooper, F.S., eds. Dynamic Aspects of Speech
Production, University of Tokyo Press, 279-304.
process for fricative consonants, Speech Communication, 2,
111-114.
orifice area: a re-examination of model experimentation, Cleft
Palate J., 17, 277-282.
articulatory synthesizer, IEEE Trans. on Acoustics, Speech and
Stålberg, E. and Young, R.R., eds. (1981) Clinical Neurophysiology,
Butterworth, Intl. Medical Review.
Stevens, K.N. (1956) Stop consonants, MIT Acoustics Laboratory,
Stevens, K.N. (1971) Airflow and turbulence noise for fricative and
stop consonants: static considerations, JASA, 50, 1180-1192.
Stevens, K.N. (1972) The quantal nature of speech: evidence from
articulatory-acoustic data, in Denes, P.B. and David, E.E. Jr.,
eds. Human Communication, A Unified View, McGraw-Hill, New York,
51-66.
Stevens, K.N. (1977) Physics of laryngeal behaviour and larynx modes,
in Fant, G. and Scully, C., eds. The Larynx and Language,
Phonetica, 34, 264-279.


Westbury, J.R. (1979) Aspects of the temporal control of voicing in consonant clusters in English, Texas Linguistics Forum, 14, Department of Linguistics, the Univ. of Texas at Austin.


APPENDIX 1

SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

Celia Scully
Dept. of Linguistics & Phonetics, University of Leeds, Leeds, U.K.

INTRODUCTION

Where different speakers show small differences of acoustic structure for the same broadly defined auditory goals, it may be supposed that these arise, in part, from their different individual patterns of articulator kinematics. One approach to the characterisation of rules for covarying acoustic pattern features across different speakers saying the same words of English is with analysis-by-synthesis, using a model of the physics of speech production. There is a need to go beyond linear acoustic theory, so that interactions between one acoustic source and another and between source and filter may be taken into account.

Flanagan et al. [1] employed a parametrically controlled model of speech production for the economic description of an [ai] diphthong said by a single speaker. Parameters in the model were adapted using criteria of minimum errors in the acoustic domain, comparing the model's output with the natural speech to be matched. The computation time required was very great, even though the analysis was limited to vowel configurations, with a rather long sample time of 12.8 ms. The capability of their speech production model to synthesise consonants was not used in this study. As the authors point out, a full understanding of articulatory constraints and acoustic behaviour of the system is lacking, especially for consonant articulations. Our aim is to gain insight into speaker-to-speaker variations in acoustic signal, rather than to investigate new methods for low bit-rate coding of speech. A composite model of speech production processes is used in which sequences containing consonants as well as vowels can be synthesised and identified by listeners. The functional models for voice and turbulence noise source generation and the kinematic descriptions of articulation are flexible, so that different speaker types may be modelled. The usual values of the parameters which determine the time paths of articulators are based on data from natural speech. The aerodynamic stage of speech production is accessible in the model, so that comparisons between synthetic and natural speech are not limited to those for acoustic outputs. A descriptive framework for the organisation of timing of articulatory events and their coordination across quasi-independent articulators has been developed [2,3]. This last simulates the planning stage of natural speech. It exhibits a rather high level of complexity. Currently about 7 or 8 expressions of coordination are needed for each phonetic unit, whether this is a vowel element or a consonant. Figure 1 shows...
SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

the events (E) and coordinations (D) proposed for voiced or voiceless fricatives. In its formal rules, this planning stage is at present underconstrained. Constraints on vocal tract configurations are introduced informally, based on data from real speech and principles such as constant tongue volume.

[Diagram with annotations]

Fig. 1 Articulatory transitions and coordination for a fricative consonant.

Simplifications that invoke concepts of natural phonetic classes [4, 5] need to be introduced, to characterise a number of different real speakers of English.

The model has a phonetic orientation, with the main turbulence noise and transient source generating constrictions specified. These are regions of small cross-section area across which a significant pressure drop develops. Aerodynamic data from real speech can give estimates for part of the articulatory descriptions required in the modelling, the time function of specified constriction areas. The orifice equation is used, with an empirical constant [6], viz.

\[ A = 0.00076 U^{0.5} \]

A = cross-sectional area of constriction in cm²
U = volume flow rate of air through the constriction in cm³/s
\( \Delta P \) = pressure drop across the constriction in cm H₂O

Constriction areas AG, AC and AV, as shown in Figure 1, together
with cavity volumes for the lungs and the vocal tract, link the articulatory and aerodynamic blocks of the model. Estimates from real speech, input to the model, give resynthesised aerodynamic traces as outputs. These can be compared with real speech traces, generalising beyond the original contexts. The data here were obtained as in a previous study, using an earlier form of the model [7]. A trace proportional to area A was obtained by means of the Aerodynamic Speech Analyser (Electronic Instrument Design, Leeds). The traces were all LP filtered at 50 Hz to partially remove the a.c. components. Undoubtedly, the methods used are only approximate and there are many sources of error [8]. Undoubtedly, also, the model is a highly simplistic representation of the complexities of speech production. However, sharing the hope of Bridle et al. [9] that "Good, robust solutions to dramatic simplifications of a real problem can be more useful than weak solutions to a more accurate idealisation of the problem", we want to try to use in the articulatory domain something comparable to their speaker-adaptive procedures in the acoustic domain. It is hoped that, with sufficient data from natural speech, solutions to many simultaneous equations might be optimised and the results used to characterise a particular speaker, for the purpose of articulatory synthesis. Some examples of results obtained follow.

**ARTICULATORY PATTERNS OF NATURAL SPEECH**

Tongue tip-blade articulation for [s] and [z] in 3 vowel contexts

The phonetic class considered here is alveolar fricatives, both voiced and voiceless. Can voiced and voiceless fricatives be lumped together? Does the vowel context affect the time course of the main vocal tract constriction A_C? Are the patterns the same for different speakers? Here four adult English speakers with near-RP accents are analysed: two women A, B; two men C, D. Figure 2 shows traces of alveolar constriction A_C articulation. They exhibit the kind of speaker-specific complexity that might be anticipated on the basis of other analyses [10]. Speaker A seems to use similar articulatory paths regardless of voicing or vowel context. The other speakers could be modelled as having a long static occlusion for [s] in some contexts; for [z] also in the case of speaker D. Some of the traces for [s] have double troughs, which suggests that the actions may be more complex than simple closure-occlusion-release. There may be oscillatory paths, possibly mediated by feedback control. But alternative interpretations associated with sources of error in the experimental techniques [8] need to be explored first, before the tentative explanations offered here are considered more carefully.

**Invariance across a change of speaking rate**

As a first step to considering what different speakers might maintain constant across different styles of speech, traces of words in isolation said at medium and fast rates by two different speakers E and F, both women, are shown in Figure 3. There seems to be overshoot at the fast rate for both speakers. At the medium rate, speaker E makes a noticeably shorter occlusion for "eyes" than for...
SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

Fig. 2 Tongue tip-blade articulation \( A_c \) for words in the frame "A...it said," for 4 speakers. Solid lines [s]; dashed lines [z].

100 ms

Fig. 3 Tongue tip-blade articulation for words in isolation, for 2 speakers E and F. Solid lines medium rate; dashed fast.
"ice", but her patterns for the two words seem identical at the fast rate. For speaker F at the medium rate, most but not all tokens of "eyes" have shorter occlusions than those for "ice"; a small difference is maintained even at the fast rate.

Larynx articulations for vowels

It is clear that larynx actions are of central importance in speech production. Unfortunately, only invasive techniques are currently available for transducing vocal fold articulation and subglottal pressure. Pursuing the philosophy of dramatic simplification, subglottal pressure is estimated here from the more accessible variable of oral pressure. The phonetic element whose larynx articulation is to be estimated is embedded in voiceless fricatives or aspirated plosives, preferably in a high vowel context, where cavity volume changes due to jaw movements are minimised [11]. The reliability of the method has been discussed [12, 13]. Subglottal pressure in a vowel is assumed to equal oral pressure in an adjacent voiceless consonant. This method has been used to characterise different kinds of singing voices [14]. Combined with an airflow measure, it has been used to derive glottal area during vowels for dysphonic speakers before and after treatment [15]. For the 4 speakers A, B, C and D, peak oral pressure for [s] in "peace" or "pass" in the frame sentence "A ... I said" was taken to indicate subglottal pressure $P_{sg}$ in the preceding [i] or [a] vowel. The minimum value of oral airflow during the vowel (nasal airflow was zero) $U_0 \text{ min}$ gave an estimate for transglottal airflow; here oral pressure was close to zero, so that glottal area was estimated as:

$$\hat{A}_g = \frac{0.00076 U_0 \text{ min}}{P_{sg} 0.5}$$

Table 1 shows the minimum and maximum for 3 tokens of each item for each speaker.

Table 1. Estimates of the articulatory (d.c.) component of glottal area during 2 vowels for 4 speakers

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Speaker</th>
<th>$\hat{A}_g$ range (3 tokens) in cm$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>A (F)</td>
<td>0.024 to 0.030</td>
</tr>
<tr>
<td></td>
<td>B (F)</td>
<td>0.038 to 0.040</td>
</tr>
<tr>
<td></td>
<td>C (H)</td>
<td>0.088 to 0.098</td>
</tr>
<tr>
<td></td>
<td>D (M)</td>
<td>0.069 to 0.085</td>
</tr>
<tr>
<td>[i]</td>
<td>A (F)</td>
<td>0.016 to 0.030</td>
</tr>
<tr>
<td></td>
<td>B (F)</td>
<td>0.021 to 0.036</td>
</tr>
<tr>
<td></td>
<td>C (M)</td>
<td>0.044 to 0.062</td>
</tr>
<tr>
<td></td>
<td>D (M)</td>
<td>0.042 to 0.054</td>
</tr>
</tbody>
</table>

The difference for each speaker across vowel context may be spurious but, if enough different vowels were analysed, an overall value for each speaker might be stated. From this admittedly very small sample there is a hint that men may operate with a larger glottal area than women. Given the likely differences in vocal fold length, this suggests that glottal width is what is controlled. A possible...
SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

reason might be offered. The Bernoulli force may operate over only a very narrow range of glottal width and is crucial for the maintenance of vocal fold oscillation [16]. Estimates of subglottal pressure are useful in their own right for characterising the respiratory component of articulation for each speaker.

Larynx articulations for consonants

The errors in estimating subglottal pressure are likely to be more serious in the final articulatory component to be discussed here, that is the abduction-adduction of the vocal folds, shown in Figure 3 for fricatives. The following data are offered very tentatively, see Table 2 and Figure 4, for speaker B only. The articulatory time path of glottal area \( A_g(t) \) is estimated from

\[
A_g(t) = \frac{0.00076 U_r(t)}{(P_g(t) - P_0(t))^{0.5}}
\]

Table 2 is for voiced fricatives (individual tokens) in the frame “Say [spa\textsuperscript{3}c\textsuperscript{3}sp] again”; an articulatory path \( A_g \) is shown in Figure 4, for one token of [v] in the frame “Just [av\textsuperscript{3}st] again”. An attempt is made to consider the effect upon \( A_g \) of a likely error in \( P_g \).

Table 2. Estimates of maximum glottal area \( A_g \max \) during voiced fricatives and during the following vowels

<table>
<thead>
<tr>
<th>Fricative</th>
<th>( A_g \max ) in cm(^2)</th>
<th>( A_g ) near mid-{3} following in cm(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>0.110</td>
<td>0.055</td>
</tr>
<tr>
<td>[z]</td>
<td>0.073</td>
<td>0.044</td>
</tr>
<tr>
<td>[s]</td>
<td>0.083</td>
<td>0.040</td>
</tr>
<tr>
<td>[v]</td>
<td>0.110</td>
<td>0.053</td>
</tr>
<tr>
<td>[\textsuperscript{3}]</td>
<td>0.087</td>
<td>0.048</td>
</tr>
</tbody>
</table>

These preliminary figures are compatible with fixed patterns of vocal fold articulation for voiced fricatives regardless of place of articulation.

Analysis-by-synthesis for improved matching

The next step will be to try out the articulatory patterns for each speaker in the model. Where the match between aerodynamic outputs from the model and the corresponding traces from the real speech is poor, the model itself can be used to indicate the correction needed. For example, in estimating glottal area during a voiced fricative it was assumed that essentially all the transglottal airflow appears at the mouth outlet. Clearly, this is not the case in general: airflow is absorbed by vocal tract cavity enlargement, both active and passive, and by the build-up of air pressure behind the vocal tract constriction. The aerodynamic equations of the model quantify these processes. It is hoped that a limited number of analysis-by-synthesis cycles will yield a converging solution to the optimisation problem.

422
**SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS**

![Diagram](Image)

**Fig. 4** Estimated glottal articulation \( \theta_g \) and its coordination with the main vocal tract articulator \( A_C \) during \( [v] \), for speaker B.

**REFERENCES**


Proceedings of The Institute of Acoustics

SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS


Acknowledgements
The assistance of Mr. E. Brearley in the experiments on real speech is gratefully acknowledged.
The computer modelling of speech production is supported by the SERC.
Session 3

Speech production simulated with a functional model of the larynx and the vocal tract

Celia Scully

Department of Linguistics & Phonetics, University of Leeds, Leeds LS2 9JT, U.K.

A composite model of speech production has been used to try to match the articulatory, aerodynamic and acoustic patterns for some real speakers of British English. The aim is to characterize particular speakers or speaker types acoustically through a match to their individual articulatory actions. In the model, the larynx contributes two independent articulatory actions; in addition it generates the acoustic source for voicing. The problems of direct matching of acoustic patterns are very great. The particular advantage of the analysis-by-synthesis methods described here is that the aerodynamic stage, which links actions to sounds, is included. As a result, the articulatory match can be made with some confidence before acoustic source and filter properties are varied; acoustic sources and acoustic pattern features of the synthetic speech co-vary in ways similar to those observed in natural speech. The model has been used to simulate disordered as well as normal larynx behaviour.

1. Introduction

The way the larynx is used, both for articulatory actions and as the generator of the voice source, may be characteristic of an individual speaker or of a speaker type. This paper outlines some attempts to match a model of speech production processes to particular speakers of English in their productions of so-called voiced and voiceless fricatives. The vocal folds need to be abducted for [s]; here voicing ceases for two reasons at least: the vocal folds move away from an ideal voicing configuration and the transglottal pressure drop is very severely reduced. The maintenance of voicing throughout the whole of [z] is a very unstable acoustic pattern feature for English speakers. For five adults saying "A hiss it said" and "A his it said", spectrograms showed a wider range of duration for this than for other segmental durations, both within and across speakers; this variation is, clearly, acceptable within the broadly-defined auditory goals for English. It is well known that the relative durations of the preceding vocoid and the fricative contoid may be very important for an English listener's judgement of contrasts such as "hiss" versus "his", but this need not always imply a contrast of articulatory timing and co-ordination (Scully, 1979).

A model of speech production processes has been made to generate sequences perceived as containing "hiss" or "his" by English listeners, in a forced-choice test, by means of a single articulatory contrast, the degree of abduction of the vocal folds,
represented in the model by the size of maximum glottal area for the fricative. In the model’s version of “his” and for four of the five speakers there was a noticeably voiceless segment for [z] (Scully & Allwood, 1985b).

The modelling of “hiss” and “his” in a frame sentence included specification of the voice source with a phenomenological model. Three voice source waveshape defining parameters, provided in Fant (1980), are made to vary with three controlling parameters. These are: (1) a slowly changing component of glottal area, \( A_g \), representing the articulatory action of vocal fold abduction and adduction; (2) the effective stiffness and mass of the vocal folds, called \( Q \); (3) the transglottal pressure drop. Factors (1) and (2) are taken to be articulatory actions, under a speaker’s control and specified as inputs to the model; the third factor is derived in the aerodynamic block of the model. The two larynx articulations are assumed to be independent, so that vocal fold stiffness does not vary with vocal fold abduction; this refinement, if shown by data from natural speech or excised larynx measurements to be appropriate, could be added. The parameter values in the table can be altered so as to simulate different larynx types, for example trained singing versus speech (Scully & Allwood, 1985a), here for a man or a woman speaker type.

2. Speakers and speech material

The results presented here are matches to two speakers of British English recorded in previous studies: speaker D (male) and speaker B (female, the present author), see Scully (1984) and Scully & Allwood (1985b). Speaker D is unusual, in that he maintains voicing throughout the voiced fricative [z]. The modelling described here attempts to match his production of six tokens of each of “A peace it said” and “A peas it said”. Speaker B produced a series [pazˈapazˈa . . . ] on a single expiratory breath, with little emphasis and without a pitch fall on the stressed syllables, the aim being to maintain subglottal pressure as constant as possible. Generally the modelling takes as its starting state the end of an inspiratory breath, ending at the end of a speech-bearing expiratory breath, but in this case it is the middle portion [VsV] or [VzV] of an utterance which is to be simulated.

3. Steps in the matching process

3.1. Articulation

The movement path for the tip and blade region of the tongue into and out of the vocal tract occlusion needed for [s] or [z] in [VsV] or [VzV] was obtained for the real speakers as a time-varying cross-section area of a constriction of the vocal tract tube in the region of the alveolar ridge. Volume flow rate of air out of the speaker’s mouth (\( U_0 \)) was measured using a mask and a pneumotachograph (Mercury flow-head with Gaeltec pressure transducer). This was taken as an indication of flow through the tongue-alveolar ridge constriction (\( U_c \)). Pressure drop across the alveolar constriction (\( P_c \)) was recorded using a polyethylene tube inserted via the mouth with its open end in the oral cavity pointing across the air stream. The reference pressure was that inside the mouth space of the mask and a Gaeltec pressure transducer was used.

The cross-sectional area of the alveolar constriction \( A_c \) has been estimated from the
Simulated speech production

409

kinetic equation

\[ \Delta P = \frac{gKU^2}{2A^2} \text{ (dyne/cm}^2) \]  

(1)

with \( K = 1 \) and \( \Delta P \) in dyne/cm\( ^2 \) (1 cm H\( _2 \)O = 980 dyne/cm\( ^2 \)), and the density of air \( \rho = 1.14 \times 10^{-3} \text{g/cm}^3 \). As discussed by Gauffin, Binh, Ananthapadmanabha & Fant (1983), the value of \( K \) varies between 0.9 and 1.5; for simplicity and consistency between investigators, a value of \( K = 1 \) is used. Substituting \( \Delta P \), Equation (1), for the alveolar pressure drop \( P_c \), we arrive at an empirical equation

\[ A_c = 0.00076 U_c/(P_c)^{0.5} \]  

(2)

with \( A_c \) in cm\( ^2 \), \( U_c \) in cm\( ^3 \)/s, and \( P_c \) in cm H\( _2 \)O. Traces proportional to \( A_c \) were obtained with an aerodynamic speech analyser (Electronic Instrument Design, Leeds).

The speaker's area-time movement path was matched in the model as closely as allowed by the constraint of an S-shaped form of area transition of variable duration and slope. This alveolar constriction of the vocal tract together with that of the glottis constituted the orifices included in the aerodynamic block of the model. Other components were the two cavities of the lungs and the vocal tract tube behind the alveolar constriction, as described elsewhere (Scully & Allwood, 1983).

The articulatory component of glottal area was estimated in the real speech, from the oral pressure and oral airflow traces, both for the vowels and for [z], as previously described (Scully, 1984). A larger value of glottal area was used for [s]. A first guess for vocal fold articulatory transition type was made, based on evidence in the literature, since a reliable indication is not available for the speakers simulated.

The vocal tract shapes used were based on previous modelling of [i] and [s] type vowels with [s] or [z] intervocally (Allwood & Scully, 1982), but the vocal tract was shortened to 14.5 cm for the simulation of the female speaker. Peak oral air pressure for preceding or surrounding [p] allophones was used to estimate pressure of air in the lungs, see, for example, Rothenberg (1982). This procedure was likely to be more reliable in the case of [psz\'z . . .].

3.2. Aerodynamics

Aerodynamic traces generated in the model were compared with those for the speaker. Timings of articulatory events, the lip release movement path for [p] and maximum glottal area for the fricatives were modified to improve the aerodynamic match. Figure 1 shows the third and fourth attempts at matching the aerodynamics of [psz\'zp . . .] together with the target real speech data. When first estimating a glottal area articulatory path for [. . . az\'z . . .], mouth mask airflow \( U_0 \) (an indication of \( U_c \)) from the real speech is used. The model itself can assist in improving the estimate of larynx behaviour: it generates volume flow rate through the glottis \( U_g \) as well as through the alveolar constriction \( U_c \). These differ slightly when oral pressure \( P_c \) or oral cavity volume \( V_c \) or both are changing. \( U_c \) can be used from one simulation to get a better estimate of larynx articulation for the next simulation attempt. In Figure 1 the synthetic \( U_c \) traces both fall too low at the end of the oral pressure \( P_c \) fall for [z]. This suggests that vocal tract volume in the model, shown as the trace \( V_c \) in Figure 1, has been made to increase too much here. The timing of the vocal fold abduction for the final [p] has been improved in the fourth synthesis compared with the third attempt, but the fact that volume flow rate \( U_c \) is still too high here suggests that the modelled vocal fold abduction is still too early.
Figure 1. Aerodynamic traces for [. . . paz'ap . . .] (a) for natural speech, speaker B (F); (b) third modelling attempt; (c) fourth modelling attempt. $U_0$ is the volume flow rate of air through the speaker's mouth, measured at a flow-head, in cm$^3$/s; $U_a$ is the volume flow rate of air through the alveolar constriction, in cm$^3$/s; $U_t$ is the volume flow rate of air through the glottis, in cm$^3$/s; $P_c$ is oral air pressure, in cm H$_2$O; $V_c$ is oral cavity volume, in cm$^3$. $U_a$ is shown as a solid line, $U_t$ as a dashed line where $U_a$ and $U_t$ are noticeably different.

Figure 2(a) shows some aspects of the model's aerodynamics for attempts at matching the real speaker's "... peace ..." and "... peas ...". Contrasting patterns for airflow $U_a$ and oral pressure $P_t$ have been obtained. The match with the real speaker's airflow traces is not yet very good, but the oral pressure patterns for ['tsz] and ['izl] are quite close to those for the real speaker.
3.3. Acoustic sources

The aerodynamic block of the model generates acoustic sources of voice, aspiration noise (at the glottis) and fricative noise and transient (injected in front of the tongue–alveolar ridge constriction). The various sources depend upon local geometry and aerodynamic conditions, including oral pressure, which is common to both glottal and supraglottal sources; therefore they covary, in forming acoustic contrasts for “... peace ...” and “... peas ...”. This is illustrated for voice and fricative noise in Figure 2(a). The right kind of voicing contrast has been obtained for this particular speaker: “peas” has no loss of voice source and 80 ms of fricative noise source compared with 80 ms of fricative noise (range 16 ms) in 6 tokens of natural speech. For “peace” the model gave a fricative source segment of 155 ms compared with, in the natural speech, 140 ms (range 11 ms).
3.4. **Acoustic output**

The final test of the simulation (apart from formal listening tests) is to compare spectrograms of the real and synthetic speech. In the final stage of the model the acoustic sources are injected at appropriate places as excitations in the reflected pressure wave method for the acoustic filtering (see Allwood & Scully, 1982). The synthetic speech lacks the full richness and complexity of the natural speech, but some pattern features, including some segment durations and some formant transitions, match quite well.

Spectrograms for “... peace ...” show voicing ending and beginning again more suddenly in the real than in the synthetic speech, at either side of the [s] frication segment. When all other aspects of the modelling have been adjusted to give a good match, the parameter values used in the table to describe the larynx as a voice source can be altered. Few data are available, especially for female speakers, but several different larynx types have been constructed in seeking a better match to the fall-off of voice amplitude and reduced high frequency associated with reduced transglottal pressure drop and increased abduction of the vocal folds for the female speaker’s [z]. Data on which to base future voice model tables are becoming available, for example through the research of Karlsson (1985).

4. **Future developments**

With a larger number of iterations, something closer to optimization of the parameters could be achieved. Some aspects of articulatory control are accessible; others, such as the total vocal tract area function for an individual speaker’s vowels, are extremely difficult to obtain from real speech. Matches for a single speaker need to be extended to different vowel contexts and different places of articulation for the fricatives, to see whether some invariance holds within particular phonetic classes for a given speaker. More real speakers need to be simulated in order to assess the capabilities of the voicing model as a realistic and internally consistent analysis-by-synthesis tool. The whole model is under-constrained and it may perhaps be useful for characterizing disordered as well as normal speech. Insufficient vocal fold closure has been modelled for a speaker with recurrent nerve palsy studied by Fritzell, Hammarberg, Gauffin, Karlsson & Sundberg (see this issue). For pre- and post-teflon injection, different values of average glottal area have been modelled for a static [a] vowel. The model generated contrasting patterns of aspiration and frication noise as well as contrasting waveshapes for voice, which yielded different spectral balances in the acoustic output.

Against the advantages of flexibility of control must be set the limitations of this voice source model. It is not a true physical model of the processes and it may prove too simplistic in concept. It has the disadvantage of a large number of parameters to be optimized when a match is attempted.

The chief point about the approach described here is that the matching procedure is broken down into several stages. Comparisons at the aerodynamic stage permit some confidence in the articulatory match, before the properties of the acoustic stage, specifically the voice source parameters, are varied.

Thanks to Miss Marion Shirt, Dr Peter Roach and Mr David Barber who acted as Subjects; and to Dr Gustav Clark and Mr Eric Brearley for assistance with the computer modelling and the experiments on natural speech. The computer modelling of speech production processes is supported by the Science and Engineering Research Council.
References


LINGUISTIC UNITS AND UNITS OF SPEECH PRODUCTION

Celia SCULLY
Department of Linguistics and Phonetics, University of Leeds, Leeds LS2 9JT, United Kingdom

Received 19 September 1986
Revised 14 January 1987

Abstract. Links are needed to bridge the gap between the analysis of speech as a set of discrete, ordered but durationless linguistic units and analyses of the continuously changing acoustic signals, defined along a time axis. Current recognition and synthesis devices do not make good use of the structure imposed by speech production processes on the mapping between an allophone sequence and the many possible associated speech signals. A quantitative, flexible articulatory time framework has been developed as a contribution to the new kinds of phonetic descriptions needed. Units of articulation for allophones of the phonemes of British English and methods for linking adjacent allophones are proposed. Tentative specifications for a sub-set are offered, based on a review of published findings for natural speech.

Articulatory schemes are taken to be organised with reference to particular events $E$. Pairs of events need to be appropriately coordinated in time. The two events may relate to inter-articulator coordination between two different quasi-independent articulators or to the durational extent of a statically maintained state for a single articulator. The coordination between the two events is expressed through the duration $D$ of the time interval between them. Six examples are given of the construction of a complete articulatory time plan for an English sequence. This forms the first stage for a computer-implemented model of the articulatory, aerodynamic and acoustic processes of speech production. The synthetic speech output from the model is given acoustic variations intended to mimic those arising in natural speech due to a speaker's choice of options, including a change in rate of speech. This is achieved in the modelling by altering one or more $D$ values in the articulatory time plan and by dispensing with some optional actions. The variability of multiple repetitions by a real speaker can be introduced into the synthetic speech by perturbing the $D$ values. The model needs to be matched to specific real speakers in order to assess the extent to which it is realistic in its simulation of the variation and variability of acoustic pattern features for natural speech and the extent to which covariations can be predicted with it.


Résumé. Des liens sont nécessaires pour surmonter le fossé entre l'analyse de la parole en tant qu'ensemble d'unités linguistiques discrètes, ordonnées mais atemporelles et les analyses de signaux acoustiques continûment variables dans le temps. Les dispositifs actuels de reconnaissance et de synthèse usent mal de la structure imposée par les processus de production de la parole lors de l'application d'une séquence allophonique sur la pluralité des signaux de parole y associés. Nous avons développé en tant qu'aide à la description phonétique un cadre de référence articulatoire à la fois quantitatif et souple. Nous proposons des unités articulatoires pour quelques allophones de phonèmes anglais ainsi que des méthodes de liaison avec des allophones adjacents. Des spécifications préliminaires pour un sous-ensemble d'allophones sont proposées qui sont basées sur des résultats déjà publiés.

Les schémes articulatoires sont organisés par rapport à des événements saillants $E$. Il s'agit alors de coordonner dans le temps des couples d'événements. Ces derniers peuvent concerner la coordination inter-articulateurs entre deux articulateurs différents et quasi autonomes ou concerner un seul articulateur maintenu dans un état statique. La coordination entre deux événements est exprimée à travers la durée $D$ de l'intervalle temporel qui les sépare. Nous donnons six exemples de construction d'un plan articulatoire complet pour une séquence d'anglais. Ce plan forme le premier module d'un modèle numérique des processus articulatoires, aérodynamiques et acoustiques de la production de la parole. Le signal ainsi synthétisé est acoustiquement modifié afin de simuler les variations observables en parole naturelle et dues aux options des locuteurs, y compris les modifications du débit d'élocution. Ceci est obtenu en altérant une ou plusieurs valeurs $D$ dans le plan temporel et en négligeant quelques actions optionnelles. La variabilité observée lors de multiples répétitions par un locuteur réel peut être simulée en perturbant les valeurs $D$. Le modèle réclame d'être confronté à des locuteurs réels afin d'évaluer le réalisme de sa simulation de la variation et de la variabilité des traits acoustiques en parole naturelle ainsi que son degré de prédictabilité des covariations.

Key words. Articulatory synthesis, units of speech production, linguistic units, coordination, rate of speech, temporal organisation, variation, covariation, variability, composite model of speech production processors.

CONTENTS

ORGANISATION OF THE PAPER 79
MEANINGS OF THE SYMBOLS 80
INTRODUCTION: LINGUISTIC UNITS AND UNITS OF SPEECH PRODUCTION 82
CHAPTER I. THE MODEL OF SPEECH PRODUCTION 86
1.1. Aerodynamics 87
1.2. Acoustic sources 87
1.3. Acoustic filtering 88
1.4. The articulators 88
1.5. Instructions for movement paths with conventional time scales 89
1.6. Instructions for movement paths with events $E$ 90
1.7. Coordination expressions: $D$ and $E$ 91
1.8. Articulatory transitions: basic types 92
1.9. Complex articulatory transitions 92

CHAPTER II. THE ARTICULATORY MODEL AND THE DATA BASES IN RELATION TO DATA FROM NATURAL SPEECH AND OTHER MOTOR SKILLS 93
II.1. 'Icebergs' in natural speech and transitions in the model 94
II.2. General constraints in speech communication 95
II.3. Rate effects 95
II.4. Articulators and degrees of freedom: motor skills 96
II.5. Tongue body constraints 97
II.6. Movement path dynamics and kinematics 97
II.7. Transition end points and articulatory undershoot 98
II.8. Variability in speech and other motor skills 99
II.9. Variability in speech: end point states and mid transition portions 99
II.10. Variability in speech: end point states for different articulators 100
II.11. Variability in speech: end point states for different regions of the vocal tract in vowels 100

Speech Communication
Organisation of the paper

The articulatory time plan framework described here arose from our synthesis of British English speech, by means of a computer-implemented mathematical model of speech production processes. It became necessary to introduce a generalised time description in order to avoid laborious recalculation of the whole articulatory plan whenever one or two aspects of timing were to be changed.

The model is a means of generating speech-like sounds from a defined set of actions. The articulatory phonetic knowledge needed to devise time plans is set out in phonetic data bases. The rules for defining al-
lophone onset and offset and for linking pairs of allophones are applied by hand at present, but could benefit from being expressed in PROLOG or LISP algorithms. Application of the rules to a given linguistic string results in coordination expressions of the form $E_a := E_b + D_c$ which, together with specified $D$ values, are entered into the first, articulatory block of the composite model of speech production.

The paper is divided into four chapters as shown in the Contents list. Chapter I, Sections 1.1, 1.2 and 1.3 describe in outline the later stages of the model. There are references to previous publications containing fuller descriptions, but the details of these automatic physical processes and their treatment in the model are not important in the context of this paper. Section 1.9 may be omitted, since complex articulatory transitions are not used in any of the examples given.

Chapter II, Sections II.1-II.5, II.7 and II.19 include discussion of a number of issues concerned with the organisation of speech production and the representation of articulatory processes in the model and the data bases. The remaining sections, which are mostly concerned with some of the supporting evidence from natural speech for the choices of end point states and coordination $D$ values in the data bases for some allophones of English, may be omitted.

Chapter III, Sections III.1—III.6, III.9 and III.10 are needed for a general understanding of the structure of the data bases, the kinds of phonetic knowledge expressed in them and the procedures for linking allophones to form a completely specified articulatory time plan. The remaining sections in this chapter are detailed data bases specifications and may be omitted. Many of the values in these data bases are speculative, being based on insufficient data for even one speaker of R.P. English. The detailed descriptions given in these sections are not essential for an understanding of the articulatory time plans. Instead, these are intended to be set out in and explained by the following sketches: Fig. 8 for stressed vowels; Fig. 9 for non-nasal plosives, fricatives and affricates and for approximants; Fig. 10 for nasal consonants; Fig. 11 for a glottal stop and for [h]: Fig. 12 for speech-initial and speech-final.

Chapter IV contains six examples of the construction of an articulatory time plan. All except Example 4 have been used to generate synthetic speech-like sounds. Example 4 is a theoretical discussion and exemplification of one of the many ways in which phonetic reorganisation might take place in changing from a full formal style to an informal style. The first example, in Section IV.2, goes through the procedures in full. As in the case of the articulatory time plans for individual allophones, Figs. 13–16 are intended to stand on their own as descriptions of articulatory time plans appropriate for generating the target words or sequences of English. Many different acoustic signals could be generated from each of these time plans, simply by perturbing some or all of the various $D$ values shown in the sketches. More variations still could be generated by perturbing the end point states and the transition types.

Only a few mathematical symbols and very simple mathematical expressions are used. These are defined in the next section. The intention has been to use symbols consistently in the articulatory time plan sketches, Figs. 8–16. The $D$ and $E$ labels have specific meanings which may not be self-evident from the figures. For this reason, although they are defined in various sections of the text, it seems useful to list them next. With the system of symbols, the series of articulatory time plan sketches should be able to express in diagrams what the text says in words.

**Meanings of the symbols**

**Phonetic symbols**

Labels and classifications for some vowels and consonants of English are given in Tables 2, 3 and 5. IPA notation is used as far as possible. The vowel and consonant symbols for English conform to a set recommended for the United Kingdom (Wells, 1986).

**Mathematical symbols**

- $>$ is greater than.
- $>=$ is greater than or equal to.
- $<$ is less than.
- $\leq$ is less than or equal to.
- $==$ is set equal to.
- $(+/−x)$ means timing perturbations of as much as $+x$ or $−x$ in 5 ms time units.

**Axes**

In Figs. 3–16 the horizontal axis, the abscissa, always represents time. Although each figure is intended as a sketch the durations shown are meant to be realistic. Figures 8–16 show how the movement paths of up
to six articulators are coordinated. The articulators are listed in Section 1.4. For each articulator’s graph the vertical axis, the ordinate, sketches the parameter values, with positive values above the base line. Except in the case of the tongue body, jaw and lips (TB, AJ, AL), the direction of change of the parameter values is shown correctly and the base line has value zero.

Transitions

The start of each articulatory transition is marked by a circular dot • and its end by a line I. In Figs. 8, 9, 10 and 12, which are generalised sketches applying to several members of phonetic classes relevant to English, the type of transition is not specified since this is likely to differ for different members of the class, for example for different places of articulation. In Figs. 3, 4, 6 and 11 and in the examples which show the construction of a time plan for a sequence of allophones, Figs. 13–16, specific transition types have been selected from the data bases. In this case each transition has an arrow and a label, as follows: S (Slow), M (Medium), F (Fast) or, for example (16.100) (a transition taking 16 × 5 ms with a slope factor of 100). These are described in Section 1.8 and Fig. 5.

The completion of a particular transition may constrain the total time nearby. In that case it is called the Duration Determining Factor DDF.

Articulatory events E

Each event label includes E or .E.

After E comes the name of the articulator, for example AG for glottal area: thus this particular event is labelled EAG. Events are at precisely defined time points, marked on the time plan sketch with a short line parallel to the ordinate. Events are located either at the beginning of a transition of the articulator concerned, for example EPLU, or at the end of a transition, in which case the label ends with X, for example EPLUX. The event label can begin with the phonetic symbol for the specific allophone described, for example d.EAFX and d.EAF in Fig. 6 for [d] and s.EAG in the phonetic data base for [s]. Alternatively the label can begin with the symbol for a phonetic class [V] (vowel) [C] (consonant) or [#] (speech-initial and speech-final).

In Figs. 8–11 abbreviated E labels are used; they refer to the vowels or consonants characterised in that sketch. In Figs. 13–16, where several such phonetic elements are strung together, each is labelled sequentially [#1], [#2], [C1], [C2] and so on. [V1], [V2] and so on: the associated event labels are similarly named, for example C1.EAG or V2.EQ.

Event labels always appear just under the base line for the articulator concerned. The names always agree, except that the generalised articulator name AC is used in labels for the main articulator for consonants. In Figs. 8, 9, 10 and 12 the trace is also labelled AC for generality. AC stands for different articulators for the various places of articulation here: AL for labial, AF for alveolar or AN for velar. In Fig. 11 and Figs. 13–16 the trace for the main articulator is given both names. Thus in Fig. 13, AG is the main articulator for [C1] [2]: Fig. 14 has no main articulator since there is no consonant; in Fig. 15(b) AF is the main articulator for both [C1] [s] and [C2] [t]: in Fig. 16(a) [C1] and [C2] have two different main articulators AG and AF respectively, so their event labels are, correspondingly, C1.EAC and C1.EACX for the first consonant [2] and C2.EAC and C2.EACX for the second consonant [s].

Here AG switches its role from main articulator for [C1] to the larynx articulator for [V2] and its event labels switch names correspondingly.

Of the events shown on the articulatory time plan, one and only one is taken to be the onset of each allophone included there EON; one other event defines the offset of each allophone EOFF. The choice of events is discussed in Sections III.3, III.6 and III.18. For consonants, onset EON is always C.EACX and offset EOFF is always C.EAC. For vowels, onset and offset may each be defined by different articulators.

Onset and offset events are marked by lines extending from top to bottom of the sketch across all the movement paths for the independent articulators. They are labelled #1.EON (always at the first time point ET0) #2.EOFF C1.EON C1.EOFF V1.EON V1.EOFF and so on. #2.EOFF is always the final time point in the plan ETMAX. On the figures, these onset and offset labels are placed so that the . is just before (to the left of) the event time point, with E just after it.

Articulatory durations D

A D value Dc gives quantitative expression to the coordination between two articulatory events Ea and Eb, through an expression of the form

\[ Ea := Eb + Dc \]

This means that event Ea is dependent on event Eb. Ea must be correctly timed relative to Eb. Dc is the duration of the time interval between the two events. If event Ea lags behind event Eb, this time interval is
given a positive duration value, so that \( D_c \) is positive; if \( E_a \) occurs before \( E_b \), \( D_c \) is negative. The time point of event \( E_a \) is set equal to the time point of \( E_b \) plus the time interval specified by \( D_c \).

These time interval durations are shown stippled on the articulatory time plans as part of the trace for the articulator which contributes the dependent event of the pair. Their labels appear just above or to the side of this stippled block. The full \( D \) label is used in the phonetic data bases. This contains \( D \) preceded by an allophone label and followed by the independent and then the dependent articulator in brackets. For example in Section III.11 in the data base for \([t]\) the \( D \) label t.D1 (AC-AG) expresses coordination of the dependent articulator AG with the independent articulator AC, the main articulator for \([t]\). The form t.D4 is used here also.

Figures 8-12 use abbreviated \( D \) labels D1-D8, understood to refer to the allophone or phonetic class characterised in the figure.

In Figs. 13-16 each \( D \) label shows the allophone to which it belongs. For example in Fig. 13, V1.D1 belongs to \([V1]\); C1.D4 belongs to \([C1]\). This form of label is used in the text of Chapter IV.

\( D \) has a particular significance: both vowels and consonants continue for \( D_4 \) in the sense that this is the time interval between the onset event \( E_{ON} \) and the offset event \( E_{OFF} \). All the \( D \) labels refer to a time interval between two events, but \( D_4 \) has an additional meaning in that it is also a measure of the articulatory duration for the allophone element to which it belongs. Other \( D \) labels are numbered consistently: in one way for all vowels, in a different way for all consonants and in another way for speech-initial and speech-final.

For vowels, all coordination is expressed relative to vowel onset \( V.E_{ON} \) whichever articulator's event defines onset, as shown in Fig. 8. Here D1 and D5 describe timing for Q; D2 and D6 for PLU; D3 and D7 for AG in those cases where AC or TB defines onset; D8 and D9 for AV.

For consonants, durations D1, D2 and D3 relate events for specific articulators, AG, TB and AV respectively, with consonant offset \( C.E_{OFF} \) (C.EAC) while D5, D6 and D7 are the corresponding coordinations with consonant onset \( C.E_{ON} \) (C.EACX). as shown in Figs. 9 and 10.

For speech-initial \([#i]\) and speech-final \([#f]\), D1 and D2 relate events for AG and AV respectively to events \#i.EOFF or to \#f.EON. Linking durations express the meshing together of successive allophones in the string as described in Section III.4. Here \( D \) means the duration of the time interval between the offset event \( E_{OFF} \) of the preceding element and the onset event \( E_{ON} \) of the following element. The \( D \) value is positive if \( E_{ON} \) lags behind \( E_{OFF} \) and is negative if \( E_{ON} \) occurs before \( E_{OFF} \). These linking \( D \) labels contain two element names, the two which the \( D \) value links together. These durations are shown hatched near the top of the time plans.

In Fig. 13 for example #1C1.D links \([#1]\) \([#i]\) to C1 \([#2]\). It is negative because the offset event \#1.EPLUX for \([#1]\) comes after the onset event C1.EACX for \([C1]\). C1V1.D linking \([C1]\) and \([V1]\) is positive and equal to a Fast (F) transition of AG (AC1) in this homorganic link.

End point states

The labels for articulatory end point states are intended to be self-explanatory. They are defined with tentative values in Section III.2 and Table 4. Phonetic labels are used for tongue body, jaw and lips (TB, AJ, AL), for example \([V]\) in Fig. 8, \([c]\) in Figs. 9 and 10, \([\alpha:\]\) in Fig. 13. TB, AJ and AL comprise five or six parameter points whose individual values are given in the vowel and consonant data bases and in Tables 4 and 5.

Introduction: linguistic units and units of speech production

Spoken language is flexible and appropriate for the circumstances of the discourse. Variety, both across speakers and for a single speaker, may be discerned at all levels of analysis, from the choice of lexical items and grammatical structures for sentences to subphonemic details of phonetic structure. The term 'speech style' will be used here to denote a totality of phonetic variations.

Style is taken as a continuum ranging from formal, which is likely to have a slower speech rate, to informal, which is likely to have a faster one.

Different styles are likely to be characterised phonetically by changes in intonation pattern, in numbers of stressed syllables and weak forms, in the frequency and regularity of elisions, in frequency of assimilation, in pace (rate) fluctuations...
and by the audibility of plosive releases (Ramsaran, 1983). A change of speech style relates to the information which a speaker needs to convey. Speech production is a goal-directed activity; the auditory goal is adjusted to suit what is already known by the listener and other aspects of the communication context (Lindblom, 1983). Change of style includes the reorganisation of the phonetic material.

This richness and complexity is notably lacking in the currently available speech synthesis systems. Although articulatory synthesis is not necessarily a viable approach to take for improved naturalness in the next generation of synthetic voices, quantitative consideration of articulatory processes may, at the least, help to explain and formulate rules for some of the complexity found in natural speech signals, with applications for speech recognition and for acoustic-based synthesis by rule.

It seems possible also that acoustic variation from one token to another may contribute to the intelligibility of consonants in natural speech (Clarke, Dermody and Palethorpe, 1982). In one study, repetition increased intelligibility of natural [CV] stimuli much more than that of synthetic ones, suggesting that the richness of structure of natural speech is useful for intelligibility (Pollock, 1959). It seems probable that, if phonetic variety could be introduced into artificial voices, not in an ad-hoc way, but in ways which correspond to articulatory, aerodynamic and acoustic processes in natural speech production, the resultant synthetic speech would probably be more intelligible as well as more acceptable. The view adopted here is that, while the goal of speech production is a broadly defined sequence of sounds, details of acoustic structure and, in particular, covarying bundles of acoustic pattern features are rather strongly constrained by the processes and requirements of speech production. Our modelling has demonstrated that a single articulatory contrast generates a bundle of acoustic pattern feature contrasts and that small repertoires of simple contrasting actions for a few independent articulators combine to give numerous acoustic variations having an appealingly lifelike complexity. It is possible, therefore, that the characterisation of a particular speaker type, whether for speech recognition or synthesis purposes, can be stated more economically in articulatory terms than in the acoustic domain.

The articulatory framework proposed here may perhaps appear unduly complex, with about seven events needed for each consonant and up to nine for some vowels. But there is good reason to believe that greater complexity of analysis will be required if future human-machine communication systems are to capture to any significant extent the richness of natural speech (Lindblom, 1983).

In an earlier paper (Scully and Allwood, 1983) the basis for constructing articulatory time schemes for input to the first, articulatory, stage of a composite model of speech production processes was outlined. A severe problem was identified: that of trying to match up incompatible values of articulatory components across time-domain boundaries between successive allophones for the string of phonemes representing the word of English to be modelled. This paper offers a solution to that problem. It describes a theoretical framework for a description of speech production units and their linking to form utterances. The description becomes simpler if it is assumed that the members of natural phonetic classes (IPA, 1979; Stevens, 1979) share some common articulatory features. It is hoped that the principle may apply in natural speech, although the units may well be differently grouped by different speakers. Some aspects of articulatory coordination for a few speakers of English are being investigated to see whether their speech production can be represented by the framework described here, in a not too Procrustean manner, and whether notions of natural phonetic classes produce simpler characterisations for individual speakers.

At one level of description speech is commonly viewed as a string of discrete, ordered but durationless linguistic units. Some new kind of phonetic description is needed to portray the flexible and varied relationships between these linguistic units and a speaker's articulatory plans, from which appropriate speech signals, continuously changing and defined with reference to a time axis, may be generated. Articulatory descriptions must be multidimensional and the various ar-
ticulators do not all move simultaneously from one target configuration to the next. Therefore, if articulatory segments were to be established, they would lack clearly defined boundaries in time; they would have 'ragged edges' because of this asynchrony of articulatory gestures. Abutting units of speech production would have to accommodate to each other somehow.

We attempt to establish automatic procedures for the mapping from a linguistic description of a word or short phrase onto a rather large number of possible articulatory plans for a synthetic utterance. Several kinds of phonetic variation are intended to be simulated: (a) allophonic variation within one style of speech for a single speaker; (b) inter-speaker variation, the different options selected from those available by different individuals having the same accent of English; (c) different options selected by one speaker for different styles of speech; (d) variability which is the lack of perfect precision in human actions, the apparently unavoidable perturbations of timing and of end points for speech production gestures.

The word or words of English to be spoken will be described as a sequence of phonetic units called allophones. In the data bases that form part of this scheme, the description of the allophones for a phoneme of English includes some phonetic attributes that are assumed to be context-free; others that are dependent upon immediate phonetic context, defined by the immediately adjacent allophones. A wider domain for context will be required to give appropriate relative prominence for nearby vowel allophones. This will be considered in a preliminary way here for [3:n:'s:].

As an example of the kinds of phonetic knowledge to be represented in the data bases, consider [t], the allophones for the /t/ phoneme of English. [t] is assumed to be always an alveolar plosive, but it will be aspirated [t, h] in some contexts and unaspirated [t, o] in others. This is a simplified phonetic description for some common allophones.

We leave for later the consideration of assimilation and other processes. Examples might be: in "eighth" [t] before [e] becomes dental [t, n]; in "that case" /t/ may be realised as closer to [k] than [t] due to assimilation of place across a word boundary (Gimson, 1980); in very informal conversational speech assimilation of manner may occur, with plosives 'weakening' to fricatives and fricatives becoming approximants (Roach. 1983). At this stage, when an utterance of more than one word is to be simulated, word boundaries are represented in the allophone string, so that the kinds of effects mentioned above may be incorporated later. In each simulation the style of speech is specified as formal, rather informal, very informal and so on.

It is assumed that individual articulator gestures are constructed around particular events and that individual movement paths and inter-articulator coordination can be expressed by the way in which pairs of these events are related to each other in timing as defined by a single duration. Since the duration may be varied the movement paths and their coordination are flexible. The statements are applied first to each individual allophone in the string and then to successive allophones in order to link them together. Six quasi-independent articulators are used to construct articulatory time plans. An event $E$ is associated with one particular articulator and a precise point in time. The choice of what in the movement path to label as an event must be to some extent arbitrary at the present. Here events are taken to be located at the start or termination of an articulatory transition. These end points will not be referred to as 'targets', since that term seems to imply that end points are more invariant, in some sense more important and salient, than the transitions themselves. This may well be far from the truth. Analysis of tongue and lip movement paths suggests that variability is lower for the middle portions of some transitions than for their end points. Some mid-transitional sections, called 'icebergs', may remain constant as phrase structure varies, while the coordination between icebergs across different articulators changes significantly (Fujimura, 1981).

Two consecutive events for an articulator $AY$, labelled $EAY$ at the start of a transition and $EAYX$ at its end, define the transition between. Coordination between different articulators is clearly an important aspect of speech production and needs to be precisely defined. It needs to be flexible also, in a precisely quantifiable way. Here inter-articulator timing is expressed by the time
separation between two events, one for each of the two articulators concerned, labelled as a duration, \( D \). Once an appropriate articulatory time plan has been devised, through trial and error with auditory monitoring, it can be perturbed in the modelling by varying the \( D \) values and the transition end point configurations, between limits consistent with the variability observed in natural speech.

For each allophone, one of its events defines articulatory onset and is called \( E_{ON} \); a second event, defining offset, is called \( E_{OFF} \). Other events needed for that allophone are coordinated with respect to either the onset or the offset event, not the other way round. These onset and offset events are used for linking successive allophones. A single duration value, the linking \( D \), coordinates the offset event for the earlier allophone with the onset event for the following one. If the link is homorganic it has a fixed positive \( D \) value. A non-homorganic link is not limited in this way: its linking \( D \) may take negative as well as positive values, with the result that events for adjacent allophones can be more or less tightly meshed together.

Successive allophone units do not abutt each other in the time domain. The linking need not be sequential in the way suggested by the ordering of the allophones in the string. A time located boundary between allophones is never defined, either in the articulatory domain or in the acoustic output. It may be helpful to segment output spectrograms, or other acoustic displays, distinguishing, for example, generally vowel-associated portions called vocoids and generally consonant-associated portions called contoids. However, perceptually we are not concerned here with which portion of the acoustic signal conveys to an English listener the presence of a particular vowel or consonant. What matters is whether the whole synthetic utterance is acceptable to listeners as the target word or words of English for a stated speech style.

Events \( E \) and durations \( D \) constitute the articulatory timing parameters. A set of expressions, each containing two \( E \) labels and one \( D \) label, with allophone data bases specifying appropriate transition types and \( D \) values, and rules for linking between successive allophones, together define an articulatory time plan and thus map an ordered string of linguistic – allophone – units onto the time domain in a defined quantitative way.

The descriptions and rules are based on previous experience with speech production modelling for a small subset of English words and from phonetic knowledge communicated in the literature. Some of the relevant publications will be cited in Chapter II. The evidence from natural speech includes other languages besides English and other accents besides R.P. It is assumed that some basic articulatory constraints, such as temporal perturbations and some transition durations, are likely to be independent of the particular language, as are, also, perceptual constraints such as the ability to perceive a stop or burst phase. The information will be most relevant to our modelling where the phonetic characteristics match allophones of phonemes for an R.P. accent of British English. Particular values found in natural speech are to be taken as a guide only; they can at least provide a plausible starting point for processes of trial and error with auditory monitoring and with acoustic comparisons. Useful ranges of values need to be established for the style being modelled. Apparent disagreements and discrepancies in the speech literature may well reflect differences of speaker type or of linguistic and phonetic context that are not apparent in broad phonetic transcriptions. We intend to be cautious about accepting apparent constraints as universal ones.

The notion of 'coarticulation' used as a blanket term to describe the influence exerted by adjacent allophones on each other, and embodied in terms such as forward (anticipatory) versus backward (perseverative) coarticulation is rejected here as being insufficiently specific. Instead coarticulation will be defined as the freedom that a real speaker, or a modeller, has to vary the way in which the articulatory events mesh together. This concept of coarticulation is given quantitative expression by the actual \( D \) duration values selected from the allophone data bases and the linking \( D \) values chosen. The former might be loosely termed internal coarticulation for a given allophone; the latter, external or linking coarticulation.
Flexibility is not confined to manipulation of these two types of D values. Some actions are considered to be optional. They are likely to be present in a formal speech style, but may be omitted altogether in an informal style. This will be illustrated with the sequence [əː].

Chapter I. The model of speech production

Introduction

The phonetic knowledge needed for events and durations descriptions of simulations with a model must, of course, come from the production of natural speech. Much of the information needed is not available; it will perhaps prove extremely difficult to obtain, even with the new X-ray microbeam and other techniques being developed. While the model does need to be made consistent with known facts about speech, it can contribute to understanding on the basis of incomplete specifications. Where coordination is known, the model can be driven outside the natural speech range, producing intermediate articulations which would be extremely difficult, if not impossible, for a speaker with habitual, well-established contrasting sets of actions to produce. Simulation of deliberately wrong timing can give insight into reasons for the right timing. The model may be used, also, as an analysis-by-synthesis tool. Articulatory parameters which are not known for the natural speech can be varied in the model until its aerodynamics traces and its acoustic outputs match the natural speech patterns well (Scully, 1984a. 1986 in press; Scully and Clark, 1986).

Automated optimisation procedures have been applied to articulatory patterns for fricatives (Shirai and Masaki, 1983) and to a diphthong (Flanagan, Ishizaka and Shipley, 1980).

An apparent weakness of these methods is that, with many parameters under the experimenter’s control, too many good matches can be obtained. Our experience so far does not confirm these fears: it is extremely easy to get things wrong and very difficult to achieve a lifelike simulation except through repeated trials, but in working towards one, some of the problems solved by real speakers become apparent.

The end product of the computation of a complete events and durations specification for an articulatory plan is a set of input instructions to drive the first, articulatory, block of a composite model of speech production processes. The stages of speech production modelled and the block structure of the model are shown in Fig. 1. The modelling for each stage has been described elsewhere (Allwood and Scully, 1982; Scully and Allwood, 1983. 1984. 1985a. 1985b). The stages following articulation will be outlined first.

---

**Fig. 1.** Block structure for the model of speech production processes, implemented on a VAX 11/780 computer. The sample time at the output of each block is given.
I.1. Aerodynamics

The aerodynamic block of the model, which links actions to sounds, ensures that acoustic sources and filters covary automatically in ways similar to that seen in natural speech. This is a highly simplified, phonetically oriented representation of the fluid mechanics processes of speech production. The term used here 'aerodynamic' will mean the low frequency components of air flow and air pressure, excluding acoustic frequencies. When traces of air pressure and volume flow rate of air computed in the model are compared with corresponding ones for natural speech, aerodynamic reasons for particular articulatory options may become apparent.

Figure 2 shows the parameters for articulatory shapes and movement paths; these are described below. It includes the aerodynamic parameters also. The lungs are represented by a single cavity, linked by a lung volume-dependent flow resistance (conductance GSG) to the glottal orifice AG, which connects the subglottal resistance to the second cavity, the air enclosed in the vocal tract, of volume VC. This cavity has two outlet orifices in parallel: AV, the velopharyngeal port and AC, the main constriction in the oral vocal tract. It is assumed that the speaker controls the pressure of air in the lungs PLU.

A single vocal tract constriction cannot realistically represent in one run constrictions for all combinations of vowels and consonants of English. But for two nearby places of articulation, as in words containing both labial and alveolar allophones, for example, it is considered reasonable to lump the two constrictions together, forming an equivalent composite constriction AC for use in the aerodynamic block. AC is a composite of AL and AF in this example. Some preset parameters define aerodynamically relevant aspects of speaker type. These are: initial lung volume VLU0, vital capacity, maximum subglottal airways conductance and the compliance of the walls of the vocal tract CW. This last controls how the cavity volume changes with changes in oral pressure. The volume of the vocal tract cavity at atmospheric pressure is obtained from the articulatory block of the model by simple summation of all the cross-section areas enclosed between the glottis and the main constriction of the vocal tract. Variables computed, through numerical solution of the differential equations, include low frequency components of subglottal pressure PSG and oral pressure PC, and of volume flow rate of air through the glottis UG, the velopharyngeal port UN, and the oral vocal tract constriction UC.

I.2. Acoustic sources

Acoustic sources for voice and aspiration noise just above the glottis and for frication noise and a transient located in front of the vocal tract constriction are derived from the local articulatory and aerodynamic conditions. The turbulence noise sources are modelled as white noise, with envelope amplitude defined by the constriction area and the pressure drop across it. The glottal noise source is filtered to give a spectrum similar
to that in real speech. A pole pair with centre frequency 1000 Hz and bandwidth 2000 Hz is used as a preliminary attempt to take Strouhal number into account (Stevens, 1972a). The magnitude of the transient source associated with the sudden linking together or separation of two regions of different air pressure (Fant 1960, pp. 279–280) is the value of the time derivative of oral pressure. The voice source model derives the acoustic component of volume flow rate of air through the glottis, a quasi-periodic waveform, from a linear representation of the combined effects of three controlling factors: the low frequency (articulatory) component of glottal area AG, the pressure drop across the glottis (PSG − PC), and the effective stiffness and mass of the vocal folds Q (Scully and Allwood, 1985a).

1.3. Acoustic filtering

The response of the vocal tract acoustic tube to each source is obtained using the reflected pressure wave method (Kelly and Lochbaum, 1962) with added losses (Scully and Allwood, 1982). The output of the whole model is synthetic speech or singing.

Each block of the model contains parameters which are preset for a particular run of the model. By changing their values different speaker types can be modelled. For the present purpose it is assumed that all parameter values needed to run the model have been specified, so that the independent variables are purely those of the articulatory time plan.

1.4. The articulators

The first stage of the model requires an input in which are specified, along a time base, the actions of a number (about 6 to 8) of quasi-independent articulators. Figure 2 shows the set used. They are:

**AG.** The slow, articulatory, action of altering the glottal area AG, by abducting and adducting the vocal folds. This does not include the acoustic component, the vocal fold vibration involved in the creation of the voice source.

**AC.** The main oral vocal tract articulator for a consonant. AC is the general name for the cross-section area of the constriction which defines the place of articulation for the consonant. AC can represent nearby but different parameter points during the course of an articulatory time plan. For example, in "purse" AC1 is AL, the lip outlet area for [p], while AC2, for [s], is AF, a constriction created between the tip-blade region of the tongue and the alveolar ridge in the roof of the mouth.

**TB.** The tongue body. This comprises several parameter points whose constriction cross-section area transitions coincide in time. These are generally AE, at the entry to the larynx cross-section area. AP in the pharynx and AB near the back of the tongue in the oral cavity. For [u:]-type vowels AN is required in addition, located near the nasal entry point of the vocal tract.

**AJ.** The cross-section area AJ most affected by jaw actions and located just behind the teeth. For simplicity this is usually given the same transition time as the tongue body points and its actions coincide with theirs. The cross-section area at the teeth AT is tied to the value of AJ.

**AL.** The lip outlet area AL in its general vocal tract shaping role; made to perform movements in synchrony with the tongue body.

**Q.** The laryngeal control factor for fundamental frequency $F_0$. This is a label for the effective tension and mass of the vocal folds. Its transitions are taken to be associated with stressed vowels, but they may need to extend across consonantal portions of an articulatory time plan, also.

**AV.** The area of the velopharyngeal port, indicating the position of the velum.

**PLU.** The pressure of air in the lungs. This is taken to represent the subglottal articulator, controlled by respiratory muscles.

The supraglottal parameter points are intended to define regions of the vocal tract. They execute changes of cross-section area which define articulatory transitions. The total changing area function is derived by linking the parameter point values together. A logarithmic curve is used for the interpolation, so that area $A$ as a function of
distance \( d_{fg} \) has gradient \( dA/d(d_{fg}) \) zero at each parameter point, with a maximum gradient value halfway between each pair of adjacent points and a smoothly changing resultant area function. The position of each vocal tract parameter point (distance from the glottis, \( d_{fg} \)) is preset, but can be varied during a simulation by adding instructions to control \( d_{fg} \) for individual parameter points. Within one simulated utterance an individual articulator can change its role. In "purse", the example given for AC above, AL functions as the main articulator for [p] at first, then assumes its general vocal tract shaping role for the rest of the actions for "purse". In Example 1 in Section IV.2, AG functions as the main articulator for [ə] at first, then resumes its more usual role as a laryngeal articulator, needed for vowels, consonants and speech offset or onset.

1.5. Instructions for movement paths with conventional time scales

For a monosyllabic word of English each articulator is likely to require from 2 to about 6 lines of instructions, specifying starting state, transition, and next transition end point state. The transition end points are given labels to which numerical values will be attached when a particular set of options is chosen. Each value may be tightly defined in the relevant phonetic data base or it may represent a choice from an allowable range. For example, if lung air pressure is simply to be raised to a suitable value for speech and lowered at the end of the utterance, then instructions for PLU could be (see Fig. 3):

\[
\begin{array}{|c|c|c|}
\hline
\text{Line} & \text{End point} & \text{Time allocated} \\
\hline
0 & \text{zero} & 0 \\
1 & \text{normal} & 104 \\
2 & \text{zero} & 24 \\
\hline
\end{array}
\]

(For time, 5 millisecond (ms) time units are used in all the articulatory plans.)

Starting from PLU value 'zero' (line 0), PLU is to be increased to 'normal' (say 8 cm H\(_2\)O) by means of a medium transition (M, i.e., 24 x 5 ms) and then stay at this 'normal' level for a further 80 x 5 ms beyond the termination of the transition time, that is, until 104 x 5 ms have elapsed from the starting time. At time 104 x 5 ms PLU is to begin to fall to 'zero' (with value 0) by means of a medium (M) transition (line 2), reaching 'zero' 128 x 5 ms after the start. The end of any single breath, articulatory plan has been reached when PLU has just reached 'zero' value. However, the start of each time plan is usually defined by an articulator other than PLU.

The articulatory plan occupies a longer time span than the output acoustic duration, since no acoustic source will be generated unless lung air

![Fig. 3. An articulatory time plan for lung air pressure PLU, expressed in 5 ms time units and with event labels. The final fall in PLU is given a transition of type -M if its timing is defined by an event, here ETMAX, at the end of the transition.](image)
pressure is somewhat above zero. The exact duration acoustically cannot be predicted from the articulatory plan, but will probably be more than 400 ms (80 time units) and will be less than 640 ms (128 time units) in this case.

For a single pitch fall, the laryngeal component of $F_0$ control might be instructed as follows (see Fig. 4):

<table>
<thead>
<tr>
<th>Line</th>
<th>End point</th>
<th>Time allocated (5 ms units)</th>
<th>Transition type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>high</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>high</td>
<td>70</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>low</td>
<td>58</td>
<td>S</td>
</tr>
</tbody>
</table>

$Q$ is to start ‘high’ (say, at 90 Hz), remain ‘high’ until $70 \times 5$ ms from the start, then fall with a slow (S, i.e., 38 time units) transition, then remain ‘low’ (at, say, 50 Hz) until the end of the action plan (128 time units from the start).

The $Q$ fall needs to be correctly timed with respect to the action or actions associated with a stressed vowel onset, so as to achieve the auditory goals. Since acoustic events do not, in general, coincide with one of the articulatory events causing them, trial and error is required. Small modifications are made, including changes to inter-articulator coordination, until, for example, the pitch fall is judged auditorily by British English listeners to be appropriately timed. When this criterion has been satisfied a suitable articulatory plan has been found. Our experience with informal listening shows that auditory pattern features can be assessed in parallel. For example, even if the phonetic quality of the vowel is poor, an apparently reliable judgement can be made about the timing of the pitch fall.

The actual $F_0$ contour generated by the model will usually be more complex than the $Q$ contour as shown in Fig. 4 for example, since an aerodynamic $F_0$-controlling factor is added to this laryngeal muscle control, as follows

$$F_0 = Q + 4(PSG - PC).$$

where $(PSG - PC)$ is the low frequency aerodynamic component of pressure drop across the glottis. This very simple expression nevertheless produces complex $F_0$ contours (Scully and Allwood, 1985b). The value of 4 is based on published data (see, for example, Ladefoged, 1963).

1.6. Instructions for movement paths with events $E$

A rigid time base of the kind used in the examples above for PLU and $Q$ creates very great

Fig. 4. An articulatory time plan for $Q$, the laryngeal component of fundamental frequency $F_0$ control, expressed in 5 ms time units and with event labels. An event for $Q$, $V1.EQ$, is coordinated with a vowel onset event $V1.EON$ for a different articulator, via the time interval duration $V1.D1$. 

Speech Communication
problems during the process of trial and error with auditory monitoring. When small changes of timing need to be made for a single articulatory action the whole time plan has to be recalculated. Often, two or more articulators need to be functionally linked for particular actions. In this case a small change of timing entails a whole set of changes. The same recalculation problems arise when variations are to be played on a successful basic time plan.

These time manipulation difficulties led to a change of focus. Instead of considering time as going forward from \( T = 0 \) to \( T = T_{\text{max}} \) (a Physicist's view!) the temporal organisation is now viewed as being organised around a number of articulatory time points. These are the events \( E \) of the scheme described here. Each event is for a particular articulator and for a particular vowel or consonant allophone; each articulator may have several labelled events, or none. The event name begins with a 'V' or 'C' label, or with a '#' label for speech-initial and speech-final units, or with a specific allophone label. For example, Fig. 6 in Section 1.9. Thus the instructions for PLU and Q given above might be expressed as (see Figs. 3 and 4):

\[
\begin{array}{ccc}
\text{PLU} & \text{Q} \\
0 \text{ zero} & 0 \text{ high} \\
1 \text{ normal} & 24 & 1 \text{ high} \text{ V1.EQ} \\
2 \text{ zero} & \text{ETMAX} & -M & 2 \text{ low} \text{ ETMAX} \\
\end{array}
\]

Lung air pressure PLU is to begin to be lowered M (24) time units before ETMAX, so that its 'zero' value is reached precisely at ETMAX and not before. This is the meaning of the negative sign for that transition. Transitions with positive signs (understood if there is no negative sign) begin at the specified event label. Generally, this final transition for PLU would not have a negative code, since ETMAX is defined by PLU reaching zero, not the other way round as in the example given here. ETMAX and V1.EQ are time points, not the time allocated for a single action. V1.EQ is an event for the Q articulator, located at the start of a Q transition; it is associated with a stressed vowel allophone [V1]. The time allocated for each instruction line can be either a simple duration in 5 ms units, or an event such as V1.EQ, or an arithmetical combination of an event with durations added or subtracted.

1.7. **Coordination expressions: D and E**

Now, as discussed above, it is important that a fall in Q should be correctly aligned relative to an event for the onset of a stressed vowel. A rule is needed to express this dependency of the Q event V1.EQ on the vowel onset V1.EON. A time interval duration \( D \), labelled D1, relates the two events, as follows (see Fig. 4):

\[
\text{V1.EQ} := \text{V1.EON} + \text{V1.D1}
\]

(\( := \) means that the left hand side is to be set equal to the right hand side).

V1.D1 is the duration of a time interval expressing an aspect of inter-articulator coordination which is relevant for the stressed vowel allophone [V1]. The expression is left-right asymmetrical: EQ depends on the timing of EON but not vice versa. From previous modelling, a suitable value may be known for D1; possibly this value will be applicable to all vowels of English in a similar context, or perhaps to a subset: to all 'long' vowels in similar contexts, for example. It is necessary to define a suitable vowel allophone onset event V.EON; this takes different forms, with different articulators referenced, depending on the phonetic context for the vowel, as discussed in Section III.6.

When many of the time points in an articulatory plan are labelled and related to each other in this way, it becomes extremely easy to make modifications to any portion of the plan, often by simply altering a single \( D \) value. All the required events for all the articulators for the whole plan will be recomputed automatically. In Section IV.2 a complete coordination description will be derived for an isolated vowel. Before that, in Chapter II, the ways in which events and durations in the model relate to some effects observed in natural speech production will be made explicit and some evidence about timing and coordination in natural speech will be cited.
I.8. Articulatory transitions: basic types

Transitions have the general form shown in Fig. 5 with, generally, an acceleration phase, maximum velocity midway and a deceleration phase. Besides transition types S (slow), M (medium) and F (fast) with transition durations (TR) of 190, 120 and 40 ms (38, 24 and 8 time units) respectively and slope factors near 100, pairs of values (TR, slope factor) define other types. For example, (16,100) means a transition duration of $16 \times 5 = 80$ ms, with slope factor of 100, like the intermediate transition in Fig. 5. A positive value for TR in the instructions for an articulator means that the transition starts at the event indicated in that line; a negative TR value means that the transition is to be delayed if necessary, so as to end at but not before the given event. Except when the model is being matched to a particular speaker as accurately as possible, a limited set of transition types is used. For an explanation of the terms used to label end point states, see Table 4 in Section III.2 and Table 5 in Section III.7. The transition types generally used are as follows:

For AG: M type, except when going from closed to phonation or closed to pressed, which is F.

For AC: This depends on which articulator is used. For the tongue tip-blade actions needed in alveolar stops and fricatives (16,100) is generally used; for stops a complex release is sometimes used, as shown in Fig. 6 in Section 1.9. Closing up from [s] to [t] the Fast 40 ms transition F is used.

For TB: M if the tongue body is already at its front state and is being moved into or out of a consonant state; also for vowel to vowel transitions if both vowels have front configurations or both are back vowels. S if the tongue body moves between back and front states for vowel to vowel, vowel to consonant or consonant to vowel.

For Q: S for $F_0$ falls in a stressed vowel ["V"]; M for $F_0$ rises preceding this fall.

For AV: M.

For PLU: M.

I.9. Complex articulatory transitions

Two part transitions, for example to give an initially fast release for a plosive, can be constructed, as shown in Fig. 6. This illustrates an additional feature. To simulate the interconnection between tongue tip, blade and dorsum the AF value can be first computed as part of the tongue body linking, probably as a changing value, with the tongue body moving through an appropriate region for an intervocalic consonant. Instructions for AF are then applied, of the following form:

For AF in a sequence [V d V]
(a) instruction number,
(b) end point (cm$^2$) and transition specification,
(c) time point event $E$ or time allocation for the transition (5 ms units),
(d) transition duration (5 ms units),
(e) transition slope factor (arbitrary units).

<table>
<thead>
<tr>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AT 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>TO 0</td>
<td>d.EAFX</td>
<td>-16</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>TO 0</td>
<td>d.EAF</td>
<td>16</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>TO 2</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>BY 0</td>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>AT 0</td>
<td>ETMAX</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The instructions mean that AF remains at its previously defined tongue body value (AT 0) at first, then moves, with a $16 \times 5$ ms transition, to the value 0 cm$^2$, reaching this end point at d.EAFX, not earlier. AF stays at 0 cm$^2$ until d.EAF, then moves with a two part transition back to the previously defined tongue body link value. The first part of the release is linear and occupies 6 time units; the second part has its end point at the previously defined tongue body link value (BY 0) and takes 10 time units with a linear transition. Since 10 time units are allowed for this, the whole transition is just completed 16 time units after the release d.EAF. From then on, AF remains at its previously defined tongue body link value. This is labelled as AT TB value in the examples of articulatory time plans described later. The end point value for the stop closure, given as AF = zero here, can take negative values to simulate a firm contact; the values are set to zero later, for the acoustic block. An advantage of this double calculation for the main articulator of a consonant is that its actions can be executed from a suitable region of the tongue as a whole, without requiring the whole tongue shape to be specified exactly. This approach is consistent with evidence from real speech for some consonants, as discussed in Section II.16.

Chapter II. The articulatory model and the data bases in relation to data from natural speech and other motor skills

Introduction

Our modelling framework is underconstrained so as to permit application of speaker type specific rules, but, at this early stage, many simplifications are made, in the hope that this will enable us to address some of the more obvious questions of articulatory organisation without becoming overwhelmed by the subtle detail of patterning which is clearly present in real speech production. Thus, in considering the articulatory evidence from natural speech, particular attention needs to be paid to those articulators and those portions of their gestures which appear to be most salient for particular output sound patterns. The actions essential for plosives or fricatives appear to impose stronger time plan constraints than those needed for vowels (Scully and Allwood, 1985a). It is not
II.1. 'Icebergs' in natural speech and transitions in the model

X-ray microbeam derived traces show transitional aspects of the tongue body and other articulators. There appear to be 'iceberg' central portions of some transitions for specific articulators which remain stable and invariant across changes in context or style (Fujimura, 1981, 1986). Figure 7 shows how these characteristics observed for real speakers may be approximated within the framework of our model. Portions of the vertical displacements of the lower lip (above) and tongue (below) are shown for the word "memory" said (a) within a sentence and (b) in isolation. Below the X-ray microbeam traces are shown simplified time paths of the kinds used in the modelling which approximate those above. In natural speech the transitions are complex: for example the lower lip movement path has a longer, deeper valley between icebergs when the word is said in isolation. The tongue blade too continues to move slowly upwards after the iceberg transitions (shown bracketed) in both versions; the next tongue blade transition located in the [ri] portion of the utterance begins later when the word is said in isolation. The following representation is proposed: a mid-transition, iceberg region of a transition in natural speech becomes nearly the whole transition of a parameter point in the model. Articulatory movements between the icebergs of natural speech are replaced by steady state portions in the modelling. The model's paths for AF in Fig. 7 show a static segment for the tongue tip-blade when "memory" is spoken more slowly in isolation, and none when it is spoken more rapidly, in a sentence. As a first approximation, large and small distance transitions for the same articulator are traversed in a fixed time. The basis for this assumption has been discussed earlier (Scully, 1975). Recent publications (for example Kuehn and Moll, 1976; Kiritani, Imagawa, Takahashi, Masaki and Shirai, 1982; Ostry and Munhall, 1985) show that this is not unreasonable as a first step, although they show a rather strong positive correlation between peak velocity and distance moved in the transition, rather than constant time.

Since the relatively rigid iceberg or island portions of natural speech gestures seem to be associated with information-bearing transitional aspects of formant structure which are likely to provide stable perceptual cues for consonants (Fujimura, 1986), it seems possible, at least, that by mimicking these portions while relaxing the matching criteria for intermediate portions of movement paths, we may be able to achieve accuracy where it is judged to be most needed, without unnecessary complexity in the less crucial aspects; these appear as sensible tactics for modelling a complex system (Bridle, pers. comm.). Similarly, one approach to fricative articulation takes contact for a small number of points on an
artificial palate to define obstruction onset and offset (Löfqvist and Yoshioka, 1984). This may be used for a reasonable approximation to paths obtained from aerodynamic traces of natural speech by means of the orifice equation \( A = \text{const.} \frac{U}{\Delta P^{\frac{1}{2}}} \). They show constriction cross-section area apparently changing continuously, without a truly static obstruction phase (Scully, 1984a).

A pair of events in the model, labelled .EA and .EA X, define, for an articulatory parameter A, the start and end of a transition, respectively. Invariant iceberg transitions are likely to form links across successive allophone units [U_i] and [U_j] in many cases. Then the end point events in the model are U_i.EA and U_j.EAX. Suppose that invariant transitions themselves come nearest to being the true articulatory 'targets' in speech production: then one or both of these events could be said to represent the crucially important features of speech production. The particular articulatory events selected for use may, of course, be very different from those referenced by a speaker's actual nervous system. The region of maximum velocity half way through a transition, for example, seems just as likely a candidate as the end points, in view of the kinaesthetic information available to a speaker. Feedback paths which provide an already skilful speaker with knowledge about on-going articulatory and aerodynamic conditions are not represented in the present scheme, so that the role of trial and error with auditory monitoring after the event may well play an unduly strong role in our modelling.

II.2. General constraints in speech communication

Speech production, transmission and reception are subject to a variety of constraints. Goal-related language-specific auditory requirements may impose constraints on acoustic features such as relative and absolute durations of vocoid and contoid segments, timing and coordination of acoustic events such as a plosive release and voice onset, the time course of the \( F_0 \) contour, extent and timing of formant transitions, and so on. All of these are subsumed under the general requirement that the executed motor programmes should generate auditory patterns acceptable to native British listeners as examples of the target words, in the prescribed style. These perceptual constraints need not therefore be considered in more detail here. Robustness of the acoustic signal in the presence of noise will not be considered either. Basic psychoacoustic limitations and information provided by context will interact in complex ways (Lindblom, 1983). General properties of communication systems and of discourse in particular do, no doubt, influence a speaker's actions. There is very probably a need to speak as fast as conveniently possible and in a way appropriate to the surroundings and context, so as to avoid boredom on the listener's part. But there are costs to be paid as rate of speech increases: presumably a fast tempo is more tiring and less sustainable than a slow one. Neuromuscular processes and the mechanical properties of the moving structures must surely condition speech production, but in ways that are not yet understood (Scully, 1984b).

II.3. Rate effects

Overall duration seems to be one parameter for motor programmes. Entire sequences as long as 3 or 4 seconds can be run faster or slower while maintaining the relative timings of the component parts (Schmidt, 1982). On the other hand, some evidence from animal locomotion studies suggests that within a cycle of actions, some portions are unaltered by an increase in running speed, while other portions are performed in a shorter time (Tuller, Fitch and Turvey, 1982). Although speech production is not cyclical in this way, possibly some parallels are to be found in speech production at different speech rates. Maintenance of durational properties across a change in speech rate has been suggested for some muscle activities (Tuller, Kelso and Harris, 1981). Cross-speaker differences in jaw transitions found in another study suggest individual options for modifying muscle force programmes to increase speaking rate (Nelson, Perkell and Westbury, 1984).

There appear to be different ways of reducing
transition durations for a faster rate of speech. In one study the distance moved was always reduced, but the distance reduction interacted with peak velocity of the movement. Some speakers kept maximum velocity the same or reduced it; they reduced the distance more than other speakers who increased peak velocity (Kuehn and Moll, 1976). In another study, one speaker actually increased the distance moved by the jaw at a faster speaking rate (Hughes and Abbs, 1976). Other studies have found cases where peak velocity of an articulator is higher at the slower speaking rate, associated with larger distances traversed at the slower rate (Chistovich, Kozhevnikov, Aliakhrinski, Bondarko and Golzin, 1965; Kent and Moll, 1972). Peak velocities of the tongue tip associated with [l] were found to be essentially the same at different speech rates (Giles and Moll, 1975). Speech rate had no significant effect on upper lip protrusion as regards distance or duration: time was saved by reducing the durations of the static portions having maximum lip protrusion (Benguerel and Cowan, 1974).

Clearly, complex reorganisation of gestures is possible, with a choice of options available. Although articulatory transitions do not remain completely invariant, it seems that one mechanism for saving time is to keep the central portion of the transition relatively unaltered, while reducing in duration the lower velocity or nearly static end portions of the transitions or actually static end points. The various reductions are all represented in the modelling by reductions in static end points for invariant transitions.

II.4. Articulators and degrees of freedom; motor skills

The assumption of a small number of quasi-independent articulators is consistent with a general requirement in motor skills to reduce the number of degrees of freedom. This would be of the order of 40 if each muscle used in speech were considered as independently controlled and greater still if motor units were taken as the appropriate level of description. Sets of independent muscles appear to work cooperatively to achieve changes in shape of one region of the vocal tract. An example of this is the motor equivalence found between lower lip and jaw in closing the lips for a plosive (Hughes and Abbs, 1976). In modelling these actions functional synergy is assumed: the lip area–time function can be described by a single transition, even though different amounts of jaw and lip muscle forces may be used on different occasions. Even repetition of a given task may employ different compositions of cooperating units for motor activity (Ludlow, pers. comm., 1985). The emphasis in the modelling is on the shaping of portions of the vocal tract rather than on the solid structures themselves. In contrast to the term functional synergy applicable to the control of one part of the respiratory tract, cooperation between articulators each of which controls more or less distinct portions or properties of the respiratory tract will be referred to as inter-articulator coordination. This usage appears to be compatible with the concept of development of a generalised motor programme for a particular task: a schema formation (Schmidt, 1982). The task in speech production is likely to be specific to a single allophone or to a group of allophones constituting a natural phonetic class, in a specific linguistically-determined context.

Inter-articulator coordination is a manifestation of the previous successful operation of processes which in one analysis are characterised as coordination, control and skill. Coordination has been defined as the discovery of a function relating some independent muscle forces; control is the process of assigning values to the variables; skill acquisition is the process of optimising those values (Kelso, 1982). The number of independent articulators functionally linked for a particular speech action might be small or large. An unskilled performer (perhaps including a young child) might well keep the number of degrees of freedom small so as to simplify the problem. On the other hand, it seems possible that the poorly coordinated use of too many degrees of freedom might be a characteristic of some unskilled speakers. It has been suggested that, as skill increases, additional degrees of freedom are released, so that the actions can be made more efficient and can exploit the available forces better (Turvey, Fitch and Tuller, 1982). The articulatory plans discussed here are intended to refer to normal adult
speech, so it may be assumed that the number of degrees of freedom is rather large. An estimate for this number in speech or other highly skilled motor activity does not seem to be readily available. It is to be hoped that the number of articulators used here, up to about eight, is plausible for skillful, well practised human actions.

II.5. Tongue body constraints

In our model the number of parameter points is greater than the number of articulators, but some of these points are functionally linked to define the tongue body configuration and its transitions. Except where one of them is the main articulatory constriction for a consonant, the parameter points E, P, N and B coincide in the timing, but not in general the extent, of their transitions. This is consistent with X-ray traces (Houde, 1968). For simplicity, lip actions also, other than cases where AL is the main articulator for labial consonants, and actions of the jaw, AJ, in general, are assumed to coincide with those of the tongue body. It is recognised that jaw height and tongue front height are in reality independent variables with different dynamic characteristics and that jaw height variations are associated with suprasegmental effects such as a change in the stress for a vowel, probably independently from tongue body shape (Fujimura, 1981; Macchi, 1985). Although only one consonantal event is allotted to describe the timing of the start of a tongue body, jaw and lip release and transition towards a new vowel or consonant end point, it is possible to give the jaw region a different transition time from the tongue body regions of the vocal tract. It is possible also in the modelling to dissociate tongue body, jaw and lip movement onset events if necessary.

Tongue factor constraints found for vowels in natural speech are applied informally in the model. As the tongue body moves from back to front, corresponding approximately to Kiritani’s factors J1 and T1 (Kiritani, 1977) or to the front raising of Harshman et al. (Harshman, Ladefoged and Goldstein, 1977), pharynx areas AE and AP increase while the palatal region area AB decreases. This produces an effect observed in natural speech, whereby vocal tract area hardly alters in a small region of the soft palate as if it were a pivot point. It is assumed that, within a front posture, AB can be varied to control the constriction size at the hard palate. Similarly, the magnitudes of AP and AE can be controlled within a back posture. Kiritani’s factor T2 is represented by a continuum velar–not velar, which is taken to affect mainly AN. To be consistent with the back raising factor of Harshman et al. (1977), the pharynx tube size is made to increase slightly by increasing AE and AP as AN is reduced for velar states. The slight decrease needed in the volume of the front cavity is modelled by a slight decrease in AJ, with AB left unchanged. Thus the tongue body TB variables are given plausible values compatible with the basic physical constraint of incompressibility of the tongue volume. Appropriate transition times can be used where a change of configuration is needed.

A transition end point description may need to contain negative as well as positive statements. Implicit in IPA (IPA, 1979) descriptions of place of articulation for consonants is the requirement to avoid a narrowing of the vocal tract elsewhere. Modelling makes the effect explicit: unwanted, intrusive sounds and a surprisingly large number of 'extra' perceived phonemes can be generated unless care is taken to avoid making significant constrictions of the vocal tract besides that of the main articulator for the intended consonant. Thus, for bilabial and alveolar consonants, for example, the tongue body is explicitly characterised in the data base as 'not palatalised'.

II.6. Movement path dynamics and kinematics

Ease of articulation is a concept which appears very difficult to quantify, but assessments of the time and maximum force costs have been made by theoretical consideration of the dynamics of unimodal movement paths in general; and for related costs of maximum impulse, energy, acceleration and jerk (rate of change of acceleration dx^2/dt^2, Nelson, 1983). Jaw transitions for speech-like [sa:] sequences at increasing rates have been measured for three American English subjects. For two out of the three subjects, the distance moved
by the jaw more than halved as rate increased from 2 to 6 repetitions per second. Peak velocity decreased slightly while peak acceleration remained nearly constant across rate changes, suggesting that these subjects hardly altered their muscle forces. The third subject increased muscle forces as rate increased. It seems that each speaker was constrained by the peak force per unit mass developed during this repetitive movement. The upper limit may be very different for different speakers. This may constitute a constraint on all speech producing actions, since the jaw is more massive than the other articulators (Nelson, Perkell and Westbury, 1984).

Kinematic studies of articulation show how rather than why actions are performed, but commonly observed features may be taken as a guide to realistic modelling. One of the most powerful constraints built into the model is the time required for an individual articulator to perform a movement path gesture. Most such transitions are expressed in the model as changing values of cross-section area of a portion of the vocal tract. The parameter value–duration graph is given a particular $S$ shape, as shown in Fig. 5 in Section 1.8, based mainly upon cinefluorographic data from real speech. This is an approximation only, partly because distance in the mid-sagittal plane, as seen on X-ray frames, and tube cross-section area, needed in the model, are two different domains. Mapping from one to the other of these is not yet a well-established process.

The available data for real speech are perhaps not sufficient for consideration of inter-speaker differences falling within a normal range as applied to all the articulators, but patterns of velum activity may be mentioned as an example of the complexity, subtlety and individuality found in real speech. An analysis of X-ray microbeam data showed that speakers of American English could differ in the variety of their velum transition types, in the number of end point states and in the influence exerted by contextual features (Vaissière, 1983). Clearly, the transition types specified in our modelling, although based on results for real speakers, mainly cinefluorographic studies of rather limited numbers of speech-like sequences, are over-rigidly defined. A few context effects are taken into account but, in the main, it is assumed that each articulator has a constant transition time and this set of transitions is taken to characterise one simulated speaker, at a medium rate of speech.

II.7. Transition end points and articulatory undershoot

Articulatory transitions are assumed, as discussed above with reference to other motor skills, to be components of sets of prelearned motor schemata, stored in the central nervous system. Schmidt (1982) distinguishes between fast and slow movement paths. Since articulatory transitions are generally completed in 150 ms or less, they are taken to be fast actions which cannot be halted once initiated. On-line correction in response to applied perturbations has been demonstrated for speech (Folkins and Abbs, 1975). It is to be supposed that feedback can operate to correct minor errors of execution during a transition, but not so as to initiate a new motor programme while a current transition is in progress. We assume, therefore, that each movement path is performed essentially perfectly, given feasible limits of accuracy. Within this conceptual framework it does not seem appropriate to invoke a concept of articulatory undershoot, in the sense of a transition end point being aimed at but not reached because of the intervention of contradictory ‘commands’ or for any other reason. It seems perfectly reasonable to suppose, however, that a speaker may modify the movement path parameter values in respect of muscle forces and their relative timings so as to reduce in distance or in duration a particular transition at a faster speech rate, for example. But this is not considered to be undershoot, in the sense of failure to reach an intended end point ‘target’. Indeed, the notion of a single fixed articulatory target is rejected here, in favour of consideration of tightness of constraints on transition end points and on transition mid portions, for particular articulators and in the production of particular allophones. Thus the view of an end point as a target is weakened to that of an end point as a region.

The constraints on end points may be relatively tight, as in the case of the main articulator for a
fricative consonant, or looser, for an articulator whose position is neither critical for the sounds to be generated nor closely determined by anatomical and physiological constraints. The tongue may be relatively free during a bilabial consonant, it may move within a small region for consonants which employ the tongue tip or tip-blade as the main articulator, and it is likely to be rather precisely positioned for velar consonants where the tongue body itself is the main articulator. These kinds of constraints have been included in other articulatory models (Mermelstein, 1973).

II.8. Variability in speech and other motor skills

All quantitative studies of speech production and speech acoustics show variations across tokens, even when an attempt is made to say the same thing in the same way during one recording session. Relative precision in individual transition paths and in inter-articulator coordination seems likely to be a feature of skilled performance. This view is supported by a comparison of jaw transitions in speech-like [sa: sa: sa:...] actions and non-speech, jaw-wagging actions by the same subjects. Variability for each of the jaw movement parameters – distance, peak velocity and peak acceleration – was much lower in the speech than in the other, unfamiliar, task, for all three subjects (Nelson, Perkell and Westbury, 1984).

Evidence from other motor skills suggests that variability in transition end point position should be proportional to the distance moved and inversely proportional to the time taken (Fitts, 1954; Schmidt, 1982). These relationships can be derived from consideration of the muscle force-time programmes used. For articulatory transitions having one acceleration phase followed by one deceleration phase, increased distance can be achieved while keeping transition time constant by raising the level of muscle forces but maintaining constant the relative timings of accelerating and decelerating forces. These tactics would increase the end point variability in proportion to the increased variability in the forces, that is, in proportion to the force increase itself. This might be an appropriate control method at a fixed rate of speech, as represented in our model, with its generally constant transition durations. To traverse the same transition distance in a reduced time, two changes need to be made to the execution of the motor programme: the muscle forces need to be increased and the time interval between agonist (accelerating) forces and antagonist (decelerating) forces needs to be decreased. If a constant accuracy of end point is required at a faster rate of speech, then it seems that movement distance and movement time would need to be reduced in proportion.

II.9. Variability in speech: end point states and mid transition portions

When a hand-held stylus was moved to a dot drawn on paper, target width was proportional to distance $d$ divided by duration $t$ for the action. Extrapolating from these results to lower values of $d/t$ suggests that minimum variability would be 2 mm standard deviation in the direction of the movement and 1–1.5 mm in the perpendicular direction (Schmidt, 1982). Perhaps this order of accuracy may be expected in speech production, where $d/t$ values are very small compared with these stylus actions.

For tongue blade transitions in [VC] and [CV] sequences of American English, variability of tongue height was found to be smaller in the iceberg middle portion, approximately one third of the total transition, than at the lower end point, by a factor of about 1 to 8. Results were closely similar for upwards and downwards movements and appeared to be independent of associated emphasis. In this study, [se] in “seven” had lower tongue pellet height variability at the high starting point for the fricative (range less than 1.5 mm) than at the end point for the vowel (range about 8 mm). Greater accuracy of tongue blade height seems to be needed here for [s] than for [e] (Fujimura and Spencer, 1983).

In another study of tongue blade movement paths produced by an American English speaker, for pairs of words having minimal stress location contrasts such as the [ai] transitions for “insight” and for “incite”, central portions of the transitions could be matched regardless of the stress context. Positional variances were up to about 6 times...
greater at the lower tongue blade end point than at its higher position at the other end of the transition. Both these end points varied much more than the matched central portions of the transitions (Beckman, Fujimura and Spencer, 1984).

Similarly, for repeated [s:克莱] sequences produced by three American English speakers at increasing repetition rate up to about 5.5 Hz, the variability in the height of the jaw shown for one subject was less at its highest position, for [s], than at its lower end, for [a:]. Height ranges were about 1 mm and 6 mm respectively. At repetition rates below 4 Hz greater end point precision was associated with transitions of longer duration for this subject; however, the relative accuracies were maintained at slightly higher repetition rates. The other two subjects maintained their peak jaw velocities approximately constant as repetition rate increased, suggesting that central portions of their jaw transitions may have remained invariant as icebergs (Nelson, Perkell and Westbury, 1984).

Variability in the end point for a vocal fold abduction gesture in speech can be estimated from multiple tokens of glottal width for different Hindi plosives (Kagaya and Hirose, 1975). The units are arbitrary, but fractional variation in width can be estimated as being from 0.15 to 0.42 across 6 different plosives. Minimum cross-section area for [s] and [z] in multiple tokens of 6 English words in carrier sentences have been estimated by us. from aerodynamic traces (unpublished data). These transition end points have somewhat different ranges for the 4 speakers, but, in general, the inferred minimum area can apparently vary by a factor of two or more. However, variability for this single parameter probably compounds true minimum cross-section area variability and constriction length variability.

II.10. Variability in speech: end point states for different articulators

Whether articulators are of primary or secondary importance for a particular allophone has been assessed through measurement of their end point variabilities (Bothorel, 1983). If successful production of a given sound pattern is sensitive to only one or two articulatory actions, then some of the problems of the degrees of freedom are simplified: acoustic stability with respect to articulatory perturbations becomes less multidimensional and the achievement of stable configurations (Stevens, 1972b) seems likely to be aided. Stability with respect to most articulatory controlling variables but sensitivity to one or two of them seems to be a plausible means of increasing control reliability. Inference of variability in vocal tract area function from articulator dispersion is complicated by synergisms such as the cooperation between lower lip and jaw for achieving a vertical lip opening which is much less variable than that of the lower lip and jaw positions taken separately (Hughes and Abbs, 1976).

II.11. Variability in speech: end point states for different regions of the vocal tract in vowels

A speaker seems to know which portions of the vocal tract are crucial for a particular vowel. For example, it seems likely that the palatal portion of the vocal tract area function is more crucial than other regions for [i:]. Under bite-block conditions, this part of the vocal tract remained very close to normal, while other portions varied (Gay, Lindblom and Lubker, 1981). Position variability ellipses for tongue pellets were obtained for two subjects who produced many tokens of [i:], [a] and [a:]]. In the constriction region of the vocal tract for [i:], minor and major axis lengths were about 3–4 mm and 6–8 mm respectively for one speaker; and about 5 mm and 9–10 mm for the other speaker. It was suggested that the speakers were achieving greatest precision of tongue position in the acoustically critical direction here, aided by saturation effects. Similarly shaped but differently oriented dispersion ellipses were shown for the rear portion of the tongue body in [i:]. The low variability along the vocal tract here seems unlikely to be related to precise control of, for example, back cavity volume for [i:]; it may perhaps indicate instead the direction of the muscle forces acting on a portion of the tongue where saturation effects do not operate (Perkell and Nelson, 1982).
II.12. Variability in speech: inter-articulator coordination

Multiple tokens for inter-articulator coordination in natural speech give indications of the appropriate values for perturbations of the D values in our modelling. For example, the coordination between the event of the main articulator release and the event of maximum glottal width showed total ranges of 20–60 ms for one speaker's Hindi stops (Kagaya and Hirose, 1975); 20–55 ms for two speakers of Mandarin (Iwata and Hirose, 1976); and 25–45 ms for [t] produced by two American English speakers (Lofqvist and Yoshioka, 1984), suggesting that D2 for stops in the model (see Fig. 9 in Section III.9: coordination between AC and AG) should be perturbed by up to about +/- 30 ms. Table 1, which gives some D values for natural speech, includes ranges of values where these are available.

II.13. Interpretation of natural speech data on inter-articulator coordination

In the modelling, an event E is at a precisely defined time point, whereas a duration D is subject to the variability discussed above and also to context-dependent variation. The next sections consider some of the basis in real speech for choices of particular D values. Complex results have been summarised and simplified; values are quoted to the nearest 5 ms in Table 1. Where not otherwise indicated the speech rate is assumed to be normal rather than fast. Reference will be made to the acoustic pattern feature of voice onset time (VOT) (the time, interval between a plosive burst and the onset of voice), for categorising stops as aspirated or unaspirated during the release phase. Systems of phonological opposition need to be considered also, when assessing the relevance of analyses of other languages or accents for the modelling of an R.P. accent of British English. These data can provide only a guide to reasonable duration values: apart from differences due to language and accent, different speakers have been recorded and measurement methods vary. Definitions of rate may differ also.

II.14. Timing and coordination for stops and fricatives: the main articulator of the vocal tract and the vocal folds

In Fig. 9 and Table 1, D4 is the closure or obstruction duration for the main vocal tract articulator, the time interval from EACX to EAC. D4 will be considered in conjunction with D1, the coordination between the release of the main articulator at EAC and the start of vocal fold adduction for a following vowel, EAG.

Investigators have generally found that there is no static phase of maintained open glottis; thus EAG and EAGX generally coincide in natural speech and D5 = D1 = D4 even when consonant clusters are included (Löfqvist, 1980; Löfqvist and Yoshioka, 1980, 1981, 1984; Pétursson, 1976). D1 is zero if the event EAG coincides with the release of the main articulator EAC; it has a negative value if EAG falls within the closure or obstruction phase.

Where different places of articulation have been included for the consonants studied, the results suggest that vocal fold actions and vocal tract articulator-vocal fold (AC-AG) coordination does not change with place of articulation (Kagaya and Hirose, 1975; Löfqvist, 1980). It seems well motivated therefore to assume for the present that D1 values cited here for one place of articulation are likely to apply to others also.

As far as D1, for AC-AG coordination is concerned, a rather clear and consistent picture emerges for aspirated and unaspirated voiceless plosives. D4, the closure duration, appears to vary more, but some trends are apparent in the published data.

Closure duration D4 is often longer for a voiceless unaspirated than for the corresponding aspirated plosive, as indicated in Table 1. This is consistent with the suggestion of Löfqvist (Löfqvist, 1980), also supported by our modelling and analysis of some natural speech (Scully, 1975, 1976), that for an unaspirated voiceless plosive a double action – abduction and adduction of the vocal folds – must be contained almost completely within the stop closure, so as to avoid preaspiration and release phase aspiration. The vocal fold actions constitute the DDF (duration-determining factor) in that case. However, the action may not
Table 1
A summary of some articulatory coordination duration values in natural speech. D1, D2 and D4 refer to time interval durations in Fig. 9 in Section III.9. D4 is the closure or obstruction duration of the consonant allophone; D1 is the time interval from the consonant release event to the onset of adduction of the vocal folds for a following vowel; D2 describes, similarly, main articulator-tongue body coordination near the release for the consonant. S1, S2 and S3 refer to different subjects in the study cited. Durations are given as ranges, or as mean values followed by ranges in brackets; values are quoted to the nearest 5 ms. VOT is voice onset time, the interval between a plosive burst and the onset of voice. It has a positive value if voice begins after the burst.

<table>
<thead>
<tr>
<th>Language and reference</th>
<th>Consonant, Rate, Context</th>
<th>VOT (ms)</th>
<th>D1 (ms)</th>
<th>D2 (ms)</th>
<th>Consonant, Rate, Context</th>
<th>VOT (ms)</th>
<th>D1 (ms)</th>
<th>D1/D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>American English</td>
<td>[i1], normal, S1</td>
<td>25 to 70</td>
<td>30 to 5</td>
<td></td>
<td>[i1], normal, S1</td>
<td>90 to 155</td>
<td></td>
<td>-0.6 to -0.5</td>
</tr>
<tr>
<td>Lofqvist and Yoshioka, 1984</td>
<td>[i1], fast, S1</td>
<td>20 to 45</td>
<td>40 to 5</td>
<td></td>
<td>[i1], fast, S1</td>
<td>65 to 110</td>
<td></td>
<td>-0.6 to -0.5</td>
</tr>
<tr>
<td></td>
<td>[i1], normal, S2</td>
<td>60 to 95</td>
<td>15 to 25</td>
<td></td>
<td>[i1], normal, S2</td>
<td>90 to 190</td>
<td></td>
<td>-0.6 to -0.5</td>
</tr>
<tr>
<td></td>
<td>[i1], fast, S2</td>
<td>55 to 70</td>
<td>10 to 15</td>
<td></td>
<td>[i1], fast, S2</td>
<td>75 to 140</td>
<td></td>
<td>-0.6 to -0.5</td>
</tr>
<tr>
<td>American English Gay, 1977</td>
<td>[vpv], [i1v], S1</td>
<td>50, 70</td>
<td>-35 to -5</td>
<td></td>
<td>[vpv], [i1v], S1</td>
<td>70, 70</td>
<td>-50 to 20</td>
<td></td>
</tr>
<tr>
<td>American English Gay et al., 1974</td>
<td>[vpv], S1 and S2</td>
<td>90 (70 to 120)</td>
<td>50 to 80</td>
<td></td>
<td>[vpv], fast, S1 and S2</td>
<td>70 (60 to 100)</td>
<td>45 to 60</td>
<td></td>
</tr>
<tr>
<td>Hindi Kagaya and Hirose, 1975</td>
<td>[p, h v]</td>
<td>5 (0 to 20)</td>
<td></td>
<td></td>
<td>[p, o v]</td>
<td>35 (10 to 65)</td>
<td></td>
<td>-60 (70 to 50)</td>
</tr>
<tr>
<td>Mandarin Iwata and Hirose, 1976</td>
<td>[t, h v]</td>
<td>0 (0 to 20)</td>
<td></td>
<td></td>
<td>[t, o v]</td>
<td>35 (10 to 65)</td>
<td></td>
<td>-60 (70 to 50)</td>
</tr>
<tr>
<td>Danish Frekjaer-Jensen et al., 1971</td>
<td>[p, h v] normal rate, S1</td>
<td>110 (100 to 120)</td>
<td>10 (0 to 20)</td>
<td></td>
<td>[p, o v] normal rate, S1</td>
<td>130 (110 to 150)</td>
<td>-80 (95 to 65)</td>
<td>-0.6</td>
</tr>
<tr>
<td></td>
<td>[p, h v] normal rate, S2</td>
<td>95 (85 to 105)</td>
<td>5 (0 to 10)</td>
<td></td>
<td>[p, o v] normal rate, S2</td>
<td>115 (100 to 130)</td>
<td>-65 (75 to 65)</td>
<td>-0.6</td>
</tr>
<tr>
<td></td>
<td>[p, h v] normal rate, S3</td>
<td>75 (70 to 80)</td>
<td>5 (0 to 10)</td>
<td></td>
<td>[p, o v] normal rate, S3</td>
<td>85 (75 to 95)</td>
<td>-50 (80 to 40)</td>
<td>-0.6</td>
</tr>
<tr>
<td>Swedish Lofqvist, 1980</td>
<td>[t, h v] normal rate, S1</td>
<td>20 to 50</td>
<td>75 to 100</td>
<td>-15</td>
<td>[t, o v] normal rate, S1</td>
<td>10</td>
<td>105 to 160</td>
<td>-75 to 40</td>
</tr>
<tr>
<td></td>
<td>[t, h v] normal rate, S2</td>
<td>30 to 60</td>
<td>90 to 120</td>
<td>-25</td>
<td>[t, o v] normal rate, S2</td>
<td>20 to 25</td>
<td>135 to 160</td>
<td>-90 to 70</td>
</tr>
<tr>
<td>Icelandic Petursson, 1976</td>
<td>Aspirated stops, word-initial, normal rate</td>
<td>35 to 50</td>
<td>105 to 130</td>
<td>50 to -25</td>
<td>Unaspirated stops, word-initial, normal rate</td>
<td>10 to 15</td>
<td>105 to 160</td>
<td>-0.7 to -0.6</td>
</tr>
</tbody>
</table>
be needed at all in English, since it has no phonological opposition between a voiceless unaspirated plosive, with voicing almost completely absent in the acoustic closure contrast and a voiced unaspirated plosive, in which voicing continues throughout nearly all the acoustic closure. For aspirated voiceless stops, only a single abductive action need be accommodated within the closure. Among the languages referred to here, only Icelandic appears to control D4 for phonological reasons of consonant length (Pétursson, 1976).

Several factors will contribute to the acoustic pattern feature of VOT. Maximum glottal area and the vocal fold–vocal tract coordination defined by D1 are important ones. Maximum glottal area is more difficult to measure than coordination. In Icelandic, vocal fold abduction was probably slightly greater for aspirated than for unaspirated stops in word-initial position; and noticeably greater for preaspirated than for unaspirated stops in word-medial position: in the ratio 2.3–1.4 on a relative scale (Pétursson, 1976). For strongly aspirated voiceless stops, with a very long VOT, D1 is close to zero or even positive (as shown at the top of Fig. 9), for example as in the Danish data cited in Table 1 (Frøkjaer-Jensen, Ludvigsen and Rischel, 1971; Löfqvist, 1980; Pétursson, 1976).

This kind of coordination, with D1/D4 near −0.6, applies to voiceless fricatives also; and where speech rate has been altered this ratio seems to remain approximately invariant for fricatives. Obstruction duration D4 for a voiceless fricative [s] in American English varied with stress context and speech rate in a similar fashion to stop closures for [t] by the same speakers, but corresponding D4 values were higher for [s] than for [t]. Obstruction onset and offset for [s] were defined here by six-point contact between the tongue and the alveolar ridge region in dynamic palatography (Löfqvist and Yoshioka, 1984).

Aspiration is perhaps not closely controlled for voiceless fricatives in English and many other languages; Korean seems to be an exception (Kagaya, 1974). Preaspiration is probably tolerated to some extent in English stops: our modeling predicts transient and turbulence noise excitations near stop closure as well as at stop release and indications of this may be seen on many spectrograms of natural speech. In Icelandic preaspiration must be controlled. In word-medial position three types of voiceless stops with unaspirated releases are contrasted: short or long unaspirated, and preaspirated. For a speaker of Icelandic vocal fold abduction came earlier, with D1 15–45 ms more negative, and glottal area reached a noticeably larger value, for preaspirated stops than for unaspirated ones (Pétursson, 1976). This would be inappropriate timing for English: by shifting earlier an increase in glottal area, needed for [it] in “plight”, an intrusive palatal fricative [s] was added, due to the effect of high airflow through and [i]-shaped vocal tract, as shown in Fig. 17(d) in Section IV.7.
II.15. Vocal tract coordination for some consonants of English: the main articulator and the tongue body

Postural and movement control for the tongue body and the tongue tip or tip-blade regions must interact because of their anatomical and physiological links. The general name AC is given to the cross-section area for the region of the vocal tract forming the significant constriction for a consonant. In the model, AC is the lip parameter AL for bilabials and labiodentals, AT or AF is used for dentals, AF for alveolars. AB or AN for palatals and velars.

The release of a plosive or affricate, or the obstruction release for a fricative or approximant, does not appear to be a single event, synchronised across the main articulator AC, the lip region AL, the jaw region AJ and the tongue body region TB. The simplifying assumption is made that there are only two release events: one for the main articulator EAC and another for the rest of the vocal tract, called ETB. Similarly, there are taken to be only two events EACX and ETBX associated with achieving the vocal tract closure for plosives, nasals and affricates or the vocal tract obstruction for fricatives and approximants. These are shown in Fig. 9, Section III.9, and in Fig. 10, Section III.14. The consonantal end point states for TB, AJ and AL, labelled [c], are not intended to represent significant obstructions of the vocal tract; indeed, secondary obstructions besides AC need to be avoided as discussed in Section II.5. Here we consider the choice of reasonable values for D2 and D6 to express the coordination between the main articulator AC and the rest of the vocal tract TB, near the offset and onset for a plosive respectively, based on evidence from natural speech.

As seen in Figs. 9 and 10, D2 is negative if ETB precedes EAC; D6 is positive if ETBX comes after EACX. Where different timings appear to be necessary for the jaw region and other portions of the vocal tract jaw actions can be tied to AC instead of to TB.

Two studies on American English provide information on coordination between different portions of the tongue in plosives and on the amount of constancy found in their transitions and end points. In the first study, two speakers produced [p], [t] or [k] in [V1-V2] context, with equal stress for both syllables and at a comfortable speech rate. Tongue body and jaw movements towards their [V2] states began between the stop closure EACX and its release EAC, so that D2 was always negative and D6 always positive, for all three places of articulation (Gay, 1977). The coordination may well have been chosen by the speakers for auditory purposes, perhaps specifically in this equal stress context where the quality of the first vowel had to be preserved. The second study examined [p], [t] or [k] in a different context: preceded by [s] and followed by [a:]. Movement of the tongue dorsum and jaw towards the following [a:] began at or even before the lip closure for [p]. Here –D2 is D4 or even larger; and, with a vocal tract obstruction preceding in this case, again an auditory rather than a physiological constraint on the timing of ETB seems plausible. Movement onsets for the plosive release by tip, blade and dorsum appear to approximately coincide for [p] and [t]: the main articulator, the dorsum, perhaps lags a little in the [k] release. End points of transitions towards [a:] are reached approximately together by tip, blade and dorsum (Borden and Gay, 1979). A synchronised tongue body action, but with ETB slightly preceding EAC for [k], seems to be justified in the modelling.

II.16. End points states for different regions of the vocal tract in consonants

There have been several X-ray studies of the movements and positions of the front part of the vocal tract for plosives and fricatives in different phonetic contexts, produced by speakers of American and British English.

The first American English study showed that, when [k] was followed by different vowels [i:], [a:] or [u:] and preceded by [a:], the tongue dorsum moved through closely similar paths for the closing actions. When [a:] followed the consonant and different vowels preceded it, the tongue dorsum closed up from different starting positions, but all the release phases coincided. When [t] was followed by [a:] and preceded by [i:], [a:] or [u:],
the tongue tip showed almost identical paths for both the closing and the releasing actions. These data suggest fixed paths for the main articulator AC near EACX and EAC. The jaw and the tongue tip moved to fixed positions, with the jaw high. for [t]; the dorsum was lower for [a:–:] than for [i:–:], suggesting a region of the tongue body for [t], rather than an invariant posture. During [p], with fixed lip position and a rather precise jaw position, lower than for [t], the dorsum position varied more than for [t], but did not appear to be executing an unconstrained [V1] to [V2] transition through the intervening [p] (Gay, 1977). These findings seem to be consistent with the vowel-dependent effects on formant transitions in [vowel-plosive-vowel] sequences for Swedish and American English speakers analysed by Öhman (1966). Whether the tongue body is best represented as moving in a diphthong-like vowel-to-vowel transition, or as moving through a specific but context-dependent intermediate region for the plosive, the complex transitions available in our modelling, described in Section 1.9, permit tongue tip-blade or lip actions for an intervening plosive to be superimposed upon appropriate tongue body transitions for the whole [VCV] sequence.

In the second study, with an [a:] vowel following a consonant cluster [sp], [st] or [sk], tip, blade and jaw appeared to move to specific positions for [t] and [k] following [s]: for the [s] itself their positions were rather constant too (Borden and Gay, 1979). Jaw and lip heights for two speakers of American English have been analysed. For one speaker jaw height was found to be higher for [p] than for [k]; for the other speaker [p] and [t] had similar jaw heights. Total lip height and the lip proper contribution was greater for [p] than for [k] or [t] (Macchi, 1985).

Precise configuration of the tip-blade region of the tongue for [s] and [z], regardless of phonetic context, has been demonstrated for British English. Most of the speakers used the blade as the main articulator for [s] and [z] and the tip for [n] and [l], while [t] and [d] varied between tip and blade across speakers. In consonant clusters, or even with a vowel intervening between a plosive and a fricative, [t] and [d] shifted towards blade articulation, while [s] and [z] maintained their blade posture almost invariant (Bladon and Nolan, 1977).

In a study of American English labiodental and alveolar fricatives also, precise positioning of the tip and blade was found for [s] and [z]. The jaw was high and the tongue root advanced in all three vowel contexts studied. There were slight differences of tongue body position during the fricatives, related to the vowel context: the dorsum a little higher for [u:–u:] and the tongue root less advanced for [a:–a:] than for [i:–i:]. (Carney and Moll, 1971). These are complications which it is probably reasonable to omit in our modelling of fricatives, as long as the tongue body is made to move to a front region for these consonants.

According to results for the same speaker in the study just discussed, the tongue body should not be made to execute an underlying vowel-to-vowel transition during an intervening [s] or [z], but this may perhaps be a suitable action when [f] or [v] intervenes. In this case the lip and jaw regions of the vocal tract differed from the tongue body in that they had fixed postures for [f] and [v], not intermediate between surrounding vowel states (Carney and Moll, 1971). This tongue body freedom for labio-dental fricatives seems at variance with the lack of it during bilabial stops in a different study already discussed (Gay, 1977).

Clear and dark [l] for American English differ in their tip-blade and tongue root regions. Lip shape and velum position appeared to be unspecified for [l] and the jaw height seemed to vary a little depending on context, being lower if near a labial or velar consonant and higher if near a dental or alveolar one (Giles and Moll, 1975; Fujimura, Miller and Escolar, 1977).

From the studies cited here, it seems that the end point state for the tongue tip-blade region of the vocal tract must be specified rather precisely for alveolar plosives and perhaps even more precisely for alveolar fricatives in English. The jaw seems to be less strongly constrained for [l] than for fricatives and plosives.

Some constraints on the lip outlet region of the vocal tract for consonants are apparent also. Analyses for some French speakers suggest that the changes in lip outlet shape between a close front vowel [i] or [e] and an adjacent fricative proceeds in a different direction for [s] and [z]
from that for [i] and [u]. These fricatives seem to require specific lip configurations or perhaps gestures (Abry and Boë, 1986).

In a study of lip protrusion under bite block conditions, the reduction in lip protrusion associated with [s] between two [u] vowels for two French speakers was found to be greater than for some other dental or alveolar consonants. For an American English speaker, [s] was not thus contrasted with other consonants, but they all seemed to require a less protruded lip configuration than that for the surrounding [u:] vowels (Perkell, 1986).

II.17. Jaw transitions and their coordination with the main articulator for consonants

Lip separation movement paths for [p] in English are complex in form (Fujimura, 1961). The individual contributions of lip and jaw muscle forces need not concern us as long as we approximate to the resultant path in the model. Nevertheless, if jaw raising and lowering appear to be necessary actions for particular sound pattern sequences, then jaw transition time ought to be incorporated as a constraint on timing plans, even for labial consonants where the tongue body may be in other respects relatively free to execute something close to a [V-V] gesture across the consonant. Jaw raising to a specific height for each of the plosives [p], [t] and [k] is suggested by the evidence cited in Section II.16, even though the ordering of jaw height across different plosives seems to vary between speakers. Therefore, even though the tongue body may be already in its appropriate region for a plosive, a jaw transition seems likely to form part of the actions needed for the consonant.

The main articulator and the jaw do not seem to coincide in their movements into and out of consonant states. In a study of [VbV] sequences, the jaw reached its maximum height after lip closure for the plosive and was in advance of the lip opening in its downward movement for the plosive release (Lindblom, 1967). Similarly, for [p], [t], [k], [s], [f] and [l] preceded and followed by vowels [i:], [a:] and [u:] at a slow speaking rate in American English, closure of the main articulator for the consonant often came before the jaw reached its maximum height (Kuehn and Moll, 1976).

Two studies of American English speakers suggest that there seems to be no really static raised jaw posture for [p] in [V-V] context, at either a slower or a faster rate of speech. There may be a static portion in which the jaw stays lowered for a stressed vowel, seen at the slower speech rate but not at the faster one (Gay, Ushijama, Hirose and Cooper, 1974; Tuller and Kelso, 1984). In one case maximum jaw height was reached at lip closure for both speech rates. For the [p] release the jaw began to move down about 60 ms (80-50 ms) before the lip release for both speakers at both rates (Gay, Ushijama, Hirose and Cooper, 1974). In the other study, the jaw action was more centrally placed relative to the closure for [p], with a closure duration of about 80 ms (Tuller and Kelso, 1984). Different stress contexts, with the vowel preceding the [p] unstressed in the first study cited but stressed in the second, may perhaps be one reason for the different lip-jaw coordinations by the two groups of speakers.

It is possible that the coordination of lip and jaw actions may remain invariant when rate of speaking is increased. One study showed closely similar timing for two speakers and at two rates, as shown by D1 values in Table 1 (Gay, Ushijama, Hirose and Cooper, 1974).

II.18. Coordination across successive allophone units

Some evidence from natural speech suggests that inter-articulator coordination may be very tightly controlled indeed when successive allophone units are linked. For three speakers of American English, lip closure for [p] and the start of the opening action for [s] in consonant clusters containing [sp] were observed by cinelurography. These two events were generally separated by as little as 10 ms but their order was never reversed: the [s] release always came at or after the [p] closure (Kent and Moll, 1975). In contrast to this precision, our own cross-section area traces for tongue blade and lips combined, inferred from
aerodynamic variables, indicate, for multiple tokens of [spaː], a separation between minimum area for [s] and closure onset for [p] ranging between 35 and 90 ms (unpublished data). The discrepancy may be related to the general problem of what constitutes a pair of articulatory events.

In another study of adjacent consonants, separation of the main articulator’s release event for a [k] and the event of closure of the main articulator for [l] following it were found to vary across four English speakers.Mean values for this linking duration. C1C2.D in our framework, took positive values up to 25 ms for one speaker and negative values, implying overlap of the two closure phases, for the other three speakers, of as much as —25 ms for one of them. The separation gap was reduced or the overlap increased at a faster speaking rate (Hardcastle, 1975). This is one of the mechanisms explored in our modelling for a change from formal to informal speech style.

The question of whether particular inter-allophone linking durations are fixed (time-locked) for a given speaker has been posed, for example in the case of lip rounding and protrusion needed for a vowel following a variety of other allophone units. (see, for example. Daniloff and Moll, 1968; Benguerel and Cowan, 1974; Bell-Berti and Harris, 1982; Perkell, 1986). It is expected that the auditory effects of varying some linking D values in the modelling, where the links are non-homorganic and so are articulatorily unconstrained, will point to reasons why particular D values are likely to occur in natural speech, and will suggest the freedom that speakers have to choose their own linking durations.

II.19. Change and invariance for different speech styles

It seems premature to make general statements about what a speaker may alter and what remains invariant as speech style shifts from formal to informal or from slower to faster rate; a great deal more information on natural speech is needed. But the following assumptions, based on hints about real speech, seem to be worth trying out in the modelling, to see whether they will produce appropriate acoustic pattern changes.

Some actions, at least, must be shortened for an overall reduction in duration. The evidence for reduced transition times but for invariance of ‘iceberg’ portions of some transitions at faster rate will be reconciled, combined and represented in the modelling by invariant transitions but with static end point approximations to the portions between icebergs. The durations of these end point states in the model can be reduced at a faster rate. Included in this process is the reduction of static configurations for vowels at a faster rate.

Even more reduction of acoustic durations can be achieved by a closer meshing together of allophones for which the links are non-homorganic.

In contrast to these apparently alterable linking D values, some of the evidence from natural speech suggests that coordination internal to consonant allophones, although dependent upon linguistic factors, may remain invariant as rate increases. This might possibly include closure or obstruction duration D4 for some speakers; more generally perhaps, main articulator–vocal fold coordination D1 for voiceless aspirated stops and D1/D4 for fricatives; perhaps also lip–jaw coordination D2 for bilabial plosive releases and main articulator–tongue body coordination D2 for consonants in general.

Our modifications to articulatory schemes for a change from formal to informal style set aside consideration of reductions in the durations of the central portions of transitions, although they could easily be simulated in the model. We concentrate instead on other possible reorganisation mechanisms, including reduction or maintenance of D values as discussed above and the abolition of optional actions.
Chapter III. The phonetic data bases

Introduction

The function of the data bases is to provide necessary, sufficient and appropriate information for production of some of the common allophones of the phonemes of English. The notation used for the phonetic categorisation of allophones of English phonemes is given first. Since these data bases and the linking procedures concern phonetic prescriptions for simulated speech production, square brackets are used throughout. The terms vowel and consonant refer to one or more allophones of the phoneme concerned. The values given are tentative and even speculative in many cases, in the absence of measured values from natural speech. The contexts specified here are selections from the much larger set that would be needed in prescriptions for the production of polysyllabic words in isolation or of a sequence of connected words.

III.1. Notation

Allophones are categorised as \([V]\) vowel, \([C]\) consonant, or \([#]\) speech-initial and speech-final. Word boundaries are shown in an allophonic string by a space or \([\n\n]\). The terms vocoid and contoid are reserved for descriptions of acoustic segment types. Each allophone or group of allophones is further characterised by the phonetic labels given in Tables 2 and 3.

These labels are a convenient way of referring to the main articulatory contrasts required. They are not intended as acoustic features. For example, it is not to be supposed that a fricative consonant always has frication noise in the output signal, or that an approximant has no frication noise, or that a voiced consonant necessarily has voicing continuing throughout, or that a voiceless consonant cannot have voicing throughout much or even all of the acoustic contoid segment associated with it.

\([?]\) is included, to allow for the phonetic, but not phonological, opposition 'hard' versus 'soft' attack for vowels. A glottal stop is needed for some allophones of English plosives also.

Phonetic classes for \([#]\) are \([#i]\) speech-initial and \([#f]\) speech-final. These units are considered to be nasal \([+N]\) because the velum is assumed to be down at the start and end of an expiratory breath group. They are considered to be voiceless \([-V]\) because the vocal folds are abducted for the breathing state surrounding the speech.

Stressed vowels \([\n\n]\) are considered to be non-nasal \([-N]\); unstressed vowels \([\n\n]\) are considered to be \([ON]\) for informal style at least.

A non-nasal stop \([C, -N, STOP]\) is usually described as a plosive.

III.2. Transition end point labels and values

For each articulator, contrasting points along a continuum are given labels. Actual values are selected from the available range for quantitative prescriptions of articulation. Within one utterance the rank order is as shown in Table 4, or the values may be equalised, in which case an articulatory contrast is lost. Thus, for example, within one simulation:

\[
AG \leq AG \leq AG \leq AG \leq AG
\]

closed pressed phonation breathy apart

An articulator's states need not be – and probably should not be – limited to the rather small

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Some phonologically relevant phonetic classes for vowels of English</th>
</tr>
</thead>
<tbody>
<tr>
<td>([V])</td>
<td>vowel</td>
</tr>
<tr>
<td>([V'])</td>
<td>unstressed vowel</td>
</tr>
<tr>
<td>([V'')]</td>
<td>unstressed and reduced schwa-like vowel</td>
</tr>
<tr>
<td>([V])</td>
<td>stressed vowel</td>
</tr>
<tr>
<td>([V'])</td>
<td>stressed vowel which carries a pitch change in a nuclear syllable</td>
</tr>
<tr>
<td>([V, L])</td>
<td>long vowel</td>
</tr>
<tr>
<td>([V, S])</td>
<td>short vowel</td>
</tr>
<tr>
<td>([V, O])</td>
<td>open vowel</td>
</tr>
<tr>
<td>([V, C])</td>
<td>close vowel</td>
</tr>
<tr>
<td>([V, B])</td>
<td>back vowel</td>
</tr>
<tr>
<td>([V, F])</td>
<td>front vowel</td>
</tr>
<tr>
<td>([V, R])</td>
<td>lip rounded vowel</td>
</tr>
<tr>
<td>([V, U])</td>
<td>lip unrounded vowel</td>
</tr>
</tbody>
</table>
Table 3
Some phonologically relevant phonetic classes for consonants of English, with speech-initial and speech-final.
For example: [t] is [+(i)] C, -V, -N, STOP, ALV ("V") with the phonetic context included; [@i] is the preceding allophone unit and ["V"] the following one.
Consonant allophones may be more fully described phonetically by the addition of diacritics written as features, for example: [C, h] an aspirated consonant allophone; [C, o] an unaspirated consonant allophone; [C, v] a fully voiced consonant allophone; [C, n] a dental allophone for a normally alveolar consonant.

<table>
<thead>
<tr>
<th>IPA features</th>
<th>Phonetic label for Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voicing</td>
<td></td>
</tr>
<tr>
<td>voiced</td>
<td>+ V b d g d3 v ð z ʒ 3</td>
</tr>
<tr>
<td>voiceless</td>
<td>- V p t k tʃ f ø s ʃ #i #f</td>
</tr>
<tr>
<td>no opposition dependent upon voicing</td>
<td>0 V ʔ h w l r j m n ɲ</td>
</tr>
<tr>
<td>Nasality</td>
<td></td>
</tr>
<tr>
<td>nasal</td>
<td>+ N m n ɲ #i ʃ #f</td>
</tr>
<tr>
<td>non-nasal</td>
<td>- N p b t d k g tʃ dʒ f v ʔ ð s z ʒ ʒ ʒ 3</td>
</tr>
<tr>
<td>no opposition dependent upon nasality</td>
<td>0 N ʔ h w l r j</td>
</tr>
<tr>
<td>Manner</td>
<td></td>
</tr>
<tr>
<td>stop</td>
<td>STOP p b t d k g m n</td>
</tr>
<tr>
<td>affricative</td>
<td>AFFR tʃ dʒ</td>
</tr>
<tr>
<td>fricative</td>
<td>FRIC f v ø ð s z ʒ ʒ ʒ ʒ 3 h</td>
</tr>
<tr>
<td>approximant</td>
<td>APPR w l r j</td>
</tr>
<tr>
<td>Place</td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>LAB p b f v m w</td>
</tr>
<tr>
<td>dental</td>
<td>DEN ø ð l</td>
</tr>
<tr>
<td>alveolar</td>
<td>ALV t d s z n</td>
</tr>
<tr>
<td>palatal</td>
<td>PAL ʃ ʒ tʃ dʒ r j</td>
</tr>
<tr>
<td>velar</td>
<td>VEL k g ɲ w</td>
</tr>
<tr>
<td>glottal</td>
<td>GLO ʔ h</td>
</tr>
</tbody>
</table>

number of oppositions shown here. Actual values may be made to depend upon context and speaker type. For example, velum up need not always imply that AV is zero; it might be zero for close vowels [V, C] such as [i:] and [u:]; but greater than zero for open vowels [V, O] such as [æ] and [α:]. One end point is constrained for technical, not phonetic reasons: lung air pressure PLU must begin at or very close to atmospheric pressure to give a suitable set of starting conditions for the 'time marching' computations for the solution of the simultaneous differential equations in the aerodynamic block of the model.

III.3. Time plans for the allophone units

Each allophone has an onset event EON and an offset event EOFF. Section III.6 describes the selection of these events in order of priority for vowels. The various alternatives are indicated in Fig. 8 in Section III.6. Onsets and offsets for consonants are all defined by the path of the main articulator AC. Onset C.EON is the event EACX at the end of the transition into the constricted end point state for that consonant; C.EOFF is the event EAC at the start of the transition away from the consonant state. C.EON and C.EOFF are
Table 4
Transition end point labels and values for the articulators

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Units of speech production</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vocal folds, AG (in cm²)</strong></td>
<td>adducted</td>
</tr>
<tr>
<td>closed</td>
<td>0 (0 - 0.02)</td>
</tr>
<tr>
<td>stop</td>
<td>-0.10 (-0.20 - 0.10)</td>
</tr>
</tbody>
</table>

**Main articulator, AC (in cm²)**

<table>
<thead>
<tr>
<th>Tongue body, TB (in cm²) (see Harshman et al., 1977)</th>
<th>stop</th>
<th>fricative (fric)</th>
<th>approximant (appr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>front (front raising +)</td>
<td>velar and neutral (back raising + to -)</td>
<td>back (front raising -)</td>
<td></td>
</tr>
<tr>
<td>AE = 7.5 (5.0 - 10.0)</td>
<td>vCL = 0.5 (0.2 - 0.8)</td>
<td>vCL = 4.0 (2.0 - 10.0)</td>
<td></td>
</tr>
<tr>
<td>AP = 7.5 (5.0 - 10.0)</td>
<td>velar</td>
<td>velar</td>
<td></td>
</tr>
<tr>
<td>AB = 0.5 (0.2 - 0.8)</td>
<td>back</td>
<td>vCL = 0.5 (0.2 - 0.8)</td>
<td>pharyngeal</td>
</tr>
<tr>
<td>nCL = 1.0 (0.8 - 2.0)</td>
<td>palatal</td>
<td>i.e., the value of</td>
<td></td>
</tr>
<tr>
<td>&gt; 2.0</td>
<td>not palatal</td>
<td>&gt; 2.0</td>
<td>not pharyngeal</td>
</tr>
<tr>
<td>neutral: not palatal, not pharyngeal, not velar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Jaw, AJ (in cm²)**

<table>
<thead>
<tr>
<th>Jaw</th>
<th>very high</th>
<th>high</th>
<th>mid</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 (1.0 - 2.0)</td>
<td>2.5 (2.0 - 3.0)</td>
<td>4.0 (3.0 - 5.0)</td>
<td>8.0 (5.0 - 10.0)</td>
<td></td>
</tr>
</tbody>
</table>

**Lips, AL (in cm²)**

<table>
<thead>
<tr>
<th>Lips</th>
<th>very small</th>
<th>small</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 (0.3 - 0.8)</td>
<td>2.0 (0.8 - 3.0)</td>
<td>4.0 (3.0 - 5.0)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Laryngeal component of F₀ control, O (in Hz)**

<table>
<thead>
<tr>
<th>Larynx</th>
<th>low</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 (40 - 80)</td>
<td>100 (50 - 120)</td>
<td></td>
</tr>
</tbody>
</table>

**Velum, AV (in cm²)**

<table>
<thead>
<tr>
<th>Velum</th>
<th>up</th>
<th>down</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (0 - 0.2)</td>
<td>3.0 (1.0 - 5.0)</td>
<td></td>
</tr>
</tbody>
</table>

**Pressure of air in the lungs, PLU (in cm H₂O)**

<table>
<thead>
<tr>
<th>Pressure</th>
<th>zero</th>
<th>lowered</th>
<th>normal</th>
<th>raised</th>
</tr>
</thead>
<tbody>
<tr>
<td>= 0</td>
<td>5 (3 - 7)</td>
<td>8 (5 - 12)</td>
<td>15 (10 - 20)</td>
<td></td>
</tr>
</tbody>
</table>
labelled in Figs. 9, 10 and 11, in Sections III.9, III.14 and III.16, respectively. Definitions for offset of speech-initial #i.EOFF and for onset of speech-final #f.EON are more difficult to select. Some of the possible choices of events and associated coordinations are discussed in Section III.18 and sketched in Fig. 12 in that section.

The events shown have been selected as plausible and as constituting an internally consistent framework. It is not implied that these are the only possible ones. However, in describing inter-articulator coordination, it is helpful to decide, on whatever evidence is available, the direction of the dependency between the two events. Different speakers may perhaps construct time plans by means of dependencies between different sets of events. Fixed timing perturbations are assumed, but these could be modified if real speakers are shown to vary in their timing accuracy. Small dispersions could allow a simulated speaker to operate more ‘dangerously’ close to satisfactory articulatory limits. In the representation of a speaker with high variability it might be necessary to allow a bigger margin of error. An alternative or additional possibility would be that these speaker types might fulfil auditory requirements less stringently: extreme cases of this solution could take the modelling into the domain of pathological or intoxicated speech production.

The data base information given below relates to a single speaker type as regards respiratory tract dimensions and properties and articulatory transition types. An adult male speaker of English, with near-R.P. accent is described. A vocal tract length of 17.5 cm is assumed, with parameter points (see Fig. 2) at the following distances from the glottis (dfg):

<table>
<thead>
<tr>
<th>point</th>
<th>S</th>
<th>E</th>
<th>P</th>
<th>V</th>
<th>N</th>
<th>B</th>
<th>F</th>
<th>J</th>
<th>T</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>dfg in cm</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>9.5</td>
<td>11</td>
<td>15.5</td>
<td>16</td>
<td>16.5</td>
<td>17.5</td>
</tr>
</tbody>
</table>

In the modelling, these distances can be given transition instructions similar to those for parameter values such as AF. However, for simplicity, in the examples presented dfg values all remain constant throughout. Different combinations of duration D values could be taken to represent different options available to this one speaker.

The articulatory information, which is stored in unordered data files, is partly context-dependent and partly context-free. The file for each phoneme contains:

1. A classification label [V], [C] or [#], followed by a set of phonetic descriptive labels, as set out in Table 2 for vowels and in Table 3 for consonants with speech-initial and speech-final, both in Section III.2;

2. For consonants, the name of the main articulator, with its distance from the glottis and the transition type, to supplement transition information given above;

3. The location of the frication noise source for consonants. Aspiration noise is always injected into the first section of the vocal tract acoustic tube, immediately above the glottis;

4. A set of duration D values for each allophone included and their associated contexts. These D parameters either relate two events for a single articulator, in which case they define static segments, or they express inter-articulator coordination internal to that allophone. Allowable ranges and constraints on D values are included;

5. Articulator states at the end points of transitions, expressed as labels having numerical values associated with them, as given in Table 4 in Section III.3. As the framework is refined, the precision needed for each articulator will be incorporated. Paired with each label are usually two, or occasionally only one, event E labels for that articulator. When the two events are linked by ‘AND’, the state is maintained from the first event until the second one.

### III.4. Linking procedures

Allophones in a phonetic string are taken in adjacent (sequential) pairs [Ui-Uj]. A single linking duration UiUj.D is to be found. This associates the offset event of allophone Ui with the onset event of allophone Uj. The value of UiUj.D is determined by a single articulatory transition if the link is between the same articulator, but can be varied if the two linked events are for two different articulators. In the latter case the onset of the second allophone Uj.EON can come before the offset event of the first unit Ui.EOFF. It is possible also for the offset of a vowel allophone...
to come before its onset, if the two events involve different articulators. This does not, of course, mean that a negative acoustic vocoid duration results. This effect is seen in the example [an‘a], with informal style, Fig. 16(b).

III.5 Constraints on D values

The requirements of the auditory goal impose many auditory and so acoustic constraints. Perceptually relevant acoustic constraints for English may include aspects such as:

1. Relative durations of nearest neighbour vocoids, probably contributing to appropriate rhythm;
2. Relative durations of vocoid segments for phonologically long and short vowels and for utterance-final lengthening;
3. Relative durations of vocoids and the contoids following them in the case of the voiced-voiceless ([+ V] versus [— V]) distinction for consonants of English;
4. Appropriate location for a pitch movement relative to vocoid onset for the nuclear syllable of a tone group.

Absolute values of duration may be important also.

Many other acoustic features must, of course, be appropriate; for example, so as to give phonetic qualities for both vowels and consonants of the accent and style simulated and as determined by linguistic context. Plosives may require a sufficiently long contoid segment for the acoustic ‘closure’ phase, followed by a sufficiently strong transient for the release. Even where an absolute phonological requirement for acoustic contrast is absent, the constraint of normality is likely to affect features such as the nasalisation of vowels. This kind of constraint is expected to be style-dependent: nasality might be allowed to extend across more allophones in a less formal style, as modelled in the [Vn′V] series here; in informal style plosives might be acceptable even though ‘weakened’ into fricatives, fricatives into approximants, and so on. These are processes found in conversational speech (Roach, 1983).

In the detailed data bases which follow, the suggested ranges of D values are intended to express articulatory requirements for the allophones described in that data base. Requirements of adjacent allophones are likely to constrain values further. For example, non-nasal consonants, sketched in Fig. 9 in Section III.9, and specifically [t], described in Section III.11, require the velum to be nearly up through at least part of the closure as shown by paths AV(a) and AV(b) in Fig. 9. From the point of view of [t] itself D7 can have any positive value near D4 or more; the velum can start to be lowered for a following nasal consonant from a little before up to any time after the [t] release t.EA C. But for the purposes of the following nasal consonant the velum must be lowered early enough to give a nasal contoid in the acoustic output which is auditorily perceptible, without an intervening vowel. Velum-main articulator coordination must be constrained to suit both allophones. One solution in this case is a nasally released plosive.

III.6. Data base for vowels

An articulatory time plan sketch for any stressed vowel in a nuclear syllable ["V"] is shown in Fig. 8. This assumes maximal articulatory contrast between nuclear and other vowels. The actions of the vocal folds which control glottal area AG, and PLU, the respiratory control of pressure of air in the lungs, can both contribute to the prominence of the resulting vocoid. Raised lung air pressure increases the intensity and raises the fundamental frequency of the voice source. More adduction of the vocal folds to a pressed state (see Table 4 in Section III.3 for an explanation of the state labels) will give it greater intensity and more high frequency emphasis, so long as near closure of the vocal folds is avoided. When glottal area AG has a value near to closed, the acoustic intensity of the voice source falls and F0 is reduced to a very low frequency. A transition of Q, the laryngeal component of F0 control, determines the overall contour of F0 in the output (Scully and Allwood, 1985a).

Up to nine events are assumed to be coordinated with a single event EON for a very formal, full, and probably rather slow tempo, speaking style. In some contexts and for other speech
styles, many or most of these events will be omitted and the coordination plan will be much simpler. For example, air pressure PLU can be kept at normal level throughout the utterance, in which case events V.EPLUX and V.EPLU are omitted. Unless a nasal unit precedes or follows the vowel, events V.EAVX and V.EAV are not needed. Q can be already high at the start of the utterance, in which case event V.EQX disappears, although event V.EQ remains if a pitch fall is required. If another vowel precedes or follows, then the events associated with the main articulator AC of a consonant are not relevant for the vowel in question. If the vowel is speech-initial so that [#i] precedes it, then the vocal folds move from apart to a more adducted state for phonation, arriving there at event V.EAGX, in path AG(a). It is not necessarily the case that a contrast is made between a strong pressed and an ordinary phonation setting for the vocal folds. When this contrast is omitted, events V.EAGX and V.EAG disappear in path AG(b). Some effects of these kinds are shown in Example 4 Section IV.5.

Only one of the 3 events shown as coincident for the purposes of Fig. 8 defines the vowel allophone onset V.EON and only one event defines its offset V.EOFF. Onset is the main consonant articulator event V.EACX if any consonant precedes the vowel; it is the tongue body event V.ETBX if any vowel precedes; and it is V.EAGX if speech-initial [#i] precedes. In Fig. 8 these are alternatives and the other events shown there need not, in general, coincide with the single event that defines vowel onset. In the coordination expressions for the vowel, all other events are related to the one onset event and are of the form:

\[
V.EQ := V.EON + V.D1 \\
V.EOFF := V.EON + V.D4
\]

and so on.

In speech-final position PLU defines V.EOFF; elsewhere this event is associated with other articulators: V.EOFF is a tongue body event ETB if another vowel follows, as for [VI] in the diphthong of Example 2 in Section IV.3; it is the start of a transition of the main articulator AC if a consonant follows, as for [VI] in Example 4 in

---

**Fig. 8.** Articulatory time plans showing controllable articulatory parameters as functions of time for ["V"], a stressed vowel carrying a pitch fall. All the labeled events are for this vowel allophone, with full labels V.EQ, V.EPLUX and so on, except where labels for the unit following the vowel are shown. A time interval duration is positive if the event which it links to the reference onset even EON occurs later in time than EON. Here D1, D2, D3, D4, D8 and D9 are shown as positive. D5, D6 and D7 as negative. D7 is the coordination of EAGX, the moment when the vocal folds reach their adducted state for the vowel, with vowel onset V.EON. When some articulator other than AG defines the vowel onset, D4 is the articulatory duration for the vowel. The velum path AV(a) or AV(b) is needed when a nasal consonant precedes or follows the vowel, respectively. Both paths are needed if a nasal consonant both precedes and follows the vowel.
Section IV.5. For each articulator, its V.E. event comes after or at its V.E.X event. V.D4 is an articulatory duration parameter connected with vocoid duration in the acoustic output. It can take negative values only in the case that vowel onset and offset are defined by different articulators, since articulatory undershoot is not invoked, as discussed in Section II.7. Acoustically, vowel duration is dependent upon many linguistic factors; correspondingly, the value given to V.D4 is likely to be context-dependent. In the absence of a large body of relevant measurements, and also because the mapping from articulatory timing and coordination onto acoustic structures for the particular simulated speaker must be learnt through processes of trial and error with auditory monitoring, only first guesses at a set of values for V.D4 and the other articulatory durations can be offered here. In particular, vocoid durations are determined by the particular properties of the voice source modelled as well as by vocal tract shape.

III.7. Information needed for vowels: transition end point states

See Fig. 8 in Section III.6.

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at V.ETBX and V.ETB</td>
<td>TB ~ see Tables 4 and 5 in Sections III.2 and III.7.</td>
</tr>
<tr>
<td>at V.EQ and V.EQ</td>
<td>Q = high</td>
</tr>
<tr>
<td>at V.EQ + QTR</td>
<td>Q = low</td>
</tr>
</tbody>
</table>

When Q events are not included. Q has the value low.

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at V.EPLUX and V.EPLU</td>
<td>PLU = raised</td>
</tr>
<tr>
<td>at V.EPLU + PLUTR</td>
<td>PLU = normal</td>
</tr>
</tbody>
</table>

When PLU events are not included. PLU has the value normal.

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at V.EAGX and V.EAG</td>
<td>AG = pressed</td>
</tr>
<tr>
<td>at V.EAG + AGTR</td>
<td>AG = phonation</td>
</tr>
</tbody>
</table>

When AG events are not included. AG has the value phonation.

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at V.EAVX and V.EAV</td>
<td>AV = up</td>
</tr>
</tbody>
</table>

Notes

(a) Alternative options for vowels in speech-initial position:
For a vowel with ‘soft’ attack use [#V...];
For a vowel with ‘hard’ attack use [#?V...].

(b) at V.EQX and at V.EQ

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at V.EQ + QTR</td>
<td>Q = low</td>
</tr>
</tbody>
</table>

means that Q remains at a constant high value from time point V.EQX to time point V.EQ.

(c) Only one kind of nuclear tone is included here, with a falling pitch. The coordination given is intended to be appropriate for a monosyllabic word. In simulations of a bisyllabic word “toffee” it was found that Q transition needed to begin later on relative to events for the first vowel, so that the F1 fall took place mainly during the intervocalic [f].

Table 5 defines the total vocal tract area function for some vowels of English. A number of vowel qualities have been satisfactorily produced with the model, but the labels in Table 5 are to be taken as a guide to the number of contrasting states needed, rather than as a precise prescription for vowels produced by an adult male speaker with an R.P. accent of English.

III.8. Information needed for vowels: D values

See Fig. 8 in Section III.6.

The D label shows which articulatory event is to be coordinated with respect to event V.EON. D values are positive if the first-named (left hand side) event in the full D label as shown here precedes in time the second-named (right hand side) dependent event. For example V.D1 (ON-Q) concerns event V.EQ as related to event V.EON by V.EQ := V.EON + V.D1.
**Table:**

<table>
<thead>
<tr>
<th>$D$ label</th>
<th>Value (range) in 5 ms units</th>
<th>Contexts and constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.D1 (ON-Q)</td>
<td>0 (-6 to +18)</td>
<td>['V] nuclear syllable carrying a pitch movement; obligatory</td>
</tr>
<tr>
<td>V.D2 (ON-PLU)</td>
<td>0 (-6 to +18)</td>
<td>['V, ['V] stressed vowel; nuclear or non-nuclear; optional for both types</td>
</tr>
<tr>
<td>V.D3 (ON-AG)</td>
<td>0 (-6 to +18)</td>
<td>['V, ['V] stressed vowel; nuclear or non-nuclear; optional for both types</td>
</tr>
<tr>
<td>V.D4 (ON-OFF)</td>
<td>0 (-10 to +10)</td>
<td>['V] unstressed, reduced vowel</td>
</tr>
<tr>
<td></td>
<td>+8 (+2 to +14)</td>
<td>['V, S], ['V, S] stressed short vowel</td>
</tr>
<tr>
<td></td>
<td>+16 (+10 to +22)</td>
<td>['V, S (#f)], ['V, S (#f)] utterance-final stressed short vowel</td>
</tr>
<tr>
<td></td>
<td>+8 (+2 to +14)</td>
<td>['V, L], ['V, S] unstressed vowel, long or short</td>
</tr>
<tr>
<td></td>
<td>+20 (+14 to +26)</td>
<td>['V, L], ['V, L] stressed long vowel</td>
</tr>
<tr>
<td></td>
<td>+30 (+24 to +36)</td>
<td>['V, L (#f)], ['V, L (#f)] utterance-final stressed long vowel</td>
</tr>
<tr>
<td>V.D5 (ON-QX)</td>
<td>0 (-6 to +18)</td>
<td>['V] optional; D1–D5 &gt; 0</td>
</tr>
<tr>
<td>V.D6 (ON-PLUX)</td>
<td>0 (-6 to +18)</td>
<td>['V], ['V] optional for both; D2–D6 &gt; 0</td>
</tr>
<tr>
<td>V.D7 (ON-AGX)</td>
<td>0 (-6 to +18)</td>
<td>['V], ['V] optional for both; D3–D7 &gt; 0</td>
</tr>
<tr>
<td>V.D8 (ON-AV) AV(b) in Fig. 8</td>
<td>D4 ((D4-6) to any larger positive value)</td>
<td>['V], ['V] nuclear syllable, stressed vowel</td>
</tr>
<tr>
<td>V.D9 (ON-AVX) AV(a) in Fig. 8</td>
<td>0 (any negative value to +6)</td>
<td>['V], ['V] nuclear syllable, stressed vowel</td>
</tr>
</tbody>
</table>

**Notes**

(a) The allowable upper limits for D1, D2 and D3 may depend on whether [V] is a long vowel or a short vowel and also upon the particular values of D4 for the vowel.

(b) The higher positive values for V.D1 are likely to apply if [V] is the first element of a diphthong or if it is the first, stressed, vowel in a bisyllabic word.

(c) The upper end of the range for D4 is used for a slow, formal speech style; the lower end for fast, informal style.

(d) The value of V.D4 needs to be considered in relation to C.D4 for a following consonant in the same word. The D4 values given here might be said to apply when a voiced consonant [C, +V] follows. D4 durations would be shorter for a following voiceless consonant [C, −V].

(e) If there are two or more vowels in the utterance, the value of D4 for each vowel relative to that for both the preceding and the following vowel is likely to be important in many cases where, for 'rhythmical' reasons, some vowels need to be more auditorily prominent than others.

(f) Allowable values for D8, relating to the nasalisation of a vowel when a nasal consonant follows, may depend on the vowel quality and length, besides being perhaps dependent on speaker type and speech style.
Table 5  
Vocal tract shapes for some of the vowels of English and a diphthong. The end point state labels are defined in Table 4 in Section III.2

<table>
<thead>
<tr>
<th>Vowel</th>
<th>As in Phonetic Class</th>
<th>Tongue body (TB)</th>
<th>Jaw (AJ)</th>
<th>Lips (AL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i:]</td>
<td>heed [V, L]</td>
<td>front, very palatal</td>
<td>very(?) high</td>
<td>medium</td>
</tr>
<tr>
<td>[ə:]</td>
<td>heard [V, L]</td>
<td>front, not palatal</td>
<td>mid</td>
<td>large</td>
</tr>
<tr>
<td>[u:]</td>
<td>who [V, L]</td>
<td>neutral, very velar</td>
<td>high</td>
<td>(very) small</td>
</tr>
<tr>
<td>[ɑ:]</td>
<td>hard [V, L]</td>
<td>back, pharyngeal</td>
<td>low</td>
<td>large</td>
</tr>
<tr>
<td>[ɪ]</td>
<td>hid [V, S]</td>
<td>front, palatal</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>head [V, S]</td>
<td>neutral, not velar</td>
<td>mid</td>
<td>large</td>
</tr>
<tr>
<td>[o]</td>
<td>hod [V, S]</td>
<td>back, very pharyngeal</td>
<td>high</td>
<td>(very) small</td>
</tr>
<tr>
<td>[ɑ]</td>
<td>hut [V, S]</td>
<td>back, very pharyngeal</td>
<td>mid</td>
<td>small</td>
</tr>
<tr>
<td>[ʌ]</td>
<td>hood [V, S]</td>
<td>neutral, very velar</td>
<td>mid</td>
<td>small</td>
</tr>
<tr>
<td>[æ]</td>
<td>high [VV]</td>
<td>back, pharyngeal to mid to high</td>
<td>large to medium</td>
<td></td>
</tr>
<tr>
<td>[ə]</td>
<td>gway [V]</td>
<td>not palatal, not velar, not pharyngeal</td>
<td>unspecified</td>
<td>not very small</td>
</tr>
</tbody>
</table>

Note

It is assumed that a reduced vowel for an unstressed syllable of English can assume a vocal tract shape similar to that of adjacent consonants but without any extreme narrowing of the vocal tract.

III.9. Data base for consonants

Articulatory time plans for different phonetic classes of English consonants are sketched in Figs. 9, 10 and 11 in Sections III.9, III.14 and III.16, respectively. The maximum number of events needed for a given consonant type is 7 and the numbering for the $D$ labels is intended to be consistent across all consonants of English. The main articulator AC, forming an obstruction or closure of the respiratory airway, always defines onset C.ERO and offset C.ERF. Some events for AG, TB and AV are linked to the offset event C.ERF by durations $D_1$, $D_2$ and $D_3$ respectively. Other events for AG, TB and AV are related to onset C.ERF by durations $D_5$, $D_6$ and $D_7$ respectively.

Complete consistency across all consonants has not been achieved for time descriptions which include the velum. For the non-nasal consonant allophones in Figs. 9 and 15, the velum event EAVX, where the velum reaches its up, raised, state, is related to the consonant offset event EAC by C.D3; the other velum event EAV is also at velum up and relates to onset EACX by C.D7. But for the nasal consonants in Figs. 10 and 16, the corresponding velum events have the velum down; here, instead of EAVX, event EAV is related to event EAC by C.D3, with EAVX instead of EAV related to event EACX. This inconsistency in the structuring of the label names arose because priority was given to the crucially relevant end point state of the articulator in each case: that is, up for some portion of a non-nasal allophone, but down for part of a nasal one.

A similar inconsistency across allophones exists for AG path descriptions, as between Figs. 9 and 10. For the non-nasal consonants in Fig. 9 the abducted state of the vocal folds is considered crucial for AG; but in Fig. 10, for a partially voiced nasal consonant, vocal fold adduction is chosen instead. It could equally well be argued that the crucial state of the vocal folds for [d] or an unaspirated [t] [t, o] is adducted, not abducted as is suggested by Fig. 9, which includes these allophones. Here, crucialness of the event chosen...
has been sacrificed to labelling consistency across a class of consonants.

In Fig. 15(b) in Section IV.4, this labelling consistency results in compatible duration names C1.D1 for [s] and C2.D1 for [t, o]. But the alternative approach would be manageable also: then C1.D1 for [s] would relate event C1.EAG at the start of the vocal fold adduction gesture to [s] offset, while C2.D1 for [t, o] would relate event C2.EAGX at the end of the vocal fold adduction transition to [t] offset. Each choice of labelling carries its implications for the linking procedures and there are several ways in which they could be handled.

All durations except D4 describe inter-articulator coordination and so may be positive, zero or negative. D4 defines the static obstruction or closure phase between onset C.EACX and offset C.EAC. This duration cannot be negative since articulatory undershoot is not allowed, as discussed in Section II.7. The event C.EAC defines the time point of the consonant’s release. Voice onset time and voice persistence time (the duration of the continuance of voicing into the acoustic obstruction or closure contempt segment) cannot be predicted exactly from the articulatory time plan, since each of these acoustic pattern features depends on combinations of several articulatory factors, as well as the particular properties of the voice source model and the acoustic behaviour of the vocal tract.

Duration parameter values for consonants are subject to constraints on the ordering of C.E.X and C.E.. as are the corresponding events for vowels, to avoid articulatory undershoot by any of the articulators. A first guess at some D values for consonants is given below.

Each articulator has the possibility of maintaining a static posture, for an approximation to the small distance movements occurring between mid-transition icebergs in natural speech, as discussed in Sections II.1 and II.9. It is probably the case that one particular articulator is the duration determining factor (DDF) for a given portion of a total time plan; probably, for naturalness in modelling a fairly informal conversational style of speech, this articulator should be made to move continuously, without a static segment. For example, as shown in Fig. 9, an intervocally placed voiceless fricative or aspirated plosive of English requires an abduction-adduction gesture of the vocal folds. This may strongly constrain the time plan for an utterance (Scully and Allwood, 1985b). There seems to be no good reason why, in this case, an abducted state of the vocal folds should be maintained. Published data show that in natural speech the actions of the vocal folds are generally continuous in this context, as discussed.

---

Vol. 6, No. 2, June 1987
in Section II.14. Therefore, in the modelling, the two events C.EAGX and C.EAG are generally made to coincide, giving \( C.D5 - C.D1 = C.D4 \) as a generally operating constraint. However, for completeness, the possibility of static abducted postures of the vocal folds is allowed, as for other articulators. Similarly, where the velum is the DDF, as, for example, in words such as "tent" and "cinder", there is unlikely to be a static segment with a lowered velum between n.EAVX and n.EAV, two events shown generalised across all three nasal consonants of English in Fig. 10.

III.10. Information needed for some consonants

A few alveolar and dental consonants are used for illustration, as well as [h] and the glottal stop. Some of the major allophones are considered, but this is not intended to be a comprehensive exposition. Words are assumed to be monosyllabic or bisyllabic. Phenomena of connected speech such as assimilation and elision are not considered. The immediately preceding and following units, including word boundary and speech-initial or speech-final are, in nearly every case, the only ones taken to define phonetic context. In all cases it is assumed that lung air pressure PLU is high enough for the appropriate sounds to be generated. Larynx pitch control Q actions are considered to be not relevant to consonants. The duration values given are first guesses in most cases, based on natural speech where possible. Some of the relevant data are discussed in Chapter II. [t] is considered in some detail; [d], [s], [z], [n], [l], [?] and [h] in outline.

III.11. Data base for [t]

1. \([t] = [C, -V, -N, \text{STOP, ALV}]\).
2. The main articulator (AC) is the tongue tip-blade AF with distance from the glottis DF of 15.5 cm. AF transitions are of type \((16,100)\) requiring 80 ms, but more complex transitions may be appropriate for the release.
3. Frication noise (determined by AF and aerodynamic conditions at the constriction) is injected into the acoustic tube at the teeth, that is at DT = 16.5 cm.
4. D values (see Fig. 9).

The \( D \) label shows which articulatory event is to be coordinated with respect to event EAC or event EACX. \( D \) values are positive if the first-named (left hand side) independent event in the full \( D \) label as shown here precedes in time the second-named (right hand side) dependent event. For example t.D1(AC-AG) concerns event t.EAG as related to event t.EAC by

\[
\text{t.EAG} = \text{t.EAC} + \text{t.D1}
\]
C. Scully / Units of speech production

<table>
<thead>
<tr>
<th>$D$ label</th>
<th>Value (range) in 5 ms units</th>
<th>Allophones and contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t.D1$ (AC-AG)</td>
<td>0 (−6 to +6)</td>
<td>[t, h] aspirated, where ['V] or [APPR] ['V] follows and [s] does not precede in the same word</td>
</tr>
<tr>
<td></td>
<td>−AGTR = −24 (−30 to −18)</td>
<td>[t, (h)] weakly aspirated, where ['V] follows</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[t, o] unaspirated where [s] precedes in the same word</td>
</tr>
<tr>
<td>$t.D2$ (AC-TB)</td>
<td>−5 (−11 to +1)</td>
<td>generally</td>
</tr>
<tr>
<td></td>
<td>&gt; −1.D4</td>
<td>[V−V] context</td>
</tr>
<tr>
<td></td>
<td>&gt; −(t.D4 + s.D4)</td>
<td>[s−V] context</td>
</tr>
<tr>
<td>$t.D3$ (AC-AVX)</td>
<td>0 (+6 to any negative value)</td>
<td>where [C, +N] or [#] precedes</td>
</tr>
<tr>
<td>AV(a) in Fig. 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t.D4$ (ACX-AC)</td>
<td>+12 (+6 to +18)</td>
<td>[t, h] and [t, (h)] including word-initial where there is no junctural lengthening</td>
</tr>
<tr>
<td></td>
<td>+24 (+18 to +32)</td>
<td>to emphasise the word</td>
</tr>
<tr>
<td></td>
<td>related to V.D4 for</td>
<td>[t, h] word-initial, where there is junctural lengthening for emphasis</td>
</tr>
<tr>
<td></td>
<td>the preceding vowel;</td>
<td>where ['V] precedes in the same word</td>
</tr>
<tr>
<td></td>
<td>the ratios V.D4/C.D4 help</td>
<td></td>
</tr>
<tr>
<td></td>
<td>to distinguish [t] and [d]</td>
<td></td>
</tr>
<tr>
<td>$t.D5$ (ACX-AGX)</td>
<td>constraint: $t.D5 ≤ t.D4 + t.D1$</td>
<td>equality avoids a static AG segment with vocal folds abducted</td>
</tr>
<tr>
<td>$t.D6$ (ACX-TBX)</td>
<td>+5 (−1 to +11)</td>
<td>constraint: $t.D6 ≤ t.D4 + t.D2$</td>
</tr>
<tr>
<td>$t.D7$ (ACX-AV)</td>
<td>$t.D4 (t.D4−6) to any positive value</td>
<td>where [C, +N] or [#] follows</td>
</tr>
<tr>
<td>AV(b) in Fig. 9</td>
<td>constraint: $t.D7 ≥ t.D4 + t.D3$</td>
<td>The [t] is nasally released if $D7$ is much less than $D4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>where a [+N] unit both precedes and follows</td>
</tr>
</tbody>
</table>

Notes

(a) $D4$ values probably vary with linguistic context and with speech style.

(b) A closure $D4$ as short as 6 time units (30 ms) is probably below the minimum value required for perception of a contoid segment, but might be admissible in a very informal, casual style of speech. In some contexts and for a formal style, unacceptable preaspiration will result if the closure is made too short.

(c) Where [#i] precedes it may be appropriate to allow PLU to reach its normal level before the [t] release. In this case C1.D4 needs to be made greater than or equal to −#iC1.D (see Fig. 12(a) in Section III.18). Then the minimum duration for the [t] closure is determined by the respiratory-vocal tract coordination for speech onset. Similar arguments apply to other consonants in speech-initial position.

(d) Event AGX is not needed if for example [s] or [#i] precedes, elements for which the vocal folds are already apart. This reduction in the number of required events for an allophone unit because of its immediate phonetic context is a widely applicable phenomenon which some of the examples of derived articulatory plans try to demonstrate.

(e) For single words, AV(a) and AV(b) are alternatives, but where [t] is both preceded and followed by a nasal allophone [+N] both coordinations $t.D3$ and $t.D7$ need to apply. Then the two AV transitions become DDFs and impose time constraints. $D4$ now has to be long enough to accommodate them so as to avoid intrusive sounds between the nasals and the plosive, or, if a short $D4$ closure is required on rhythmical grounds, an alternative strategy may be used: one which avoids making [t] non-nasal. For example, in conversational style ...sent me... most English speakers seem to use [?] which is [ON] in the place of [t] (unpublished laboratory data).
(5) Transition end points:

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at t.EACX</td>
<td>AC = stop</td>
</tr>
<tr>
<td>\text{AND} at t.EAC</td>
<td></td>
</tr>
<tr>
<td>at t.EAGX</td>
<td>AG = abducted (apart for {t, h}, breathy for {t, (h)}) and for {t, o} also, if the AG event is needed at all for the unaspirated plosive)</td>
</tr>
<tr>
<td>\text{AND} at t.EAG</td>
<td></td>
</tr>
<tr>
<td>at t.ETBX</td>
<td>TB = front, not palatal (region)</td>
</tr>
<tr>
<td>\text{AND} at t.ETB</td>
<td></td>
</tr>
<tr>
<td>at t.EAVX</td>
<td>AV = up</td>
</tr>
<tr>
<td>\text{AND} at t.EAV</td>
<td></td>
</tr>
</tbody>
</table>

Notes

(a) Whereas jaw position AJ is precisely defined for [t], the tongue body TB and lips AL have only to be within a region or range of values.

(b) As discussed in Sections II.14 and III.12, vocal fold abduction may not be needed for English voiceless unaspirated plosives.

III.12. Data base for [d]

(1) \([d] = [C, + V, - N, \text{STOP, ALV}].\)
(2) Main articulator as for [t].
(3) Frication noise injected as for [l].
(4) \(D\) values (see Fig. 9):

<table>
<thead>
<tr>
<th>(D) label</th>
<th>Value (range) in (5) ms units</th>
<th>Allophones and contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.D1 (AC-AG)</td>
<td>(-D4/2 (+/- 6))</td>
<td>all contexts for which vocal fold abduction is used</td>
</tr>
<tr>
<td>d.D5 (ACX-AGX)</td>
<td>(+D4/2 (+/- 6)) constraint: (d.D5 \leq d.D4 + d.D1)</td>
<td>equality avoids a static AG segment with vocal folds abducted</td>
</tr>
</tbody>
</table>

For other aspects of coordination see [t]. \(D4\) values may be different however, and these will probably depend partly on whether the speaker abducts the vocal folds for voiced plosives or not. It seems possible that a single speaker of English might abduct the vocal folds for [d] (and for other voiced plosives) in some phonetic contexts and some speech styles, but not in others for which \(D4\) needs to be short on rhythmical grounds. Where vocal fold abduction is used for [d] a large enough \(D4\) value is needed to accommodate the DDFs of the two vocal fold transitions, avoiding both aspiration at the release and preaspiration at the closure, associated with abducted vocal folds at these times. Preaspiration may be admissible in some contexts, for example if an unstressed reduced vowel precedes the [d].

(5) Transition end points.

As for [t], except that, if the vocal folds are abducted for [d], AG may go to breathy rather than to apart at EAG. If not, then this event and its associated AG state are not relevant.
III.13. Data bases for [s] and [z]

1. \( [s] = [C, -V, -N, F R I C, A L V] \).
   \( [z] = [C, +V, -N, F R I C, A L V] \).
2. Main articulator as for [t], with transitions of type (16,100) but these have also been varied to match specific real speakers (Scully and Clark, 1986).
3. Frication noise injected as for [t].
4. \( D \) values for [s] (see Fig. 9):

<table>
<thead>
<tr>
<th>Label</th>
<th>Value (range) in 5 ms units</th>
<th>Allophones and contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.D1(AC-AG)</td>
<td>(-D4/2 (+/- 6))</td>
<td>all</td>
</tr>
<tr>
<td>s.D2(AC-TB)</td>
<td>as for [t]</td>
<td>all</td>
</tr>
<tr>
<td>s.D3(AC-AVX)</td>
<td>(-D4/2 (+/- 6))</td>
<td>where ([C. +N]) precedes</td>
</tr>
<tr>
<td>Av(a) in Fig. 9</td>
<td>or any more negative value</td>
<td></td>
</tr>
<tr>
<td>s.D4(ACX-AC)</td>
<td>(+12 (0 to 24)) related to V.D4 for the preceding vowel; the ratios V.D4/C.D4 help to distinguish [s] and [z]</td>
<td></td>
</tr>
<tr>
<td>s.D5(ACX-AGX)</td>
<td>(+D4/2 (+/- 6)) constraint: ( s.D5 \leq s.D4 + s.D1 )</td>
<td>all</td>
</tr>
<tr>
<td>s.D6(ACX-TBX)</td>
<td>as for [t]</td>
<td></td>
</tr>
<tr>
<td>s.D7(ACX-AV)</td>
<td>(+D4/2 (+/- 6)) or any more positive value constraint: ( s.D7 \geq s.D4 + s.D3 )</td>
<td></td>
</tr>
</tbody>
</table>

Values for [z]: similar to those for [s] but \( D \) values may differ.

(5) Transition end points for [s]:

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at s.EACX AND at s.EAC</td>
<td>AC = fric</td>
</tr>
<tr>
<td>at s.EAGX AND at s.EAG</td>
<td>AG = apart</td>
</tr>
<tr>
<td>at s.ETBX AND at s.ETB</td>
<td>TB = front. not palatal (precise)</td>
</tr>
<tr>
<td>at s.EAVX AND at s.EAV</td>
<td>AV = up</td>
</tr>
</tbody>
</table>

Values for [z]: similar to those for [s] except that at z.EAGX and at z.EAG AG = breathy.

Note

Tongue body position needs to be more precisely defined for [s] and [z] than for [t], [d], [n] and [l], as discussed in section II.16.

III.14. Data base for [n]

1. \( [n] = [C, 0V, +N, STOP, A L V] \).
2. Main articulator as for [t] with similar transitions.
3. Frication noise injected as for [t].
Fig. 10. Articulatory time plans for some allophones of the nasal consonants of English [m, n, r]. [C, +N]. D1, D2 and D3 relate to consonant offset C.EAC. D5, D6 and D7 relate to onset C.EACX. D4 is the closure duration. Although labelled [0V] in Table 3 Section III.2 because of the lack of a phonological opposition related to [+V] versus [−V] for nasal consonants in English, these units are treated as [+V] in their articulatory requirements on grounds of audibility. Paths AG(a) or AG(b) are needed if an allophone unit requiring vocal fold abduction precedes or follows respectively. Both may be needed.

(4) D values (see Fig. 10):

<table>
<thead>
<tr>
<th>D label</th>
<th>Value (range)</th>
<th>Allophones and contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>n.D1 (AC-AGX)</td>
<td>0 (+6 to any negative value)</td>
<td>where [C, −V] or [#f] precedes</td>
</tr>
<tr>
<td>AG(a) in Fig. 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n.D2 (AC-TB))</td>
<td>as for [t]</td>
<td>all</td>
</tr>
<tr>
<td>n.D3 (AC-AV)</td>
<td>D4/2 (+/− 6)</td>
<td>where [C, −N] or [V, −N] follows</td>
</tr>
<tr>
<td>n.D4 (ACX-AC)</td>
<td>similar to [t] or [d]</td>
<td></td>
</tr>
<tr>
<td>n.D5 (ACX-AG)</td>
<td>0 (−6 to any positive value)</td>
<td>where [C, −V] or [#f] follows</td>
</tr>
<tr>
<td>AG(b) in Fig. 10</td>
<td>constraint:</td>
<td></td>
</tr>
<tr>
<td>n.D5 □ n.D4 + n.D1</td>
<td>where a [−V] element both precedes and follows</td>
<td></td>
</tr>
<tr>
<td>n.D6 (ACX-TBX)</td>
<td>similar to [t] or [d]</td>
<td>all</td>
</tr>
<tr>
<td>n.D7 (ACX-AVX)</td>
<td>D4/2 (+/− 6)</td>
<td>where [C, −N] or [V, −N] precedes</td>
</tr>
<tr>
<td>constraint:</td>
<td>where a [−N] element both precedes and follows</td>
<td></td>
</tr>
<tr>
<td>n.D7 □ n.D4 + n.D3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note

From the point of view of phonological opposition, voicing does not matter for nasal consonants of English, but for reasons of audibility the nasal contoid in the acoustic output probably needs to be at least partially voiced.

(5) Transition end points:

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at n.EACX AND at n.EAC</td>
<td>AC = stop</td>
</tr>
<tr>
<td>at n.EAGX AND at n.EAG</td>
<td>AG = adducted</td>
</tr>
<tr>
<td>at n.ETBX AND at n.ETB</td>
<td>TB = front, not palatal (region)</td>
</tr>
<tr>
<td>at n.EAVX AND at n.EAV</td>
<td>AL = medium (region)</td>
</tr>
<tr>
<td></td>
<td>AV = down</td>
</tr>
</tbody>
</table>

III.15. Data base for [l]

1. [l] = [C, 0V, 0N, APPR, ALV].
2. The main articulator is the tongue tip, represented by AF as for [t], but with a Fast (F = 8 × 5 = 40 ms) transition.
(3) Frication noise injected as for [t].

(4) $D$ values:
If [l] is auditorily acceptable when its nasality and voicing are uncontrolled, then only the coordination of $\text{AC}$ and $\text{TB}$ is relevant, as shown in Fig. 9. Otherwise velum events $\text{EAVX}$ and $\text{EAV}$ ($\text{AV(a)}$ and $\text{AV(b)}$ in Fig. 9) and vocal fold events $\text{EAGX}$ and $\text{EAG}$ ($\text{AG(a)}$ and $\text{AG(b)}$ in Fig. 10) need to be included, for [l] to be at least partially non-nasal and partially voiced.

(5) Transition end points:

<table>
<thead>
<tr>
<th>Time description</th>
<th>End point labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>at 1.EACX AND at 1.EAC</td>
<td>$\text{AC} = \text{approximant}$</td>
</tr>
<tr>
<td>at 1.ETBX AND at 1.ETB</td>
<td>$\text{TB}$: for 'clear' [l]: $\text{TB} = \text{front, not palatal (region)}$ for 'dark' [l]: $\text{TB} = \text{front, velar (region)}$ $\text{AJ} = \text{high (region)}$ $\text{AL} = \text{medium (region)}$</td>
</tr>
</tbody>
</table>

Fig. 11. Articulatory time plans for: (a) [2], and (b) [h]. No vocal fold-vocal tract coordination is required, except possibly AG-AV coordination for velum lowering in some examples of [h], as discussed in Section III.17.


(1) [2] = [C, 0V, 0N, STOP, GLO].
(2) The main articulator is AG.
(3) Here noise generated at the main articulator is aspiration noise, injected as usual, just above the glottis. A significant amount of frication noise, generated at vocal tract constrictions for adjacent allophones, is possible but unlikely.
(4) $D$ values: see Fig. 11(a). No inter-articulator coordination is needed.

III.17. Data base for [h]

(1) [h] = [C, 0V, 0N, FRIC, GLO].
(2) The main articulator is AG.
(3) Noise generated at the main articulator is aspiration noise. Frication noise generated at vocal tract constrictions will probably be important also. Some speakers of English lower the velum for [h] in non-nasal contexts, probably in order to prevent oral pressure from building up above 1 or 2 cm H$_2$O. Higher oral pressure would reduce aspiration noise and result in a vocal tract fricative (unpublished data). A comparable effect may be seen in Fig. 17(d) where a vowel shaped vocal tract has produced a fricative because of abduction of the vocal folds.

III.18. Data base for speech-initial and speech-final units

Besides the familiar phonological classes of vowels and consonants and their allophones, additional phonetic units, labelled [#], are needed for initiation and termination of an utterance. Each simulation is limited at present to one expiratory breath, so that eachallophonic string begins and ends with [#]. [#] (initial) is the start of an expiratory breath; [#f] (final) is its end. These units have different articulatory properties in different contexts and are incorporated in the phonetic data base as if they were allophones of a phonological element.

Appropriate inter-articulator coordination is
important at speech initiation and termination, just as elsewhere, for the generation of the intended sounds and avoidance of extra, intrusive ones. Different speakers seem to make the switch from respiration to speech in different ways (Sawashima, Hirose, Ushijima and Niimi, 1975). Several plausible alternative forms of coordination may be suggested in the modelling. These need to be tested against natural speech data. The organisation of [#] units is sketched in Fig. 12.

Each simulation begins and ends at a respiratory state, with velum AV down, vocal folds AG apart and atmospheric pressure throughout the respiratory tract. Tongue body, lips and jaw are assumed to be already at their values for the first allophone for which they need to be specified. There is assumed to be no consonantal obstruction or closure by AC. The value of Q. for $F_i$ control, may be high or low.

It is assumed that an utterance ends with coordinated gestures of AG, AV and PLU for [#f], as indicated in Fig. 12(b). These might be characteristic of a speaker type, but might also vary with phonetic context. There may be, similarly, internal coordination between AG, AV and PLU for speech-initial [#i], shown in Fig. 12(a). For example, velum raising might be tied to the first non-nasal [—N] allophone in the utterance, or it might perhaps have a fixed timing relative to event #i.EPLUX. Similarly, if raising of lung air pressure PLU is taken to be an independently controlled action, as here, then vocal fold adduction could be governed solely by the requirements of the first vowel or consonant in the string. However, in a more realistic respiratory model the rise of air pressure in the lungs would be determined partly by the actions of the vocal folds AG, the velum AV and the main vocal tract articulator AC. One important aspect of the modelling of speech onset and offset seems to be avoidance of

---

(a) [#i]

---

Fig. 12(a). Articulatory time plans for speech-initial [#i]. The whole action plan is initiated at ET0, by AC if the first allophone is a consonant and by AG if it is a vowel. There may be specific lung-vocal fold coordination for speech onset, shown by path AG(a) with coordination duration #i.D1; and, correspondingly, specific lung-velum coordination, with coordination duration #i.D2. Alternatively, vocal fold and velum actions may be linked to an allophone following [#i] in the string. This is exemplified for the case where AG actions are linked to [V1] by path AG(b).

---

(b) [#f]

---

Fig. 12(b). Articulatory time plans for speech-final [#f]. The end of the whole action is ETMAX, where PLU has just reached zero. Actions of AG and AV are assumed to be related to an event for PLU by durations #f.D1 and #f.D2 respectively. The possibilities may be more complex than this, as discussed for [#i].
intrusive sounds (Scully and Allwood, 1984). For the respiratory model employed here this can be achieved at speech onset if the whole action plan is initiated by the main articulatory action required for the first allophone in the utterance, that is, AG for [V1] or AC for [C1]. PLU should be raised late enough with respect to this action to avoid pre-speech acoustic sources. Therefore a linking duration needs to be specified:

\[ #1V1.D (PLUX-AGX) \]

or

\[ #1C1.D (PLUX-ACX) \]

for which reasonable values might be \(-10 (\text{to} -4)\). Similarly, suitable values of the linking \( D \) between the last allophone and [\#f] need to be established. All the examples in Chapter IV end with a vowel. Here the vowel offset is taken to be [\#f.EPLU], the start of the final fall in lung air pressure. This is the onset event for [\#f], so the linking duration is zero. For consonant endings, the linking duration and the [\#f] coordination must be such as to avoid generating an extra vowel element after the consonant release, while generating the required sounds for the consonant.

The time plan ends when lung air pressure PLU has reached zero again. Pairs of [\#i] and [\#f] define expiratory breath groups; for example [\#1...\#2 \#3...\#4 \#5...\#6] defines 3 breath groups. Odd numbers are [\#i]; even numbers are [\#f].

Chapter IV. The construction of an articulatory time plan: examples, results and conclusions

Introduction

Even a static vowel produced in isolation requires appropriate articulatory timing and coordination between the larynx and the lungs (Scully and Allwood, 1983b). To this is added, for a diphthong, coordination between the vocal tract and the other articulators. Isolated "Ah" [\#a:] and "I" [\#i] will be used as the first two examples to demonstrate this. "Tar" [\#t:z\#a:] and "Star" [\#st:z\#a:] will be used as illustrations of consonantal onsets. All these examples will be considered to be in a formal style. Some ways in which timing might be reorganised in changing from a formal to an informal style will be suggested, using [\#s:n\#3:] and related sequences. Use of the model and its flexible timing scheme to match a particular speaker will be outlined next, for a sequence [\#ps:z\#3:p]. Although these last two examples do not form real words of English, the portions of interest are found in real speech extending across word boundaries. Changes of inter-articulator coordination can result in a shift to a new word of English, or can introduce a non-English sound, or can produce a perceived change in the speaker type and speech style. This is illustrated in the final example by spectrograms of synthetic speech generated with the model.

IV.1. Procedures used

The procedures to be described in detail or outlined for each example are as follows:

1. The utterance is expressed as a string of symbols, for a broad phonetic transcription; and also as a sequence of [\#V], [\#C] and [\#] allophone units. The phonetic context of each allophone is defined; thus onset and offset events for all the allophones are known.

2. The appropriate allophones are selected from the individual data files. Context-dependent rules are applied and selections are made where options are available, both for the events \( E \) to be included and for the duration \( D \) values to be used in the coordination expressions defined by the events selected.

3. The allophone units are taken in sequential pairs [\#U_i-U_j]. A single linking duration [\#U_iU_j.D] is found. This associates one and only one event of [\#U_i] with one and only one event of [\#U_j], specifically, the offset event of [\#U_i], [\#U_i.EOFF], and the onset event of [\#U_j], [\#U_j.EON]. The value of [\#U_iU_j.D] is defined by a transition duration if the link is homorganic or is selected otherwise. In the latter case, [\#U_iU_j.D] is allowed to be negative as well as positive.

4. A sketch of the time plan is drawn. It includes event \( E \) and duration \( D \) labels, with appro-
appropriate transition types, as given in the data base.

(5) Based on this sketch, the instructions for each articulator are written, with "E" and "D" labels for the timing aspects and phonetic articulatory description labels for the end points. Precise values chosen from the data base are substituted for these end point labels.

(6) Also based on the sketch, the coordination expressions are written. AC., the main articulator for each consonant, is now given its true articulator name. "D" labels are given precise values.

The written outputs from (5) and (6) fully define the input command file for the first, articulatory, block of the model of speech production processes. Additional information supplied to the model at this point describes the speaker type to be simulated: distances from the glottis of the parameter points for the vocal tract articulators, initial lung volume, wall compliances, subglottal conductance constants and so on. When this part of the model is run, the internal consistency of the coordination expressions is checked and actual time points are assigned to all the event, "E", labels. ETMAX, which is one of them, defines the total duration of the run. To assess the suitability of the computed articulatory plan, other portions of the model need to be run, so that aerodynamic, acoustic or auditory comparisons with the target utterance in natural speech may be made.

In the examples below, time is always expressed in 5 ms units. Tables 1-5 define the phonetic symbols and suggest end point states and "D" values.

IV.2. Example 1: “Ah”, a long vowel [a:] produced in isolation, with a falling pitch (Fig. 13)

(1) Phonetic strings

“Ah”, [a:] in isolation can be either:

[#"a:"#] = [#1 "V1 #2]

or

[#2"a:"#] = [#1 C1 "V1 #2]

The second option, for ‘hard’ attack, will be chosen here. Data files needed are: [#1], [#2], ["a:"], and [#f].
Since [Cl] is [ON] the velum AV is free to coordinate with PLU in [#1] with coordination expression

\[ \text{#1.EAVX} := \text{#1.EPLUX} + \text{#1.D2} \]

D1 does not apply since AG is the main articulator for [Cl] and is not made to coordinate with PLU in [#1].

Choose \( D2 := 0 \)

\[
\text{[C1]. C1.EAC} := \text{C1.EACX} + \text{C1.D4}
\]

Choose \( D4 := 18 \)

AC is AG (going from apart to closed) so ACTR is \( M = 24 \), so

\[ \text{C1.EACX} \text{ is 24 time units after } \text{ETO} \]

[#1]. For ["V". event EQ must be included, but events EAGX, EAG, EPLUX, EPLU and EOX are optional. Choose to include only EQ here. It may be seen in the coordination expressions, below, that there is an event V1.EAGX but this is derived from an event V1.EACX. AG happens to be the main articulator AC there. The events EAGX and EAG omitted here concern instead the vocal folds in their vowel-related role. AG(b) in Fig. 8.

\[ V1.EQ := V1.EON + V1.D1 \]

Choose \( D1 := 6 \)

\[ \text{[a:] is ["V, L (#f)] so D4 should be 30 (range 24–36).} \]

Choose \( D4 := 30 \)

\[ \text{[#2]. #2. EPLU is V1.EOFF} \]

\[ \text{#2.EAG} := \text{#2.EPLU} + \text{#2.D1} \]

Choose \( D1 := 0 \)

\[ \text{#2.EAV} := \text{#2.EPLU} + \text{#2.D2} \]

Choose \( D2 := 0 \)

\[ \text{ETMAX} := \text{#2.EPLU} + \text{PLUTR} \]

PLUTR is \( M = 24 \)

(3) Links between allophone units

\[ \text{[#1–C1]. #1C1.D(PLUX–ACX)} \]

This linking \( D \) can be varied since the link is not homorganic. Choose a value of \(-10\) from the [#1] file.

\[ \text{[C1–V1]. C1V1.D(AC–ACX)} \]

This is a homorganic link, so the \( D \) value is the duration of the appropriate transition. in this case that of AG. Here AG moves from closed to phonation, so it is given a fast (\( F \)) transition of 8 time units.

\[ \text{[V1–#2]. V1#2.D(PLUX–PLUX)} \]

This linking \( D \) has the value 0 since the PLU event, PLUX, forms both the offset event for [V1] and the onset event for [#2].

(4) Sketch of the time plan Fig. 13

### Instructions for the articulators, based on the time plan, Fig. 13

<table>
<thead>
<tr>
<th>Articulator (units)</th>
<th>Line</th>
<th>End point label</th>
<th>Till time</th>
<th>Transition type</th>
<th>Label values</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG = AC1 (cm³)</td>
<td>0</td>
<td>apart</td>
<td>C1.EAC</td>
<td>M</td>
<td>apart = 0.4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>closed</td>
<td>2.EAG</td>
<td>F</td>
<td>closed = 0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>apart</td>
<td>ETMAX</td>
<td>M</td>
<td>phonation = 0.05</td>
</tr>
<tr>
<td>TB (cm³)</td>
<td>0</td>
<td>[a:]</td>
<td>ETMAX</td>
<td>S</td>
<td>[a:] TB back, pharyngeal</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>vi</td>
<td></td>
<td></td>
<td>AJ low, AL large</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>low</td>
<td>ETMAX</td>
<td>S</td>
<td>AE AP AB 'AJ AL</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>up</td>
<td>ETMAX</td>
<td>M</td>
<td>5 1 10 5 10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>high</td>
<td>V1.EQ</td>
<td>S</td>
<td>high = 90</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>low</td>
<td>ETMAX</td>
<td>S</td>
<td>low = 50</td>
</tr>
<tr>
<td>Q (Hz)</td>
<td>0</td>
<td>high</td>
<td>V1.EQ</td>
<td>S</td>
<td>high = 90</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>low</td>
<td>ETMAX</td>
<td>S</td>
<td>low = 50</td>
</tr>
<tr>
<td>AV (cm³)</td>
<td>0</td>
<td>down</td>
<td>#1.EAVX</td>
<td>-M</td>
<td>down = 2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>up</td>
<td>#2.EAV</td>
<td>M</td>
<td>up = 0.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>down</td>
<td>#2.EAV</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>up</td>
<td>ETMAX</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>PLU (cm H₂O)</td>
<td>0</td>
<td>zero</td>
<td>#1.EPLUX</td>
<td>-M</td>
<td>zero = 0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>normal</td>
<td>#2.EPLU</td>
<td>M</td>
<td>normal = 8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>zero</td>
<td>ETMAX</td>
<td>M</td>
<td></td>
</tr>
</tbody>
</table>
Note

Delayed transitions, shown with a negative sign preceding the transition type label, end at events of type E...X and not before. (ETMAX is a special name, not of this general type E...X.) Transitions without a negative label begin at specified events of type E...

(6) Coordination expressions with the main articulator for the consonant given its true name

Onset and offset events are given their specific articulatory E labels. This is based on the time plan sketch, see Fig. 13.

\[
\begin{align*}
\text{ETO} & := 0 \\
\text{C1.EAGX} & := \text{ETO} + \text{C1.AGTR} \\
\#1.EPLUX & := \text{C1.EAGX} - \#1.C1.D \quad (*) \\
\#1.EAVX & := \#1.EPLUX + \#1.D2 \\
\text{C1.EAG} & := \text{C1.EAGX} + \text{C1.D4} \\
\text{V1.EAGX} & := \text{C1.EAG} + \text{C1V1.D} \\
\text{V1.EQ} & := \text{V1.EAGX} + \text{V1.D1} \\
\#2.EPLU & := \text{V1.EAGX} + \text{V1.D4} \\
\#2.EAG & := \#2.EPLU + \#2.D1 \\
\#2.EAV & := \#2.EPLU + \#2.D2 \\
\text{ETMAX} & := \#2.EPLU + \#2.PLUTR
\end{align*}
\]

(*) This expression is a special case, arranged differently, with the event for the preceding unit [#1] defined in terms of that for the following unit [C1]. This is because [C1], rather than [#1], relates to ETO.

\[
\begin{align*}
\text{C1.AGTR} & := 24 \\
\#1.C1.D & := -10 \\
\#1.D2 & := 0 \\
\text{C1.D4} & := 18 \\
\text{C1V1.D} & := 8 \\
\text{V1.D1} & := 6 \\
\text{V1.D4} & := 30 \\
\#2.D1 & := 0 \\
\#2.D2 & := 0 \\
\#2.PLUTR & := 24
\end{align*}
\]

Besides the two transition times shown here, other transitions have the durations shown on Fig. 13: 24 units for M or \(\neg M\), 38 for S and 8 for \(\bar{F}\). The input command file for the articulatory block of the model is now fully defined. By substituting the \(D\) values, above, in the coordination expressions, the \(E\) time point values can be checked. This gives ETMAX, the total duration, as 104 time units, or 520 ms. Here the manual procedures end. When the computer programme for the articulatory block of the model is run, the actual \(D\) values are substituted in the coordination expressions and all the time values for the event labels are computed.

Diagram 1 shows the time structure of the chain of inter-dependent articulatory events for this version of "Ah". Arrows lead towards the dependent event of each pair in a coordination expression.

A single change of one \(D\) value, above, will result in the whole time scheme being recomputed. In this way, many variations on a basic time structure can be simulated. Here, the auditory effect of varying the timing of the pitch fall, by changing V1.D1, could be studied, for example. With different coordinations towards the end of the time plan, different amounts of breathy voicing would be expected; with inappropriate coordination at the beginning, extra intrusive sounds are likely to be generated, as shown earlier for a series of [i] vowels (Scully and Allwood, 1984).

IV.3. Example 2: "I", a diphthong [ai] produced in isolation, with a falling pitch (Fig. 14)

Inter-articulator coordination has been simplified here for onset and offset, leaving the coordination between the tongue body transition for [ai] and the pitch fall, controlled by Q, to be described. The instructions for the articulators are ex-
pressed first in terms of simple end point, duration and transition specifications; then with event labels that give the same durations.

\[
[ # "a1# ] = [ #1 "V1 V2 #2 ]
\]

![Diagram of articulatory time plan for Example 2: “I” said in isolation with a falling pitch.](image)

**Fig. 14.** Articulatory time plan for Example 2: “I” said in isolation with a falling pitch.

**Phonetic strings:**

“I”, [a1] in isolation can be either \[#"a1#\] = [#1 "V1 V2 #2] or \[ #2"a1#\] = [#1 C1 "V1 V2 #2]

The first option for ‘soft’ attack will be chosen here.

**Instructions for the articulators**

<table>
<thead>
<tr>
<th>Articulator Line</th>
<th>End point label</th>
<th>Duration (5 ms)</th>
<th>Transition type</th>
<th>Label values</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG (cm)</td>
<td>0</td>
<td>apart</td>
<td>0.5</td>
<td>phonation = 0.05</td>
</tr>
<tr>
<td>TB (cm)</td>
<td>[a]</td>
<td>1</td>
<td>34 S</td>
<td>[a] TB back, pharyngeal.</td>
</tr>
<tr>
<td>O (Hz)</td>
<td>0</td>
<td>high</td>
<td>high = 100</td>
<td>[a] TB back, pharyngeal.</td>
</tr>
<tr>
<td>AV (cm)</td>
<td>0</td>
<td>down</td>
<td>down = 2</td>
<td>[a] TB front, palatal.</td>
</tr>
<tr>
<td>PLU (cm H₂O)</td>
<td>0</td>
<td>zero</td>
<td>zero = 0</td>
<td>[a] TB front, palatal.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>zero</td>
<td>normal = 8</td>
<td>[a] TB front, palatal.</td>
</tr>
</tbody>
</table>

Notes

(a) The TB transition is S (slow) here because the tongue body is moving from back to front.

(b) If the phonetic data bases were more detailed as regards context effects, velum up would probably mean different AV values for the two elements of the diphthong, V1 and V2.

Next, these same timing instructions are expressed as event labels, as shown in Fig. 14. The specific end point values chosen replace their labels, except for the TB instructions.
The coordination expressions and $D$ values chosen follow:

\begin{align*}
\text{ET0} & := 0 \\
\text{#1.EPLUX} & := \text{ET0} + 24 \\
\text{V1.EAGX} & := \text{ET0} + 24 \\
\text{V1.EQ} & := \text{V1.EON} + \text{V1.D1} \\
\text{V1.D1} & := 16 \\
\text{V1.EOFF} & := \text{V1.EON} + \text{V1.D4} \\
\text{V1.D4} & := 10 \\
\text{V2.EON} & := \text{V1.EOFF} + \text{V1V2.D} \\
\text{V1V2.D} & := 38 \\
\text{V2.EOFF} & := \text{V2.EON} + \text{V2.D4} \\
\text{V2.D4} & := 21 \\
\text{#2.EPLU} & := \text{V2.EOFF} \\
\text{ETMAX} & := \text{#2.EPLU} + 24
\end{align*}

Note

In this case V1.EAGX is an event of the vocal folds as larynx articulator, not the main articulator as in Example 1. Here V1.EAGX defines V1.EON because [#1] precedes.

IV.4. Example 3: “Tar” and “Star”, [tɔː] and [ʃtɔː], two words of English, each produced in isolation, with a falling pitch (Fig. 15)

Some of the coordination in the first part of each articulatory time plan is shown in Figs. 15(a) and 15(b). The derivation will not be described in detail. Attention will be focussed on the ways in which these two words differ from each other phonetically and articulatorily, given that they differ linguistically only in the extra initial phoneme for “Star”. In the sketches, some of the coordination has been simplified for visual clarity. The point at issue here is the coordination between the tongue tip-blade articulator and the vocal folds for [t] and [st] in word-initial position [...\text{\textit{t}V}... ] and [...\text{\textit{s}tV}... ].

Allophone strings:

“Tar” is [\texttt{#t'\textalphaː#}] = [#1 C1 "V1 #2]

“Star” is [#\texttt{s't'\textalphaː#}] = [#1 C1 C2 "V1 #2]

Allophones of [t] and inter-articulator coordination.

The first consonant in each case is non-nasal [−N]. Velum actions. AV. might be tied to the build up of lung air pressure PLU for [#1] by some speakers, or to the actions for the first consonant in the string, by others. AV actions are here shown tied to [C1] actions. The velum AV needs to move to the up position early enough to give a C1.D3 value within the allowed range for [t] in one case, that is about +6 time units to any negative value; and for [s] in the other, that is, a negative value of about half C1.D4, the obstruction duration, or more negative.

The data base for [t] expresses the well-known fact that in word-initial position followed by a vowel [t] is aspirated [t, h]; when [s] precedes in the same word however [t] is unaspirated [t, ə]. Coordination between AC and AG, expressed by the value of C1.D1 for “Tar” and C2.D1 for “Star”, is crucial for the phonetic feature contrast aspirated/unaspirated. For “Tar” C1.D1 must be near zero. This kind of coordination, with vocal fold adduction beginning 2 time units (10 ms) after the [t] release, is shown in Fig. 15(a). The tongue body is shown moving away from its [t]
state 5 time units (25 ms) before the release.

For “Star” C2.D1 needs to have a negative value of about 24. the duration of the vocal folds adduction gesture, AGTR. so that AG is at or nearly at its phonation state by the time of the [t] release event C2.EAC. This can be made compatible with the requirement for [s] in “Star” that the AG transition from apart to phonation should begin approximately half way through the [s] obstruction. As Fig. 15(b) shows, from mid obstruction for [s], half way between C1.EACX and C1.EAC, to the [t] release, C2.EAC. the tongue tip-blade articulator AF has to perform a Fast (8 time units) closing followed by a stop posture for the [t] closure. It may maintain this stop configuration for a few time units. so as to ensure that a sufficiently long stop closure is perceived by the listener. The particular durations illustrated in Fig. 15(b) are (in 5 ms time units):

AGTR = 24 [s] obstruction C1.D4 = 20 [t, o] closure C2.D4 = 8 (a bit above the probable minimum value needed to satisfy perceptual constraints).

Very probably a shorter C2.D4 would be acceptable in many speech styles. especially since frication noise will perhaps weaken significantly a little before complete closure has been achieved.

C1.D1 = -10 = -C1.D4/2
C2.D1 = -26
C2.D2 = -5

This D2 value is within the range given in the data base for [t] and other consonants for tongue body release timing relative to the release of the main articulator. This internal coordination for [t] has been maintained constant across the change of phonetic context. C1.D2 is not needed since the TB (tongue body, jaw and lips) state is the same for [t] as for [s]. although more precision is required for [s]. With C2.D1 as shown here at -26, the vocal folds have reached their adducted state just before the release EAC. Preliminary
syntheses have suggested that a more auditorily acceptable effect is produced when the vocal folds reach their phonation state a little later than this.

General experience with the modelling gives the impression that for a naturalistic effect, articulatory actions should be tightly meshed together with much overlap (or coarticulation). The scheme shown in Fig. 15(b) may be too spaced out (insufficiently coarticulated) for all but a very formal style of speech. For a series of simulations representing multiple tokens of natural speech, durations C1.D1, C1.D4, C2.D1 and C2.D4 should all be perturbed by about +/- 6 time units (+/- 30 ms) about their 'optimal' values. In the resultant acoustic output, segment durations for [s] contoid frication and [t] contoid acoustic closure are likely to covary. Perceptual experiments (Carlson and Granström, 1975) showing listeners' sensitivity to the total acoustic duration for some consonant clusters of Swedish, but not to compensatory changes within the whole, may perhaps indicate that listeners make use of their knowledge of the acoustic implications of articulatory perturbations of the kind discussed here.

For an English word-initial consonant cluster [\textit{\textsc{st}}...\textit{]} there is no phonological contrast /t/ versus /d/. The speaker's freedom to overlap the actions for [s] and [t] as shown here, indeed the obligation to produce a sound that is phonetically closer to English word-initial /d/ than to [t, h], is, of course, a reflection of the lack of phonological contrast. If speakers tend to speak faster and mesh actions together more with increasing perceptual-motor and linguistic maturity, then any required contrast /st/ versus /sd/ for word onset types might tend to get lost. A shift towards a later C2.EAG timing with respect to C2.EAC would give [s t, h] and thus imply a word boundary between the [s] and the [t].

In implementing a rule-based mapping from a sequence of linguistic, ordered units onto specific articulatory schemes, different speakers might well operate with different rules, maintaining different durations invariant across a change of phonetic context. Thus, one kind of speaker might keep C1.D4, the obstruction duration for [s], constant, regardless of whether [t] follows in the word or not; another speaker type might shorten the [s] obstruction when [t] follows. Similarly, the [t] closure duration C1.D4 or C2.D4 might remain unaltered in these two word onsets, or might instead be shortened when [s] precedes in the same word. Considerations such as these may be a matter of speech style also; perhaps the two factors of speaker type and style interact.

IV.5. Example 4: \textit{[Vn'V]} sequences with articulatory reorganisation for different speech styles (Fig. 16)

Two out of many possible versions of a \textit{[Vn'V]} sequence are shown here, in Figs. 16(a) and 16(b). The allophone string for version (a) is \textit{[#23:n'3:#]} = \textit{[#1 C1 V1 C2 V2 #2]}

This version has the full complement of articulatory oppositions for the stressed vowel, involving AG and PLU as well as Q. Probably each of the relevant articulatory events (V2.EAGX, V2.EAG, V2.EQX, V2.EQ, V2.EPLUX and V2.EPLU) should be located somewhere near the vowel onset event V2.EON. Timing perturbations from zero D values are shown in the sketch. The perturbations differ, since they are assumed to be controlled by independent neuromuscular processes.

An alternative vowel could be \textit{[i:] for either [V1] or [V2] or both without changes to the timing plan, since both are front vowels with medium (M) tongue body transitions in these [\textit{-(C)}] and [(\textit{C}--) contexts. The sequences could form portions of real English speech, for example:

\textit{[3:n'3:]} "... her nerve ..." \textit{[i:n'i:]} "... we need ..."
\textit{[3:n'i:]} "... her knee ..." \textit{[i:n'3:]} "... we nurse ...".

However, in this simulation the sequence is treated as if produced in isolation, beginning and ending at a respiratory state. In this version there are many static segments between the transitions. These may well be unreasonably long for most speech styles. But from this time plan, many variations can be produced, simply by reducing one or more of the D4 values.

Although acoustic vocoid duration is not to be equated with V.D4 duration, the time from articulatory onset to offset for each vowel is a rough guide to the durational balance between the two

Speech Communication
Fig. 16. Articulatory time plans for Example 4: reorganisation for different speech styles. \[Vn^7V\] sequences produced with (a) formal, and (b) informal style. The transitions shown as double lines in version (a) are optional and are omitted in version (b). Inter-articulator coordinates for which \(D\) remains invariant across the change in style are labelled with an asterisk \(*\). In version (a), the offset of \([V1]\) is defined by the start of the AF transition into its stop state for \([n]\). This is at time point (C2.EACX-AFTR).

In version (b), alternative tongue body paths are shown: the dashed line gives a nasalised \([\tilde{s}]\) for \([V1]\); the solid line path, which begins at the \([n]\) state, results in a nasalised \([s]\) vowel for \([V1]\).
vowels. The sketch in Fig. 16(a) is intended only as an aid to visualisation of the effects of particular selections from the data base files, rules and options. But if the relative values of V1.D4 and V2.D4 are approximately as shown here, then the second vocoid is likely to be longer than the first. In addition, as well as carrying a fall in fundamental frequency $F_n$, it will be of higher intensity with more high frequency emphasis, because of the pressed state of the vocal folds and the raised lung air pressure associated with this vowel, so that the 'rhythmic' and prominence balance between the two vowels will probably be appropriate.

The second version, shown in Fig. 16(b) is:

\[ \#3n'3:# \] = \[ #1 V1 C1 "V2 #2 \]

The changes made in going from the formal version (a) to one kind of informal version (b) are not ad-hoc shortenings; instead, version (b) represents a different lawful choice of options and selections available in the data bases. Two optional articulatory contrasts relevant for stressed vowels are dispensed with: pressed versus phonation for AG and raised versus normal for PLU. The Q transition from high to low must be retained, however, if the second vowel is to carry a falling $F_n$, as is appropriate for its speech-final position. V2.D4 is kept long enough for the Q fall to be completed before the utterance comes to an end.

Two static segments of version (a), the AF closure phase for [n], and the accompanying static phase for the tongue body, are reduced in version (b) to the minimum within which jaw raising and lowering can be accomplished. The time required for a jaw raising transition (AJTR) now becomes a duration-determining factor (DDF) for the interval from VI offset to C1 offset. It is assumed that jaw raising is necessary in order to carry the tongue tip-blade up in going from [a] to [n], even though the body of the tongue need not move. For some speaker types or some speech styles this jaw movement might be unnecessary. In that case the DDFs are the raising and lowering of the tip-blade, which are likely to be quicker actions.

Besides the reductions in D4 durations for the second vowel and the consonants, it may be seen that in version (b) there is a much closer meshing together of the articulatory actions near the beginning of the utterance. This is achieved by several different means. The 'soft' vowel onset, without [2], is selected and an AG event V1.EAGX defines vowel onset V1.EON. A different articulator AF is responsible for the vowel offset V1.EOFF and, in the example shown, V1 offset comes before V1 onset. The first vowel can be very much shortened, but not to the point where it disappears, audioritely, except perhaps for an extremely rapid informal style of speech. This vowel shortening process would run into a limitation however, due to the time required for a tongue body transition from its [3:] or [i:] vowel state to its [n] state. That TB transition is shown as a dashed line in Fig. 16(b). This particular constraint can be removed if the vowel quality criterion for [V1] is relaxed. If [a] a schwa-like vowel is acceptable for [V1] then no TB transition is required here and a DDF disappears. Time has been saved also by the decision to let the first vowel be [0N] instead of [–N], as it is in the full version (a). In this way two time-consuming transitions of AV, from down to up for [V1] and back to down again for [n], are avoided. As far as the phonological system of English is concerned, omitted AV transitions and retained TB transitions can be combined. However, this results in [V1] becoming a fully nasalised [i:] or [3:], which may well be unacceptable to listeners on other than phonological grounds. It seems plausible that requirements of non-nasalisation may be less stringent in the case of reduced [a]-type vowels; in Fig. 16(b) the solid line transition for TB is probably more acceptable as normal adult English speech than the alternative dashed line transition. The former can be exemplified as English sequences:

\[ an's:] "...a nerve..." \] and \[ an'i:] "...a knee...".

Two principles have been applied in the shortening processes. First, the model's articulatory transitions have been assumed to include invariant 'icebergs' and the durational requirements and extents of these transitions have been preserved. Second, inter-articulator coordination internal to each allophone unit has been maintained invariant. These coordinations are indicated by asterisks in Figs. 16(a) and 16(b). When real speakers of English speak faster, both vowels and
consonants may become acoustically shorter, though not necessarily in equal proportion (Bridger, 1982). These two time plan sketches suggest the possibility at least of considerable complexity in the reorganisation of the underlying articulatory processes responsible for changes of rate or style in the output signal.

**IV.6. Example 5: Matches to [vowel-fricative-vowel] sequences produced by a real speaker**

Here, instead of considering a range of theoretical possibilities for productions of [VCV] sequences, matches to particular speakers are intended; the model is used for analysis by synthesis. The aim is to characterise the options selected by different speakers in their productions of a class of sounds, for example voiced fricatives.

A flexible articulatory time plan is drawn up, for example for [p3:z3:p]. Repetitions of [p3:z3:p3:z3:]... on one expiratory breath are produced by the speaker, thus allowing subglottal pressure to be estimated in the real speech (see, for example, Rothenberg, 1981). In the articulatory synthesis end point values, transition types and D values can be varied. In the real speech, constriction area versus time paths for consonantal portions can be extracted from aerodynamic traces by using the orifice equation. After matching those paths in the modelling, aerodynamic comparisons between natural and synthetic speech allow rapid improvement in articulatory matching; then acoustic patterns can be compared and further modifications made to the modelling, including the voice source parameters. The modelling itself can assist the estimation of laryngeal articulations. The procedures have been described elsewhere (Scully, 1984a, 1986; Scully and Clark, 1986). By simulating other voiced fricatives produced by the same speaker in the same phonetic environment and by extending the utterances to include other vowel contexts also, it is hoped that specification of characteristic articulatory schemes for the synthesis of this particular speaker will be possible.

**IV.7. Example 6: Timing variations for “plight” and “polite”**

Very short schwa-type vowels have phonological significance in English. This can be illustrated by some minimal pairs. Here the words “plight” and “polite” have been produced synthetically by going through all the stages of the model, as shown in Fig. 1. The output analogue signal from the model was low-pass filtered by a 7th-order elliptical filter giving a reduction of 60 dB between 5.90 and 5.95 kHz. Spectrograms of the model’s outputs are shown in Figs. 17(a)–17(f). An adult male speaker was simulated. The first two sounds (Figs. 17(a) and 17(b)) are heard as acceptable versions of “plight” and “polite”. The different coordination required in the two cases, between AL defining the lip release for [p] and AF providing the tongue tip-alveolar ridge contact for [l], affects aerodynamic conditions at AL and AF and hence acoustic structure for both [p] and [l]. In [pl’ait] the [p] is affricated and the [l] is partially devoiced, effects found in natural speech. This is an example of the ability of articulatory synthesis to generate automatically covariations in acoustic structure with phonetic context. Many different versions of “plight” and “polite” could be produced by choosing slightly different coordination schemes; in each case intra-speaker variability could be generated by perturbing the timings of the actions by up to +/- 30 ms.

Applying larger changes of coordination, “plight” can be turned into another English word “blight” simply by advancing in time the initial adduction of the vocal folds. The effect is seen in Fig. 17(c): the [l] is now fully voiced and there is very little turbulence noise at the stop release. In the original version the amount of preaspiration for the [t], to be seen on Figs. 17(a), (b) and (c), is reasonably acceptable to English listeners. The result of abducting the vocal folds earlier for the [t] of “polite” is shown in Fig. 17(d). Now the end part of the [ai] diphthong has changed into a voiceless palatal fricative [ç]. Enlargement of the glottal area has raised the air pressure behind the constriction of the vowel-like vocal tract; voicing has ceased because of a combination of unsuitable larynx geometry and aerodynamic conditions; turbulence noise has been generated at the glottis.
Fig. 17. Example 6: Articulatory synthesis. Spectrograms (700 spectrograph, Voice Identification Inc. analyzing filter bandwidth 300 Hz) of outputs from the composite model of speech production processes, simulating an adult male speaker. Articulatory coordination is varied and the output signal is judged auditorily to be as follows: (a) "Plight"; (b) "Polite"; (c) "Blight". derived from version (a) but with earlier adduction of the vocal folds relative to the initial plosive release event there; (d) "Polite" but with an intrusive, non-English sounding palatal fricative [pʰi'æt], derived from version (b) but with earlier abduction of the vocal folds for [t]; (e) "Polite", less 'coarticulated'; (f) "Polite", more 'coarticulated'. For each spectrogram, the time marker shows 100 ms and the frequency calibration lines are at 1 kHz intervals.
Fig. 17. (c) and (d).
Fig. 17. (e) and (f).
and at the vocal tract constriction; non-English [psl"aikt] is heard. This illustrates the principle that the mapping from allophones to articulation or from articulation to perceived phonetic string is not a series of one allophone-one action connections.

For Figs. 17(e) and 17(f) some of the actions were meshed less and more tightly respectively. The 'less coarticulated' version of "polite" is excessively long: the 'more coarticulated' version sounds only slightly hurried, supporting the view that for adult speech the component actions are executed in the main rather quickly, with little in the way of static postures. Shorter versions than this could be modelled. The [a] vocoid could be reduced to only one or two cycles of voicing by means of a negative D4 value for [a], as in the [an'z:] example in Section IV.5: probably most of the static raised tongue tip AF occlusion for [l] could be removed without adverse perceptual effects; the [t] actions could be brought forward, as long as tongue tip-blade AF and vocal folds AG remained well coordinated; D4 for the [t] closure could be reduced.

IV.8. Simulations and natural speech

The output from each of the articulatory time plans discussed in these examples defines all the inputs needed to run the model of speech production processes. In most cases some acoustic outputs have been obtained. Synthetic versions need to be compared with natural speech in which the style or tempo has been varied.

Vowel qualities can be changed simply by altering their end point states. The timing plan will be the same so long as the tongue body transition durations remain constant. If a back vowel were to be substituted in Examples 4 and 5 then it would probably be appropriate to change the tongue body transition time from M (medium) to S (slow). A lengthened transition and other durational changes resulting from it can easily be accommodated within an already established articulatory plan.

IV.9. Limitations and future work

The framework proposed here, for mapping from a phonetic allophonic description of British English onto speech production schemes, is at present seriously lacking in quantitative articulatory data from a number of speakers, even for one accent. Unfortunately, much of this information, on tongue movements and larynx actions for example, is very difficult and time consuming to obtain. But some important parameters, such as stop closure duration as a function of linguistic context, or pitch control patterns as a function of vowel and word types, could be estimated from acoustic studies. Another source of optimism, when considering whether a framework such as this one might be useful for synthetic voices, is that rather few iterations at the articulatory and aerodynamic stages of the modelling seem to be capable of yielding greatly improved matches to the outputs from real speakers. It may be anticipated that a similar iterative approach to the acoustic stages of the model will yield further improvements in acoustic matching.

The articulatory time plans set out here could perhaps be more economically described if the events were taken to be located at the mid-points of articulatory transitions, instead of at one or other of their ends. Besides offering the practical advantage of reducing the number of event labels needed, such an approach might well reflect more closely the organisation used in natural speech. It would be interesting to try to determine whether, for some subset of English utterances, different speakers having the same accent could each be characterised by his or her articulatory transition types and coordination schemes. This might prove to be an economical way of predicting each speaker's acoustic pattern features, especially their variations with phonetic context and covariations with each other. The model might then perhaps be a useful tool, in conjunction with acoustic analyses of natural speech, for future speech recognition devices, which will need to handle the great variety of acoustic patterns arising from different speakers and even from a single speaker.
Acknowledgments

This work was supported by the Science and Engineering Research Council, project number GR/C/30252.

It is a pleasure to acknowledge the major contribution made by Ted Allwood to the development of the model of speech production processes and the time scheme framework, and his implementation of them, during this and two previous SERC-supported projects.

The modelling used the VAX 11/780 system, part of the Interactive Computing Facility, in the Department of Mechanical Engineering, University of Leeds.

Thanks to Eric Brearley for assistance with experiments and analyses, to Desmond Ohara, Helen Gray and Cynthia Foster for their work on the figures, to John Scully for help with the proof reading, and to the Reviewer who proposed the Diagram in Section IV.2.

I am grateful also to the many colleagues who have given us advice and help with the modelling and from whose research into speech production we have benefited.

References

J.S. Bridle, personal communication.
O. Fujimura and W. Spencer (1983). "Effects of phrasing and word emphasis on transitional movements – location and...

C.L. Ludlow (1985), personal communication.

Vol. 6, No. 2, June 1987
C. Scully / Units of speech production


GLOTTAL LOSSES IN A COMPOSITE MODEL OF SPEECH PRODUCTION

C. SCULLY and E. ALLWOOD

DEPARTMENT OF LINGUISTICS AND PHONETICS
UNIVERSITY OF LEEDS

LOSSES IN SPEECH PRODUCTION

One way in which the acoustic block of an articulatory synthesiser can be improved is by better modelling of losses. Energy is lost from the vocal tract by the following processes:

- Radiation of sound from the mouth and nostrils, also radiation from the walls;
- Vibration of the yielding walls of the head, neck and chest;
- Heat conduction through the walls;
- Viscous losses at the walls;
- Turbulent as well as viscous losses in the glottis; also in severe constrictions of the vocal tract;
- Losses in the subglottal airways.

When the vocal tract acoustic system is modelled as one-dimensional waves on electrical transmission lines, viscous losses are represented as series resistances, heat conduction as shunt resistances and wall vibration and radiation as shunt impedances. Mouth radiation impedance is represented by appropriate termination of the line. Glottal losses can be related to a vocal fold model (1). The method of reflection coefficients (2), used by us, is faster to compute but assumes that pressure and volume velocity remain in phase. It does not model the formant frequency shifts and losses due to radiation impedance and losses within each vocal tract section; however, methods of doing so have been proposed (3). Many of the loss parameters are frequency dependent, but a single value for each must be selected in time-domain simulations. This limitation applies whichever method is employed in an articulatory synthesiser.

Our modelling has demonstrated that some, at least, of these losses must be simulated in order to achieve formant bandwidths similar to those of real speech. Viscous and heat losses may reasonably be neglected in a simplified acoustic model. They are apparently small compared with other types of losses, being much less than glottal and wall losses at low frequencies and much less than glottal and radiation losses at high frequencies (1). Radiation loss is proportional to frequency squared up to about 3 kHz for vowels other than those with a very large lip outlet area. Increased radiation loss should lower formant frequencies. The main effect of wall losses is seen at low frequencies: increased loss should raise first formant frequencies; the effect is expected to be noticeable in [æ]-type vowels especially. Increased glottal losses should raise formant frequencies. Increased losses during large glottis-occluded vocal tract portions were found to be essential; otherwise, the pressure wave trapped in the vocal tract took too long to decay during a simulated [a]. It seemed worthwhile to attempt a fairly sophisticated representation of glottal losses.

LARYNGEAL PROCESSES

The larynx is an important element of the articulatory, aerodynamic and acoustic blocks of the model (4). Consistency across these stages of speech
Proceedings of The Institute of Acoustics

GLOTTAL LOSSES IN A COMPOSITE MODEL OF SPEECH PRODUCTION

Production is sought. Within each block the aim is to simulate different speaker types while conforming to the physical constraints of the system. The larynx is assumed to contain two independent articulatory actions: abduction and adduction of the vocal folds for the control of glottal area $A_g$; tension of the vocal folds as a factor, called $q$, in fundamental frequency control. Aerodynamic conditions at the glottis are determined by the changing state of the whole respiratory system. Subglottal air pressure $P_{sg}$, oral air pressure $P_c$ and volume flow of airflow through the glottis $V_g$ are obtained. Here, a variable $V_x$ is the slowly changing 'd-c' component of the total, while $V_x'$ is the rapidly changing acoustic 'a-c' component and $V_x$ is the total time varying function. $U_g'$ is the voice source in the acoustic block. It is the short circuit volume velocity, not the true volume velocity dependent upon the vocal tract filter function. $U_g'$ is derived from a functional model of the larynx. Its waveform parameters are those of Fant (5). Parameter values depend on three controlling variables: transglottal pressure drop $(P_{sg} - P_c)$, glottal area $A_g$ and vocal fold tension $q$.

It is predicted that the glottal contribution to formant bandwidths is approximated by

$$BW_{1}(gl) = \frac{c^2 A_g^2}{2.11 k_1 U_g V_c}$$

assuming a Helmholtz resonator, or twice this, for all formants, if the vocal tract is modelled as a tube (6). $V_c$ is the volume of the cavity above the glottis and $k_1 = 0.875$ is an empirical constant. $c$ is the velocity of sound in the vocal tract. The expression for glottal differential signal resistance assumes that turbulent losses dominate, although viscous losses are included also in the aerodynamic block of the model.

In the reflected pressure wave model of vocal tract acoustics, glottal losses are included by making the reflection coefficient $R_{cg}$, at the glottal end of the first section, less than 1. Similarly the reflection coefficient $R_{cm}$ at the mouth termination is made greater than -1 ($|R_{cm}| < 1$) for simulation of radiation losses. Increased glottal area $A_g$ results in increased losses at the glottis since $R_{cg}$ is a function of $A_g$, viz:-

$$R_{cg} = \frac{a}{(R_{cg})^n}$$

$a$ and $R$ are constants chosen such that at $A_g = 0.05$ ('normal' phonation value), $R_{cg} = R_{cg1}$ and at $A_g = 0.15$ ('breathy' voicing), $R_{cg} = R_{cg2}$. $R_{cg1}$ and $R_{cg2}$ are parameters whose values may be changed, but which are generally 0.8 and 0.5 respectively. An upper limit for $R_{cg}$ is specified, usually 0.8, as a representation of glottal losses during a voicing cycle when the vocal folds are tightly adducted.

A STUDY OF BANDWIDTHS USING THE MODEL

To investigate the influence of cavity volume $V_c$ upon losses in the model two vocal tract shapes, as shown in Figure 1, were compared. With the aim of ensuring that equal fractions of the pressure wave were reflected at the exit of the larynx tube in both cases, the ratio of larynx tube cross-section area to maximum cross-section area in the cavity was made 0.25 in each case. The area ratios were not equal at the forward outlet of the pharynx cavity, since tongue constriction area was 0.25 in both cases. Formant frequencies and auditory quality were appropriate for [k]. $F_2$ was too low in frequency for [i] and the auditory quality was not good. A suitable voice source was used for judging vowel quality.
Proceedings of The Institute of Acoustics

GLOTTAL LOSSES IN A COMPOSITE MODEL OF SPEECH PRODUCTION

Figure 1 Two contrasting vocal tract area functions

A sin(x)/x source, flat to within +/-0.5 dB up to 5 kHz, was used; the radiation block of the model was omitted. Thus the vocal tract transfer function was studied. Two values for R\textsubscript{G}, 0.8 and 0.4, were combined with two values for R\textsubscript{M}, -0.8 and -0.4. Peak frequencies and bandwidths, defined as 3 dB down from the peak, were computed from an FFT program with a Hanning window. Results, interpreting spectral peaks as specific poles of a vowel-like vocal tract, are given in Table 1.

Table 1 Frequencies and bandwidths for poles of the vocal tract transfer function as vocal tract shape, glottal losses and mouth losses are changed. Frequencies in Hz to the nearest 5 Hz.

<table>
<thead>
<tr>
<th>RC\textsubscript{G}</th>
<th>RC\textsubscript{M}</th>
<th>(glottal loss)</th>
<th>(mouth loss)</th>
<th>F\textsubscript{1}</th>
<th>F\textsubscript{2}</th>
<th>F\textsubscript{3}</th>
<th>F\textsubscript{4}</th>
<th>(BW\textsubscript{1})</th>
<th>(BW\textsubscript{2})</th>
<th>(BW\textsubscript{3})</th>
<th>(BW\textsubscript{4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8 (low)</td>
<td>-0.8 (low)</td>
<td>720 (55)</td>
<td>1195 (80)</td>
<td>3300 (105)</td>
<td>4395 (80)</td>
<td>220 (20)</td>
<td>1965 (40)</td>
<td>3780 (100)</td>
<td>4720 (135)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8 (high)</td>
<td>-0.8 (high)</td>
<td>720 (315)</td>
<td>1195 (100)</td>
<td>3300 (110)</td>
<td>4440 (145)</td>
<td>220 (45)</td>
<td>1990 (110)</td>
<td>3905 (285)*</td>
<td>4755 (180)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4 (low)</td>
<td>-0.4</td>
<td>745 (100)</td>
<td>1100 to 1200*</td>
<td>3350 (very wide)*</td>
<td>4395*</td>
<td>220 (32)</td>
<td>1965 (40)</td>
<td>3780 (100)</td>
<td>4860 (290)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4 (high)</td>
<td>-0.4</td>
<td>745 (385)</td>
<td>1100 to 1200*</td>
<td>3335 (very wide)*</td>
<td>4440*</td>
<td>220 (55)</td>
<td>1990 (115)</td>
<td>3895 (365)</td>
<td>4895 (315)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimated accuracy of frequency measurement was +/-6 Hz approximately.

Bandwidth measures seem more likely to be susceptible to error, since absence of a sample at the true peak could significantly alter the 3 dB down points.

In most cases increased losses at the glottis or mouth resulted in increased bandwidths. BW2 and BW3 for [i] were almost independent of mouth losses while BW3 for [a] was almost independent of glottal losses. Under comparable loss conditions, BW1 was always greater for the [a]-type vocal tract than for the [i] shape, as expected. BW2 and BW3 were greater for [a] in some cases and greater for [i] in others. BW4 was always greater for [i] than for [a].

Glottal and other contributions to formant bandwidths have been assessed
separately using time domain (1) or frequency domain (7) transmission line models. Quantitative comparisons are difficult to make, partly because vocal tract lengths and shapes differ. BW1 for [i] and [a] in Table 1 are both slightly higher than corresponding contributions of glottal loss alone shown by Flanagan et al., Figure 5 (1). BW1, BW2 and BW3 for [i] in Table 1 are lower than values given by Wakita and Fant, Table II-A-IV, where the subglottal system is included (7). Data from real speech are included in each of these studies: by Dunn (8) and by Fujimura and Lindqvist (7). Table 2 summarises the trends in formant bandwidths.

The range in these - and other - studies is so wide that precise guidelines for the modelling are not apparent. It seems that the modelled bandwidths are of the right order of magnitude. Losses introduced near to the middle of the vocal tract are expected to have less influence on bandwidths (7). Table 3 confirms the small effect of increasing the losses through a small opening of the velopharyngeal port.

The range in these - and other - studies is so wide that precise guidelines for the modelling are not apparent. It seems that the modelled bandwidths are of the right order of magnitude. Losses introduced near to the middle of the vocal tract are expected to have less influence on bandwidths (7). Table 3 confirms the small effect of increasing the losses through a small opening of the velopharyngeal port.

### Table 2
Bandwidths of spectral peaks in the model compared with real speech data. Formant frequencies refer to the model. Frequencies are in Hz.

<table>
<thead>
<tr>
<th>Study</th>
<th>Approximate Formant Frequency and Vowel Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>RCG = 0.8</td>
<td>20-32</td>
</tr>
<tr>
<td>RCG = 0.4</td>
<td>45-55</td>
</tr>
<tr>
<td>Dunn (8)</td>
<td>38</td>
</tr>
<tr>
<td>Table II</td>
<td></td>
</tr>
<tr>
<td>Fujimura (1) and Lindqvist:</td>
<td>-100</td>
</tr>
<tr>
<td>Mean for females</td>
<td></td>
</tr>
<tr>
<td>Mean for males</td>
<td>70</td>
</tr>
<tr>
<td>Flanagan et al (1)</td>
<td>65</td>
</tr>
<tr>
<td>Total, Figure 4</td>
<td></td>
</tr>
<tr>
<td>Wakita and Fant (7)</td>
<td>49</td>
</tr>
<tr>
<td>Table II-A-IV</td>
<td></td>
</tr>
<tr>
<td>Closed glottis</td>
<td>146</td>
</tr>
<tr>
<td>Large glottis</td>
<td>Ag = 0.16 cm</td>
</tr>
</tbody>
</table>

Table 3 The effect on bandwidths of doubling the velopharyngeal port area Av. RCG = 0.8, RCM = -0.9. Frequencies to the nearest 5 Hz.

<table>
<thead>
<tr>
<th>Av cm</th>
<th>P1 (BW1)</th>
<th>P2 (BW2)</th>
<th>P3 (BW3)</th>
<th>P4 (BW4)</th>
<th>P1 (BW1)</th>
<th>P2 (BW2)</th>
<th>P3 (BW3)</th>
<th>P4 (BW4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>(60)</td>
<td>(60)</td>
<td>(60)</td>
<td>(60)</td>
<td>230</td>
<td>1975</td>
<td>3780</td>
<td>4710</td>
</tr>
<tr>
<td>0.1</td>
<td>(60)</td>
<td>(60)</td>
<td>(60)</td>
<td>(60)</td>
<td>230</td>
<td>1975</td>
<td>3780</td>
<td>4710</td>
</tr>
</tbody>
</table>
GLOTTAL LOSSES IN A COMPOSITE MODEL OF SPEECH PRODUCTION

CYCLICAL VARIATIONS IN GLOTTAL DAMPING

Variations in glottal damping during the voice cycle are important: oscillations need to decay before the next time of excitation of the vocal tract resonances. The pattern of decay may be a characteristic of different speaker types. As a first step to making the cyclical variations in glottal losses consistent with the cyclical variations in conditions at the glottis, an a-c component of glottal area, $A'g$ was obtained from the voice acoustic source $Ug'$ as follows:

$$A'g = \frac{\bar{\bar{A}_g}}{Ug'} \cdot Ug'$$

This mapping does not take into account the difference between the shapes of the glottal area and glottal airflow curves, specifically the greater asymmetry in $Ug'$ due to inertia in and above the glottis. $A_g$ and $A'g$ were combined in three ways to give a total glottal area function $A_g$:

(1) $A_g = A'g + \bar{A}_g$

(2) $A_g = w \cdot A'g + (1-w) \frac{\bar{\bar{A}_g}}{Ug'}$

where $w = VOIA_{\max}/VOIA$ and $VOIA$ is the envelope of $Ug'$.  

(3) $A_g = \frac{\bar{\bar{A}_g}}{Ug'} \cdot (Ug' - VOIA/2) + \bar{A}_g$

Slight changes in some bandwidths were apparent on spectrograms: more detailed studies of single formant decay during the voice cycle remain to be made. The last two lines in Table 2 (7) suggest the kind of bandwidth magnitude swings which may be needed.

ACKNOWLEDGEMENTS

This research is supported by SERC grant GR/B/34874. The original version of the acoustic program was kindly supplied by colleagues at the Joint Speech Research Unit.

REFERENCES


APPENDIX 5

THE REPRESENTATION OF STORED PLANS FOR ARTICULATORY COORDINATION AND CONSTRAINTS IN A COMPOSITE MODEL OF SPEECH PRODUCTION *

C. SCULLY and E. ALLWOOD

Department of Linguistics and Phonetics, University of Leeds, Leeds LS2 9JT. U.K.

Received 28 March 1983

Introduction: the model

A composite model of speech production has been developed. The stages modelled are: articulation, aerodynamics, generation of acoustic sources voice, noise and pulse, filtering by the acoustic tube of the vocal tract and radiation loading. Inputs define, for each articulator, a sequence of target states, the types of transitions between successive targets and the time allowed for each transition. The output from the model is a pressure waveform recorded as synthetic speech, for which phonological judgements can be made and having some of the covariations in acoustic patterns of real speech (Allwood and Scully, 1982). Most of the computation required is in the simulation of the physical laws of aerodynamics and acoustics, described in highly simplified forms, applied to time-varying configurations of the respiratory tract. Traditional articulatory phonetic parametric descriptions are invoked, together with published X-ray and other data, in preliminary planning of the actions needed to attain a particular auditory goal, as implied by a word of English said in isolation. The auditory goal, as judged by native English listeners, has primacy. Speech is not viewed as a succession of distinct sounds, one for each phoneme, linked by transitions. The goal of speech production is seen as a succession of sound patterns, largely dominated by acoustic transitions, without intrusive, extraneous sounds. The requirements are tied to the phonological system of auditory contrasts for the particular dialect and language modelled, in our case, that of an R.P. (Received Pronunciation) accent of British English. Notions of coarticulation are rejected in favour of the concept of articulatory scheme options available to an individual speaker, notably the freedom to vary timing and coordination of some articulatory component gestures within the constraints imposed by the auditory goal.

The articulatory modelling

The first block of the model simulates voluntary movements of about 8 structures, the articulators. These comprise: the lung walls, vocal fold control of the d-c component of glottal area, the vocal fold effective stiffness and mass contribution to pitch control, called Q, the tongue body (described by about 4 points whose transitions are synchronised), the tip or tip-blade region of the tongue, the jaw, the lips and the soft palate. (Larynx raising and lowering, which constitutes a third laryngeal articulation, has not yet been modelled.) The set of articulatory parameters define the shapes controlled rather than the related movements of anatomical structures. The articulatory model is under-constrained. Where data from real speech are available these are incorporated as interactions between input descriptions for two or more articulators and as limitations on individual articulatory actions. The approach is similar to that of Ladefoged (1979). Fig. 1 shows the parameter points and the variables used to define the statics and kinematics of vocal tract shapes. The distances from the glottis of each parameter point, DE, DP, etc., can also be altered by means of similar transitions. Lung volume control is represented directly

* Research supported by the SERC, project no. SR/B/34874.
Parameter points to describe vocal tract shapes and transitions of cross-section area, AE, AP, etc. (called DVLU) or by air pressure in the lungs, PLU. Cross-section areas, AP, AL etc., for some or all of the vocal tract parameter points, are defined by the commands input to the articulatory block. Areas for the parameter points are obtained at each 5 ms sample time. Next, by linking between adjacent points, the whole vocal tract area function is derived for every sample. Transition times for each cross-section area, moving from one target value to another, are taken to be characteristic of each type of articulatory action and independent of the distance traversed in the transition. The evidence for this approximation was adduced by Scully (1975). Transitions are generally fast, medium or slow with times 40, 120 and 190 ms respectively. More precise matches to the kinematics of real speech can be made within the present articulatory framework, as data are published. The constraint of constant tongue volume is not modelled directly, but two factors are used for constraints on tongue body shapes, in keeping with published research (see Sawashima and Co-

Articulatory events and coordination

Coordination between articulators is of the essence in speech production. An individual speaker perhaps uses fixed patterns of inter-articulator timing, within his or her timing variability. In the modelling, particular articulatory events, such as the start of the tongue release at the F point for an alveolar plosive, are labelled. Other articulatory actions are stated in terms of these events. An advantage of modelling is that articulatory schemes can be perturbed in ways which would be very difficult to elicit from a normal adult speaker. Problems of too many degrees of freedom arise unless groups of parameter points are functionally linked as synergies or coordinative structures (Kelso and Saltzman, 1982). A large number of versions of 'purse' and 'purses' have been generated, by making the front of the tongue body AB, the jaw region AJ, the lip area AL and the con-
toid-forming constriction AF act synergistically with identical timing perturbations. A single charge of specification for glottal area AG can change listeners' phonological response, but many options with non-minimal articulatory contrasts (perhaps valuable for enhancing reliability of the auditory contrast) achieve the auditory targets also. It may perhaps prove possible to characterise different synthetic speaker types by specifying their transition times for each type of articulatory action and by the target values, timing and coordination options selected by them.

Table 1

<table>
<thead>
<tr>
<th>Rule number</th>
<th>Target area (cm²)</th>
<th>Time allowed (5 ms unit)</th>
<th>Transition type</th>
<th>Time units needed to reach the target</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP 0</td>
<td>1.0</td>
<td>7.5</td>
<td>60</td>
<td>S</td>
</tr>
<tr>
<td>AB 0</td>
<td>10.0</td>
<td>1.0</td>
<td>60</td>
<td>S</td>
</tr>
</tbody>
</table>

(The tongue remains on the second vowel target for 60 – 38 = 22 time units.)
Generalisation of procedures to synthesise words from stored information

The next task is to store, retrieve and link together successful articulatory schemes for the simulation of one speaker type. We wish to avoid the use of phonetic segment boundaries as the operational equivalents of phoneme to phoneme links in time. These do not constitute articulatory events, in our view; there is no need to pinpoint them in our articulatory modelling. Transition components will be stored for each phonetic element as well as target values and important events. Constraints of various kinds will be imposed. They are not based on concepts of dominance by higher-ranking phonetic elements, as in Holmes et al. (1964). Fig. 2, below, illustrates an articulatory scheme for the English name “Bert”.

Key to event labels (Duration numbers in 5 ms units, TB is tongue body.)

$E_1$: Moment when PLU reaches its high value, starting from zero at $T = 0$.

$E_2$: AL release for [b], 5 after $E_1$; AG reaches its vocoid value at $E_2$.


$E_4$: Q fall (slow transition) starts 5 after $E_2$.

$E_5$: Targets for long vowel, VI, reached by TB and AJ ($E_3 + 24$).

$E_6$: Start of TB movement for [t] after being on VI targets for $x$ (= 14).

$E_7$: Start of AF movement to [t] closure (4 after $E_6$).

$E_8$: AF reaches its [t] closure target (16 after $E_7$).

($E_9$): (Earliest) time for TB and AJ to start moving away from [t] positions (24 after $E_6$).

($E_{10}$): (Earliest) time for AF start of [t] release ($E_9 + 4$). $E_9$ and $E_{10}$ are delayed by 7 to make the relative durations for $x$ (for the contoid) and $y$ (for the vocoid) appropriate.

$T_{MAX}$: For a released plosive, the AF transition is to be completed by $T_{MAX}$ (16 after $E_{10}$). The PLU fall (medium transition) is timed so as to reach PLU = 0 at $T_{MAX}$.

From the events labelled $E_1$ to $E_{10}$, between $T = 0$ and $T_{MAX}$, if transition types are fully specified, commands to all articulatory components can be derived.

Articulatory constraints include the time needed to achieve essential transitions. A critical path, with different articulators functioning as the duration determining factor in different portions of the utterance, is mapped out. Phonological groupings of long versus short vowels will influence the time spent at target configurations. Auditory requirements, for English at least, include relative perceived lengths of successive vocoid and contoid segments. The prosodic feature of pitch, controlled mainly by Q, is planned in the same sort of way as segmental units. Particular options are represented here. A different speaker might perhaps make $E_3$ coincide with $E_2$ or $E_3$ might be very much earlier, so that the tongue body target for [a] was nearly reached by the time of the [p] release. Variations such as these, always constrained by the auditory goal, would constitute the equivalent of coarticulation in the synthetic speech generated. Problems of matching up incompatible values of articulatory components appear inevitable. It is hoped that unstressed syllables, specified as vocoids but without rigidly defined vocal tract shapes will be useful as interfaces.

References


1. Introduction

From different combinations of a few rather simple articulatory actions, a variety of quite complex aerodynamic conditions and acoustic outputs can be created. The most basic requirement of all for speech is the creation of voice; this is easily achieved by new-born babies. What is examined here is the building up of a repertoire of lung and larynx actions appropriate for controlled operation of the voice source. Even apparently simple speech sounds demand correct coordination. The auditory goal of the simulation described here was an [i] vowel quality with 'modal' as opposed to 'breathy' or 'pressed' ('laryngealised') phonation type and with falling pitch. The tasks of speech production are by no means clear, but one basic aim is to achieve a subglottal pressure suitable for the onset and maintenance of voice.

2. The model

A model of speech production processes implemented on a VAX 11/780 computer was used. The stages modelled are shown in Figure 1. Inputs to the model define speaker dimensions, initial conditions, larynx type for a functional model of voicing and articulatory transitions. Eight quasi-independent articulators are used, as controllers of the geometry rather than as anatomical structures. Most articulatory actions are represented by changes in cross-section area of a few constrictions of the vocal tract. Articulations of the lung walls are represented either by air pressure in the lungs $P_l$, or as in the study described here, by the rate of change of lung volume $DVLU$. Vocal fold articulations are represented by the slowly changing (d.c.) component of glottal area $A_g$ and by a variable called $Q$, for the effective stiffness and mass of the vocal folds. Vertical movements of the vocal folds are not modelled at present. The bases for the modelling have been described (Scully, 1975; Allwood and Scully, 1982).

Timing and coordination in the articulatory block determine aerodynamic conditions throughout the respiratory tract. Articulatory states and aerodynamic conditions combine to determine the magnitude of turbulence noise sources for aspiration and frication. A pulse source, derived from rate of pressure change in the oral cavity, has been introduced recently, but was not
used in this study. A parametric description of the voice source is used as shown in Figure 2 (Fant, 1980). A minimum $\Delta P$ of 2 cm H$_2$O was assumed for the onset and offset of voicing. Fundamental frequency $F_0$ was derived from $F_0 = Q + 4 \cdot \Delta P$. A voicing 'plateau' region was defined between $AG = 0.04$ cm$^2$ and $AG = 0.08$ cm$^2$. $F_0$ decreased for $Ag$ less than 0.04 cm$^2$. $K$ varied inversely with $Ag$. TCR was constant at 0.1. Aspiration and frication sources were weakened and modulated when voicing was present. In an alternative form of the voicing model the wave parameters VOIA, K and TCR can all be made to vary as linear functions of three controlling physiological variables: $Ag$, $\Delta P$ and $Q$. Using the model interdependence of vowel and consonant durations have been demonstrated for voiced and voiceless fricatives having constant supraglottal articulation and for open and close vowel contexts. The effects were similar to those of real speech and the model's outputs were intelligible and speech-like (Allwood and Scully, 1982).

3. Modelling of aerodynamic processes

The system in Figure 3. A set of first order differential equations expresses the assumptions made and the physical principles invoked in the model, which are as follows:

1. The compliance of the lung walls need not be included. It is assumed that the speaker takes the nett compliance (recoil) into account when adjusting muscle pressures at different lung volumes so as to give a pre-planned rate of lung volume decrement. Passive changes in rate of lung volume decrease are not modelled at present.
2. The walls of the subglottal airways are taken as rigid, with flow rates in speech well below limiting flow rate.

3. The supraglottal cavity has an active component of volume change due to articulatory actions, added to a passive component associated with wall compliance (Rothenberg, 1968).

4. All but 4% of the subglottal volume is located in the respiratory zone of small airways, with generations higher than 16. Subglottal flow resistance is almost totally confined, on the contrary, to the large tubes of generation less than 10. This striking separation of subglottal volume and flow resistance justifies a model with one lumped lung volume and a separate single flow resistance linking it to the glottal orifice. This contrasts with the more complex representation in the model of Rothenberg (1968).

5. Subglottal flow resistance is an 'ohmic' conductance which increases linearly with lung volume, up to a maximum value of about 2 L/cm H₂O.

6. Inertance of air and tissues may be neglected.

7. The air in the respiratory tract is assumed to be an ideal gas and to be compressible. Departures from atmospheric pressure are small. Isother-
mal conditions are assumed. The flow is taken as one-dimensional. There is continuity of mass flow for each of the two cavities.

8. For each of the two orifices (constrictions) there is conservation of energy at the inlet (the Bernoulli effect), but energy is lost in the turbulent mixing region at the outlet. This gives a turbulent, flow-dependent component of pressure drop. A laminar 'ohmic' component of pressure drop is added to this. The same empirical constants are used for both orifices.

(Space does not permit reference to the relevant respiratory literature).

Parameter values are chosen to define cavity wall compliance, subglottal properties and initial conditions, Lung volume $V_l$, lung and supraglottal air pressures $P_l$ and $P_c$ are integrated at each time step to obtain values for the next sample. Merson's method (NAG library, 1981) was used here. There were problems with numerical instabilities in the aerodynamic variables, especially when oral pressure $P_c$ was very low, in vowel-like segments. Other methods for the integration, including Gear's method for dealing with 'stiff' equations (NAG library, 1981), have recently given improved stability and much reduced computation time for the aerodynamics.

4. The modelling of lung and larynx coordination

Some articulatory plans yielded inappropriate pitches, voice qualities or vowel lengths. Two of a series of articulatory plans are shown in Figure 4

*Figure 4. Two contrasting coordinations for the lung walls (DVLU in cm$^3$/s) and the larynx (Ag in cm$^2$ and Q in Hz). Also shown: a computed aerodynamic variable $P_{sg}$ in cm H$_2$O and the envelopes of acoustic sources voice and frication noise (FRIC) in arbitrary units. (a) --- (b) ——.*
Figure 5. Spectrograms for (a) and (b) in Figure 4 and for additional runs (c) and (d).
together with some of the aerodynamic and acoustic results. Unwanted sounds were generated in both cases. (a) was an attempt at 'braethy' attack. It was transcribed auditorily as [ɛʰiʔi] with falling pitch. (b) was an attempt at 'hard' (or 'glottalised') attack and was transcribed as [hiʔi] with 'gulp' effect, sudden onset and falling pitch. Spectrograms for (a) and (b) are shown in Figure 5. Two other unsuccessful attempts at the auditory goal are shown as (c) and (d) in Figure 5. (c) gave [ breath drawn in sharply ] then [i] falling pitch. (d) gave a 'strong' [i] sound with no audible noise, but not a falling pitch. In another set of syntheses for target words 'purse' and 'purrs', unwanted vowel-like segments were often generated at the speech offset. By trial and error, combinations of lung and larynx actions could be found which avoided unwanted onset and offsets. It is suggested that auditory feedback must be of overwhelming importance for the acquisition of speech, as in our modelling. The onset and offset of speech present speakers with specific problems. The options selected by a particular speaker for the achievement of rather broadly defined auditory goals will be reflected in the details of acoustic structure. Modelling of the kind outlined here may be able to assist in defining the probable acoustic variations within one accent, with potential applications in automatic recognition of speech.

Acknowledgement

This work is supported by the Science and Engineering Research Council, Grant Gr/B/34874.

References


PRODUCTION AND PERCEPTION OF AN ARTICULATORY CONTINUUM FOR FRICATIVES OF ENGLISH

C. SCULLY and E. ALLWOOD

Department of Linguistics and Phonetics, University of Leeds, U.K.

Received 28 September 1984

Abstract. Simple articulatory contrasts for a phonological opposition generate a multiplicity of acoustic cues. Knowledge of the covariations of acoustic pattern features within and across speakers is needed for automatic speech recognition. A model of speech production processes was used to generate stimuli along an articulatory continuum, the degree of abduction of the vocal folds for the fricatives in English words “hiss” and “his”. Transition times imposed strong constraints on the articulatory plans devised as inputs to the model. Acoustic segment durations output from the model covaried in the same way as those produced by 5 real speakers; there was good quantitative agreement in most cases. Listeners' responses suggest that the articulatory dimension synthesised is a suitable one for natural speech. Future directions for the modelling of multiple articulatory dimensions and for mapping from speaker-specific patterns of articulation and their perturbations onto the stability of particular acoustic cues are discussed.

Introduction

For the phonologically significant contrast to which the name “voiced” versus “voiceless” is generally given, bundles of contrasting acoustic attributes are found. For plosives an additional feature is needed to describe the four-way contrasts found in many languages: two-way contrasts of voicing during the acoustic closure segment can be combined with two-way contrasts of release phase, with or without...
aspiration. Fricatives seem to be of simpler structure than plosives, although here, also, presence versus absence of aspiration can be significant, in Korean for example [1]. In a number of languages many contrasting acoustic features have been found to accompany this "voiced"/"voiceless" contrast for fricatives, including vocoid and contoid segment durations, strength of frication noise and $F0$ patterns [2,3]. Many of these details of acoustic pattern contrast have been shown to be useful individually as cues for the perception of the phonological contrast [4, 5, 6, 7].

Questions of how these separately manipulated acoustic features interact as perceptual cues may be related to questions of naturalness of the synthetic stimuli employed in the perception experiments. Consideration of the source of the multiplicity of acoustic cues leads to consideration of the mapping from the articulatory onto the acoustic domain. Which acoustic pattern features does a speaker deliberately control in order to achieve a required phonological contrast? Are some aspects of acoustic structure merely concomitants of the main effects, allowed to occur for the sake of articulatory simplicity, rather than being deliberately sought?

Articulatory synthesis will not, of course, provide an immediate answer to these difficult questions, but a simulation which captures some of the features of the processes of natural speech may at least suggest which acoustic cues are likely to covary; and should be able to provide stimuli for perception experiments in which the acoustic cues do so in a natural way. The extent to which this covariation is similar to that found in real speech will be a reflection of the goodness of the model of speech production. If some sense of the range of possibilities in real speech is to be obtained, the model needs to be flexible instead of operating with a fixed set of synthesis rules. Given that every speaker of the same accent of a language must generate broadly similar sets of contrasting sound patterns, there is no reason to suppose that all speakers use identical patterns of articulation to achieve them. The difficulties encountered in speech recognition are witness to the acoustic variations across and within speakers. On anatomical and physiological grounds speakers must be expected to differ. A single speaker varies the style of speech to suit the communication context, while, for each style, there may be several options available. In addition, each person's speech contains apparently unavoidable perturbations of timing and articulatory placement.

Abramson et al. [8] used an articulatory synthesiser which modelled the acoustic filtering stage of speech production to investigate nasal/non-nasal perceptual boundaries for stops in different vowel contexts. The listeners did not appear to place the nasal/non-nasal boundary at a fixed size of velopharyngeal port opening. Shifts in the phoneme boundary were consistent with the shifts in the onset of significant acoustic nasality which would be predicted by acoustic theory. The important question of what constitutes a real articulatory dimension was considered. Scales of velopharyngeal port area increments, radius increments and fractional area increments with constant Weber ratio were considered. It was not possible to conclude that one scale was better than the others in terms of a statistical model, but the radius scale gave symmetry of response about the phoneme boundary and was preferred for that reason.

This paper describes a first attempt at using an articulatory continuum to establish perceptual boundaries for the voiced/voiceless distinction in fricatives of English. In a model, as in natural speech, covarying bundles of acoustic features arise automatically if the aerodynamic stage of speech production is taken into account when generating acoustic sources. An earlier form of our model, which included articulation, aerodynamics and acoustic sources but excluded the final acoustic filtering stage, was used to simulate the main larynx and vocal tract actions needed for intervocalic [s] and [z]. Identical actions by the tongue tip-blade articulator combined with two contrasting vocal fold actions gave durations for the fricative contoid segments and the adjacent vocoid segments which were in qualitative and even quantitative agreement with data from a real speaker [9].

The modelling

The present model of speech production processes includes articulation, described parametrically...
(see Fig. 1), and a simple aerodynamic model which is suitable when all the consonants in a single utterance have the same or closely similar places of articulation. Acoustic sources of voice, aspiration noise (just above the glottis), and friction noise (just in front of a vocal tract constriction) are derived by simple functional models. In the simulation described here, acoustic pressure for turbulence noise located at the outlet of a constriction was the product of a random number sequence and an amplitude envelope proportional to

$$A_0^0.5 \cdot \Delta P^{1.5}$$

where $A$ is the constriction cross-section area and $\Delta P$ is the pressure drop across it, based on Steven's analysis [10]. The acoustic source for voicing was a volume velocity waveform defined by 4 parameters, following Fant's description [11]. Fundamental frequency was assumed to be controlled partly by the effective stiffness and mass of the vocal folds, called $Q$, and partly by the pressure drop across the glottis, $\Delta P_g$, thus:

$$F_0 = Q + 4\cdot\Delta P_g,$$

The factor 4 is based on published data from natural speech; see, for example [12,13]. $Q$ is one of the articulators (see Fig. 2). The amplitude of the voice wave was derived from

$$\frac{\Delta P^{1.5}}{A G^2}.$$

$AG$ is the slowly changing articulatory component of glottal area and $\Delta P_g$ is the aerodynamic variable of slowly changing average pressure drop across the glottis. When the glottal area $AG$ was very small the denominator was replaced by a constant AGMIN, in order to prevent the amplitude from increasing as $AG$ tended to zero. In addition, wave amplitude and fundamental frequency were both scaled down progressively as glottal area decreased from AGMIN to zero. This provided quite a realistic modelling of glottalised voicing and glottal stops, with a few cycles of very low $F0$ sometimes appearing in the output. A plateau region of glottal area was thus defined, across which the voicing mechanism operated successfully. Voice wave amplitude declined at each side of the plateau, for very large or very small glottal areas. The asymmetry factor $K$ and the closed fraction (mark-space ratio) for the wave [11] varied inversely with glottal area $AG$. By this means the volume velocity slope at vocal fold closure was made to increase in magnitude as the vocal folds were adducted, giving increasing high frequency emphasis to the voice source as the voicing continuum was traversed from "breathy" voice to "pressed" voice. Acoustic losses at the glottal termination of the vocal tract acoustic tube were made to decrease also as glottal area decreased.

The onsets and offsets of noise sources were found to be sharply defined without reference to Reynold's number, but it was necessary to modulate these acoustic sources by the amplitude of the voice source (above a threshold value) in order to simulate the cyclical variations in the aerodynamic conditions which helped to generate them. Reduction in their acoustic amplitude was felt to be appropriate during voicing also, to make some allowance for time delays expected in the onset of turbulence when aerodynamic conditions alternate between favourable and unfavourable (F. Mobbs and N.H. Binh, personal communications).

The sum of the responses of the vocal tract acoustic tube to these three sources constitutes the output waveform from the model. After D-A conversion and anti-aliasing low-pass filtering to give a bandwidth of 5.9 kHz, the signal is recorded onto audio tape [14]. The values of
parameters such as vocal tract length, initial lung volume, compliance of the vocal tract walls and so on define a particular speaker type. It is perhaps worth emphasising that, once the choice of articulatory actions which form the input to the model has been made, the aerodynamic patterns and the complexity of the acoustic output structures, including overlapping of source types, varying strengths of voice and noise components, and acoustic segment durations, arise automatically, not by copying acoustic features of natural speech.

The articulatory plan

The aim was to create a minimal pair of English words, by means of a single articulatory contrast: the degree of abduction of the vocal folds for [s] or [z]. The words “hiss” and “his” were contained within the frame sentence “A it said”, thus placing [s] or [z] in a symmetrical [i-i] context. Appropriate durations for each articulator’s transitions were based on data from natural speech [15].

An articulatory plan was devised such that the necessary actions for one or other of the sentences were performed in the least possible time (see Fig. 2). While working on this plan it was found that there was very little room for manoeuvre in the time domain. Although the tongue body and lip shape transitions for the underlying vowel to vowel gestures are the slowest of all (requiring 190 ms), the time needed for changes in glottal area seems to be the chief constraint here. Thus the vocal folds are the duration determining factor (DDF) for most of the plan. Timing constraints would be different in different vowel contexts. With [a] vowels, for example, the tongue body transition from [a] to [s] or [z] and back might well dominate as DDF.

There are only two arbitrarily set static durations. One is for the glottal area AG, kept at the phonation state for 8 time units (40 ms), in order to give extra length to the stressed [i] vowel carrying the pitch fall. The final [e] vowel is given 80 ms of static vocal folds posture, to signal the end of the utterance; this is appropriate, even for short vowels of English. The [t] in “it” is glottalised. Firm contact between the tongue tip and the alveolar ridge for [t] is simulated by a negative value for AF. The tongue tip actions are superimposed on tongue body shapes to simulate tongue tip-body links, including limitations on tongue tip stretch.

Comparisons between synthetic and real speech

Spectrograms were used to compare the two acoustic outputs from the model, corresponding to the two AG patterns in Fig. 2, with recordings of speakers who read the two sentences at a moderate tempo set by themselves. They were not copying the synthetic speech and no time marker or guide was given. The 5 speakers (3 women and 2 men) have near-RP accents of English. Segment durations for the words “hiss” and “his” are shown in Table 1. Vocoid onsets and offsets were defined by changes in energy visible at the fre-
Table 1
Segment durations and duration ratios for sentences containing (a) “hiss” and (b) “his”. Values for durations (with standard deviations) in ms. 5 real speakers: A, B, C (F) and D, E (M), with outputs from the model, heard as “hiss” or “his”. Figures for an intermediate synthesis near the “hiss”-“his” perceptual boundary are included also. Maximum glottal area is shown for the ambiguous synthetic sentence. n is the number of tokens analysed. Segmentation criteria are described in the text.

Speaker n [v] onset to [s] or [z] duration segment [v] or [z] frication segment ratio of [v] to [s] or [z] frication durations

<table>
<thead>
<tr>
<th>Speaker</th>
<th>n</th>
<th>[v] onset to [s] or [z] duration segment</th>
<th>[v] or [z] frication segment</th>
<th>[v] or [z] voiceless segment</th>
<th>ratio of [v] to [s] or [z] frication durations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) “hiss”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>355 (15)</td>
<td>75 (10)</td>
<td>160 (13)</td>
<td>185 (14)</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>320 (7)</td>
<td>75 (7)</td>
<td>140 (7)</td>
<td>140 (11)</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>280 (17)</td>
<td>65 (6)</td>
<td>105 (7)</td>
<td>125 (8)</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>345 (14)</td>
<td>90 (17)</td>
<td>140 (11)</td>
<td>165 (9)</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>290 (21)</td>
<td>60 (13)</td>
<td>135 (12)</td>
<td>170 (9)</td>
</tr>
<tr>
<td>Model</td>
<td>1</td>
<td>310</td>
<td>95</td>
<td>140/160</td>
<td>175</td>
</tr>
<tr>
<td>AG = 0.3 cm²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>1</td>
<td>285</td>
<td>70</td>
<td>65/100/115</td>
<td>70</td>
</tr>
<tr>
<td>Ambiguous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speaker</th>
<th>n</th>
<th>[v] onset to [s] or [z] duration segment</th>
<th>[v] or [z] frication segment</th>
<th>[v] or [z] voiceless segment</th>
<th>ratio of [v] to [s] or [z] frication durations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b) “his”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>330 (11)</td>
<td>140 (17)</td>
<td>100 (16)</td>
<td>95 (14)</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>315 (8)</td>
<td>135 (11)</td>
<td>70 (9)</td>
<td>30 (24)</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>290 (27)</td>
<td>100 (18)</td>
<td>75 (17)</td>
<td>65 (36)</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>345 (13)</td>
<td>155 (7)</td>
<td>80 (16)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>285 (18)</td>
<td>100 (11)</td>
<td>75 (12)</td>
<td>55 (19)</td>
</tr>
<tr>
<td>Model</td>
<td>1</td>
<td>295</td>
<td>135</td>
<td>40</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 1 suggests that the modelling is capturing some essential features of articulatory timing constraints. As expected, a multiplicity of acoustic differences characterises the “hiss” and “his” portions of the sentences in the real speech. For all 5 speakers, corresponding segment durations or ratios comparing “hiss” and “his” are significantly different at a probability level of 0.01 except for the durations of [v] onset to [s] offset. Here the differences are not significant, with p ≥ 0.5 for all the speakers (t-test, 2-tailed). Constancy of word duration is suggested, though not proved, by this finding.

Similar sets of contrasts are seen in the synthetic speech, and the actual values fall well within the ranges for the real speakers in nearly every case. The synthetic [v] vocoid preceding [s] is rather long, but the more notable discrepancy is in the frication segment for [z]. This is much shorter in the synthesis then in the natural speech. As a result, the [v]/[z] frication durations ratio is excessively high in the synthesis.

In the natural speech, it is the disappearance of voice for [z] which is the most variable feature,
both within and across speakers. The duration of the voiceless segment for [z] has a cross-speaker range of 95 ms, and a fractional range (i.e., range divided by overall mean value for the 5 speakers) of 2. This is much larger than all the other fractional ranges, which lie between 0.2 and 0.4. The synthetic “his” falls well within the natural range for voiceless duration but it differs greatly from the real speech versions in the co-occurrence of voicing and frication noise. These two source types overlap — although to very varying degrees — for most of the natural utterances. In all but 2 of the 29 tokens of the real versions of “his” the [z] frication segment is longer then the [z] voiceless segment. The opposite is true for the synthetic version of “his”, with voicelessness extending 25 ms beyond the friction segment. This unnaturalness in the modelling is probably ascribable to the reduction in amplitude of frication and aspiration noises applied when there is voicing. Figure 4 shows the frication noise source amplitude envelope before modulation and reduction, for two stimuli, one either side of the “hiss”/“his” perceptual boundary. At this stage of the modelling there is plenty of overlap between voice and frication noise sources.

The F0 traces in Fig. 4 may appear to be incompatible with findings on real speech, since F0 dips lower in the stimulus heard more as “hiss” than in the one heard more as “his”. However, most of the dip in F0 is not realised in the former case, because voice is at a very low level or absent here; for the latter, part or all of the F0 dip is likely to be preserved in the output.

The listening experiment

Twelve synthetic stimuli were made for a forced-choice listening test. These correspond to a continuum of maximum glottal area, AG, attained for the fricatives. As may be seen in Fig. 2, it was assumed that the same time was required for both smaller and larger articulatory transitions of AG (and of other articulators). This is certainly an oversimplification, but is based on many results found for natural speech, [15,16] for example. The modelling framework allows for improved descriptions of articulatory kinematics as more data for real speech are published. Four repetitions of each item were made. The 48 items were presented in random order to 20 adult native speakers of English, with a 4 seconds gap between items. A Ferrograph Model 7622H tape recorder was used for the recording of the synthetic speech and for its playback, with Koss electrostatic head-
an articulatory continuum phones model ESP 6A. The test was preceded by six practice items. Responses from 17 of the 20 listeners showed a single clear crossover from "his" to "hiss", while 3 of the listeners gave some "hiss" responses for very low values of maximum glottal area. This result may perhaps be associated with the greater difficulties which some listeners found in making "his" decisions, as compared to "hiss". Six listeners said that they heard "hid" rather than "his" on some occasions, for example. It might be expected that, if glottal area remained very small for the articoid, insufficient noise might be generated for a clear-cut "his" judgement.

Responses of all 20 listeners pooled are shown in Fig. 3. The crossover point is near 0.3 cm² for peak glottal area in this case. It seems surprising that the average "hiss" response declines for the largest values of peak glottal area. When this is 0.45 cm² or more, oral pressure approaches subglottal pressure; the transglottal pressure drop becomes less than 0.2 cm H₂O during the fricative. It might be expected, therefore, that a stable region for [s] would be attained, regardless of further increases in glottal area. One possible explanation for the falling off in the "hiss" response may be suggested. Whereas the frication noise source amplitude envelope trace has a single peak or a plateau ending in one peak for glottal area up to about 0.35 cm², for higher values of glottal area a double peak of increasing prominence appears. The start of a double peak for frication noise may be seen in Fig. 4. Two short peaks of frication in the output may give some listeners an impression partly characteristic of [z].

Other features seen on Fig. 4 may perhaps help to explain why the phoneme boundary is near to 0.3 cm² for maximum glottal area. The two sets of traces are for stimuli on either side of this boundary. Glottal area increases further, to 0.4 cm², for one and less far, to 0.2 cm², for the other. In Fig. 4 there is a clear switch across the phoneme boundary in the coordination between voice, the voice source amplitude envelope and AF, the tongue tip articulatory transition. In the case heard more as "hiss" than "his", most of the tongue tip transition occurs while the voice source is extremely weak; the formant transitions reflecting it are probably not apparent in the model's output after filtering and the radiation transform have been applied. In the other case, heard more as "his" than "hiss", the voice source is strong for most of the tongue tip transition. This is likely to be reflected in the output as F1 or F2 transitions seen in voice.

Because of the conditions inimical to voicing, the steepness of the voice wave at closure of the vocal folds is weakened here, in the region where voice and noise co-occur. The amount of shift in spectral balance towards low frequency emphasis nearer the centre of the fricative will be reflected in whether only F1 or both F1 and F2 are visible on spectrograms. This is likely to vary across speakers in natural speech. The 5 real speakers do show just this kind of contrast in formant transitions, with the kind of cross-speaker variations expected. In general it is the case that for [z] F1 or F2 or both have dipping transitions before voicing ceases, while for [s] there is little or no F1 dip seen and an F2 dip transition, if visible at all, occurs after voice has ceased. Within this framework the speakers differ. Speaker D, for example, has fully voiced [z] with F1 and F2 and even F3 transitions visible all through the frication noise segment. For his [s] an F2 dip is sometimes seen, but only after voice ceases. Speakers A and C, by contrast, have long voiceless segments for [z]. Speaker C has a dip in F2 clearly visible before voicing ceases here, while for [s] she has an F2 dip leading into the frication noise segment, but it occurs after the end of voicing. For speaker A the formant transition contrasts are less clear cut. The F1 dip during voicing seems to be a variable feature, but a contrast seems to be made between slightly falling F2 transitions seen in voice: the transition is of short duration for "hiss" and of longer duration for "his". The outputs from the model show more dip in F1 during voicing for "his" than for "hiss", with very little or no F2 transition visible on the spectrograms in either case. This is consistent with the relationships between the voice source and tongue tip articulation traces as seen on Fig. 4.

If this aspect of formant structure is indeed salient for the "hiss"-"his" contrast, then the listeners' responses to stimuli numbers 4 and 8 are predictable. The model seems to be showing some important features of natural speech, although it...
has rather more low frequency emphasis in the sounds here than does most of the real speech.

In Fig. 3, the square root of peak glottal area $AG$ is used as the abscissa scale, since this gives a more symmetrical relationship between the apparently optimum regions for "hiss" or "his" responses and the phoneme boundary region than do scales of AG or constant Weber ratio increments $\Delta AG/AG$. $\sqrt{AG}$ has the dimension of articulatory distance; Abramson et al. [8] selected radius of velopharyngeal port as their articulatory dimension for the same reason. Abduction of the vocal folds may be what a speaker controls, rather than glottal area. Even if vocal fold abduction does represent a relevant articulatory continuum, discrimination tests using it could be difficult to interpret. Several acoustic features change along this axis; one might vary smoothly but another might exhibit a very nonlinear dependence on the articulatory variable. Discrimination peaks could reflect locally large acoustic changes, or might instead signify categorical perception of a steadily changing acoustic variable which provided a salient cue for the perception of "hiss" versus "his".

It would be interesting to see whether other articulatory dimensions yield a monotonically increasing "hiss" response, or whether, in some cases, the acoustic patterns generated may give multiple or extended regions of ambiguity. The articulatory dimension chosen here does appear to be an appropriate one for the phonological contrast.

**Improvements to the model**

The weakest aspects of the model's simulation, as indicated by these and the earlier comparisons [9], seem to be in the representation of larynx behaviour. This is not surprising given the extreme simplicity of the functional model used. The strength and even the presence of voicing during the [z] contoid segment are clearly sensitive to the fall-off in the strength of the voice source as the vocal folds are abducted, for example. The coexistence of voicing and turbulence noise sources is not modelled in a very realistic way yet. Their co-occurrence and the degree of reduction of noise source amplitudes when voicing is present may depend on processes for which vocal fold length and constriction shapes are relevant variables. These are likely to vary from speaker to speaker. A more flexible and more realistic simulation of the generation of voice has been developed for the model now. In this, three factors interact to control the voice source waveshape. They are: one articulatory variable, glottal area $AG$; one myoelastic variable, effective stiffness and mass of the vocal folds $Q$; and one aerodynamic variable, pressure drop across the glottis. As their values vary during a modelling run, the acoustic properties of the voice source vary correspondingly [17]. Different larynx types, for example that of a professional singer as opposed to that of a speaker, can be specified.

Clearly, many other articulatory means, not necessarily a minimal pair, besides the single articulatory dimension of maximum glottal area used here, can be exploited by speakers to create phonological contrasts. The determination of suitable sets of articulatory schemes may become much more complex when the interactions between several dimensions need to be considered. Individuals may vary not only in the physical parameter values of their own larynx but also in the ways they choose to use the larynx and other articulators, including inter-articulator coordination. It is quite possible that some English speakers may use a voiceless contoid for [z] as well as for [s]. The vocal fold abduction gesture could even be the same for both fricatives, while the fricative and other segment durational contrasts could be achieved by contrasting tongue tip transitions, occlusion durations and occlusion onset timings. It should be possible to simulate the actions of individual speakers, arriving at an approximation to their articulations via a series of analysis-by-synthesis cycles. Perception experiments similar to the one described here should be able to establish articulatorily based phoneme boundaries for such speaker-specific simulations. By applying perturbations to the articulatory plans, the ranges within which the auditory goals are achieved could be established, together with the corresponding allowable perturbations in a number of covarying acoustic pattern features. This kind of information may be useful for multi-speaker speech recognition by computers, as well
as contributing perhaps to our understanding of how listeners' requirements for auditory pattern contrast are reconciled with speakers' probable preference for simple articulatory programmes and with their limited ability to control articulatory targets accurately.

Acknowledgement

The computer modelling of speech production is supported by the Science and Engineering Research Council.

References

ANALYSIS OF SPEECH SIGNAL VARIATION BY ARTICULATORY SYNTHESIS

C. Scully and G.C.F. Clark (SPAR group)

Department of Linguistics & Phonetics, University of Leeds, UK

INTRODUCTION

The richness and complexity of natural speech are lacking in currently available speech synthesisers. Natural speech is characterised by variation and variability. Sample-to-sample variability contributed to intelligibility for human listeners in a study by Pollack (1) and, as suggested by Clark et al. (2) that the simple, degraded acoustic structures of speech are not likely to provide poorer, less consistent acoustic cues. Yet the linguistically determined and sample-to-sample variations of natural speech constitute a problem for the recognition devices needed for speech input to computers. Human listeners can make use of a variety of acoustic cues for speech perception and it seems likely that they understand the rules governing the ways in which the acoustic pattern features covary. We believe that a speaker's goal is to output a required sequence of broadly defined auditory patterns appropriate to the linguistic message and accent of the speaker, but that details of acoustic structure in the speech signal are rather strongly constrained by the processes of speech production.

A computer-implemented composite model of speech production processes has been developed. In the modelling, actions of a small set of independent moving structures - the articulators - combine to give rather lifelike complexity in the synthetic speech output. A simple change to one articulatory action can result in a bundle of acoustic pattern feature changes. Use of the model to generate synthetic speech may be one way of gaining a greater understanding of the rules underlying the acoustic complexity of speech. Constructions purely acoustic analyses. With this overall aim, the model being used as an analysis-by-synthesis tool to examine some individual speakers of British English.

It is possible that articulatory synthesis like that from our model may be able to provide improved naturalness and acceptability for a future generation of synthetic voices, but there are very great difficulties connected with the gathering of information from natural speech to drive the model. In the application described here we try to match only certain crucial aspects of natural speech, aspects for which information can be gathered in our laboratory, combined with knowledge based on the published research of other investigators. Once a reasonably good match has been obtained to a single sample of one speaker's speech, the simulated actions in the model can be made to vary by realistic amounts as indicated by available data on natural speech, and the acoustic effects generated can be compared with those observed in multiple tokens of the natural speech. Different speaking rates or styles of speech can be simulated with the model, to investigate linguistic variation as opposed to the variability associated with repetitions said in the 'same' way.

One important way in which current terminal-analog (acoustic-based) synthesisers fail to match the complexity of natural speech is in the specification of the acoustic excitations for voice and turbulence noise. Since speech production is characterised by nearly continuous movement with resultant changes in the properties of the vocal folds and in the aerodynamic conditions at the larynx, the frequency spectrum of the voice excitation, as well as its fundamental frequency, is expected to vary almost continuously. In other speech synthesis systems these changes to the voice excitation must be explicitly stated as rules; because our type of model is a better representation of the physical processes, it automatically generates these effects. Parameter values will be extracted from waveforms of inverse-filtered natural speech under known conditions. These will be used for the voice generation model in simulations where articulatory and aerodynamic parameters are well matched to those of the real speaker. The analysis-by-synthesis procedures described in this paper are precursors to improved modelling of the voice excitation.

THE NATURAL SPEECH DATA

Vowel-consonant-vowel sequences (VCV), produced by several speakers with a closely similar accent of British English, are being examined. The subset of the vowels and consonants used at present includes the vowels [a] as in "nerve", [i] as in "see", and [e] as in "pass", with alveolar consonants [z], [z], [d], and [n]. The procedures are exemplified here by the series [p[z]'p[z]'...]. All produced on one expiratory breath. The auditory goal is satisfied: the vowel following the [z] is to be heard as stressed and the preceding vowel as unstressed; this effect is to be produced without the effect of strong emphasis and without pitch movement. The study concerns analysis-by-synthesis of the [VCV] sequence [z'z']; the surrounding [p] elements are included because they permit a reasonably reliable estimate of the pressure of air in the lungs. With the speech material controlled in this way, the number of variables that interact to generate complexity of acoustic structure is reduced and matching by the modelling can be attempted with more confidence in its reliability.

The speaker (an adult, female Londoner who lives in Leeds) recorded the speech material in an acoustically-treated studio, with gaps. This recording was played back in the laboratory to provide a cue for her to speak the same speech material, at the same speaking rate, into an airflow and pressure measuring...
device. A mask, constructed to fit that speaker's face so as to minimise air leakage, contained a hose to connect the air streams. Volume velocity of air into or out of the mouth (UC) was measured with a pneumotachograph. This was used to indicate volume velocity (UC) through the alveolar region of the vocal tract where it was narrowed or closed for the consonants being investigated. The pressure drop across this constriction or closure (PC) was measured by inserting a polyethylene tube via the mouth, with the open end of the tube inside the air cavity enclosed in the oral vocal tract behind the alveolar constriction. The open end pointed downwards, across the air stream to minimize spurious pressure readings, while the outer end was connected to a differential pressure transducer outside the disk, the reference pressure being that inside the mouth space of the mask. A miniature microphone mounted inside the mask gave acoustic information. A larynx microphone showed the extent of voicing.

Figure 1 shows the traces for the third [z] in the sequence, used for the model simulations. It is common for voicing to disappear during nominally voiced fricatives of English, as indeed happens for this speaker's [z]. The area trace, A in Figure 1, is proportional to area AC: that is, the cross-section area of the alveolar constriction for [z], and of the mouth for [p], as estimated from the orifice equation, see Warren and DuBois (3):

\[ AC = \frac{k \cdot UC}{0.5} \quad \ldots \ldots \ldots (1) \]

k is an empirical constant. The A trace in Figure 1 has k = 1 in a hardware device whose input signals are UC and PC; it indicates the movement path of the tongue tip-blade region for [z]. Actual values of AC in cm² were calculated from equation (1), using a k value of 0.0076 with UC in cm³/s and PC in N/m⁴, see (3). The equivalent of AC for [z] in the model is AF. For the analysis of natural speech, aerodynamic variables UC and PC are used to estimate an articulatory variable AC. In the model AF is input as an artifactual device that is under the simulated speaker's control. The kind of consonant chosen for study here, a voiced fricative, is such that the voice excitation waveform is likely to change during much of the voiced portions of the [CV] sequence. It is thus an interesting sequence to try to match in the model.

MODELS OF SPEECH PRODUCTION PROCESSES

Our model simulates the final stage of speech production, as shown in Figure 2. Each block in the figure represents a set of processes. A speaker almost certainly acts as a closed-loop control system, but the many sensory feedback paths available are not represented in the model. The input block does, however, include a planning scheme which allows the timing and coordination of key events in the movements of the articulators to be varied; this scheme may perhaps mirror some processes in the speaker's central nervous system.

Other models. Research has focussed on different combinations of stages of speech production. For example constraints on the moving structures and the mapping from the resultant area function to formant frequencies have been investigated by Lindblad and Sundberg (4), Ladefoged et al. (5). In modelling the movement of the articulators, the tongue body as seen on X-ray views may be represented as circular; these models may perhaps be overconstrained in their kinematics; but are able to model consonants as well as vowels, see Mermelstein (6), Coker (7), Rubin et al. (8).

The acoustic filtering processes have been represented by electrical analogues. Electrical transmission lines with variable L and C elements represented the acoustic tube of the vocal tract, in the establishment of mapping from tube shape to formant frequencies, in vowels, see Stevens et al. (9), Fant (10). Satisfactory dynamic control of this kind of hardware is difficult to achieve, see Rosen (11), Keltai (12). Digital simulation has helped to overcome some of the problems, although quantisation effects and lengthy computations are undesirable features of current computer-implemented models. Losses, lip radiation impedance and wall vibrations have been included in the electrical transmission line of Planagan et al. (13), in the acoustic transmission line of Maeda (14) and in Liljencrants (15) development of the reflected pressure wave method of Kelly and Lockbaum (16). This last method (16) is used in our model. Voice source models represent the vocal folds as mass-spring systems, see (13), as continuous tissue structures, see Titze and Talkin (17) and as oscillatory glottal area functions Ag(t), see (14). A subglottal acoustic tube has been included by Titze (18). Childers et al. (19) have modelled acoustic loading of the voice source by the vocal tract filter. Frequency domain simulations such as those of Wakita and Fant (20) give greater understanding of frequency-dependent losses, which must be included in the time-domain simulations needed for speech synthesis. The z transform offers improved representation, see (15). All the acoustic models are valid up to about 4 kHz, since they assume plane wave propagation. The full range of frequencies is included in, for example, (13) and (14), but the aerodynamic system, defined as the low-frequency components of volume velocity of air and air pressure, has been modelled separately, by Rothenberg (21) and others following him, including ourselves. The aerodynamic processes link the articulatory actions to the speech sounds, since they all originate from the severe obstructions of an airstream. Because the aerodynamic system, unlike the articulatory system, is irreducible, the various acoustic sources are affected by shared physical conditions; for this reason they are not independent of each other.

THE LEEDS MODEL

Articulation. Our articulatory model contains parameter points to define 8 quasi-independent articulators, as shown in Figure 3. The points are E, P, B, F, J, T, L, and V. Cross-section areas of the tube at these points, AE, AP and so on, change with time to simulate articulatory actions and the resultant change to the acoustic tube shape. AE, AP, and AB are synchronised and define the tongue body; AF shapes the alveolar constriction for [z]; AJ, with AT, and AL, representing jaw with teeth and lips, are usually synchronised with the tongue body. Additional articulators are AV for soft palate actions, PLU the pressure
of air in the lungs and two larynx articulations, slow changes of glottal area AG, and slow changes of vocal fold stiffness and mass Q.

Aerodynamics. The variables, subglottal air pressure PSg, oral air pressure PC, volume velocity of airflow through the glottis Ug and through the alveolar constriction UC are computed by numerical solution of the differential equations for continuity of mass flow for the two cavities and the orifice equations, equation (1), for the glottis AG and the alveolar constriction AF, where flow-dependent resistances occur. CW is the compliance of the walls, see (21).

Acoustic sources. The parametric representation of Fant (22) is used for the aerodynamic component of volume velocity of air through the glottis Utg(t). We make this waveform vary with the conditions which generate it. We use as controlling variables two articulatory parameters AG and Q, with an aerodynamic variable computed in the model, the pressure drop across the glottis A P. Some aspects of acoustic source-filter interactions are modelled: first, Utg(t) is skewed relative to oscillatory area changes AG; secondly, transglottal losses from the vocal tract filter depend on AG and can also be made to vary within each voicing cycle. The model has been described elsewhere, see for example Scully and Allwood (23), (24), Allwood and Scully (25).

ANALYSIS BY SYNTHESIS
Stages in the matching. Other articulatory models have been used for analysis by synthesis of natural speech, see Flanagan et al. (26), Ghirai and Masaki (27), Levinson and Schmidt (28). Many parameters need to be optimised. Our methods make use of aerodynamic as well as acoustic data from natural speech, thus permitting some confidence in earlier stages of the matching before the final acoustic processes are simulated.

The procedures used for matching the model to the speaker's [ps'sp...] sequence have been described elsewhere, see Scully (29), (30). They will be summarised here.

1. The movement path for AF in the model is matched to AG in the natural speech, see Figure 1.
2. Vocal tract shapes for [s] and [z] and articulatory transition paths are based on evidence in the literature and on previous modelling, see (25).
3. UC and PC, computed in the aerodynamic block of the model, are compared with Utg and PC in the natural speech. Steps (1), (2) and (3) are repeated varying timings and transition end point states, until a good match is obtained in step (3).
4. Discrepancies suggest changes to the modelling, for example if UC falls too low in the model this indicates that the vocal tract cavity volume VC is enlarged unrealistically there.
5. (4) Acoustic sources are computed using a larynx model appropriate to, in this case, a female adult speaker, see Karlson (31).
6. (5) The final stages of the model are used and the output synthetic speech signal is compared with the real speech signal, generally as spectrograms.

**References**


ACKNOWLEDGEMENTS

Part of an Alvey project, SPAR (Speech Pattern Algorithmic Representation), collaborative research between University College London, Imperial College of Science and Technology, General Electric Company, Plessey, and the University of Leeds, SERC grant GR/D/40791. It is a pleasure to thank Mr. E. Brearley for help with the experiments and Mrs. C. Foster for preparation of the manuscript.

The VAX system, in the Faculty of Engineering University of Leeds, is part of the SERC Interactive Computing Facility.

Figure 1 Natural speech data for [(p)x's(p)] Voicing (top) with UO, PC and A, explained in the text.
INPUTS  Speaker type, initial conditions, time plan, instructions for articulators

ARTICULATION  Compute movement paths, vocal tract area function, cavity volume VC, composite constriction for the aerodynamics 5 ms

AERODYNAMICS  Compute low frequency components of air pressures PSG, PC, volume velocities UG, UC, UV.

ACOUSTIC SOURCES (EXCITATIONS)  Compute waveforms for voice, frication noise (at the vocal tract constriction), aspiration noise (at the glottis), transient 28 µm

FILTERING  Reflected pressure wave method for response of the vocal tract tube (0.5 cm sections) to each acoustic source. 28 µm

RADIATION  High-pass filtering 6 dB/oct., 28 µm

OUTPUT  Digitised waveform low-pass filtered and down-sampled to 84 µs; transferred from VAX 11/780 to PDP 11/03; real-time output with clock; D to A conversion; anti-aliasing filter; analogue output waveform with bandwidth 5.9 kHz. 84 µs

Figure 2  Block structure of the composite model of speech production processes.

Figure 3  Parameters for articulation and aerodynamic variables PSG, PC, UG, UC, UV, explained in the text.

Figure 4  Three voice excitation parameters above and their three controlling variables below, explained in the text.

Figure 5  Voice excitation waveshapes (a) for the vowel ['s'], (b) for [s] where voicing ceases.