Vowel Normalization:

a perceptual-acoustic study
of Dutch vowels

Omslag - wonderworks.nl
Printed and bound by Ponsen & Looijen bv, Wageningen

ISBN: 90-9016939-3

Vowel Normalization:
a perceptual-acoustic study of Dutch vowels

Een wetenschappelijke proeve op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen
op gezag van de Rector Magnificus Prof. Dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op maandag 23 juni 2003,
des namiddags om 3.30 uur precies

door

Patricia Martine Adank

geboren op 8 december 1972
te Ommen

Promotor:	Prof. dr. R. van Hout
Co-promotor:	Dr. R. Smits (MPI)
Manuscriptcommissie:	Dr. A.C.M. Rietveld (voorzitter)
	Prof. dr. L.C.W. Pols (UvA)
	Prof. dr. V.J.J.P. van Heuven (UL)

Acknowledgments

The research described in this thesis was carried out at the Department of General Linguistics and Dialectology of the University of Nijmegen, the Netherlands, and at the Department of Linguistics of the University of Alberta, Canada. This research was carried out as a part of the Flemish-Dutch cooperation project (VNC) “The pronunciation of Standard Dutch: variation and varieties in Flanders and The Netherlands”, funded by the Netherlands Organisation for Scientific Research (NWO) and the Fund for scientific research, Flanders, Belgium (FWO), project nr. 205-41-069.

This thesis would not exist without the help, advice, moral support, and guidance of various other people. I would like to dedicate these pages to acknowledge and thank them.

First and foremost, I would like to express my thanks to my two supervisors, Roeland van Hout and Roel Smits. Thank you for your patience and for many inspiring meetings. You made a great and efficient team of supervisors, I still find the similarity of the comments you both wrote in margins of chapters after proofreading amazing (and somewhat eerie).

Many thanks are due Terry Nearey for his time, advice, and of course for programming “NeareyTracker4x” while I was at the Department of Linguistics in Edmonton. It was very useful when I was manually verifying all the 4800 vowels. I am grateful to everybody in the Linguistics Department for making my visit enjoyable. I would also like to thank Linda MacDonald and Darren Nicholls for spicing up my social life when I was in Edmonton.

Furthermore, I am particularly grateful to the members of the reading committee, for their careful reading and their useful comments. I would especially like to thank Vincent van Heuven for his help in the first two years of my research project.

For the use of their research facilities in 2000, I am indebted to the Max Planck Institute for Psycholinguistics. Thanks are especially due to John Nagengast, for programming the listening experiment, and to the research assistants of the Language Comprehension Group, for helping me carry out the experiment at the institute. Also, I would like to express my gratitude to the expert listeners for participating in possibly the most tedious (and lengthy) listening experiment they ever participated in.

Next, I would like to acknowledge the Netherlands Organisation for Scientific Research (NWO) and the University of Nijmegen for their financial support. This support allowed me

to visit several conferences and to carry out part of this research in Canada.

Mirjam Wester, Stijn Ruiter, and Dick Smakman, thanks for carefully proofreading almost every page of this thesis.

Johannes Verheijen and RadBoud de Bree, thanks for designing the wonderful 'klinkerweg' on the cover of this thesis.

Many people in the Faculty of Arts have made my time here much more enjoyable. Thanks are due to the 'AiO's' in the Department of General Linguistics and Dialectology, and my colleagues at the Department of Language & Speech, for many interesting discussions, and the relaxing lunches and coffee breaks.

Michiel Adank and Mirjam Wester, I am honored that you have agreed to be my 'paranimfs'.

Finally, I would like to thank my family and friends for their support all these years. Stijn, I thank you for your patience and love.

Contents

Chapter 1

Introduction

1.1 Background

Pols, Tromp & Plomp (1973) and Van Nierop, Pols & Plomp (1973) are considered the standard studies in which the vowels of Dutch are described acoustically (see for instance, Fant, 1975; Disner, 1980; Rietveld & Van Heuven, 2001). The 12 monophthongal vowels of Dutch are described in terms of the mean frequencies of the first three formants.

Pols et al. attributed the variance in the formant frequencies to three sources. In their Table II (on page 1095), the total variance is classified as originating from the vowel, the speaker, or measurement error (their "residual variance"). The vowel-related variance is the most substantial source, followed by the residual variance and the speaker-related variance.

A similar classification of the sources of variation in the speech signal was proposed earlier by Ladefoged & Broadbent (1957). They, too, distinguished linguistic and speaker-related variation. However, their classification is more specific than Pols et al.'s: Ladefoged & Broadbent split up the speaker-related variation into "personal" variation and "socio-linguistic" variation. They stated that the personal variation in the signal originates from differences between speakers in the shape and size of their vocal tract and larynx. The sociolinguistic variation originates from differences in social characteristics of speakers, such as regional background, educational level, or gender.

In the present research, I adopt Ladefoged & Broadbent's classification and thereby assume that the speech signal conveys three types of variation: phonemic variation (identical to Ladefoged & Broadbent's "linguistic" variation), sociolinguistic ("socio-linguistic") variation, and anatomical/physiological variation ("personal variation").

In different disciplines in speech science, the acoustic consequences of the three variation sources are considered relevant or irrelevant, depending on the discipline's goal. For instance, in an automatic speech recognition task concerning speaker-independent vowel recognition,

all phonemic variation is considered to be useful, while all speaker-related variation is considered to be noise. This is the case regardless of whether the variation in the signal originated from sociolinguistic or anatomical/physiological characteristics of the speaker. Furthermore, for a context-independent automatic speaker-verification task, all phonemic variation has to be ignored, while all speaker-related variation, anatomical/physiological as well as sociolinguistic, can be used to improve the performance. Finally, in sociolinguistics – a branch of speech science that studies the effects of sociological differences between speakers on their spoken language – the phonemic variation and the sociolinguistic speaker-related variation are considered useful, whereas variation related to the speakers' anatomical/physiological characteristics is considered to be noise.

Pols et al. (1973) aimed to cluster formant frequencies into their corresponding vowel category for improving vowel categorization. Therefore, it can be said that their goal was similar to the goal of the automatic speech recognition task described earlier: preserving all phonemic variation, while minimizing the acoustic effects of both sources of speaker-related variation. In order to minimize the speaker-related variation, they transformed the formant measurements through a procedure consisting of subtracting a scaling factor, the mean formant frequency (on a logarithmic scale) across all 12 vowels from a speaker, from each individual formant frequency of each token produced by that speaker.

In phonetics, a wide variety of studies aimed at eliminating speaker-related variation from the signal by designing procedures that can be subsumed under the heading of vowel, or speaker, normalization (for instance, Gerstman, 1968; Lobanov, 1971; Syrdal & Gopal, 1986). The approach through which Pols et al. sought to improve clustering of formant frequencies can be regarded as an example of acoustic vowel normalization.

It can be useful to apply normalization procedures for sociolinguistic purposes, especially when small groups of speakers have to be compared. One of the problems associated with comparing small groups is that the anatomical/physiological differences between individual speakers obscure the sociolinguistic differences between the groups. Suppose, for instance, that the speech of three women speaking language variety A and three men speaking language variety B have to be compared. In such a case, it is not clear what causes underlie differences between the vowels belonging to dialect A or B, sociolinguistic differences between the speaker groups (dialect A vs. dialect B) or anatomical/physiological differences: e.g., the fact that men generally have longer vocal tracts than females (e.g., Chiba & Kajiyama, 1941; Fant, 1966; Nordström, 1977). If the anatomical/physiological difference could be eliminated, it would be possible to evaluate whether the differences between vowels can be attributed to the sociolinguistic difference between the groups.

However, according to Thomas (2002), all normalization procedures have their specific drawbacks and the appropriate choice of procedure depends on which drawbacks are tolerable for the study at hand. He illustrated his statements using Nearey's (1978) procedure. According to Thomas, this procedure is, first, inappropriate for cross-dialectal comparisons,

because differences between vowel configurations bias the scaling factors for the formants. Second, Thomas argued that Nearey's procedure does not reflect human speech perception; in order to obtain the scaling factors for each formant, more than one vowel per speaker is required, while "...listeners are capable of normalizing a single vowel without hearing another vowel by the same speaker.".

Two additional disadvantages of vowel normalization procedures can be found in Disner (1980) and Hindle (1978). Both researchers claimed that the procedures for a large part minimize the variation in the speech signal related to sociolinguistic differences between speakers along with the variation related to anatomical/physiological differences. Both researchers stated furthermore that there are indications that, after transformation, the representations of acoustic vowel data display artifacts of the normalization procedures. However, as is discussed in further detail in Chapter 2, the validity of the claims made by Disner and Hindle is questionable, because Hindle's study was carried out on too small a scale to provide conclusive support for his claims, and Disner used judgments that were not based on the vowel data to which she applied the normalization procedures.

Nevertheless, other researchers have used vowel normalization procedures in sociolinguistics. For instance, Labov (1994) used a procedure proposed by Nearey (1978), a transformation similar to the one used by Pols et al. (1973). Labov stated that Nearey's procedure provides an efficient and reliable way of eliminating variation in the acoustic signal that is related to differences in vocal tract length between speakers, while preserving variation related to sociolinguistic differences between speakers. More recently, Watson, Maclagan & Harrington (2000) used a transformation by Lobanov (1971), to transform their acoustic vowel data for displaying evidence for vowel change in New Zealand English.

Summarizing, although procedures for acoustic vowel normalization can be useful for sociolinguistic research, it remains undisclosed whether these procedures preserve the sociolinguistic speaker-related variation in the transformed speech data. Therefore, in order to be able to use normalization procedures for sociolinguistic purposes, it should be investigated whether or not the normalization procedures eliminate the sociolinguistic variation along with the anatomical/physiological variation.

Acoustic vowel normalization can, generally, be defined as a transformation of the acoustic representation of vowel tokens that aims at minimizing the acoustic consequences of specific sources of variation in the acoustic representation of vowel tokens. For sociolinguistic purposes, acoustic vowel normalization is defined as a transformation of the acoustic representation that aims at minimizing the acoustic consequences of anatomical/physiological speaker-related sources of variation, while preserving the phonemic and the sociolinguistic variation.

The research presented in this thesis aims to establish which procedures for acoustic vowel normalization succeed best at separating the three types of variation conveyed in the acoustic signal: phonemic variation, sociolinguistic speaker-related variation, and a-

natomical/physiological speaker-related variation. To this end, a comparison of acoustic normalization procedures is carried out.

Some of the studies that compare normalization procedures, evaluate the procedures' success in the acoustic domain by collecting a vowel database, measuring the formant frequencies on this database, applying the normalization procedures, and by applying a pattern classification algorithm to the normalized acoustic data. If application of a procedure leads to an increase of the percentage of correctly classified vowel tokens, it is decided that it succeeds in increasing the clustering of the formant frequencies. Other studies combine this acoustic comparison with a an evaluation in the perceptual domain, by comparing the results of the analysis with judgments of phonetically trained listeners.

The comparison in the present research consists of two parts: it combines an acoustic analysis with a perceptual-acoustic analysis as was done in Hindle (1978) and Disner (1980). The acoustic comparison is consists of applying the procedures to a large database with speech from speakers of Dutch who were stratified for certain sociological and anatomical/physiological characteristics. Each procedure is evaluated on how it deals with the three variation sources in the acoustic signal. In the perceptual-acoustic comparison, the output of each procedure is compared to perceptual judgments that are obtained on (a subset of) the same vowel data used for the acoustic comparison.

My research differs in three respects from other studies that evaluate normalization procedures in the acoustic domain only (Syrdal (1984); Deterding, 1990), as well as from studies that evaluate the procedures with acoustic as well as perceptual-acoustic comparisons (Hindle 1978; Disner 1980). First, where Hindle (1978) and Deterding (1990) use vowel data produced by a small number of speakers, my vowel data was produced by a substantially larger number of speakers. Second, Disner (1980) uses perceptual judgments that were obtained using a different set of vowel data than the set of vowel data to which she applied the normalization procedures. In my research, the perceptual judgments and the output of the normalization procedures were obtained using the same set of vowel data. Third, most studies use speech produced by speakers of (language varieties of) American or British English (Syrdal, 1984; Deterding, 1990; Hindle, 1978), while I compared the procedures using speech from speakers of various varieties of standard Dutch.

The remainder of this chapter is set up as follows. Section 1.2 discusses the process of vowel normalization in more detail. Section 1.3 discusses the research goals and central question of the present research. Section 1.4 provides an outline of the entire thesis.

1.2 Vowel normalization

1.2.1 Acoustic vowel normalization

Human listeners deal seemingly effortlessly with variability in speech and classify almost all vowel tokens from any speaker of their native language correctly (e.g., Verbrugge et

al., 1976; Assmann, Hogan & Nearey, 1982). Although humans can process the phonemic variation (necessary for recognizing vowel tokens) independently of the anatomical/physiological variation with apparent ease, formant measurements show considerable variation related to anatomical/physiological differences between speakers. This variation in the measurements becomes apparent, when the frequencies of the first two formants corresponding to vowel tokens produced by different speakers are displayed in a so-called 'F_1/F_2 plane' (cf. Ainsworth, 1975) or formant plot. As a rule, it can be observed that areas representing different vowels show considerable overlap. The discrepancy between how listeners are affected by differences between speakers and how these differences affect formant measurements, is also known as the 'lack of one-to-one correspondence between acoustics and perception'. This issue is generally illustrated using the results of Peterson & Barney's (1952) classic study. Peterson & Barney found that vowel tokens from the same category produced by different speakers can have widely differing formant frequencies, while vowel tokens from different vowel categories produced by different speakers can have identical formant frequencies.

In phonetics, a wide variety of studies has been carried out that sought to improve the correspondence between the acoustic and perceptual dimensions of speech. One of the hypotheses in these studies is that listeners naturally perform normalization when perceiving speech sounds. Humans are assumed to somehow perceptually 'even out' differences between speakers. Therefore, research in the acoustic domain aims at understanding the normalizing behavior of listeners from the following two perspectives.

From the first perspective, transformations of formant frequencies are devised that aim to classify vowel tokens as efficiently as humans. Vowel normalization procedures have been developed to obtain higher percentages correctly classified vowel tokens for automatic speech recognition purposes (cf. Gerstman, 1968, or Lobanov, 1971, both discussed in Chapter 2). These procedures accomplish their goal by minimizing the dispersion of formant frequencies within vowel categories due to differences between speakers, i.e., by minimizing speaker-specific variation. Within this type of research, it is considered less important to understand the perceptual and cognitive aspects of vowel processing.

From the second perspective, the goal is to understand the perceptual and cognitive processes involved in vowel processing. To this end, various normalization procedures have been developed to serve as stages in psychological models for human vowel recognition. The primary purpose of these procedures is to model human speech perception in order to explain how listeners categorize vowel sounds. The reduction of speaker-specific variation is a secondary objective (cf. Syrdal & Gopal, 1986, and Miller, 1989, both discussed in Chapter 2).

1.2.2 Perceptual vowel normalization

In everyday conversation, listeners have to decode the same three types of variation encoded in the acoustic signal as mentioned earlier: phonemic information, sociolinguistic variation,

and anatomical/physiological variation. When recognizing vowels, listeners presumably select information in the acoustic signal that allows them to categorize vowel tokens correctly. Studies carried out within the fields of experimental phonetics and cognitive psychology can be classified into two approaches, depending on how listeners are thought to make use of the three types of variation.

The first approach involves studies that presume that listeners categorize vowels from different speakers by selecting the phonemic variation in the acoustic signal, while ignoring the anatomical/physiological as well as the sociolinguistic variation. In this approach, vowel normalization is regarded as a separate process in the perception of vowels, in which the listener selects the relatively invariant, or phonemic, information in the acoustic signal necessary to categorize a vowel token correctly. Under this guise, many studies examine the relevance of different individual acoustic sources of information for vowel categorization, such as the fundamental frequency (e.g., Traunmüller, 1981), the first two formants (e.g., Joos, 1948; Peterson & Barney, 1952), the third formant (e.g., Fant, Carlson & Granstrøm, 1974), or relations between spectrally adjacent formant frequencies (Syrdal, 1984).

The second approach includes studies suggesting that listeners, under certain circumstances, can make use of anatomical/physiological variation as well as the phonemic variation, to achieve vowel recognition (e.g., Johnson, 1990b). Under this approach, models of vowel perception[1] are formulated in which normalization is not explicitly incorporated. The ideas underlying these models result from experiments in which listeners were required to categorize vowel tokens in single-speaker vs. multi-speaker conditions (e.g., Assmann et al., 1982; Mullennix, Pisoni & Martin, 1989; Johnson, 1990b).

In order to establish how listeners deal with sociolinguistic information in the speech signal, it is necessary to investigate how phonetically-trained (expert) listeners transcribe sociolinguistic variation. Although phonetically naive listeners can perceive sociolinguistic differences between speakers, only phonetically-trained listeners are able to interpret and transcribe these differences. Trained listeners generally achieve this by making a phonetic transcription of recordings of the speech material reflecting these differences. Such a transcription entails recording the perceived realization of speech sounds using symbols from the alphabet of the International Phonetic Association (IPA, 1999). The transcription is either broad (or phonemic), involving only symbols for phonemes, or narrow (or phonetic), involving diacritical marks to specify detailed and non-contrastive aspects of the realization of speech sounds, such as palatalization and labialization, as well as phonemic symbols.

In sociolinguistics, phonetic transcription is used as a research tool to investigate language variation, because even the smallest perceived aspect of the phonetic quality of a speech sound can be indicated, thus allowing language changes in progress to be traced. For instance, when it is suspected that a process of diphthongization is taking place in a certain language or dialect, the transcriber can perceptually focus on cues in the realization of vowels by those

[1] For instance, the exemplar model of speech perception.

speakers who are expected to display variation in the speech sounds in question, and represent the perceived cues using narrow phonetic transcription.

However, the use of phonetic transcription as a research tool is not without drawbacks. It has been shown that transcriptions vary depending on the transcriber's language background (Jaberg & Judd, 1927), the type of training the transcriber received (Ladefoged, 1960), the type of speech material to be analyzed (Cucchiarini, 1993), and expectations about the type of variation in the speech utterances to be transcribed (Oller & Eilers, 1975). Another drawback is that phonetic transcription is an extremely time-consuming task, which is disadvantageous when large corpora of speech utterances have to be analyzed.

It is assumed that phonetically-trained listeners process the three variation sources conveyed in the acoustic signal as follows, when describing speech utterances in their native language. The listeners use phonemic variation that in the acoustic signal to make a broad phonetic transcription. The use the available sociolinguistic variation as well as the phonemic variation[2] to make a narrow phonetic transcription. They, finally, ignore the anatomical/physiological speaker-related variation in the acoustic signal to reliably judge the phonemic and sociolinguistic variation.

1.3 Objectives

The present research investigates the following question: which procedure for acoustic vowel normalization succeeds best at separating the three types of variation conveyed in the acoustic signal: phonemic variation, sociolinguistic speaker-related variation, and anatomical/physiological speaker-related variation?

As explained in section 1.1, I aim to answer this question for a specific research domain: sociolinguistics. Therefore, in the present research, my goal is to establish which procedure for vowel normalization is suitable for use in sociolinguistics.

A normalization procedure is considered suitable for use in sociolinguistics when it meets the following criterion. The procedure must preserve the phonemic variation and the sociolinguistic variation and, at the same time, minimize the anatomical/physiological variation in the transformed acoustic vowel data.

It was decided to limit the number of procedures to be compared in this thesis by selecting only procedures that have already been used in sociolinguistics and to those that could theoretically be used in sociolinguistics. This decision was implemented in the present research using the following two criteria. First, the procedure has to use formant frequencies and/or the fundamental frequency as input data (instead of, for instance, whole vowel spectra, or formant bandwidths). A second criterion is that the procedure was evaluated before, by someone other

[2]This is only the case when listeners are transcribing material in their native language. When making a narrow transcription of a language (or dialect) that is not native to them, they may not be able to distinguish the anatomical/physiological speaker-related variation from the sociolinguistic variation, cf. Cucchiarini (1993).

than its designer. A consequence of these two criteria is that no new procedures for acoustic vowel normalization were proposed in this research. The reasons underlying these criteria are discussed in further detail in Chapter 2.

In the present research, it is investigated for each selected normalization procedure whether it is suitable for use in sociolinguistics. This was done through an acoustic comparison of procedures, which involves comparing the normalized acoustic data of each procedure, and through a perceptual-acoustic comparison of procedures and listeners, which involves comparing the normalized acoustic data per procedure to judgments of phonetically-trained listeners.

The acoustic comparison of the normalization procedures is similar to the comparisons described in, for instance, Syrdal (1984). The only difference is that I evaluated the procedures also on how well they preserved sociolinguistic variation, as well as evaluating how well they preserved phonemic variation, while at the same time minimizing anatomical/physiological variation.

The perceptual-acoustic comparison was set up as follows. In the present research, I assume that a procedure for vowel normalization to be used in sociolinguistics should not only model vowel categorization by phonetically-naive listeners when performing vowel recognition tasks (as is done in some studies investigating acoustic and perceptual vowel normalization, e.g., Assmann et al., 1982). Instead, the procedures evaluated in the present research must meet a human benchmark consisting of the judgments made by phonetically-trained listeners of phonemic and sociolinguistic variation in vowel tokens produced by different speakers. Any method for vowel normalization that meets this human benchmark is hypothesized to model judgment behavior and therefore to succeed in preserving phonemic variation and sociolinguistic variation in the data, while minimizing anatomical/physiological variation.

Listeners are considered to be phonetically-trained when they received formal training in phonetic transcription, for instance the IPA system as described in The Handbook of the International Phonetic Association (1999) or DJCVS (Daniel Jones' Cardinal Vowel System, as described in Jones, 1917). They are further expected to have extensive experience with narrow phonetic transcriptions and therefore to be capable of judging the perceived articulatory characteristics of a vowel token, regardless of the speaker. I regard the task of categorizing vowel tokens as judging phonemic variation and judging differences between vowel tokens belonging to the same vowel category as judging sociolinguistic variation. In the present research, the expression 'phonemic variation' thus refers to variation between vowels, while the expression 'sociolinguistic variation' refers to variation within vowels.

However, as argued in section 1.2.2, the reliability of trained listeners is questionable (e.g., Ladefoged, 1960; Cucchiarini, 1993). Therefore, the performance of the listeners was also evaluated in the present research, in order to the able to establish the stability of the benchmark against which the normalization procedures were compared.

The similarity in performance at preserving phonemic and sociolinguistic variation of the normalization procedures and the listeners was established using two types of representations of the same set of vowel data: acoustic and perceptual representations. The acoustic representations consist of the values of the first three formants and the fundamental frequency for each vowel token, normalized following the normalization procedures. These measurements are considered to be related primarily to the articulatory characteristics of the vowel tokens. The perceptual representation consist of series of judgments of the articulatory characteristics of the same vowel tokens, perceived openness, tongue advancement, and lip rounding of the vowel token. The similarity of these two types of representation is used to evaluate how well each normalization procedure models the listeners' judgments.

The perceptual representation used in the perceptual-acoustic comparison had, ideally, to be of a continuous nature, because the acoustic representation was also continuous. Furthermore, discrete judgments – such as vowel category labels – usually do not display enough variation (i.e., misclassifications) that can be modeled reliably. It was expected that a continuous perceptual representation displays enough variation to be modeled reliably. A discrete perceptual representation was obtained as well as a continuous perceptual representation, using the same speech material. This was necessary in order to be able to investigate how the response category affects the judgments of the articulatory characteristics of the vowel tokens. Throughout the present research, the discrete perceptual judgments are referred to as category judgments and the continuous perceptual judgments are referred to as articulatory judgments.

1.4 Outline

Chapter 2 discusses the criteria that were used to select the normalization procedures. In addition, for each selected procedure, its original purpose, underlying idea(s), and implementation in the present research is described. This chapter further presents a literature study of formerly published studies involving comparisons of normalization procedures: Hindle (1978), Nearey (1978), Disner (1980), Syrdal (1984), Deterding (1990), and Nearey (1992). This literature study aims at obtaining a preliminary idea about how well the procedures perform in vowel categorization tasks, and to what extent the procedures model human judgments.

Chapter 3 provides an overview of experiments carried out in experimental phonetics and psychology investigating the relationship between vowel categorization and acoustic and/or speaker-related factors. The goal of this literature study is to draw a picture of how listeners preserve phonemic and sociolinguistic variation, while ignoring anatomical/physiological variation when performing tasks involving category judgments or articulatory judgments.

In Chapter 4, I describe the research design of this thesis. First, the relevance of the studies

described in Chapters 2 and 3 is evaluated for the present research. Second, I illustrate how the normalization procedures are compared to a human benchmark through an experiment involving phonetically-trained listeners.

Chapter 5 provides an overview of the setup of a larger sociolinguistic research project of which my project is a part. The goal of the larger project, is to describe language variation and change by describing the phonemes of standard Dutch in the Netherlands and Flanders. This chapter also describes the speech material that is used in Chapters 6, 7, 8, and 9.

Chapter 6 describes the raw measurement values, values of the fundamental frequency (F_0) and the first three formant frequencies (F_1, F_2, F_3), of the acoustic representation. A detailed description is given of the procedures and algorithms used to obtain these measurement values.

In Chapter 7, it is described how the normalization procedures are compared with each other. To achieve this, the raw acoustic measurements, described in Chapter 6, are transformed using the normalization procedures described in Chapter 2. I evaluate the normalization procedures on how well they preserve phonemic variation, preserve sociolinguistic variation, and minimize anatomical/physiological speaker-related variation in the transformed acoustic representations of the vowel data described in Chapter 5. Using these comparisons, it is established which normalization procedure meets the criterion proposed in the present chapter best in the acoustic domain.

Chapter 8 describes the listening experiment outlined in Chapter 4. In this experiment, phonetically-trained listeners provide a category judgment and an articulatory judgment of a set of read vowel tokens. The purpose of this experiment is twofold. Using these articulatory judgments, the articulatory perceptual representation is obtained, which is used for the comparison with the acoustic representation(s) in Chapter 9. Second, the articulatory judgments are used to evaluate the reliability of the phonetically-trained listeners, and to evaluate the role of the availability of various sources of variation on articulatory judgments made by these listeners.

Chapter 9 presents the comparison of the acoustic representations and the articulatory representation, carried out with regression analysis. This comparison is used to establish which normalization procedure models the perceptual representation best.

Chapter 10 discusses the results and presents the conclusions of the present research. An overview is presented of the procedures' performance in tasks described in Chapters 7 to 9. Given these results, it is decided which procedure is suitable for use in sociolinguistics.

Chapter 2

Acoustic vowel normalization

2.1 Introduction

This chapter deals with acoustic vowel normalization. Its purpose is threefold. First, the criteria are presented that were used to select the normalization procedures evaluated in this research. Second, the selected procedures are described in detail. Third, a literature study is carried out that discusses studies describing previous comparisons of acoustic normalization procedures.

Section 2.2 describes the selection of the normalization procedures. In section 2.3, the classification of the selected normalization procedures is presented. Section 2.4 describes the procedures themselves in detail. Section 2.5 discusses the literature study. Finally, in section 2.6, the conclusions of the literature study are given.

2.2 Selection of normalization procedures

In Chapter 1, I wrote that a normalization procedure must meet two criteria in order to be selected for comparison in the present research. First, the procedure must produce results that can be used in sociolinguistic research investigating language variation and language change. Second, the procedure was evaluated before, by someone other than its designer.

As for the first criterion, roughly speaking, two types[3] of normalization procedures are proposed in the literature: formant-based procedures and whole-spectrum procedures. Examples of formant-based procedures can be found in Gerstman (1968), Fant (1975), Syrdal & Gopal (1986), and Miller (1989). Formant-based procedures seek to improve the correspondence between the acoustic and perceptual domains of speech, for instance, by representing

[3] A third approach is to use neural nets to model vowel recognition, e.g., Weenink (1993; 1997).

the formant frequencies on auditory scales (e.g., bark-scale, ERB-scale), or by multiplying each value by a scale factor, for instance, to correct for differences in vocal tract sizes. Overall, the purpose of formant-based procedures is to minimize the variation within a set of vowel tokens spoken by different speakers, while maximizing the separation of sets of vowel tokens that belong to different vowel categories by transforming the values of (combinations of) F_0, F_1, F_2, and F_3.

Whole-spectrum approaches to vowel normalization are proposed in Klein, Plomp & Pols (1970), Pols, Tromp, & Plomp (1973), Bladon & Lindblom (1981), Bladon (1982), and Klatt (1982). Generally speaking, these approaches use spectral information beyond the center frequency of the spectral peaks used in formant-based approaches. Whole-spectrum approaches assume that all spectral information is relevant and that no speaker-specific information should be discarded. For instance, Bladon & Lindblom (1981) claimed that the perceived distance (in vowel quality) between two vowel tokens can be determined by the Euclidian distance between the bark-transformed amplitude spectra of these two vowel tokens.

It was decided to only compare formant-based procedures in the present research[4], for two reasons. First, one of the advantages of formant frequencies is that they provide a very compact (two- or three-dimensional) description of vowels. Because of this, it is possible to visually represent acoustic differences between vowels in a two-dimensional formant plot; which is a common way to display vowel tokens in phonetics and sociolinguistics (usually displaying F_1 vs. F_2, and sometimes combinations such as F_1 vs. F_0, or F_2 vs. F_3). Second, sociolinguists such as Labov (1994), who have used normalization procedures to reveal variation patterns, have exclusively used formant-based procedures[5].

The second criterion results in the selection of procedures that are discussed and compared with each other in previously published papers on acoustic vowel normalization. A variety of studies evaluate the performance of formant-based procedures, either for use in studies of language variation and change (Hindle, 1978; Disner, 1980), for a phonetic theory of vowel perception (Nearey, 1978; Syrdal, 1984; Nearey, 1992), or for automatic speech recognition (Deterding, 1990). I decided to evaluate all procedures that are described in these six studies, with the exception of those procedures that are not formant-based[6], or are not documented well enough to allow implementation[7].

Finally, by applying the second criterion, several existing formant-based procedures are

[4]It is not unimaginable that whole-spectrum procedures can be applied for sociolinguistic purposes. However, there are indications that formant-based procedures perform better at classifying vowel tokens correctly, cf. Deterding (1990).

[5]It could of course be the case that differences between vowels are represented better with procedures as proposed in Pols (1977), but my decision was to evaluate procedures that have successfully been used in, or that are very promising for sociolinguistic research, and all these procedures are formant-based.

[6]Four of the procedures that are evaluated by Deterding (1990) are not selected: the vocal-tract correction proposed by Wakita (1977), Bladon & Lindblom's distance model (1981), Perceptually-based Linear Prediction by Hermansky (1985a;1985b), and Pickering's (1986) centroid procedure.

[7]This concerns Sankoff, Shorrock & McKay's regression procedure (1974), described in Hindle (1978).

excluded: i.e., any transformation involving F_2' (the 'weighted second formant') as proposed by Carlson, Fant & Grandström (1975) and an alternate version of Gerstman's (1968) procedure as proposed by Hieronymus (1991). Although these procedures are documented well, they were not evaluated by someone other than the designer.

After applying the two criteria, 12 acoustic normalization procedures remained, the baseline procedure, hertz, and 11 procedures that transform the data in hertz. These procedures are selected for comparison in the present research. The procedures are also listed below and are described in more detail in section 2.4.

HZ	the baseline condition, formants in Hz
LOG	a log-transformation of the frequency scale
BARK	a bark-transformation of the frequency scale
MEL	a mel-transformation of the frequency scale
ERB	an ERB-transformation of the frequency scale
GERSTMAN	Gerstman's (1968) range normalization
LOBANOV	Lobanov's (1971) z-score transformation
NORDSTRÖM & LINDBLOM	Nordström & Lindblom's (1975) vocal-tract scaling
CLIH_{i4}	Nearey's (1978) individual log-mean procedure
CLIH_{s4}	Nearey's (1978) shared log-mean procedure
SYRDAL & GOPAL	Syrdal & Gopal's (1986) bark-distance model
MILLER	Miller's (1989) formant-ratio model

2.3 Classification of normalization procedures

Traditionally, formant-based normalization procedures are classified according to the type of information they employ. I describe two previously proposed ways to classify normalization procedures. The first is a two-way classification, as proposed by Ainsworth (1975), the second is a tripartite classification, proposed by Rosner & Pickering (1994).

The first classification is commonly referred to as the intrinsic/extrinsic classification. Ainsworth (1975) was the first to distinguish between extrinsic and intrinsic types of information in vowel recognition. Intrinsic procedures use only acoustic information contained within a single vowel token to categorize that vowel token. These procedures typically consist of a nonlinear transformation of the frequency scale (log, mel, bark), and/or a transformation based on a combination of formant frequencies (e.g., $F_1 - F_0$). An example of an intrinsic procedure can be found in Syrdal & Gopal (1986). Extrinsic procedures, on the other hand, assume that information is required that is distributed across more than one vowel category of a speaker; for instance, the formant frequencies of the point vowels for that speaker. Examples of extrinsic procedures can be found in Gerstman (1968), Lobanov, (1971), Nordström & Lindblom (1975), and Nearey (1978). The majority of the intrinsic procedures was developed

in the field of perceptual phonetics with the purpose of modeling human vowel perception. In general, extrinsic procedures were designed for improving vowel classification, often for automatic speech recognition.

The second classification was introduced by Rosner & Pickering (1994). In their tripartite classification, they classify the normalization procedures as category-independent, as category-specific, or as speaker-specific.

Rosner & Pickering argued that the weakness of the intrinsic/extrinsic distinction lies in the ambiguity of the role of F_0. First, F_0 can act as a factor enabling the identification of an individual vowel token uttered by an individual speaker. Here, F_0 acts as a vowel-specific factor, reflecting vowel-intrinsic F_0, independent of the type of speaker (male, female, or child). The procedures that incorporate F_0 in this manner are category-independent. Syrdal & Gopal (1986) and Miller (1989) are examples of procedures that use F_0 in this fashion[8]. Second, they argued that the F_0 of an individual vowel can provide a cue for the speaker category: listeners may create separate templates for men, women, and children. Under this hypothesis, F_0 facilitates vowel recognition by selecting an appropriate template for the speaker's category. Procedures employing F_0 in this manner are classified as category-specific[9]. The third type are the speaker-specific procedures: these procedures correspond to the extrinsic procedures in the intrinsic/extrinsic classification.

Rosner & Pickering's tripartite division thus boils down to renaming extrinsic factors into speaker-specific factors and splitting the intrinsic factors up into two categories, depending on how the procedures employ F_0; as providing information about the speaker or about the vowel. Rosner & Pickering remarked that previously proposed theories of vowel normalization are either category-independent or category-specific. They added that no explicit vowel normalization theories were developed that incorporate category-specific factors.

In the present research, I adopt an extension to the intrinsic/extrinsic classification. The intrinsic/extrinsic classification is expanded to the formants, thus creating formant-intrinsic and formant-extrinsic categories in addition to the vowel-intrinsic and vowel-extrinsic categories. The intrinsic/extrinsic scheme classifies procedures according to whether they use information within one vowel token or across vowel tokens. The added subdivision serves to classify procedures based on whether the transformed dimensions use information contained within one formant (including F_0), using 'formant-intrinsic' information, or across formants (including F_0), using 'formant-extrinsic' information. This way, a two-way classification of normalization procedures is created that has four possible combinations. My two-way classification seems more efficient than Rosner & Pickering's tripartite classification, the

[8]Rosner & Pickering (1994) argued against F_0 as a category-independent factor, because the normalization procedures proposed by Syrdal & Gopal (1986) and Miller (1989) cannot explain the recognition of whispered vowel tokens (that have no F_0).

[9]A similar hypothesis was also proposed by Johnson (1990b) and Nusbaum & Magnuson (1997). These authors claimed that listeners use auditory variables such as F_0 to learn the speaker's vowel system.

Table 2.1: *Two-way classification of the 12 normalization procedures evaluated in the present research sorted according to the type of information they employ.*

Information	*Vowel-intrinsic*	*Vowel-extrinsic*
Formant-intrinsic	HZ, LOG, BARK, MEL, ERB	GERSTMAN, LOBANOV, CLIH$_{i4}$
Formant-extrinsic	SYRDAL & GOPAL	NORDSTRÖM & LINDBLOM, MILLER, CLIH$_{s4}$

tripartite classification has one empty category, because no category-specific theories for vowel normalization were proposed.
In Table 2.1, the scale transformations LOG, BARK, MEL, ERB, and HZ are classified as vowel-intrinsic/formant-intrinsic procedures. SYRDAL & GOPAL is vowel-intrinsic because it uses information contained within a single vowel token, and formant-extrinsic because it employs information across formants: its first dimension involves F_1 and F_0 in bark, and its second dimension includes F_3 and F_2 in bark.

GERSTMAN, LOBANOV, and CLIH$_{i4}$ are classified as vowel-extrinsic/ formant-intrinsic procedures in Table 2.1. All three transformations require formant frequencies across more than one vowel per speaker, and are calculated for each formant separately[10].

NORDSTRÖM & LINDBLOM and MILLER, and CLIH$_{s4}$ are classified as vowel-extrinsic/formant-extrinsic procedures. NORDSTRÖM & LINDBLOM use a scale factor derived from the values of the mean values of F_3 for a reference speaker[11], and all formant frequencies for all vowels of a speaker are multiplied by that scale factor. MILLER uses between-formant differences on a logarithmic scale. One of these ratios includes a speaker-specific anchor point based on the mean fundamental frequency of all vowel categories of that speaker. CLIH$_{s4}$ corrects each formant frequency for a log-mean value computed using multiple formants (i.e., the mean across F_0, F_1, F_2, and F_3).

My two-way classification has consequences for the status of F_0. Some designers of the procedures treat F_0 implicitly as if it were a full formant, by using transformations like $F_1 - F_0$. Such an F_0-correction is then placed on the same level of importance as, say, a transformation of the second formant frequency. All formant-extrinsic procedures treat F_0 in this manner. Syrdal & Gopal (1986) stated that F_0 and formants may be more similar for the auditory system than they appear from their differences in production and their acoustic

[10]Nearey's original procedure (CLIH) implies using only one scale-factor based on the formant frequencies of F_1 and F_2, and is in its original form a formant-extrinsic procedure. But as is explained in section 2.4.3, in the present research I adopted two versions, one with a shared scale factor (CLIH$_{s4}$), that is therefore vowel-extrinsic/formant-extrinsic, and one with separate scale factors for each formant is adopted: (CLIH$_{i4}$), that is therefore vowel-extrinsic/formant-intrinsic.

[11]NORDSTRÖM & LINDBLOM could be seen as 'speaker-extrinsic' in addition to being vowel- and formant-extrinsic, because all values of all formants of all vowels of the speakers in a group depend on the values of a reference speaker.

definitions. Syrdal & Gopal supported their position by stating that F_0 behaves 'formant-like', because F_0 and F_1 vary systematically across vowels and by referring to Fant's (1974) and Traunmüller's (1981) studies, in which F_0 influenced vowel perception. My two-way classification thus implies treating F_0 like a formant.

2.4 Description of selected normalization procedures

This section describes the 12 normalization procedures that are compared in the present research. The description of each procedure consists of three parts: first, information is given about the hypotheses underlying the design of the procedure, second, the procedure itself is explained briefly, and third, a detailed description is given of how the procedure is implemented.

When describing each procedure, it is assumed that the procedure attempts to represent the acoustic variables in an *n*-dimensional space with axes D_0 ... D_n. Additionally, the names of the normalization procedures are displayed in superscript, e.g., D^{M} for a mel-transformation, or D^{lobanov} for Lobanov's (1971) z-score transformation. Furthermore, indices for the formant frequency (F) number are displayed in subscript (e.g., F_2). Moreover, all procedures are applied to measurements of F_0, F_1, F_2, and F_3. This has consequences for several procedures: Gerstman (1968), Lobanov (1971), as well as Nearey (1978) proposed that it suffices to apply their normalization procedures to the first two formant frequencies per vowel token. Nevertheless, as is explained in further detail in Chapter 4, these procedures are also applied to F_0 and F_3, to be able to evaluate the role of F_0 and F_3 in the normalization process as well as the role of F_1 and F_2. Finally, when a procedure's name is printed in small caps, then the procedure as implemented in the present research is referred to. For instance, 'BARK' refers to the transformation used in the present research, while 'bark' refers to the generic bark-transformation.

2.4.1 Vowel-intrinsic/formant-intrinsic procedures

HZ: the frequency scale

The frequency scale, or the hertz scale, served as the baseline scale against which all other normalization procedures were evaluated. All raw measurements of the fundamental frequency and the formant frequencies are presented in Hz. The baseline is expressed by F_0, F_1, F_2, and F_3; transformed formants are represented by D_i (e.g., D_0^{M}, D_1^{M}, D_2^{M}, and D_3^{M} for mel-transformed F_0, F_1, F_2, and F_3, respectively).

LOG: the logarithmic scale

Joos (1948) suggested using the logarithmic scale in vowel perception research. Both Miller (1989) and Nearey (1978) applied a log-transformation in their normalization procedures.

The logarithmic, or log, scale was originally presented as a scale of musical pitch: two notes whose fundamental frequencies define a particular musical interval (octave = 2:1, fifth = 3:2). The difference between log frequencies is constant for a given interval. The log scale is thought to reflect the psychological equivalent of the frequency scale (cf. Miller, 1989). In the present research, natural logarithms were used to transform a formant frequency in Hz to LOG as in equation (2.1):

$$D_i^{\mathrm{L}} = \ln(F_i) \tag{2.1}$$

MEL: the mel scale

The mel scale was derived by Stevens & Volkmann (1940) through scaling experiments that involved subjective evaluations of pitch differences by naive listeners. They asked their listeners to subdivide large pitch intervals between pairs of reference tones into four equal, smaller intervals by adjusting three variable tones. The frequencies of the reference pair varied across trials. The results of these experiments were confirmed by fractionation experiments. In these experiments, listeners had to set a variable tone to a specified fraction of the reference stimulus. The frequency of the reference stimulus was varied across trials. In the present research, a formant frequency in Hz is transformed to MEL according to equation (2.2), as described by Traunmüller (1990):

$$D_i^{\mathrm{M}} = 2595 \times \ln(1 + \frac{F_i}{700}) \tag{2.2}$$

BARK: the bark scale

Several authors selected the bark scale to represent perceived formant frequencies in their normalization procedures (such as Syrdal & Gopal, 1986). The human auditory system can be seen as composed of a series of overlapping bandpass filters, or critical bands. A critical band refers to the effective range of frequencies to which each place on the basilar membrane responds. The concept of the critical band was proposed by Fletcher (1940)[12]. Zwicker (1961) presented the relation between the frequency scale and the critical band(width) and the critical band rate in the form of a table. The critical band scale was determined from a wide variety of psychoacoustic experiments. These experiments included loudness summation, narrow-band masking, two-tone masking, phase sensitivity, threshold of complex sounds, musical consonance, and discrimination of partials in a complex tone (Scharf, 1970). The bark scale was derived from the critical band scale. The critical band scale divides the human

[12]However, Fletchers measurements were indirect and based on the (incorrect) assumption that a tone is masked by a critical band of noise of the same sound pressure level as the tone. Therefore, his results are generally referred to as critical ratios. See Moore (1977) for an overview.

auditory range into 24 units, or barks, (named after Barkhausen, who introduced the phone, the unit of loudness). One bark corresponds to 1.3 millimeter on the basilar membrane, or to approximately 150 neurons (Zwicker & Feldtkeller, 1967).

Two formulae exist for transforming Hz values to bark values. The first was proposed by Zwicker & Terhardt (1980). Their formula is used in the original version of Syrdal & Gopal's bark-difference model (1986). Zwicker & Terhardt's (1980) equation is displayed in (2.3):

$$D_i^{\mathrm{B}} = 13 \arctan \times (0.00076 F_i) + 3.5 \arctan \times (\frac{F_i}{7500})^2 \tag{2.3}$$

Traunmüller (1990) proposed the second formula: a simplified approximation to Zwicker & Terhardt's (1980). Traunmüller claimed that his formula is more accurate than equation (2.3) for frequencies between 2000 and 6700 Hz. He found that values obtained with equation (2.3) deviate from Zwicker's (1961) tabulated values as much as 0.2 bark, while values calculated with his formula deviate less than 0.05 bark. Therefore, in the present research, Traunmüller's (1990) version is adopted. Traunmüller's equation for computing BARK is displayed in (2.4):

$$D_i^{\mathrm{B}} = 26.81 \times \frac{F_i}{1960 + F_i} - 0.53 \tag{2.4}$$

ERB: the equivalent rectangular bandwidth scale

Pickering (1986) as well as Rosner & Pickering (1994) implemented an ERB-transformation in their respective vowel normalization procedure and model of vowel recognition. The ERB scale was developed by Moore & Glasberg (1983). Moore & Glasberg believed that the specifications described by Zwicker & Terhardt (1980) for the critical band are too wide and, that the critical bandwidth continues to decrease until below 500 Hz (Zwicker & Terhardt claimed that the critical bandwidth is constant below 500 Hz). Moore & Glasberg suggested a new scale based on the Equivalent Rectangular Bandwidth (ERB), which reflects their own measurements better. The ERB-transformation is similar to the bark-transformation, but it differs in two respects. First, the ERB scale resembles a purely logarithmic scale, more so than a bark scale. This holds especially for low frequencies. Second, as a consequence of the narrower critical bandwidth of the ERB rate, one ERB is smaller than one bark. In a later article, Glasberg & Moore (1990) proposed an adapted formula for the frequency-to-ERB translation, derived using another experimental methodology. This formula, displayed in equation (2.5) is used throughout the present research:

$$D_i^{\mathrm{E}} = 21.4 \times \ln(0.00437 \times F_i + 1) \tag{2.5}$$

2.4.2 Vowel-intrinsic/formant-extrinsic procedures

SYRDAL & GOPAL: Syrdal & Gopal's (1986) bark-distance transformation

Syrdal & Gopal's (1986) procedure was originally a component of their quantitative perceptual model of human vowel recognition. Syrdal & Gopal derived this model from psychoacoustics and speech perception experiments. The model is described extensively in Syrdal (1984) and Syrdal & Gopal (1986).

The model leans heavily on the spectral center of gravity effect, as described in Chistovich, Sheikin & Lublinskaya (1979). Chistovich et al. reported a critical distance between spectrally adjacent formant frequencies of approximately 3 to 3.5 bark within which the effect seems to operate. In their experiment, listeners were presented with a synthetic two-formant vowel intended to sound like one of the Russian vowels. They had to match this vowel to a single-formant synthetic vowel. If the distance was smaller than 3.5 bark, the listeners chose a frequency intermediate between the two adjacent formant frequencies. If the distance between the reference formants exceeded 3.5 bark, the listeners chose a formant that matched one of the two formants, but not an intermediate frequency. Chistovich et al. suggested that there may be a threshold or critical distance between 3 and 3.5 bark for the integration of formants, and that this critical distance between the first and second formant is constant in barks. The spectral center of gravity effect also implies that the critical distance may remain the same over a wide range of frequencies and between other spectrally adjacent formants (e.g., between the third and fourth formant). Chistovich et al. further suggested that this kind of perceptual integration may help in perceptual vowel classification, especially for the distinction between front and back vowels.

Syrdal & Gopal's transformation incorporates Chistovich et al.'s finding by calculating the formant distances in bark between F_2 and F_1, between F_3 and F_2, and F_4 and F_3. They extended Chistovich et al.'s concept by including the distance between F_1 and F_0. Subsequently, they investigated whether these distances between formants (and D_0) can be used to devise a binary feature system; distances between adjacent formants are classified on whether they exceed the critical distance (of three barks) or not.

Syrdal & Gopal investigated whether vowels can be classified using this binary feature system. They compared the three features (i.e., the three distances ($D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$, $D_2^{\mathrm{B}} - D_1^{\mathrm{B}}$, $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$)) with raw F_0, F_1, F_2, and F_3 in Hz. They used a series of linear discriminant analyses to evaluate how effectively the transformed data can be classified into the correct vowel categories. The distances between the three formant pairs served as predictors, while the percentage of vowel tokens classified into the corresponding vowel category served as the dependent variable. They found that the transformed data results in higher percentages correctly classified vowel tokens than the untransformed data.

Syrdal & Gopal designed their procedure as follows. They first transformed vowel data described in Peterson & Barney (1952) to bark (D_0^{B}, D_1^{B}, D_2^{B}, D_3^{B}), using the formula by

Zwicker & Terhardt (1980), as displayed in equation (2.3), and using the end-correction proposed by Traunmüller (1981). Second, they calculated all the distances between adjacent formants in bark for all the vowel tokens ($D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$, $D_2^{\mathrm{B}} - D_1^{\mathrm{B}}$, and $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$). Third, they classified the vowel tokens according to the distance criterion using a binary feature matrix. In this feature matrix, the vowel /a/ was, for instance, labeled as having a $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ distance larger than three bark, and with a $D_2^{\mathrm{B}} - D_1^{\mathrm{B}}$ distance smaller than three bark, and a $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$ distance that was larger than three bark.

In their normalization model, Syrdal & Gopal stated that the $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ dimension represents the open-close dimension best, and that the $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$ dimension represents the front-back dimension best. They reasoned that the $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ dimension represents a continuum of high to low vowels, in which high vowels have $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ differences less than three barks, while mid and low vowels have $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ differences greater than three barks. The $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$ dimension represents a continuum of front to back vowels; front vowels have $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$ differences less than three barks and back vowels have $D_3^{\mathrm{B}} - D_2^{\mathrm{B}}$ differences greater than three barks[13].

However, other authors suggested a combination such as $F_2 - F_1$ (Fant, 1973; 1982), or $F_2 - F_0$ (Hirahara & Kato, 1992), to account for the front-back dimension. Fant (1982) reported that all Swedish vowels can be classified into front and back vowels irrespective of the question whether the distance between the first and second formant of these vowels exceeds three bark or not. To account for this, Syrdal & Gopal claimed that the choice of the critical distance for the front-back dimension is language-specific: the $D_2 - D_1$ distance is not a language-universal measure reflecting front-back vowel distinctions. In addition, they argued that, while all Swedish vowels can be classified using $D_2 - D_1$, American English vowels cannot, because some of the American English vowel categories (/æ/ and /u/) have D_1 and D_2 values that show substantial overlap.

In the present research, the two dimensions of SYRDAL & GOPAL are calculated using equations (2.6) and (2.7). The transformation to barks is performed with equation (2.4).

$$D_1^{\mathrm{s\&g}} = D_1^{\mathrm{B}} - D_0^{\mathrm{B}} \tag{2.6}$$

$$D_2^{\mathrm{s\&g}} = D_3^{\mathrm{B}} - D_2^{\mathrm{B}} \tag{2.7}$$

[13]Note that the claim, that $D_3 - D_2$ distance is a language-universal measure reflecting front-back vowel distinctions, is not expected to be valid for languages that have rounded front vowels (such as /y/ and /ʏ/ in Dutch, French, German, or Swedish). In these languages, the distance between D_2^{B} and D_3^{B} exceed three bark. This means that vowels such as /y/ and /ʏ/ are likely to be confused using Syrdal & Gopal's binary classification.

2.4.3 Vowel-extrinsic/formant-intrinsic procedures

GERSTMAN: Gerstman's (1968) range normalization

Gerstman (1968) aimed to establish the number of dimensions necessary to classify a single speaker's vowels and their specification. He argued that such a classification is required for a (hypothetical) speech recognition system, so that this system can learn the speaker's vowel system. In addition, Gerstman wanted to establish how many calibration vowels the speaker has to produce in order to obtain a specification that can be used by the recognition system.

Gerstman described a procedure that consists of re-scaling the frequencies of the first and second formants. First, the lowest and highest formant frequencies for each speaker are measured across vowels. Second, these values are set to 0 and 999, respectively. Third, all other values are scaled linearly between the two extremes.

He designed a simple pattern classifier to compare the classification of the transformed formant frequencies with the classification of untransformed formant frequencies for the vowel data produced by the 76 speakers (33 male speakers, 28 female speakers, and 15 children) described by Peterson & Barney (1952). Gerstman concluded that two dimensions are sufficient for classifying vowels of a single speaker. These two dimensions correspond to the scaled F_1 and the scaled F_2. In addition, he stated that it is not necessary to use tokens from all vowel categories of a speaker in order to derive the maximum and minimum values for F_1 and F_2; the point vowels /i, ɑ, u/ suffice. Gerstman presumed that /i/ can provide minimum F_1 and maximum F_2, /ɑ/ can provide maximum F_1, and /u/ can provide minimum F_2.

In the present research, the minimum and maximum values for a speaker are calculated using tokens from all nine monophthongal vowel categories of Dutch for each speaker. Although Gerstman proposed to use only the standardized values for the first two formants, I used the standardized values for the fundamental frequency and the first three formants. GERSTMAN was calculated as described as in equation (2.8).

$$D_i^{\text{gerstman}} = 999 \times \frac{F_i - F_i^{min}}{F_i^{max} - F_i^{min}} \tag{2.8}$$

In equation (2.8), F_i is the formant frequency value (where $_i$ = 0, 1, 2, or 3) to be scaled. F_i^{min} is the minimum value of all the formant frequencies for all the vowels for a speaker, while F_i^{max} is the maximum frequency for all the formant frequencies across all vowels for that speaker.

LOBANOV: Lobanov's (1971) z-score transformation

Lobanov (1971), like Gerstman, aimed at vowel normalization for automatic speech recognition purposes. He suggested that inter-speaker differences could be minimized through re-scaling, using the center of each speaker's vowel space and the average dispersion from

that center. He claimed that only the first two formants per vowel are necessary for the normalization. Lobanov was not as concerned with using a minimum number of vowels for a single speaker as Gerstman (1968). Lobanov stated that, in order to transform all values per speaker to their z-scores, all formant frequencies for all the vowels of that speaker have to be included.

The acoustic variables in this thesis are transformed to Lobanov's transformation as shown in equation (2.9).

$$D_i^{\text{lobanov}} = \frac{F_i - \mu_i}{\sigma_i} \tag{2.9}$$

In equation (2.9), the mean formant frequency μ_i is calculated using all the vowels of a speaker for a formant *i*, while σ_i refers to the standard deviation.

CLIH$_{i4}$: Nearey's (1978) individual log-mean model

The rationale behind Nearey's transformation (1978) is that each speaker's vowel space is located on a logarithmic F_1 by F_2 plot by reference to one reference vowel. Nearey claimed that a minimum of two vowels for a given speaker must be known in order to use his procedure.

Nearey's procedure is also known as the Constant Log Interval Hypothesis (CLIH). This hypothesis states that corresponding formants of corresponding vowels for different speakers stand in constant ratios to each other and can be transformed using a multiplication by this ratio. After this transformation, the formant frequencies coincide. Nearey thus adopted the idea that the ratios of formant frequencies are more relevant for vowel recognition than the actual frequencies of those formants. This idea is commonly known as the 'constant ratio hypothesis', which was first formulated by Lloyd (1890a; 1890b; 1891), and was also described, for instance, by Chiba & Kajiyama (1941), Potter & Steinberg (1950), and Peterson (1961). NORDSTRÖM & LINDBLOM used this hypothesis as well, as is described in section 2.4.4.

Nearey's procedure consists of expressing each log-transformed formant frequency as a distance to a reference point, the log-mean. Nearey proposed two different ways to calculate this reference log-mean, defined in the present research as either the shared log-mean (CLIH-$_{s4}$: the index $_{s4}$ indicates that the shared log-mean is calculated over four dimensions, i.e., F_0, F_1, F_2, and F_3) or the individual log-mean (CLIH$_{i4}$: the index $_{i4}$ indicates that four individual log-means are calculated, one for each dimension). The individual log-mean is described here. The shared log-mean is described as a vowel-extrinsic/formant-extrinsic procedure.

Whenever I refer to CLIH$_{i2}$, the mean value of the (log) formant frequency for F_1 (D_1^{L}) and the mean value of F_2 (D_2^{L}) is referred to. For CLIH$_{i3}$, the log-mean per formant is based on the (log-transformed) mean values for F_1, F_2, and F_3, and for CLIH$_{i4}$, the log-mean per formant is based on the (log-transformed) mean values for F_0, F_1, F_2, and F_3. The formula for the single log-mean is displayed in (2.10).

$$D_i^{\text{clihi4}} = D_i^{\text{L}} - \mu_{D_i^{\text{L}}} \tag{2.10}$$

For the individual log-means, the second index of CLIH_{i4} is 4, because in my research, three formants plus the fundamental frequency are calculated. Each log-mean is calculated on the formant frequency values (or fundamental frequency) for all vowels per speaker.

Formally, CLIH_i as suggested in Nearey (1978) is identical to the "centered" procedure suggested five years earlier by Pols, Plomp & Tromp (1973). Pols et al. described a normalization procedure that would be written as CLIH_{i3} in my notation. However, throughout this thesis, I use Nearey's terminology (i.e., 'CLIH' instead of 'centered') to refer to the normalization procedure in (2.10), because all the studies that I describe in Section 2.5 refer to Nearey's (1978) procedure.

2.4.4 Vowel-extrinsic/formant-extrinsic procedures

CLIH: Nearey's (1978) shared log-mean model

The rationale behind the shared log-mean (CLIH_{s4}) is nearly identical to the rationale behind the individual log-mean (CLIH_{i4}). The only difference is that CLIH_{s4} uses one overall log-mean. Nearey's CLIH_{s4} thus uses a shared log-mean that is the grand mean of the log-means of all formant frequency numbers.

The shared log-mean used in the present research is based on four log-means, because the acoustic representation per vowel token consists of measurements of F_0, F_1, F_2, and F_3. Therefore CLIH_{s4} is the mean value of $\mu_{D_0^{\text{L}}}$, $\mu_{D_1^{\text{L}}}$, $\mu_{D_2^{\text{L}}}$, and $\mu_{D_3^{\text{L}}}$. Each log-transformed formant frequency value (D_0^{L}, D_1^{L}, D_2^{L}, and D_3^{L}) is then expressed as its distance to the shared mean (CLIH_{s4}).

The version of CLIH_{s4} as used in the present research is presented in formula (2.11), where $_i$ should be interpreted as if referring to either F_0, F_1, F_2, or F_3.

$$D_i^{\text{clihs4}} = D_i^{\text{L}} - (\mu_{D_0^{\text{L}}} + \mu_{D_1^{\text{L}}} + \mu_{D_2^{\text{L}}} + \mu_{D_3^{\text{L}}})/4 \tag{2.11}$$

NORDSTRÖM & LINDBLOM: Nordström & Lindblom's (1975) vocal tract transformation

Nordström & Lindblom (1975) sought to improve vowel categorization through the use of a scale factor calculated using the average F_3. Nordström & Lindblom's approximation is that speaker-specific variation originating from anatomical differences between male and female speakers can be accounted for solely in terms of vocal tract length. They argued that because the vocal tract length variations affect all formant frequency values uniformly, the same scale factor should be applied to all formants. Their procedure is therefore a uniform scaling

procedure. The procedure consists of first estimating the speaker's vocal tract length from the average of the third formant for open vowels, that is, vowels with an F_1 greater than 600 Hz for the male and the female speakers in the speaker groups to be compared. Second, because the length of a speaker's vocal tract is inversely related to formant frequency, the average F_3 can be used to estimate the average vocal tract length of the male and the female speakers in the speaker group. Third, the scale factor k in Nordström & Lindblom's model is calculated using equation (2.12). It expresses the ratio of the length of the average female vocal tract (L^{female}) to the length of the average male vocal tract (L^{male}). In equation (2.12), $\mu_{F_3}^{\text{male}}$ and $\mu_{F_3}^{\text{female}}$ are the mean values for the third formants of the open vowels for the male and female speakers, respectively. Fourth, any formant frequency for a female speaker (F_i^{female}) can be scaled by multiplying it with k as in (2.13).

In the present research, the scale factor k is calculated as in equation (2.12), using the frequency of F_3 of all vowel tokens with an F_1 greater than 600 Hz of all speakers in the speaker groups to be compared. Subsequently, the values for the female speakers were scaled as in (2.13). The values for the male speakers were left unchanged.

$$k = \frac{L^{\text{male}}}{L^{\text{female}}} = \frac{\mu_{F_3}^{\text{male}}}{\mu_{F_3}^{\text{female}}} \tag{2.12}$$

$$D_i^{\text{n\&l}} = kF_i^{\text{female}} \tag{2.13}$$

MILLER: Miller's (1989) formant ratio transformation

Miller's model (1989), like SYRDAL & GOPAL and CLIH_{i4} and CLIH_{s4}, was based on the constant ratio hypothesis. Miller proposed to use the distance between adjacent log-transformed formant frequencies to acoustically represent vowel tokens. He evaluated several scale transformations: bark, mel, log, and Koenig (a modified logarithmic scale by Koenig, 1949). He argued that, while all scale transformations perform approximately equal, the log scale can be compared directly with other types of sounds (e.g., musical). Additionally, he claimed that the log scale allows the ratios between the formants to be expressed as distances. He referred to these rations as 'sensory formants'.

Miller defined three sensory peaks (D_1, D_2, and D_3) corresponding to the log-transformed first three formants. He proposed representing a vowel in three dimensions, consisting of the distance between, first, the log-transformed first formant and an F_0-based speaker-specific anchor point, second, between the log-transformed second formant and the first log-transformed formant, and third, between the log-transformed third formant and the log-transformed second formant. Miller's model puts considerable emphasis on the first and second formants; they are both included twice.

The normalizing component of MILLER is expressed by the Sensory Reference (SR) (see equation (2.14)). SR serves as a speaker-specific anchor point. Miller hypothesized that

listeners use SR to judge the position of the first three formants. In the present research, the SR is calculated using the geometric mean of all values of the F_0 for a speaker, and corrected for a constant (k in equation (2.14)). Miller suggested to use a value for k of 168 Hz, (the geometric mean of 125 Hz and 225 Hz, which he adopts for the average male and female F_0, respectively). In the present research, I calculate the value of the constant k using the observed mean for the F_0 for the male and female speaker groups that are compared. MILLER is implemented as in equations (2.14), (2.15), (2.16), and (2.17).

$$\text{SR} = k(\mu_{D_0^{\text{L}}}/k)^{\frac{1}{3}} \tag{2.14}$$

$$D_1^{\text{miller}} = (\frac{D_1^{\text{L}}}{\text{SR}}) \tag{2.15}$$

$$D_2^{\text{miller}} = (\frac{D_2^{\text{L}}}{D_1^{\text{L}}}) \tag{2.16}$$

$$D_3^{\text{miller}} = (\frac{D_3^{\text{L}}}{D_2^{\text{L}}}) \tag{2.17}$$

2.5 Literature on normalization procedures

This section discusses the design and the results of previously published studies. This was done in order to assess how well the 12 selected procedures performed at the criterion proposed in Chapter 1 of the present research: the procedure must preserve sociolinguistic speaker-related variation and phonemic variation, while minimizing variation in the acoustic representation related to the speaker's anatomical/physiological characteristics. In addition, the literature study is carried out as well in order to get an idea about the methodologies that were previously employed to evaluate normalization procedures.

Hindle (1978)

Hindle (1978) compared three normalization procedures with raw data in Hz. His goal was to establish which procedure could reduce speaker-specific variation so that "...different speakers' versions of the same phoneme coincide in the normalized system", while the procedures must also meet a 'sociolinguistic criterion'. This sociolinguistic criterion is similar to the criterion used in the present research.

Hindle tested the vocal tract length procedure developed by Nordström & Lindblom (1975), Nearey's (1978) CLIH$_{s2}$ procedure (with a shared scale factor based on the mean

frequency values of F_1 and F_2 across all vowels for each speaker), and a procedure from Sankoff, Shorrock & McKay (1974)[14].

Hindle performed the comparison by applying the procedures to two sets of vowel data, in two tests. The first data set was obtained through a sociolinguistic interview, with speakers from Philadelphia. Hindle used conversational speech data from nineteen speakers (10 female and 9 male speakers). Hindle established how well each procedure meets his sociolinguistic criterion by evaluating how well a well-known age-difference effect is preserved in the normalized acoustic representations of his vowel data. This age-difference concerned a difference in the realization of the American English diphthong "/ay^0/"[15].

In addition, Hindle obtained perceptual descriptions of the vowel data. These descriptions consist of narrow phonetic transcriptions provided by phonetically-trained listeners. These perceptual descriptions indicate an age effect: older speakers produce the diphthong with a nucleus close to /a/, while the younger speakers' nucleus is closer to /ə/. This age effect should be reflected in the acoustic representation by a higher first formant frequency for the older speakers.

Hindle compared the performance of the three procedures through linear regression analysis. In each analysis, the transformed frequency of the first formant was regressed on the chronological age of the speakers. It was found that all three procedures normalized away some of the sociolinguistic variation, because only a very small age effect was found after normalization, as compared with his baseline (the raw data in Hz). Only the effects for Nearey's CLIH_{i2} and Nordström & Lindblom's scale transformation procedures show significant age effects. Hindle found these two effects to be slightly larger than those of the procedure developed by Sankoff et al.[16]. Finally, Nearey's procedure reveals slightly more sociolinguistic variation than Nordström & Lindblom's.

The second data set that Hindle tested was obtained through a random survey of the Philadelphia community. Telephone recordings were made of interviews with 60 informants. Because telephone speech usually does not allow a third formant to be measured reliably, Nordström & Lindblom's procedure (which uses the third formant to estimate its scale factor) was excluded. Otherwise, the same procedures were used as in the first study and the same age-difference effect (a difference in the realization of the American English diphthong '/ay^0/' by a younger and an older group of speakers) was studied. The same pattern in the results was found, Nearey's CLIH_{i2} procedure performs slightly better than Sankoff's procedure, although some sociolinguistic variation is minimized compared with the baseline. The effects for the age difference were significant for both procedures.

[14] Sankoff et al.'s procedure is not evaluated in the present research for reasons discussed in section 2.2.

[15] According to Hindle, this vowel occurs in American English words such as 'fight'.

[16] It is not clear whether Hindle found a significant effect for Sankoff et al.'s procedure.

Nearey (1978)

Nearey (1978) compared two versions of his own procedure, CLIH_{s2} and CLIH_{i2} (using F_1 and F_2) with two other procedures: Gerstman's (1968) range-normalization, and Lobanov's (1971) z-transformation. He evaluated these four procedures by applying them to the formant frequencies of Peterson & Barney's vowel data (1952).

Each normalization procedure's 'resolving power' was estimated. Nearey defined this resolving power as "...the ability to allow the separation of normalized formant frequencies into distinct groups corresponding to phonetic categories.". Nearey thus evaluated the procedures on how well they minimize all speaker-related variation, while preserving the phonemic variation. He applied several measures of resolving power. Only the first measure is discussed here: the percentage of vowel tokens that are correctly identified. Nearey calculated this percentage by drawing (curved) boundaries between the vowel categories in the scatter plots by hand. The results show that Lobanov's procedure is the most powerful, 94%, followed by Nearey's procedures CLIH_{i2} (using an individual log-mean for F_1 and F_2) and CLIH_{s2} (shared log-mean calculated across μ_{F_1} and μ_{F_2}), 93% and 90%, respectively. Gerstman's procedure performs poorest (89%). Nearey also performed a Linear Discriminant Analysis, and found the same pattern in the results, i.e., Lobanov's procedure performs best, followed by CLIH_{i2} and CLIH_{s2}, and Gerstman's procedure performs poorest.

Disner (1980)

Disner (1980) compared four procedures with raw data in Hz: Gerstman's range normalization (1968), Lobanov's z-transformation (1971), Nearey's (1978) CLIH_{s2} procedure (based on the mean value of μ_{F_1} and μ_{F_2}), and Harshman's (1970) PARAFAC model[17].

She compared the four procedures in two ways. She first applied them to recordings of vowel data from six Germanic languages: English (Peterson & Barney, 1952), Norwegian (Gamnes, 1965), Swedish (Fant, Hennigson & Stålhammer, 1969), German (Jørgensen, 1969), Danish (Fischer-Jørgensen, 1972), and Dutch (Pols, Plomp & Tromp, 1973). Disner evaluated the percentages of 'scatter reduction' per procedure for each language. The procedures were thus evaluated on how well they produced an acoustic representation in which all phonemic variation was preserved, while all speaker-related variation was eliminated. Disner accomplished this by plotting the acoustic representations of the vowel tokens in ellipses that cover 95% of the variance per vowel category (following Labov, Yaeger & Steiner, 1972, and Davis, 1976) in a formant space with F_1 along the ordinate and F_2 along the abscissa on a mel scale, before and after applying the normalization procedures. The scatter reduction was measured by comparing the surface of the areas covered by the ellipses covered before and after normalization, per vowel. The resulting improved clustering for all vowel categories

[17] Not discussed or evaluated in the present research because it was not documented well enough to implement.

was expressed relative to the baseline situation (raw data in Hz). For instance, after applying Nearey's CLIH_{s2} on the German data, 30% of the variance remained, compared with the baseline data (set at 100%). Nearey's procedure thus reduces the variance in the scatter with 70%, compared with the raw data. The results from this comparison show that, although no specific procedure is the most effective for all the languages, Nearey's CLIH_{s2} procedure is generally the most effective. Lobanov's procedure is slightly less effective than Nearey's, followed by Gerstman's range normalization.

Disner's second evaluation was a comparison of the procedures' output to impressionistic data to see how well the procedures preserve linguistic differences in the vowels of the different languages. She gathered statements from trained listeners about the linguistic characteristics of the vowels of the six languages from the literature. She argued that these statements are relatively reliable indicators for the linguistic qualities of vowels across languages, because she encountered examples where different phonetically-trained listeners agree independently on the relative qualities of Germanic vowels. An example of such an statement is: the Dutch /ɛ/ is more open than the English vowel in 'bed'; intermediate between 'set' and 'sat' (Koolhoven, 1968). She proposed that the normalized acoustic representation of the vowel token must reflect the trends in the auditory impressions.

She found that most procedures reflected the linguistic differences between the languages poorly, and in some cases even reversed them. In addition, the procedures that perform best at improving clustering, reduce differences more severely than the ones that are less effective at clustering.

Disner concluded that, in order to be able to compare procedures for vowel normalization across languages or dialects, the data sets must be fully phonologically comparable, because implicit assumptions about the underlying vowel system are made when languages are compared using the normalized vowel frequencies. These assumptions concern the means and the standard deviations for the vowel categories. Some procedures are only appropriate for cross-language comparisons when the languages have comparable means, while others demand that the languages have comparable standard deviations. Procedures as proposed by Lobanov (1971) and Gerstman (1968) use scaling factors based on all vowels of the vowel inventory of the speaker's language or dialect. However, because different language or dialects can have different vowel phonemes in their inventory, the scaling factors can be expected to differ across languages or dialects. Comparing normalized speech from two dialects can thus result in using different scaling factors to normalize vowel tokens, which can subsequently result in artificial differences, or the deletion of differences, in vowel quality between and within the vowels of speakers from two dialect groups.

Syrdal (1984)

Syrdal evaluated eight normalization procedures. These were the log-transformation, the bark-transformation, Syrdal's bark-difference model (1984), two versions of Miller: Enge-

Table 2.2: *Percent correctly classified vowel tokens for the two discriminant analyses performed on the Peterson & Barney (1952) data. Based on Syrdal's Table 2 (1984).*

%	LDA 1: Vowel (10)	LDA 2: Speaker Groups (3)
Baseline (Hz)	82.3	89.6
Log	86.8	89.0
Bark	85.9	88.0
Bark-difference	85.9	41.7
Miller 1	79.2	62.2
Miller 2 (F_0 corr.)	84.5	33.3
CLIH_{s4}	88.9	35.1
CLIH_{i4}	91.4	33.3
Gerstman	86.9	33.3

bretson & Vemula's log-ratio model (1980), as in equation (2.15) and one with a corrected male F_0, two versions of Nearey's (1978) CLIH_{s4} model (one log-mean factor per speaker; based on the mean of μ_{F_0}, μ_{F_1}, μ_{F_2}, and μ_{F_3}. In addition, she evaluated CLIH_{i4}, with one individual log-mean factor for each of the three formants and for the fundamental frequency, and finally Gerstman's range normalization (1968), which was also applied at the three formants frequencies and the fundamental frequency. All these procedures were compared with the baseline procedure (F_0, F_1, F_2, and F_3 in Hz).

Syrdal applied the procedures to the Peterson & Barney (1952) data set. She carried out two series of linear discriminant analyses (LDAs). In the first series, the procedures were evaluated on how well they produced output that allowed the vowel tokens to be classified correctly into the corresponding vowel category. The second series of LDAs can be seen as the complement of the first; in this LDA the percentage of vowel tokens that were grouped into the correct speaker group (male, female, child) was estimated. Syrdal stated that a good procedure must eliminate the speaker-specific variation. A low percentage (at chance level) for the results from this LDA indicates that certain speaker-specific variation is eliminated.

The results for both LDAs from Syrdal's Table 2 (page 129) are replicated here in Table 2.2. Nearey's CLIH_{i4} procedure meets both requirements best, with the highest score for LDA 1 and the lowest for LDA 2. Syrdal's bark-difference procedure is not as powerful as Nearey's CLIH_{i4} for LDA 1, but it is more powerful than several other procedures at normalizing away speaker-specific variation.

Deterding (1990)

Deterding (1990) studied vowel normalization from the perspective of automatic speech recognition. He evaluated normalization procedures on how well they minimize all speaker-related variation (sociolinguistic as well as anatomical/physiological) and preserve the phonemic variation.

Deterding compared a broad range of procedures, of which eight were formant-based. These procedures include Gerstman 1 (calculated using all the vowel tokens for a speaker to calculate the scale factors) and Gerstman 2 (using only vowel categories /ɑ/ and /i/ for each speaker). Furthermore, Lobanov (1971), Nearey's CLIH_{i2} (1978), Pickering (1986)), and four scale transformation procedures (log, mel, bark, ERB), as well as the baseline (Hz)[18].

Deterding used a set of vowel data for his comparisons consisting of speech of eight females, eight males, and two children, who all spoke the same variety of English (Standard Southern British). All speakers pronounced one token of the 11 monophthongal vowels of British English in a /hVd/ context. He calculated transformed values for all procedures for the first two formant frequencies for each vowel token.

Deterding used a similar approach for comparing Gerstman's (1968), Lobanov's (1971), and Nearey's (1978) CLIH_{s2} procedure. He transformed the frequencies for the first two formants of his vowel data according to each procedure. Subsequently, he derived 'template formant frequencies ' by averaging all the transformed values for each vowel category for the male speakers, the female speakers, and for the pooled adult speakers (male and female speakers, not including the children). He evaluated the transformed values for each individual speaker by comparing them to these templates. For instance, for each speaker the values for Gerstman were calculated using formula (2.8), and subsequently the mean value per transformed formant (for D_1^{gerstman} and D_2^{gerstman} separately) was calculated using all vowels of the adult speaker group, for the male speakers separately and for the female speaker separately. After computing the templates, a simple pattern classification procedure was carried out, in order to obtain the percentage of correctly classified vowel tokens. The procedure consists of determining the closest template (i.e., mean value) in a D_1 by D_2 plot for each vowel token. He carried out this classification procedure for all speaker groups (including children).

Table 2.3 shows the percentages correctly classified for *Gerstman 1*, *Gerstman 2*, *Lobanov*, and CLIH_{s2} as reported by Deterding (1990). It can be seen that, overall, tokens from male speakers can be classified best using the male templates, tokens from female speakers and

[18]Deterding evaluated the following procedures as well: auditory whole-spectrum approaches (Bladon & Lindblom, 1981; Bladon, 1985), scale factors based on vocal-tract length (Wakita, 1977), a procedure for broad-band spectral integration (Perceptually-based Linear Prediction, or PLP, Hermansky, 1985a, 1985b), and finally Pickering's (1986) centroid procedure. These procedures are not discussed here, for reasons given in section 2.2. Deterding reported that the formant-based procedures generally perform better than the whole-spectrum approaches in the tests he carried out.

Table 2.3: *Percent correctly classified vowel tokens for the five procedures compared by Deterding (1990). Based on Deterding's tables 6.5, 6.7, 6.9, and 6.11.*

%	Templates			
	Speakers	All adult	Male	Female
Gerstman 1	male	76.1	90.1	56.8
	female	73.9	59.1	80.7
	child	45.5	36.4	45.5
	overall	71.7	70.7	66.2
Gerstman 2	male	85.2	93.2	71.6
	female	79.5	80.7	75.0
	child	36.4	36.4	40.9
	overall	77.3	81.3	69.7
Lobanov	male	96.6	96.6	79.5
	female	92.0	80.7	94.3
	child	63.6	50.0	59.1
	overall	90.9	84.3	83.8
CLIH_{s2}	male	97.7	96.6	79.5
	female	88.6	81.8	88.6
	child	45.5	45.5	50.0
	overall	87.9	84.3	80.3

children are classified best using female templates. In general, the scores are highest for Lobanov, followed by CLIH_{s2}. Gerstman's data shows the lowest percentages correctly classified for the overall data (from the pooled male and female data). However, using only two vowel categories to estimate a speaker's vowel space (Gerstman 2) substantially improves the recognition percentages. Deterding concluded that Lobanov's normalization procedure yields the most successful classification of his data, followed by Nearey's CLIH_{s2} procedure, while Gerstman's (Gerstman 1) procedure is least successful.

In addition, Deterding compared five scale transformation procedures: Hz, log, mel, bark, and ERB. He applied the four procedures – Gerstman (1968), Lobanov (1971), CLIH_{s2}, Nearey (1978) – to data in bark, in log, and in ERB, in mel, and in Hz[19]. This procedure resulted in combinations such as Gerstman applied to ERB-transformed formant frequencies. Deterding used a similar classification task such as the one described earlier: he classified each transformed vowel token by selecting its nearest vowel template for the male speakers,

[19]Deterding presumably applied the scale transformations before transforming the data according to Gerstman's, Lobanov's, or Nearey's CLIH_{s2} procedure.

Table 2.4: *Percent correctly classified vowel tokens for the five scale transformation procedures compared by Deterding. Calculated using tables 7.1, 7.2, 7.3, and 7.4 in Deterding (1990).*

%	Templates			
	Speakers	All adult	Male	Female
Hz	male	79.3	90.9	57.9
	female	77.2	63.9	83.0
	child	40.0	34.6	43.6
Log	male	89.1	90.1	74.3
	female	84.1	74.1	86.1
	child	41.8	37.3	42.7
Mel	male	85.7	94.1	67.7
	female	81.4	70.0	83.9
	child	44.5	36.4	46.4
Bark	male	87.5	94.6	70.7
	female	83.4	72.7	85.7
	child	42.7	39.1	42.7
ERB	male	87.2	95.3	72.0
	female	84.3	73.9	85.9
	child	41.8	38.2	44.5

the female speakers, and the children. This was repeated for the vowel templates for the adult speakers, the male and female templates, and for each combination of scale transformation procedure and Gerstman's, Lobanov's, or Nearey's CLIH_{s2} procedures. The pooled results for all the combinations are displayed in Table 2.4. I calculated these values myself, using the results in his Tables 7.1, 7.2, 7.3, 7.4. Any value in Table 2.4 is the mean of four combinations (e.g. the mean for ERB is calculated across *Gerstman 1* in ERB, *Gerstman 2* in ERB, *Lobanov* in ERB, and CLIH_{s2} in ERB) This was necessary, because Deterding presented mean values in which the performance of procedures such as Pickering (1986), which I do not discuss in this research, were included as well. Each percentage in Table 2.4 is the average for the percentages for the combination of Gerstman's (two procedures), Lobanov's, or Nearey's CLIH_{s2} procedure and each scale transformation. For instance, in his Table 7.1 (page 158) Deterding presents the result for the following five combinations for the male speakers classified using the adult templates: no normalization×Hz: 65.2%, Gerstman (all vowels) × Hz: 76.1%, Gerstman (based on /i, ɑ/) × Hz: 85.2%, Lobanov×Hz: 96.6%, Nearey×Hz: 78.4%.

Deterding concluded that the log scale performs best at his classification task. Note that, this result cannot be observed in Table 2.4. Using my calculations, bark shows the same mean as log, and ERB shows the highest percentage. This difference between my calculations and Deterding's results is most likely due to the fact that the results for Pickering's procedure (1986) were not included in my calculations.

In my opinion, Deterding's results suggest that the performance of all normalization procedures he tested is sub-optimal. If the procedures effectively minimize the anatomical/physiological differences between speaker groups, the percentages correctly classified vowel tokens must be the same across speaker groups, across all templates (adult, male, female). Since the speakers were controlled for their sociolinguistic characteristics, no considerable sociolinguistic differences between groups could be present in the data. This indicated that all speaker-related variation is of an anatomical/physiological nature, and that some of this variation is preserved after normalization. The results show decreased percentages for nearly every procedure whenever the template (adult, female, or male) do not match the speaker data (male, female, or child)[20].

Nearey (1992)

Nearey compared three scale transformations: bark, ERB and log, Syrdal & Gopal's model (1986), Miller's (1989) transformation, and two versions of his own procedure: CLIH_{s3} and CLIH_{i3}. Nearey used Peterson & Barney's (1952) data set. He used generalized linear modeling to evaluate and model the transformations.

He restricted his evaluation to measurements of the (normalized) first formant, because the scale transformations for bark and ERB are both nearly logarithmic in the region above the highest first formant (around 700 Hz), so not much difference between the scale transformations was expected for the higher formants.

In the regression analyses, the criterion variable was the first formant of the first token of each vowel produced by the speakers in Peterson & Barney's data set (all 76 speakers had to produce two tokens of each vowel). Nearey evaluated how well this first token could be predicted through the transformed values of the second token for that vowel category for that speaker. Effectively, Nearey tested the regression equation that is displayed in equation (2.18). In equation (2.18), T_{1sv1} is the transformed measurement for the first formant for the first token (last subscript) of vowel *v* by speaker *s*. $T_{1s.2}$ is the mean transformed value for the F_1 for a speaker, averaged over all vowel categories of the *second token*. The factors $T_{1s.2}$, $T_{0s.2}$, and T_{0sv1} correspond to the extrinsic formant, extrinsic fundamental, and intrinsic fundamental information, respectively. In other words, Nearey regressed the transformed formant frequencies of one repetition of a vowel by a certain speaker, plus additional intrinsic

[20] In addition, his results may partially be due to the pattern classification procedure used. In this procedure, he did not take into account the (co)variance of the groups that he compared.

and extrinsic factors onto the same vowel uttered by the same speaker on a second occasion and evaluates the strength of the correlations between the predictor and the criterion variables.

$$T_{1sv1} = a_v + b_1 T_{1s.2} + c_1 T_{0s.2} + d_1 T_{0sv1} \tag{2.18}$$

Nearey used analysis of covariance to obtain estimates of the coefficients for the regression model. The magnitudes of these coefficients indicate the relative importance of different sources of information in the regression model. The normalization procedures of Syrdal & Gopal's (1986), Miller's (1989), and two versions of his own procedure – CLIH_{s3} and CLIH_{i3} (1978) – all make different predictions about the relative importance of the coefficients.

Certain combinations of the coefficients resulting from the analysis of covariance are compatible with predicted magnitudes of the coefficients that could be made for Syrdal & Gopal, Miller, or Nearey. For Nearey's it is predicted that b_1 = 1, c_1 = 0, d_1 = 0, for Miller's model this is b_1 = 0, c_1 = 0.333, d_1 = 0 (cf. Nearey, 1992, page 583 for Nearey's interpretation of Miller's (1989) procedure). For Syrdal & Gopal, the predicted magnitudes are b_1 = 0, c_1 = 0, d_1 = 1.

When evaluating the resulting models, Nearey found that his regression model did not allow him to decide which of the three scale transformations (log, bark, or ERB) is the best option. However, after invoking extra tests to study systematic patterns in the residuals of the bark, log, and ERB functions (Hinkley's test, as described in McCullagh et al., 1989), he found that the log scale is the best option, because its residual contains no systematic patterns. Nevertheless, the differences between the three scale transformations are small. In addition, he found that the only coefficient that had substantial weight in the analysis was the coefficient for the speaker's average mean F_1 (b_1), suggesting that Nearey's own CLIH_3 may be the best option (because this method consisted of correcting each value for D_1^{L} for the log-mean value for μD_1^{L} based on all other vowels for a speaker).

2.6 Conclusions

This chapter describes the normalization procedures that are compared in the present research in detail. In addition, six formerly published studies that compared normalization procedures were discussed. This was done to get a preliminary idea about how well the normalization procedures perform at the criterion proposed in the present research (minimizing anatomical/physiological speaker-related variation while preserving sociolinguistic speaker-related and phonemic variation). It appears that applying Lobanov's procedure or Nearey's CLIH_i[21] procedures results in an acoustic representation of vowel data that preserves phonemic variation best. Procedures by Gerstman (1968), Miller (1989), and Syrdal and Gopal (1986) perform slightly less well, but are still an improvement compared with the baseline (Hz). No

[21] CLIH_{i2} or CLIH_{i3}.

systematic or substantial differences were found between the scale transformations, and the improvement over the baseline was overall small.

However, the comparisons in the six studies that were discussed are by no means exhaustive, because not all 12 procedures evaluated in this research were directly compared with each other. It is not possible to obtain exhaustive results for all these procedures by pooling the results from the six studies, for the following three reasons.

The first reason is that the procedures were applied to different sets of vowel data in the six studies. Nearey (1978), Nearey (1992), Disner (1980), and Syrdal (1984) used Peterson & Barney's (1952) data set. Hindle (1978) used data from a sociolinguistic interview carried out with speakers from Philadelphia. Deterding (1990) used yet another data set, consisting of speech produced by his co-workers and his children.

The second reason is that the performance of procedures was evaluated differently in each study. Hindle (1978) used regression techniques, Nearey (1978) and Syrdal (1984) used linear discriminant analysis. Disner (1980) computed the relative scatter reduction by calculating the difference in the areas of each vowel category in a F_1 by F_2 plot before and after transformation. Nearey (1992) used generalized linear modeling. Finally, Deterding (1990) used his own procedure: a pattern recognition procedure that involves computing the distance between each token's transformed frequency and the mean frequency for (a subsection of) the speaker population.

The third and final reason is that in four cases (Nearey, 1978; Disner, 1980; Syrdal, 1984; Nearey 1992) the procedures were applied to data that was not controlled for the regional background of the speakers. It was therefore not possible to monitor the elimination of sociolinguistic information, because there were probably no systematic sociolinguistic differences present in the data. The two studies that compared the procedures on how well they preserve sociolinguistic variation used either indirect judgments (Disner, 1980, the part of the study involving the linguistic validity), or did not consider preserving sociolinguistic variation in the data (Deterding, 1990).

In order to get a complete overview of how well the procedures preserve all phonemic and sociolinguistic speaker-related variation in the acoustic representation of vowel data, while minimizing the anatomical/physiological speaker-related variation, the following comparison must be carried out. All 12 procedures must be compared with each other by applying them to the same set of vowel data. This data set must be produced by speakers that are controlled for certain sociolinguistic variables, e.g., regional background. When speech from such a controlled data set is used, systematic differences in phonemic and sociolinguistic speaker-related variation are most likely present in the data set. When systematic variation is present, it is possible to evaluate how the normalization procedures deal with this variation.

Chapter 3

Perceptual vowel normalization

3.1 Introduction

This chapter presents a discussion of studies that focus on vowel perception by phonetically-naive and phonetically-trained listeners. I discuss these studies in the context of the three sources of variation discussed in Chapter 1: how good are listeners at preserving phonemic and sociolinguistic variation, while ignoring anatomical/physiological speaker-related variation? In this chapter, studies are discussed that evaluate vowel processing tasks involving vowel normalization: category judgment tasks and articulatory judgments tasks. It can be expected that studies involving category judgment tasks give insight into how listeners preserve phonemic variation while ignoring sociolinguistic and anatomical/physiological speaker-related variation. Studies involving articulatory judgment tasks are expected to provide insight into how listeners preserve both phonemic as well as sociolinguistic variation, while ignoring anatomical/physiological variation.

In Section 3.2, I discuss studies in which listeners were required to categorize synthetic speech stimuli. Here, the role of the four acoustic variables (F_0, F_1, F_2, and F_3) in category judgment tasks is evaluated. These studies aim to establish the role of speaker-specific information in vowel categorization. I describe several studies in which the performance of listeners at tasks involving category judgments in speaker-blocked and in speaker-mixed conditions is evaluated (Strange et al., 1976; Verbrugge et al., 1976; Macchi, 1980; Assmann, Hogan & Nearey, 1982; Mullennix, Pisoni & Martin, 1989).

In section 3.3, I turn to articulatory judgments of vowel tokens. Three studies are described: Ladefoged (1960), Laver (1965), and Assmann (1979). These studies evaluate the performance of phonetically-trained listeners at tasks involving articulatory judgments of vowel tokens. Section 3.4 presents the conclusions of the literature study.

3.2 Category judgments

3.2.1 Synthetic speech

Effects of vowel-intrinsic information

Various studies evaluate the effects of vowel-intrinsic F_0, F_1, F_2, and F_3 on vowel categorization. Overall, these studies aim to gain a better understanding of the mapping of the acoustical dimension onto the perceptual dimension of speech. The majority of these studies use synthetic speech material, to be able to systematically vary one specific acoustic factor (for assessing the relative relevance of that factor) while keeping other factors constant. In all the experiments discussed in this section, the listeners were required to categorize the stimuli. These stimuli generally consisted of nonsense syllables or words in isolation.

Some researchers found a systematic effect of intrinsic F_0 on vowel categorization (Miller, 1953; Fujisaki & Kawashima, 1968; Traunmüller 1981; Ryalls & Lieberman, 1982). The majority of these experiments involved establishing how much the frequency of F_1 should be raised or lowered to preserve the perceived category of the vowel token, given variation in the frequency of F_0.

For instance, Traunmüller (1981) carried out a series of categorization experiments using synthetic one-formant vowels with various vowel heights to explore the relation between F_1 and F_0 in vowel perception. In his experiments 2, 3, and 4, the listeners were required to categorize synthetic vowels by circling one of the following 13 symbols on their answer sheet: “u”, “ü”, “i”, “o”, “ö”, “e”, “å”, “ɷ”, “ɐ”, “au”, “äul”, “ai”, “a”[22]. He found that the category boundaries for F_1 were strongly affected by the frequency of F_0. Traunmüller found that vowel-intrinsic F_0 and vowel-intrinsic F_1 appear to show a ‘cue-trading’ relationship (see also Repp, 1982): listeners can perceptually compensate for a change in F_0 when this change can be offset by a smaller change in F_1.

It is widely accepted that a strong relationship exists between the perceived vowel category and the first two formants. The frequencies of F_1 and F_2 are seen as the primary acoustic correlates of perceived vowel identity (e.g., Joos, 1948; Potter & Peterson, 1948; Peterson & Barney, 1952; Stevens, 1998). Some researchers even suggested that vowel categorization depends entirely on the first two formants (e.g., Delattre, 1952).

The role of F_3 in vowel categorization is less well understood than the role of the F_1 and F_2 (or even F_0). The relevance of F_3 for vowel classification appears to vary depending on the language and on the vowel. First, there are indications that F_3 is necessary for distinguishing between certain classes of rhotacized vowels in American English (Peterson & Barney, 1952, Lehiste & Peterson, 1959; Miller, 1989, Ladefoged, 2001). Second, the results of some studies show that F_3 may help listeners distinguishing between certain classes of unrounded front vowels (Fujimura, 1967; Fant, Carlson & Granstrøm, 1974; Aaltonen, 1985; Schwartz

[22] Orthographic symbols used in the Bavarian dialect spoken by the listeners.

& Escudier, 1987). For instance, Fujimura (1967) found for Swedish that F_3 is necessary to explain the categorization behavior of his listeners. However, studies for other languages report conflicting results on whether F_3 is necessary for the identification of rounded front vowels; listeners identify these vowels well using only the first two formant frequencies (for American English: Delattre, 1952, for Dutch: Cohen, Slis & 't Hart, 1963; 1967, and for German: Fischer-Jørgensen, 1967).

Effects of vowel-extrinsic information

It has been hypothesized that F_0, F_1, F_2, and F_3 do not only play a role as vowel-intrinsic factors in vowel categorization, but can also influence human vowel processing as vowel-extrinsic factors[23]. Here, I provide a short overview of studies advocating this hypothesis[24].

The experiments described in this section, evaluated how the presentation of a stimulus token in various acoustic contexts (e.g., in carrier phrases) affects the categorization of that stimulus token. These experiments aimed to investigate the role of intrinsic and extrinsic information sources on vowel categorization. In some of the experiments, the carrier phrase was constructed as if produced by a different (synthetic) speaker than the (synthetic) speaker of stimulus word. To this end, vowel-intrinsic F_0, F_1, F_2, or F_3 of the stimulus word and of the F_0, F_1, F_2, or F_3 of the carrier phrase (vowel-extrinsic) were varied independently.

Ladefoged & Broadbent (1957) presented listeners with stimulus vowel tokens in isolation and preceded by six sentences. These precursors sentences were designed to sound as if produced by six different speakers, by varying the values of the first two formant frequencies of the vowels in each sentence. The listeners were required to categorize the vowel stimuli. Their results show that the perceived identity of the stimulus vowel token depends in part on the precursor sentence. The perceived category label of a vowel token thus appears to be determined by its vowel-intrinsic characteristics (the values of the F_1 and F_2 for that stimulus token), as well as by the (vowel-extrinsic) characteristics of the vowels preceding the stimulus token.

However, Broadbent & Ladefoged (1960) showed, in a follow-up study, that the effect found in the 1957 study vanished when the word to be identified was presented before the carrier sentence instead of after. They also showed that the effect increased with the length of the preceding carrier sentence.

The study by Ainsworth (1975) resembles Ladefoged & Broadbent (1957), but differs in one respect. Ainsworth did not only evaluate the effects of varying vowel-extrinsic F_1 and F_2 on the perceived identity of the stimulus vowel, but he also varied the frequency of F_0 of the carried phrase (vowel-extrinsic F_0). As in Ladefoged & Broadbent's experiment,

[23] As mentioned in Chapter 1, whenever I refer to 'vowel-intrinsic', I refer to characteristics of the (stimulus) vowel token itself, and whenever I refer to 'vowel-extrinsic', I refer to characteristics outside that token.

[24] For further details on these studies, I refer to Rosner & Pickering (1994).

listeners were required to categorize the stimulus vowels, once in isolation and once preceded by each of 10 precursor sentences. Ainsworth found that the perceived vowel identity of each stimulus vowel varies depending on the values of vowel-intrinsic F_1 and F_2 of the stimulus token. Second, this perceived identity shifts depending on the values of extrinsic F_1, F_2, and of extrinsic F_0 in the precursor sentence. The effects of extrinsic F_1 and F_2 are roughly twice the size of the effect of vowel-extrinsic F_0. Ainsworth's study thus corroborates the results of Ladefoged & Broadbent (1957).

Nearey (1989) carried out an experiment similar to Ainsworth's (1975) and Ladefoged & Broadbent's (1957) and generally found the same results.

Johnson (1990b) hypothesized that vowel-intrinsic F_0 is used directly or indirectly in vowel categorization[25]. When vowel-intrinsic F_0 is used directly, all normalizing information serves as a cue to the vowel token's identity (as is done in Syrdal & Gopal's procedure, 1986). When vowel-intrinsic F_0 is used indirectly, it serves as a cue to the speaker's identity (i.e., male, female, or a child), as is done in Nearey's (1978) procedure. Johnson tested whether vowel-intrinsic F_0 is used directly or indirectly by carrying out an experiment in which listeners were required to categorize vowel stimuli preceded by a precursor phrase. By varying the difference in frequency between vowel-extrinsic F_0 and vowel-intrinsic F_0, the perceived speaker identity was varied. One half of the listeners had to categorize the stimulus word in isolation, while the other half had to categorize the stimulus when preceded by the precursor sentence. The results showed that the effect of varying the perceived speaker identity on the categorization of the vowel stimuli is larger for the vowels presented in isolation than for the vowel tokens in the precursor sentence. Furthermore, Johnson concluded that when differences in speaker identity are reduced for the stimuli in the precursor phrases, the effect of varying F_0 on the categorization of the vowel tokens decreases as well. These results were consistent with the hypothesis that the F_0 acts as an indirect cue, a finding that lends support to indirect theories of vowel categorization. Johnson concluded that users process vowel-intrinsic F_0 primarily as a cue to the speaker's identity, instead of a cue to the vowel's category.

The results discussed in this section show that, when judging a vowel token's category, listeners interpret the vowel-intrinsic characteristics of a vowel token relative to the vowel-extrinsic characteristics of other vowel tokens in the utterance.

3.2.2 Natural speech

Several past studies used natural speech to investigate how listeners adapt to a speaker. These studies evaluated the effect of the availability of more or less information about a speaker. However, only a few hypotheses were formulated about how much speech material or what

[25] A similar suggestion was made in Weenink (1986) and in Van Bergem, Pols, & Koopmans-van Beinum (1988).

type of phonetic information is necessary. In general, three different issues/hypotheses have been investigated.

First, it was repeatedly suggested that the 'point vowels' /i, ɑ, u/[26] can serve as the primary calibrators of the vowel system (e.g., Joos, 1948; Lieberman, Crelin & Klatt, 1972). These researchers argued that a listener can benefit from hearing these point vowels when categorizing vowel tokens produced by a new speaker. They furthermore argued that point vowels are the most likely candidates of all vowels, because they represent the extreme positions in a speaker's articulatory vowel space and therefore represent the extreme formant frequencies in that speaker's acoustic vowel space. Finally, point vowels are argued to be relatively stable for small changes in articulation (Stevens, 1972). In this section, I discuss one study in which the effect of the presentation of point vowels is evaluated, Verbrugge et al. (1976).

Second, some researchers, for instance Strange et al. (1976), hypothesized that the consonantal environment of a vowel token may facilitate vowel identification in two ways. Strange et al. first suggested that formant transitions may provide the listener with information about vocal tract differences between speakers. Second, they argued that vowels presented in isolation are more difficult to perceive, because it can be hypothesized that listeners usually rely upon information distributed throughout the entire syllable for categorization. I discuss three studies that deal with the effect of consonantal context: Strange et al. (1976), Macchi (1980), and Assmann, Nearey & Hogan (1982).

A third issue was concerned with how vowel categorization was affected by presenting vowel stimuli in speaker-mixed or speaker-blocked conditions. For speaker-mixed conditions, it has been hypothesized that the categorization performance decreases because the listeners have to recalibrate every time they categorize a token, because they have to deal with a new speaker in every trial. I discuss five studies in which this procedure is implemented, Verbrugge, et al. (1976), Strange et al. (1976), Macchi (1980), Assmann, Nearey & Hogan (1982), and finally, Mullennix (1989).

Verbrugge et al. (1976)

Verbrugge et al. (1976) described three experiments, of which the first two are discussed here. In these two experiments, they investigated the influence of point vowels, as well as the effect of speaker-mixed versus speaker-blocked presentation of stimuli on the categorization of vowel tokens.

The stimuli that were used in the first experiment had the following characteristics. Thirty speakers produced 10 American English monophthongal vowels and five diphthongal vowels in a /hVd/ context (/i, ɪ, ɛ, æ, a, ɔ, ʌ, ɚ, u, ʊ, eɪ, oʊ, aʊ, ɔɪ/). In addition, each speaker pronounced a precursor string that contained the point vowels /i, ɑ, u/, each in a /kVp/ context.

[26] For English; for Dutch this should be /i, a, u/.

The stimuli were presented in two experimental conditions. In the first condition, the stimuli were preceded by the precursor string produced by the speaker of the test stimulus. In the second condition, the stimuli were presented in isolation. In both conditions, the stimuli were presented randomized by speaker and vowel.

The listeners in this experiment were phonetically naive. Under both experimental conditions, they were required to categorize each stimulus by checking choosing one of the options on the response sheet. The responses were arranged as follows: 'hood', 'head', 'hoed', 'heard', 'who'd', 'hide', 'heed', 'how'd', 'hud', 'hayed', 'hod', 'hoyed', 'had', 'hid', or 'howed'.

First, Verbrugge et al. calculated the errors (e.g., the response vowel category differed from the intended vowel category). Subsequently, they compared error scores across experimental conditions. The results of this analysis revealed no differences between the identification rates of the two conditions. Furthermore, when they looked at the results for the individual vowels, generally no significant effects were found either. Finally, several vowels, /i/, /u/ (two of the three point vowels), /ɝ/, /eɪ/, /aɪ/, and /ɔɪ/, were identified with high accuracy, even in the no-precursor condition.

Verbrugge et al. remarked that the high level of identification (around 87%) points to a ceiling effect in the no-precursor condition. They argued that the failure to find an effect of the precursor string may indicate that point vowels do not play a role as calibrators of a speaker's vowel system, or that the listener does not need additional information about a speaker's vowel system to perform categorization tasks, and that categorization errors were due to uncertainties in normalization (i.e., the listener did not have enough information about the speaker to make a correct judgment).

Subsequently, Verbrugge et al. established what proportion of the errors made in the no-precursor condition is due to listeners' uncertainty about the vocal tracts of the speakers. This proportion would then define the maximum improvement that could be obtained by presenting precursor phrases, compared with a condition in which no precursor was presented. They investigated this hypothesis in their second experiment.

In the second experiment, categorization of vowel stimuli in a /pVp/ context was evaluated by combining the presentation of precursor sentences with the presentation without precursors in speaker-blocked as well as speaker-mixed conditions. Thus a total of four conditions was evaluated: speaker-blocked without precursor sentences, speaker-mixed with precursor sentences, speaker-mixed with point-vowel (/hi, hɑ, hu/) precursor speaker-mixed with a central-vowel (/hɪ, hæ, hʌ/) precursor. The vowels in the precursor sentence of the fourth condition were chosen to represent point vowels. If point vowels are really "privileged carriers of information for normalization", then the condition with the point vowels should produce lower error scores than the central-vowel-precursor condition. Finally, Verbrugge et al. predicted that if the information available in point vowels is gained by the listener during extended familiarization with a speaker's vocal tract, then the performance in the

point-vowel-precursor condition should resemble that of the speaker-blocked no-precursor condition.

The speech material of the second experiment consisted of natural speech, read aloud by 15 speakers (five men, five women, and five children) who were selected to represent "a wide variety of vocal tract sizes and fundamental frequencies". The speakers were selected to represent a fairly homogeneous "dialect group". Each of the talkers pronounced one token of each of nine monophthongal vowel categories (/i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/) in a /pVp/ context. In this experiment, listeners were required to identify the stimuli as one of these nine categories: 'peep', 'pip', 'pep', 'pawp', 'pap', 'pop', 'pup', 'puup', or 'poop'.

Verbrugge et al. reported error rates for the speaker-blocked and the speaker-mixed conditions of 9.5% and 17.0%, respectively. They concluded that familiarity with a speaker improves the accuracy of vowel categorization. They further found that the effect of speaker familiarity accounts for only half of the errors made. In addition, they remarked that 9.5% is rather high error rate for a condition in which the speaker's characteristics and the consonantal context are entirely predictable from trial to trial, while 17.0% seems rather low given the unfamiliarity with the speaker's voice from trial to trial. In order to investigate these percentages further, they compared the error percentages in the two speaker-mixed conditions: the central-vowel-precursor condition and the point-vowel-precursor condition. Here, no difference could be found between the two conditions. Finally, they concluded that neither set of vowel precursors appears to be efficient carriers of information that is available in more prolonged exposure to a speaker's voice.

Based on their two experiments, Verbrugge et al. concluded that a single (CVC) syllable contains enough information about its vowel, regardless of whether the listener is familiar with the speaker's voice and that familiarity with a speaker's voice plays only a secondary role in vowel identification. They finally concluded that point vowels play no major role as calibrators of the speaker-specific vowel space.

Strange et al. (1976)

Strange et al. (1976) compared the identifiability of natural vowel tokens produced in a consonantal environment and in isolation, in speaker-blocked and speaker-mixed conditions. Their goal was to establish the relative effect of the vowel token's consonantal context and of information about the vowel token's speaker on the accuracy of vowel categorization.

They formulated two hypotheses. First, if the presence of a consonantal context aids vowel categorization by serving as a calibrating signal for vocal tract normalization, an interaction should occur between the context and speaker information. In this case, the loss in identifiability of the vowel tokens that occurred due to the absence of consonantal transitions is expected to be more severe in the speaker-mixed condition, because the listener is required to re-calibrate before each trial. No such disadvantage was expected for the

speaker-blocked conditions. The second hypothesis stated that, if the consonantal transitions provide information that specifies vowel identity independently of speaker normalization, no interaction should occur between context and speaker information. Here, the categorization of vowel tokens in isolation should be less accurate than for the vowel tokens in the consonantal context for the speaker-blocked and speaker-mixed conditions.

The speech material was produced by the same set of speakers as described in Verbrugge et al. (1976) second experiment. These speakers had produced the same nine monophthongal vowels in isolation as well as in a /pVp/ context. For the speaker-mixed conditions, three out of nine vowels per speaker were selected (one of which was a point vowel). For the speaker-blocked condition, one man, one woman and one child produced a list of 45 items that contained five different tokens of each of the nine vowel categories.

Four experimental groups of listeners were asked to label each stimulus as one of the nine vowel categories in four experiments. By orthogonally combining the factors blocking (blocked or mixed presentation) and context (presentation in isolation or in context) four experimental conditions were obtained.

The results showed a higher error percentage for the isolated vowels than for the context vowels, and a higher error percentage for the mixed condition than for the blocked condition (speaker-blocked/isolation 31.2% errors, speaker-blocked/context 9.5%, speaker-mixed/isolation 42.6%, speaker-mixed/context 17.0%). No significant interaction between the variables was found. Therefore, the hypothesis that consonantal environment contributes to vowel perception by providing cues for vowel normalization, was not supported by the results. Instead, the results supported the second hypothesis: consonantal transitions provides the listener with information about the vowel token's category. Furthermore, the results indicated that variation in the speakers increases the number of identification errors, but that the presence or absence of context is the most influential factor. Strange et al. concluded that the presence of a consonantal context is much more relevant to listeners than the familiarity with the speaker. In addition, they stated that isolated vowel stimuli may be poor stimulus material, because the error percentages for isolated vowels are substantially higher than those for vowel tokens in a consonantal context.

Strange et al. performed a second experiment to investigate whether a consonantal context that varied from trial to trial provides relevant information for vowel identification. They asked a subset of the speakers from the first experiment to produce the nine vowel tokens in CVC contexts consisting of symmetrical and asymmetrical combinations of six stop consonants (/p, t, k, b, d, g/). For the speaker-blocked condition, the same three speakers were selected as in the first experiment. Listeners had to categorize the CVC stimuli in a speaker-mixed and a speaker-blocked condition.

For the speaker-mixed condition for the vowel tokens in the varying contexts, Strange et al. found 21.7% errors; a percentage not significantly higher than the percentage found for the /pVp/ contexts in the first experiment (17.0%). For the speaker-mixed condition, vowels

in CVC-contexts were identified considerably better than vowel tokens in isolation. They concluded that presenting vowel tokens in a consonantal context gives listeners an advantage over stimuli in isolation even when the consonantal context is not known beforehand. However, they could not find an advantage for the speaker-blocked condition over the mixed condition. Finally, the results revealed 22.9% errors for the speaker-blocked condition for the CVC stimuli. This percentage is lower than the percentage found for vowels in isolation (31.2%), but not significantly higher than the error percentage for the vowels in the /pVp/ context (9.5%).

Given their results, Strange et al. concluded that 'consonantal' cues contribute to vowel perception; many of these cues are contained in the formant transitions. They further suggested that no temporal cross-section of a syllable conveys as much vowel information as is present in the dynamic contour of the formants. They finally stated that the categorization of vowels in isolation can be viewed as a rather unnatural task, and that processes during this task are not the same as those typically used in speech perception. Strange et al.'s conclusions can be summarized as consonantal context being of major importance for vowel categorization and that information about the speaker is of minor importance.

Macchi (1980)

Macchi (1980) aimed to falsify Strange et al. 's (1976) hypothesis that vowels in isolation are impoverished stimuli. Macchi's experiment differs from Strange et al.'s in three respects. In Macchi's experiment, the listening tasks were carried out under high-quality conditions and the speakers and the listeners were closely matched for regional accent. Finally, the possible response alternatives for the vowel tokens were minimized to ensure that listeners did not have difficulty pairing stimuli with orthographic symbols, and that the symbols themselves would not bias the listeners' performance towards either isolated vowels or vowels in contexts, by having listeners rhyming the stimuli with English words.

Macchi's speech material consisted of recordings of 11 American English vowels read in isolation and in /tVt/ contexts (i, ɪ, e, ɛ, æ, u, ʊ, o, ɔ, ɑ, ʌ/). These stimuli were presented in the same four conditions as in Strange et al.'s (1976) study (speaker-blocked/isolation, speaker-blocked/context, speaker-mixed/isolation, speaker-mixed/context). The listeners were required to label these stimuli as 'teet', 'tit', 'Tate', 'tet', 'tat', 'toot', 'Toot'[27], 'tote', 'taut', 'tot', or 'tut'.

The resulting percentages of identification errors (vowel tokens classified as another vowel than their intended vowel) were as follows: speaker-blocked/isolation: 1.5%, speaker-blocked/context: 2.0%, speaker-mixed/isolation: 7.8%, mixed-speaker/context: 8.6%. A main effect was found for the factor 'presentation type': for the two blocked conditions, the mean error rate was 1.8%, while for the speaker-mixed conditions the mean was 8.2%. Second, no effect was found for the factor 'context' (isolation vs. context).

[27] As in "Tootsie roll" (cf. Macchi), for /ʊ/.

Macchi reported differing results across vowels. The percentages of misclassified vowel tokens was highest for the vowel /ɷ/ (21.4%) in the speaker-blocked/isolation condition, while for the three other conditions the percentage misclassifications was highest for vowel /ɑ/ (speaker-blocked/ context: 10.6%, speaker-blocked/isolation: 3.3%, speaker-mixed/context: 42.5%).

When Macchi's results are compared with the results reported by Strange et al., it can first be observed that Macchi's error percentages are considerably lower across the four experimental conditions. For instance, the overall error percentages for the blocked conditions reported by Strange et al. show an increase of 10% compared to Macchi's results. Second, while Strange et al. found differences between the presentations of vowel tokens in context and vowel tokens in isolation, Macchi reported no differences between the error rates for the vowels in isolation and vowels in the /tVt/ context.

Macchi concluded that vowel identification is a function of the familiarity of the listener with the voice of a speaker. Second, she concluded that vowel tokens presented in isolation are not poor stimulus material and that in general, naive listeners can identify vowels in isolation satisfactorily when listening conditions, accent of the speakers, and the response alternatives are carefully controlled.

Assmann, Nearey & Hogan (1982)

Assmann, Nearey & Hogan (1982) performed two experiments in which vowel stimuli were presented in a speaker-blocked or a speaker-mixed condition. Their aim was to investigate if familiarization with several vowels from a single speaker may facilitate vowel identification.

In their first experiment, 100 vowel stimuli were recorded: 10 vowel categories of Canadian English: /i, ɪ, e, ɛ, æ, ʌ, ɒ, o, ʊ, u/, produced in isolation by five male and five female speakers of Canadian English. These stimuli were presented in a speaker-blocked and speaker-mixed condition. All listeners were phonetically-trained. They had to categorize the vowel tokens as one of the 10 IPA symbols listed above. The results from this experiment showed error rates of 5.4% for the mixed condition, 4.1% for the blocked condition. The differences between responses for the two presentation types were therefore minimal.

Assmann et al. concluded that context is not essential for vowel identification, because very few errors were made overall with the isolated vowels. They offered three possible explanations. First, they suggested that there exists less overlap between formant frequencies of different vowel categories in the F_1/F_2 plane than expected. Second, they posed that English vowels could be distinguished along other dimensions than tongue advancement and openness, such as the tense and lax distinction. Third, (vowel-intrinsic) dynamic properties of the vowels such as duration and diphthongization can be used to disambiguate vowel categories with overlapping formant frequencies.

A second experiment was designed to investigate the hypothesis that an increase in errors is expected if the intrinsic dynamic properties are eliminated from the stimuli. They varied

presentation type (blocked versus mixed) independently of the dynamic specification of the stimuli. The stimulus vowels were shortened by gating out part of the waveform, in order to eliminate the effects of dynamic characteristics such as diphthongization and duration. The gated vowel tokens were presented in isolation in speaker-blocked and speaker-mixed conditions. Assmann et al. hypothesized that the error rate for the gated speaker-blocked presentations should be lower than for the mixed speaker condition, if errors in the speaker-mixed condition are the result of overlap in formant frequencies, and if exposure to more vowel tokens from one speaker aids the identification process.

The results showed that the overall error rate for the gated vowel tokens in the speaker-mixed condition was 13.8%, while the percentage for the gated speaker-blocked vowel tokens was 9.5%. According to Assmann et al., this result indicated that steady-state segments contain speaker-specific information that can be used by the listener to improve vowel identification. In addition, they remarked that the error rates were still remarkably low, even though the gated vowels showed overall higher error rates than the full vowels in isolation that were investigated in the first experiment.

Based on these two experiments, they concluded that vowels are well identified, even in the absence of consonantal context and that their results indicate that the overlap between vowel categories may not be as severe as suggested by, for instance, Joos (1948).

Mullennix, Pisoni & Martin (1989)

Mullennix, Pisoni & Martin (1989) investigated the effect of speaker variability on spoken word recognition in four experiments. In the first experiment, listeners were asked to identify CVC words with varying consonantal contexts in three different signal-to-noise ratios (an S/N ratio of +10, 0, and -10). The CVC syllables were presented in a speaker-mixed and a speaker-blocked condition. In the blocked condition, all the stimulus words were produced by a single speaker (a different speaker for each listener). In the mixed condition, syllables produced by 15 speakers were mixed (eight females and seven male speakers, who all spoke a midwestern variety of American English).

Mullennix et al. found effects of condition (blocked-mixed) and of the S/N level. The identification scores were 40.6% correct for the mixed-condition and 33.6% for the blocked-condition (pooled for the three S/N levels). The percentages for the three S/N ratios were 63.6% for the +10 S/N condition, 42.2% for the 0 S/N condition, and 5.9% correct for the -10 S/N ratio. Mullennix et al. concluded that speaker variability has a substantial effect on the perception of words degraded by noise.

Mullennix et al. performed two additional experiments in order to investigate the robustness of these effects with non-degraded stimuli. These experiments involved naming tasks; the listeners had to repeat the word they had just heard. In all experiments, response latencies were measured in addition to the identification responses. These two naming experiments

replicated the results from the first experiment. Higher percentages correctly classified vowel tokens were found for the speaker-blocked conditions (95.8% for experiment 2 and 97.8% for experiment 3) than for the speaker-mixed conditions (91.4% for experiment 2 and 92.9% for experiment 3). The listeners could furthermore repeat the stimuli in the speaker-blocked condition overall 50 ms faster than in the speaker-mixed condition[28].

Finally, Mullennix et al. suggested that listeners do not evoke normalization processes for every sample of speech they hear. Instead, some exposure to the speech of a new speaker is required in order to process speech of that new speaker as efficiently as the speech from speakers that the listener is already familiar with.

3.2.3 Summary category judgments

When listeners are required to categorize tasks that involve preserving the phonemic information in the acoustic signal, while discarding, or ignoring, anatomical/physiological speaker-related information, the following results were reported in the literature on category judgments for the following vowel-extrinsic and vowel-extrinsic sources of information.

First, in section 3.2.1, I discussed the role of the acoustic variables F_0, F_1, F_2, and F_3 in vowel categorization as vowel-intrinsic stimuli for the categorization of synthetic one-syllable stimuli. For this type of stimuli, the most relevant acoustic correlates of the perceived vowel category appear to be vowel-intrinsic F_1 and F_2. The relevance of vowel-intrinsic F_0 for vowel categorization is smaller than F_1 and F_2. Finally, the results for the studies that investigated the role of F_3 show that its relevance for vowel categorization of isolated syllables is unclear.

Second, I discussed the finding that F_0, F_1, F_2, and F_3 affect the categorization of vowel tokens as vowel-extrinsic factors. When the F_1 and F_2 of precursor sentences are varied independently of the F_1 and F_2 in the stimulus words, the (vowel-extrinsic) F_1 and F_2 affects the categorization of the vowel tokens, albeit to a lesser extent than varying vowel-intrinsic F_1 and F_2. In addition, a similar effect was found for vowel-extrinsic F_0. No such effects were found for vowel-extrinsic F_3.

Third, in section 3.2.2, I discussed the effect of vowel-intrinsic and vowel-extrinsic information on the categorization vowel tokens produced by real speakers. In order to give an overview of the general pattern in the results found in this section, the percentages of misclassified vowel tokens per experimental condition for the studies that were discussed in this section, are listed in Table 3.1. For each study only the significant effects for the speaker-blocked and speaker-mixed conditions (following the significance levels of that study) are included.

In the studies listed in Table 3.1, the type of information available to the listener was varied, the listener was provided with only vowel-intrinsic information or with vowel-ex-

[28] Mullennix et al. carried out one last experiment, which is not discussed here, because its emphasis is on the effects of word frequency on word recognition, which is beyond the scope of my research.

Table 3.1: *Error percentages for the studies discussed in this section, per presentation type.*

% Error	Study	Speaker-blocked	Speaker-mixed
Isolation	Strange et al.	31.2%	42.6%
	Macchi	1.5%	7.8%
	Assmann et al.	4.1%	5.4%
Context	Verbrugge et al.	9.5%	17.0%
	Strange et al.	9.5%	17.0%
	Macchi	2.0%	8.6%
Context: gating	Assmann et al.	9.5%	13.8%
Noise	Mullennix et al.	66.4%	59.4%

trinsic information in addition to vowel-intrinsic information. In speaker-mixed conditions, the listeners have only intrinsic information and to base their judgment on one single vowel token. In speaker-blocked conditions, the listener is provided with more information about speaker's vowel system. As can be seen in Table 3.1, it was generally found that listeners make less errors in speaker-blocked conditions.

In addition, the consonantal context was varied: the vowel tokens were presented in isolation and in consonantal contexts. Verbrugge et al. (1976) and Strange et al.'s (1976) results indicated that vowels in isolation are categorized with considerably more difficulty than vowels in context. However, results found by Macchi (1980) and Assmann et al. (1982) (listed in Table 3.1) did not confirm Verbrugge et al. and Strange et al. 's findings. It thus seems plausible that isolated vowels are not poor stimulus material, as stated by Strange et al., compared with vowels in context. Finally, all studies described in this section, except for Strange et al. (1976) and Mullennix et al. (1989), report ceiling effects.

3.3 Articulatory judgments

Phonetic and articulatory dimensions of vowel tokens

Although it can be presumed that trained listeners are able to distinguish a large number of points on the three articulatory dimensions tongue height, tongue advancement, and lip rounding (resulting from different configurations of the vowel tract), only a limited number of points per dimension are defined in phonetic theory. In the IPA chart (IPA, 1999) the following four levels of tongue height are distinguished: close, close-mid, open-mid, and open. For tongue advancement, three levels are distinguished: front, central, and back. For lip rounding, two levels are defined: rounded and unrounded (sometimes also referred to as 'spread'). Using these levels on each dimension, vowels are described in terms such as 'a close rounded back vowel' (/u/).

The acoustic consequences of different configurations are extensively documented (e.g., Lindblom & Sundberg, 1971; Ladefoged et al., 1978; Stevens, 1998). It is generally assumed that strong correlations exist between F_1 and articulatory tongue height and between F_2 and articulatory tongue advancement.

The relationship between articulatory lip rounding and acoustic variables is less transparent than for tongue height and tongue advancement. Although lip rounding is generally accepted as one of the three articulatory dimensions that are necessary for describing differences between vowels, no straightforward relations between lip rounding and the acoustic dimensions of speech have been reported in the literature. For instance, in speech production, the effect of lip rounding on the values of F_2 is confounded with the effects of tongue advancement (e.g., Fant, 1960; Lindblom & Sundberg, 1971, Stevens, 1998). Stevens predicted a different effect of lip rounding on the acoustic characteristics of different types of vowel categories. For instance, when the lips were rounded as opposed to not rounded, Stevens reported a lowering of the frequency of F_1 for low vowels, a lowering of F_1 and F_2 for low back vowels and a lowering of F_2 and F_3 for low front vowels. Furthermore, other researchers even suggested that, at least for English, lip rounding is an articulatory maneuver without a contrastive function distinct from backing the tongue (Bloomfield, 1933; Ladefoged, 1975). Finally, others put forward that lip position is not conveyed reliably in the acoustic signal at all (Abercrombie, 1985; Lisker & Rossi, 1992).

Although the knowledge about the relationship between articulatory and acoustic characteristics is far from complete, even less has been published about the relationship between the perceived articulatory characteristics of vowel tokens and the acoustic characteristics of those vowel tokens. To my knowledge, only one study described the relationship between F_1 and perceived tongue height and between F_2 and perceived tongue advancement, namely Assmann (1979). This study is described in section 3.3.1. In section 3.3.2, two studies are discussed that describe experiments in which articulatory judgments were obtained.

3.3.1 Articulatory judgments

Assmann (1979) described five experiments, four of which are discussed in Assmann, Nearey & Hogan (1982). Here, the fifth experiment is discussed (see section 3.2.2 for a description of two of the remaining four experiments). The fifth experiment used the same vowel stimuli as the second experiment: 100 stimuli with gated vowel centers, produced in isolation by five male and five female speakers of Canadian English, belonging to the vowel categories /i, ɪ, e, ɛ, æ, ʌ, ɒ, o, ʊ, u/.

Assmann (1979) aimed to find answers to the following three questions. First, do judgments of perceived vowel height and perceived vowel advancement show overlap in cases where vowels are confused by listeners? Second, do these judgments correlate best with F_1 and F_2 or do F_3 and F_0 play as role as well? Third, do phonetically-trained listener use

information about the speaker in their judgments?

Two phonetically-trained listeners provided a phonemic label (i.e., a category judgment) of each of the 100 vowel tokens, and they had to judge the vowel token's height and advancement (an articulatory judgment). The stimuli were presented blocked by speaker in a fixed order. The listeners could listen to each token as often as they wanted. All category and articulatory judgments were based on the consensus of the two listeners.

In the transcription task, the listeners could use one of the 10 phonetic symbols, with optional diacritical marks ('>>' for fronting, '<' for retraction, '/' for raising and '\' for lowering). In the judgment task, each stimulus token was to be located on a two-dimensional vowel diagram. The diagram was partitioned into 80 cells. The 10 vowel categories (/i, ɪ, e, ɛ, æ, ʌ, ɒ, o, ʊ, u/) were placed in the vowel diagram as reference points. The coordinates of the chosen location were used as the perceived height and advancement for each stimulus vowel.

Assmann first calculated the mean coordinates per vowel and observed that the listeners did not use the 10 key point as anchors, because in some cases the mean locations per vowel differed considerably from the locations of the 10 reference points. When he carried out a discriminant analysis on the coordinates for height and advancement, he found that 95% of the vowel tokens could be assigned to the correct (intended) vowel category on the basis of the height and advancement coordinates. This result indicates that there appears to be little overlap between vowel categories in the judgments of height and advancement. Therefore, it seems plausible that the listeners were influenced strongly by the vowel token's (perceived) category. Nevertheless, it should be mentioned that the vowel token's perceived category does not necessarily have to be located at the same point in the vowel diagram as the reference point.

Second, Assmann correlated the judgment coordinates with acoustic measurements. He correlated the judgments of height and advancement for the judgments pooled for all vowel categories with log-transformed F_0, F_1, F_2, F_3 (his G0, G1, G2, G3) and mean log-transformed F_1 (G1AV) and mean log-transformed F_2 (G2AV). The two parameters G1AV and G2AV served as 'speaker parameters', additional information about the speaker. For height, he found a high correlation between height and G1 (88%). When G1AV was added to the model for height, the fit of the model improved. Adding G0, G2, and G3 to the model did not lead to a improvement of the correlation. However, when G0 and G2 were both entered as predictors, a significant effect was found. The same was found for G0 and G3, and when G0 was entered as the sole predictor. Assmann found a significant correlation between advancement and G2 (95%), adding G2AV improved the model significantly. When G0, G1, and G3 were entered together as predictors for advancement, a significant effect was found, as was the case for the combination of G0 and G3. In addition, he tested a model for both dependent variables with CLIH_{i2} (Nearey, 1978) normalized values of G1 and G2 (i.e., the log-transformed values of the mean formant frequencies per speaker for each formant were

subtracted from the frequency of G1 or G2). He found that the correlations were stronger after transforming the values of G1 and G2 (91% for CLIH_{i2}-transformed G1 and 96% for CLIH_{i2}-transformed G2). Finally, to correct for the shape of the vowel diagram (which may have influenced the judgments) he subtracted the mean judgment value per vowel category from the height and advancement judgments for each judged vowel token. Using these values, he carried out the analyses again for height and G1 and for advancement and G2 again and found significant correlations: 26% for G1 and height and 54% for G2 and advancement.

Assmann's results suggest, first, that listeners' height judgments can be modeled using the log-transformed vales of F_1, while the judgment of advancement can be modeled through log-transformed F_2. Second, the correlation scores increase when modeled with acoustic parameters that reflect more information about the speaker, such as the speaker parameters G1AV and G2AV or formant frequencies transformed using CLIH_{i2}.

3.3.2 Articulatory judgments of cardinal vowels

In this section, I describe two studies: Ladefoged (1960) and Laver (1965). These two studies discuss the relative merits of Daniel Jones' Cardinal Vowel System or DJCVS (Jones, 1917), a system for vowel classification. Daniel Jones claimed that training with this system provided phoneticians with fixed perceptual reference points in a phonetic space for a set of 18 vowels, the primary and secondary cardinal vowels. Ladefoged and Laver both investigated the validity of this claim.

Ladefoged (1960)

Ladefoged (1960)[29] compared the variability in, and agreement between, judgments made by 15 British phoneticians trained using DJCVS and three phoneticians who were not trained to use DJCVS.

Ladefoged selected 10 Gaelic stimulus words that contained vowel tokens that strongly resembled 10 of the Cardinal Vowels defined by Daniel Jones (1917): *beid, sgòl, cùl, reub, lon, big, fál, laochan, stagh,* and *gaoth*[30]. The vowels in these words were also selected because they were thought to be fairly monophthongal and differ greatly in their phonetic quality. A native speaker of Gaelic pronounced the 10 words.

The phoneticians were asked to plot the 10 stimulus words on cardinal vowel diagrams. When locating each stimulus word this way, the listeners could listen to it as often as they thought was necessary. The listeners were not familiar with the selected variety of Gaelic. Ladefoged hypothesized that his experimental procedure was sufficiently standardized to ensure that each one of the listeners "was assessing the same phonetic data and presenting

[29] Also published in Ladefoged (1967).

[30] Ladefoged did not explicitly provide corresponding IPA symbols.

his results in the same way", and that the procedure resembles a typical task of describing vowels phonetically.

Ladefoged's results showed that, first, the judgments of the phoneticians trained in DJCVS show a high degree of agreement compared with the judgments of the phoneticians who were not trained to use DJCVS. However, he found there was some disagreement among the phoneticians trained in DJCVS as well; differences were found between the phoneticians from London and from Edinburgh. The phoneticians from London showed a greater tendency to consider vowel tokens as peripheral. Second, he reported greater agreement in the judgment of some words than for other words in the data from the DJCVS phoneticians. Furthermore, he found that all 15 phoneticians had more difficulty with the judgment of lip-rounding, compared with the open-close and front-back dimensions. he inferred from this last result that lip-rounding is not easy to perceive through listening alone.

Ladefoged concluded that training with DJCVS allows phoneticians to make adequate judgments of vowel tokens that were judged to have articulatory positions like those of similar primary cardinal vowels. In addition, he concluded that phoneticians trained in DJCVS are in substantially closer agreement than phoneticians with other types of training.

Laver (1965)

Laver (1965) aimed to establish to what extent a listener's judgments of the same stimulus vowel token varied over a given period of time. Laver designed an experiment, resembling Ladefoged's (1960) experiment, in which phonetically-trained listeners were asked to judge the same stimuli on different occasions. Laver stated that Ladefoged's experiment was useful in assessing the phoneticians' variability and agreement, but that its results lost some of their relevance and validity because Ladefoged used only one localization per vowel stimulus per listener.

Laver synthesized thirty three-formant vowel-like sounds on the PAT synthesizer (Lawrence, 1953): 10 test vowels (labeled A to J), designed to correspond to 10 cardinal vowels, plus 20 filler stimuli (labeled 1-20). The filler vowels were, first, intended to obscure the fact that the same 10 test vowels were localized three times and, second, to reduce learning effects. The F_0 of each stimulus was set at 120 Hz and the frequency of the fourth formant was 3800 Hz. The 20 filler vowels were synthesized at intermediate frequency values between the values for the 10 cardinal vowels.

The listeners were phonetically-trained (with DJCVS). They were asked to locate the thirty stimulus vowels on a vowel diagram in six separate tests on six different occasions. The first test (test A) consisted of 50 stimuli; the 20 filler stimuli each occurred once, and the 10 test vowels each occurred three times in a randomized sequence. In the next three tests (1, 2, and 3), each of the 10 test vowels occurred only once and were interspersed with 10 of the filler vowels in a randomized sequence. The last two tests (4 and 5) were

identical to tests 2 and 3. The first two tests (A and 1) were held on the same day, a few hours apart, the remaining four tests were held one per day in the following four days. The listeners were allowed to listen to the vowel stimuli as often as they thought necessary. They were instructed to locate each vowel on the provided vowel-quadrilaterals (paper), using one diagram per vowel, thus obtaining judgments of tongue height and tongue advancement. They were expected to judge the position of the lips (rounded or spread) on the chart as well, but only if the lip-position corresponded to that of the nearest primary cardinal vowel. In addition, if the degree of perceived lip-rounding seemed 'inappropriate' to its position on the chart in relation to the nearest Primary Cardinal Vowel, the listeners had to indicate the appropriate degree of rounding or spreading for that stimulus, using the scale 'close-rounding, open-rounding, neutral, spread' as reference points. The tests were taken by five experts, two of whom had also participated in Ladefoged's (1960) experiment.

Laver restricted his analyses to the locations of the test vowels. He analyzed the dispersion of the individual locations for each listener and of each test vowel, the average locations, and the consistency (i.e., the variability of the locations in the five tests). He calculated the dispersion of the locations by calculating the mean locations for each of the 10 stimulus vowels for each listener, once for Test A and once for Tests 1-5. Using these mean values, it was calculated for each vowel what percentage of the entire response area was occupied by responses for that test vowel. For instance, for the test vowel labeled A, it was found that 95% or the responses that were closest to test vowel A enclosed 3.4% of the total response area for Test A (cf. Laver's Table 3). Laver noted that the locations of lip-position are, overall, the most divergent of the judgments of his three dimensions (i.e., tongue height, tongue advancement, and lip position).

He proceeded with the position of the average locations. He analyzed the shifts of the average positions for the test vowels in Test A and Test 1-5. He found only random movements of the shift over the period of a week (the period between Test A and Test 5) in the locations for each test vowel.

Furthermore, he aimed to find the limits of the variability of the location of a given stimulus within the most extended test (Test A) and across a number of short tests (tests 1-5). For each data point (i.e., a localization by one of the listeners), Laver calculated its distance to each location of the 10 test vowels. In order to limit the number of calculations, Laver grouped the 10 test vowels into eight 'divisions' based on articulatory criteria, such as 'Close front', 'Open back', or 'Front'. He used the standard deviation for each division to represent each test vowel within that division. Since one of the assumptions of DJCVS is that the intervals between the Cardinal Vowels are 'auditorily equal', the standard deviation should not vary significantly from test vowel to test vowel within a division. Laver compared the standard deviations of the test vowels in the divisions for test A with those in Tests 1-5. He found significant results for only one listener, indicating that the overall variability was low. Laver stated that no overall significant variability in different areas of the vowel quadrilateral

(i.e., the divisions) indicates that the relative inter-area variability is constant, although the inter-listener variability may vary.

Given all his results, Laver concluded that phonetically-trained listeners vary in their judgments over a short period of time. Laver found no systematic patterns in the variations. Laver concluded that the Cardinal Vowel Theory has an important weakness: the cardinal reference points are relatively unstable, and show some variation with time[31].

3.3.3 Summary articulatory judgments

In this section, literature study was carried out to investigate the influence of acoustic factors on articulatory judgments by trained listeners to get an idea about how listeners succeed in preserving phonemic and sociolinguistic variation in the acoustic signal, while ignoring the anatomical/physiological variation. I started by describing studies investigating the relationship between articulatory and acoustic characteristics of vowels. Subsequently, I discussed studies about the relation between the acoustic and the perceived articulatory characteristics of those vowels.

The first discussion can be summarized as that the relationship between the position of the tongue during articulation (height and advancement) and the acoustic variables is reasonably well understood: F_1 correlates with articulatory tongue height and F_2 with articulatory tongue advancement. The relation between lip rounding and the acoustic variables appeared to be less clear. Regarding the second set of studies, it was first found that there exists a (strong) correlation between perceived articulatory vowel height and (log-transformed) F_1 and between perceived articulatory tongue advancement and (log-transformed) F_2. Furthermore, listeners' behavior can be modeled better when the acoustic predictor variables incorporate more information about the speaker.

The results of the studies discussed in section 3.3.2 can be summarized as follows. Both Ladefoged and Laver found that the Cardinal Vowels in DJCVS cannot serve as stable and fixed universal reference vowels and that localizations vary with time and across listeners. Nevertheless, the two studies are limited in that they do not provide insight in how reliably phonetically-trained listeners normalize speech from different speakers, because the studies used either speech from only one speaker (Ladefoged) or synthetic speech (Laver).

3.4 Conclusions

The literature study described in the present chapter was carried out to get a better understanding of how (phonetically-trained) listeners perform at the task of preserving phonemic

[31] It is not clear from Laver's results whether this variation was substantial.

and sociolinguistic information in the acoustic signal, while ignoring, anatomical/physiological speaker-related information when making category and articulatory judgments of vowel tokens.

For the experiments involving category judgments, for which it can be said that the listeners were required to preserve phonemic variation, I found the following. For studies that used synthetic speech material, it was concluded that the perceived vowel category is determined primarily by vowel-intrinsic F_1 and F_2, and to a lesser extent by vowel-intrinsic F_0 and F_3. In addition, F_0, F_1, F_2 influence vowel categorization as vowel-extrinsic factors. There are also indications that listeners interpret the vowel-intrinsic characteristics of vowel tokens relative to the vowel-extrinsic characteristics of other vowel tokens in the utterance, when judging a vowel token's category.

For studies involving vowel categorization tasks that employ natural speech, the following results were found. Listeners' performance decreases when listeners categorize vowel tokens in a condition in which the speaker changes from trial to trial (speaker-mixed condition) compared with a condition in which the speaker is kept constant across trials. Two compatible interpretations were provided in the literature for this result. First, the listeners' performance decreases because it is necessary to 'recalibrate' for each new speaker in the mixed condition. Second, the listeners' performance increases in a speaker-blocked condition compared with the speaker-mixed condition, because the listener is provided with more information about the speaker's vowel system, which facilitates the judgment process.

For studies describing articulatory judgments, for which it can be said that listeners are required to preserve phonemic and sociolinguistic variation, while anatomical/physiological variation must be ignored, the results are as follows. First, it was found that perceived vowel height and perceived vowel advancement correlate with F_1 and F_2, respectively. This issue must be further investigated in order to establish how acoustic measurements relate to judgments by phonetically-trained listeners. Second, the results of two studies that investigate the performance of phonetically-trained listeners at tasks involving articulatory judgment indicate that these judgments show questionable variability. It must be concluded that it is not clear how reliably phonetically-trained listeners judge speech material produced by multiple speakers, because in both of the two studies speech from one (synthetic) speaker was judged.

Chapter 4

Research design

4.1 Introduction

The present research aims to establish which procedure for vowel normalization is most appropriate for use in sociolinguistics. In Chapter 1, I formulated a criterion that must be met in order for the normalization procedures to be considered appropriate. The normalization procedure must preserve the phonemic variation and the sociolinguistic speaker-related variation, while minimizing the anatomical/physiological speaker-related variation in the transformed acoustic vowel data. In this chapter, I describe how it was evaluated how well the procedures met this criterion.

The setup of this chapter is as follows. The conclusions from the two literature studies described in Chapter 2 and Chapter 3 that are relevant for the research design, are put together in section 4.2. Section 4.3 presents and discusses the research design and its components.

4.2 Previous research

In the two previous chapters, I discussed studies that evaluate the performance of normalization procedures (Chapter 2) and listeners (Chapter 3) on tasks involving vowel normalization. The results in Chapter 2 show that it was not feasible to establish which procedure performs best at preserving phonemic variation and reducing anatomical/physiological variation, because the 12 procedures evaluated in the present research were never exhaustively compared. Moreover, although it can be said that the studies by Hindle (1978) and by Disner (1980) compare normalization procedures on how well they preserve sociolinguistic variation, no definitive conclusions can be drawn from these studies. This is because Hindle's study is carried out on too small a scale and because Disner's argumentation does not seem very

convincing: she based her conclusions on comparisons between acoustic patterns in the vowel data and indirect judgments by phonetically-trained listeners.

Given the literature study in Chapter 2, I concluded that it is necessary to carry out an exhaustive comparison of the procedures on well they preserve phonemic variation; no such comparison was carried out before. In addition, in order to evaluate how well the procedures preserve sociolinguistic variation, the procedures must be applied to a database in which systematic sociolinguistic variation is present.

The conclusions of the literature study described in Chapter 3, in which perceptual vowel normalization was studied, are as follows. It was concluded that the effect of F_0, F_1, F_2, and F_3 as intrinsic and as extrinsic factors on tasks involving vowel categorization seems to be established reasonably well, for phonetically naive listeners as well as for phonetically-trained listeners. Furthermore, it must be concluded that the performance of listeners at tasks involving vowel judgment was studied less extensively: only two studies deal with this issue: Ladefoged (1960) and Laver (1965). These two studies show that judgments by phonetically-trained listeners vary over time and show variability depending on the type of training that the listener received. It must be concluded that the performance of phonetically-trained listeners at tasks involving judgment of speech produced by multiple speakers has not been investigated in great depth. For instance, both of the studies presented the listeners with speech from only one speaker. Finally, the results of the literature study show that the relationship between the acoustic factors (F_0, F_1, F_2, and F_3) and perceived vowel height, vowel advancement, and lip rounding has not been not studied extensively: only one paper examined this relationship (Assmann, 1979a). It was concluded that it is necessary to further investigate this relationship.

4.3 General scheme

The research design is depicted in Figure 4.1. The boxes in Figure 4.1 refer to processes in which information is generated and ellipses refer to that information itself. The research design is described as follows.

Instruction $\times$ Articulation

The box at the top of Figure 4.1 that contains the word 'Instruction' refers to the instruction to the speaker to pronounce the word that contains the target vowel token. This target vowel is one of the nine monophthongal vowels of Dutch: the 'intended' vowel category. 'Articulation' refers to the speaker's response to this instruction.

In analogy with Pols, Tromp & Plomp (1973) and Van Nierop, Pols & Plomp (1973) the present research focuses on read monophthongal vowels in a fixed consonantal context. This means that the three diphthongal vowels of Dutch /au, ɛi, œy/) are not taken into account in the present research, due to the dynamic character of their formant frequencies.

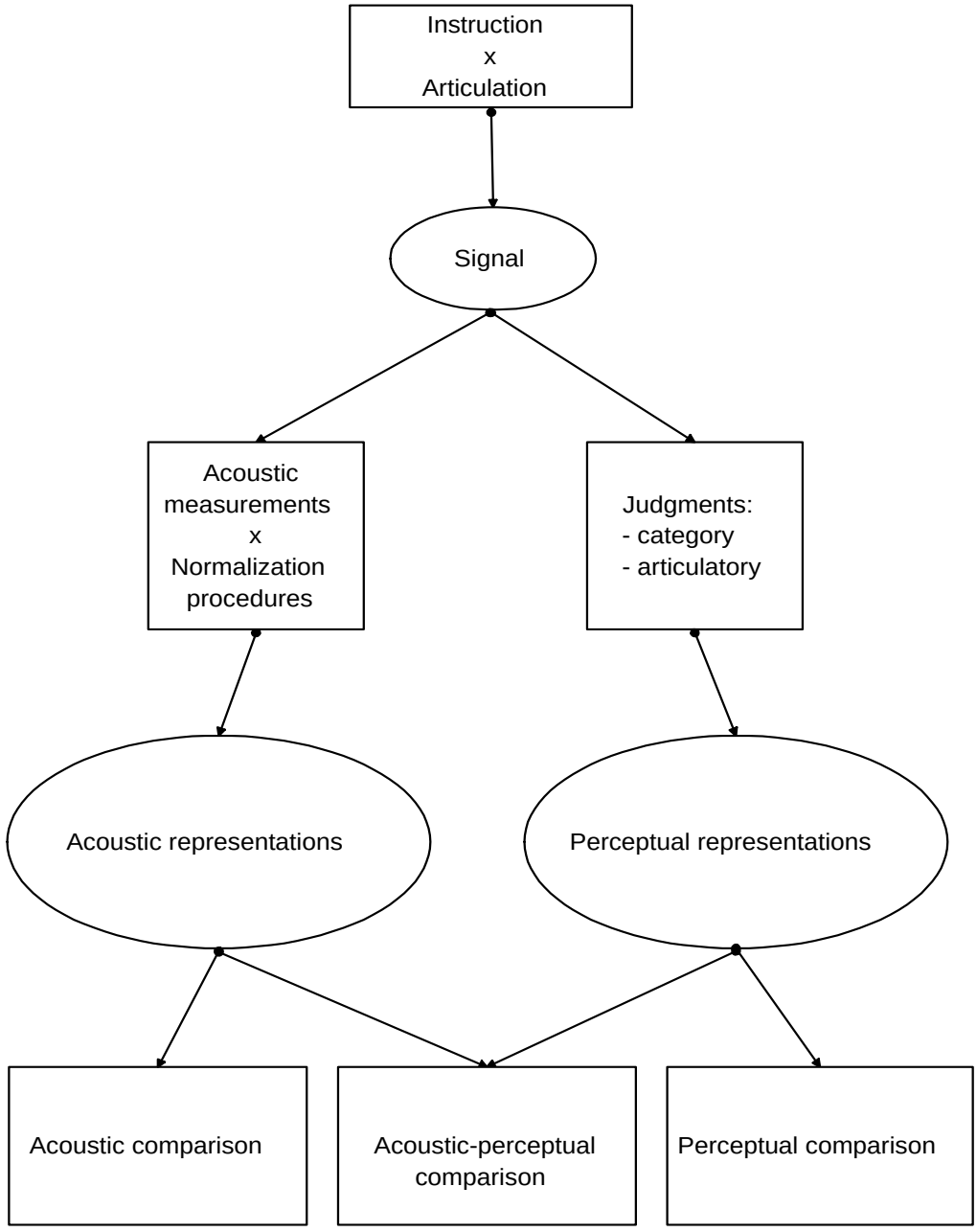

Figure 4.1: *The design of the present research.*

The three long mid vowels (/e, o, ø/) are excluded for the same reason: they show more diphthongization than the nine remaining monophthongal vowels of Dutch in some regional varieties of Northern Standard Dutch (cf. Van de Velde, 1996). The three long mid-vowels are regarded as semi-diphthongal and therefore excluded from the research design. The dynamic specification of the six (semi)-diphthongal vowels is considered to be an additional (and complicating) source of variation. If it is assumed that diphthongization results in variation in the values of the formant frequencies at different points in the vowel duration, then this variation is of a phonemic nature. If a normalization procedure preserves phonemic variation, then it can be expected that the dynamic variation in the formant frequencies is present in the normalized formant frequencies as well. The same can be argued for the consonantal context if it is assumed that the formant frequencies are affected systematically by the context.

The speakers are selected on their anatomical/physiological (e.g., sex, age) and sociological (e.g., regional background) characteristics. This is necessary to obtain a sociolinguistically balanced database. The speakers in this database must differ as much as possible in their anatomical/physiological characteristics[32], in order to be able to evaluate the normalization procedures on how well they minimized the acoustic consequences of these characteristics. Some of the sociological characteristics are kept constant (such as education level and profession) across all speakers in the database and others (such as regional background) are varied. This is necessary in order to obtain a database in which systematic sociolinguistic variation is present that can be attributed to a specific sociological factor. In total, speech from 160 speakers, 80 females and 80 males, is used. I describe the setup of this database in Chapter 5.

Signal

The ellipse in Figure 4.1 labeled 'Signal' refers to the acoustic signal, which was argued in Chapter 1 to transfer three types of variation: phonemic, sociolinguistic speaker-related, and anatomical/physiological speaker-related. This acoustic signal is used to generate the perceptual and the acoustic representations.

Acoustic measurements×normalization procedures

This box in Figure 4.1 denotes the process in which the raw values of F_0, F_1, F_2, and F_3 are measured. These measurements are taken at one point in the duration of the vowel token. Special attention is given to the formants: the aim is to automatically generate reliable formant frequencies and to minimize measurement error. To this end, two programs for formant estimation are evaluated, as is described in Chapter 6. The raw measurements of the four acoustic variables F_0, F_1, F_2, and F_3 are transformed through each procedure for normalization, described in Chapter 7.

[32] In order to maximize this variation, children should also be included, but no recordings of children were included in the data set that was used.

F_0, F_1, F_2, and F_3 are measured and subsequently transformed following each of the 12 selected normalization procedures. This way, 12 different acoustic representations are generated for each vowel token in the sociolinguistically balanced database. The normalized values of F_0, F_1, F_2, and F_3 are referred to as D_0, D_1, D_2, and D_3 throughout the present research, using the notation system described in Chapter 2.

Judgments: category and articulatory

This box in Figure 4.1 refers to the judgments of the stimulus vowel tokens, which are obtained through an experiment with phonetically-trained experts. Two types of perceptual representations are obtained: the vowel token's category label, the category judgment, and a judgment of the vowel tokens' perceived tongue height, tongue advancement, and lip rounding or spreading, i.e., the articulatory judgments. The category judgment is of a discrete nature and the articulatory judgment are considered to be of a continuous nature.

In the experiment, phonetically-trained are required to perform two tasks. They have to categorize each stimulus vowel token as one of the nine monophthongal Dutch vowels: /a, ɑ, ɛ, ɪ, i, ɔ, u, ʏ, y/. Second, they have to judge the articulatory characteristics of each stimulus vowel token by locating that vowel token in a phonetic space. The stimuli are read vowel tokens, produced by 10 male speakers and 10 female speakers of Dutch, a subset of the 160 speakers of the sociolinguistically balanced database.

The coordinates of the locations of the vowel tokens are regarded as the perceptual counterparts of the acoustic variables D_0, D_1, D_2, and D_3. The coordinates of the perceived articulatory characteristics of each stimulus token are obtained as follows. For each token the perceived tongue height, tongue advancement, and lip rounding has to be judged by locating the vowel token in a phonetic space, as was done in Ladefoged (1960) and Laver (1965). This phonetic space consists of a vowel quadrilateral in which the abscissa represents the perceived advancement and the ordinate represented the perceived height. However, in contrast with Ladefoged and Laver's studies, rounding is judged on a scale outside the quadrilateral to be able to obtain a (continuous) articulatory judgment of perceived rounding[33] A second reason to use a separate scale for rounding is to prevent confounding of judgments of advancement with rounding judgments; this is explained in Chapter 8 when the experimental procedure is discussed. Throughout the present research, I refer to the three articulatory variables, perceived tongue height, perceived tongue advancement, and perceived lip rounding, as Height, Advancement, and Rounding, respectively.

In the listening experiment, articulatory judgments are obtained as well as category judgments, because it was expected that the variation in the category judgments is small. This can be expected, given the results from the studies by Strange et al. (1976), Verbrugge et al. (1976), Macchi (1980), and Assmann et al. (1982), in which low error rates are reported

[33] Both Ladefoged (1960) and Laver (1965) used discrete responses for rounding.

in speaker-blocked as well as in speaker-mixed conditions. Another reason for obtaining articulatory judgments is to be able to establish whether the normalization procedures are able to model the judgment of articulatory differences within vowel categories.

In my opinion, phonetically-trained listeners can project the perceived articulation of a vowel token as one point in an three-dimensional phonetic space. Disner (1980, page 253) worded a similar view: "in some real sense, there is a level of representation, comprised of all but the speaker-particular aspects of the speech signal, in which each vowel is effectively a point in phonetic space.".

My choice to use an experimental approach comparable to Ladefoged's (1960), Laver's (1965), and Assmann's (1979) has two implications. In the first place, it means that I do not use phonetic transcriptions to represent the speech material. Narrow phonetic transcription can be said to be the common procedure in phonetics and sociolinguistics for describing differences in articulation. However, it seems justified to exclude narrow phonetic transcription as a possible way to represent vowel tokens perceptually, because I expect that allowing the listeners to use diacritical marks will introduce unwanted variation into the data. It may, for instance, be the case that some listeners have different notions about some of the diacritical marks than other listeners, or use different marks for perceived height, advancement, or rounding. Second, I expect that trained listeners are able to perceive and reliably record differences in the articulation of vowel tokens that are considerably more fine-grained than can be expressed through narrow phonetic transcription.

Acoustic representations

Two sets of acoustic representations are used. First, the normalization procedures are applied to the vowels from all 160 speakers. Second, a subset is distinguished; vowels produced by the 20 speakers that are also used to obtain the perceptual representations. The first data set is used in the acoustic comparisons, which are carried out in the process referred to by the box 'Acoustic comparison'. The second data set is used to carry out the comparison with the perceptual representation (based on the same speech material). This comparison is described in the box in Figure 4.1 labeled 'Perceptual-acoustic comparison'.

For each data set, the 12 acoustic representations are calculated separately. This is necessary, because two of the normalization procedures, MILLER and NORDSTRÖM & LINDBLOM, use scaling factors calculated using measurements from the speakers in the speaker groups that are compared.

Perceptual representations

The perceptual representation is obtained using speech material with the following characteristics. To limit the number of stimuli that are judged in the experiment, a subset of 20 speakers of the sociolinguistically balanced database is used in the experiment. The speech material of this subset must meet the two following requirements. The first requirement is

that the anatomical/physiological differences between the speakers are maximal, in order to be able to investigate whether (and how) the listeners are able to perceptually process a considerable amount of anatomical/physiological speaker-related variation in the speech signal. The second requirement is that the sociolinguistic speaker-related variation present in the speech material should be moderate. This is essential, because one of the goals of the present research is to evaluate whether the normalization procedures are able to preserve subtle sociolinguistic differences between vowel tokens. These differences can be indications of possible language changes. Furthermore, presenting listeners with vowel stimuli reflecting moderate to fine-grained sociolinguistic differences between and within vowel tokens may encourage them to use their perceptual scale optimally. If listeners are presented with stimuli in which large sociolinguistic differences between vowel tokens are present, it is not possible to establish whether phonetically-trained listeners can perceive, and reliably represent, subtle sociolinguistic differences.

Acoustic comparison

The process described in this box involves the comparison of the 12 normalization procedures on how well they preserve phonemic variation and sociolinguistic variation, while minimizing anatomical/physiological variation in the acoustic domain. This process is described in Chapter 7. The 12 normalization procedures are applied to the vowel data produced by the 160 speakers of the sociolinguistically balanced database.

It is evaluated how well the phonemic variation is preserved in the transformed acoustic representations by assessing how well the transformed data can be grouped into the intended vowel category as described in Nearey (1978), Syrdal (1984), and Deterding (1990). The procedure that delivers the highest percentage correctly classified vowel tokens, is considered to preserve phonemic variation best. Furthermore, it is evaluated how well the normalization procedure reduces anatomical/physiological speaker-specific variation, this time using multivariate analysis of variance. This is done to establish how much of the variance in the data of each transformed data set is related to characteristics such as the speaker's sex and chronological age. If a normalization procedure shows little anatomical/physiological variation, then this procedure is considered to efficiently minimize the anatomical/physiological speaker-related variation. Finally, the same set of transformed vowel data is used to evaluate how well sociolinguistic variation is preserved in the data. Here, the same procedure is used as for determining which procedure minimizes anatomical/physiological variation best. The only difference is that it is evaluated how much variation in the multivariate analysis can be attributed to sociological differences between speaker groups, such as regional background. The procedure that shows the relatively highest variance component, is considered to be the best option for preserving sociolinguistic variation.

Perceptual comparison

The following is concluded on the basis of the literature study described in Chapter 3. It is necessary to assess the relative importance of vowel-intrinsic and vowel-extrinsic sources of information for vowel categorization and judgment behavior by phonetically-trained listeners, to be able to carry out the comparison between the acoustic and perceptual representations.

In my opinion, the results from the literature study described in Chapter 3 show that is not feasible to directly establish which normalization procedure models human judgment behavior best, for three reasons. First, it is not clear how the judgment behavior of listeners is influenced by vowel-intrinsic or vowel-extrinsic and formant-intrinsic or formant-extrinsic sources of information in the speech signal, because no studies were published that investigate this matter. Second, although one study describes the effect of the availability of additional vowel-extrinsic information on vowel categorization by phonetically-trained listeners (Assmann, Nearey & Hogan, 1982), no studies have been published that investigated the relationship between the category that a vowel token was assigned to and the judgment of articulatory characteristics of that vowel token. It is necessary to establish whether the availability of a vowel token affects the judgments, because it can be hypothesized that the reliability of the judgments is influenced by the presence or absence of the vowel token's category label. Third, studies by Ladefoged (1960) and Laver (1965) show that judgments of vowel stimuli that are judged on different occasions show considerable variation. Consequently, it is not clear how reliably phonetically-trained listeners judge the articulatory characteristics of vowel tokens.

The listeners are required to judge a set of vowel stimuli under three different experimental conditions to be able to study the effect of two additional factors, i.e., the presence or absence of information about the vowel category's label and of information about the speaker. In the first condition, the stimuli are presented in a random order (speaker-mixed condition) and the listeners have to categorize each vowel token. In the second condition, the stimuli are presented blocked by vowel category, but mixed by speaker. In this condition, the listeners are not required to categorize the vowel token; the category label is provided (vowel-blocked condition). In the third condition, the stimuli are presented blocked per speaker (speaker-blocked condition). In this task, information about other tokens produced by the same speaker as available. In this condition, the listeners are required to categorize the vowel tokens.

By varying the information available to the listeners across conditions, differences in the judgments of the three variables Height, Advancement, and Rounding can be examined. For instance, it is expected that the variance of the judgments would vary across experimental conditions. All the differences that are expected between the three sub-experiments are discussed in Chapter 8.

Finally, it is decided to examine the reliability of the phonetically-trained listeners, because studies from Ladefoged (1960) and Laver (1965) show that listeners vary in their judgments over time, both between and within listeners. To examine whether this is the case

in my experiment, listeners are required judge a subset of the stimuli twice within each of the three experimental conditions, to allow the intra-rater as well as the inter-rater reliability to be established.

Perceptual-acoustic comparison

The perceptual representation consists of the judgments of Height, Advancement, and Rounding, while the acoustic data consists of the measurements of the fundamental frequency and the first three formant frequencies at one point in the vowel duration, transformed following each of the 12 selected normalization procedures. Both types of representations are generated using the same subset of the sociolinguistically balanced database.

The comparison of the perceptual and acoustic representation is carried out using regression techniques. Using linear regression analysis, it can be evaluated how well the perceptual data, the coordinates of Height, Advancement, and Rounding, are modeled using the (transformed) values of D_0, D_1, D_2, and D_3.

This comparison (described in Chapter 9) serves two purposes. It is used to determine which normalization procedure models judgment behavior by phonetically-trained listeners best. The perceptual representation is regarded as a 'human benchmark'. The procedure that generates an acoustic representation that showed the best fit with this human benchmark is considered to model the judgments of phonetically-trained listeners best. The comparison of acoustic and the perceptual representation is further used to obtain a mapping of the four acoustic variables, D_0, D_1, D_2, and D_4 onto the perceptual variables Height, Advancement, and Rounding.

Chapter 5

Speech Material

5.1 Introduction

This chapter describes the speech material used in the present research. The present research is part of a larger sociolinguistic project[34](cf. Van Hout et al., 1999), on pronunciation variation in the Netherlands and in Flanders, the Dutch-speaking part of Belgium. This project aims to generate more insight into how certain sociological characteristics (e.g., gender, age, regional background) of speakers affect variation and change in the articulation of standard Dutch (SD) as spoken in the Netherlands and in Flanders. To achieve this goal, realizations of all phonemes of Dutch were collected through a so-called 'sociolinguistic interview' and described in terms of their acoustic characteristics. These recordings constitute a sociolinguistically balanced data set. Although monophthongal and diphthongal vowels of Dutch were collected, I concentrate solely on the monophthongal vowel phonemes.

Section 5.2 describes the setup of the (sociolinguistically balanced) data set. In section 5.3, it is described which task in the sociolinguistic interview was used to obtain the target vowels used in the present research.

5.2 Design of the data set

The speech data was obtained from 160 speakers of standard Dutch (SD) who were stratified for the following sociological variables: speech community, regional background, gender, and age. All 160 speakers were teachers at secondary education institutes at the time the

[34]Also known as the 'VNC-project'.

interview was recorded. The majority of them were teachers of Dutch[35]. They were selected for the three following reasons. First, Dutch teachers can be considered professional language users, because they are expected to speak standard Dutch on a daily basis. Second, they are instructors of the standard language and can thus be regarded as having a normative role (Van de Velde & Houtermans, 1999). Third, Dutch teachers' speech is expected to show more variation than that of broadcasters, whose speech is used in most other pronunciation studies of variation and change in standard Dutch (Van Hout et al., 1999; Van de Velde & Van Hout, 2000).

Speech community

Two speech communities are distinguished: the Netherlands and Flanders. There are 80 Dutch speakers and 80 Belgian speakers.

The data set was split into a Dutch and a Belgian component, because the pronunciation of Dutch spoken in Flanders differs considerably from that of Dutch spoken in the Netherlands. Two different varieties are identified: Northern Standard Dutch as spoken in the Netherlands, and Southern Standard Dutch as spoken in Flanders. The differences in the pronunciation of the two varieties have evolved differently from the time the Dutch area was split up into two parts in the 19th century. See Van de Velde (1996) for a detailed description.

Regional background

The 160 speakers were sampled across four regions per speech community. In each community, four regions were appointed: a central region, an intermediate region and two peripheral regions (peripheral I and II). These are described below.

The central region is the economically and culturally dominant region in each of the speech communities. For the Netherlands, the central region is the west, consisting of the provinces of Northern-Holland, Southern-Holland and Utrecht, also known as "De Randstad" and referred to as "N-R" (Netherlands- Randstad) in the present research. The cities Amsterdam, Rotterdam, Utrecht, and The Hague are part of the Randstad. In Flanders, the central region is "Brabant", denoted as 'F-B' (Flanders-Brabant). Brabant contains the provinces Antwerpen and Flemish-Brabant, with the cities of Antwerpen and Leuven, respectively.

The intermediate region in the Netherlands encloses the southern part of the province of Gelderland, named "South-Gelderland" by Daan & Blok (1967), and a part of the province Utrecht. This region is referred to as "N-M" (Netherlands-Middle). The intermediate region in Flanders is the province of East-Flanders, referred to as "F-E" (Flanders-East). The intermediate region was thought to be a transitional region between the central region and the peripheral regions. Dutch as spoken in the intermediate selected regions was thought

[35]The original design aimed at 160 teachers of Dutch, but it turned out not to be feasible to find enough teachers to fill the design.

to show only moderate regional influences and to resemble Dutch as spoken in the central region.
The choice for the peripheral regions was made using criteria for geographical distance and linguistic distance. The linguistic distance is the distance between language varieties that are considered standard and language varieties that are considered less standard, such as regional dialects. Generally, dialects are spoken more in peripheral regions than in the central region. The linguistic distance is therefore expected to be larger between the varieties of Dutch spoken in the central region and the peripheral regions than between the central region and the intermediate region. In the Netherlands, the two peripheral regions are the province of Limburg, or "N-S" (Netherlands-South), in the south of the Netherlands, and the province of Groningen, or "N-N" (Netherlands-North), in the north of the Netherlands. The two peripheral regions for Flanders are the provinces of (Belgian) Limburg, or "F-L" (Flanders-Limburg), and of West-Flanders, denoted by "F-W" (Flanders-West).

Several towns were selected per region. The criteria for the selection of the towns were as follows. First, the selected towns in all regions had to have a comparable socioeconomic profile. Second, the towns within a region had to belong to the same dialect group. Third, the Dutch spoken in that town had to be regarded as characteristic of that region. No major cities were selected, because it was expected that the Dutch spoken in these cities is influenced by other dialects (or languages) as well as by the dialects spoken in the surrounding region, due to migration. And finally, there had to be enough teachers of Dutch at the schools to meet the requirements of the research design. Table 5.1 shows the towns that were selected per region and per speech community.

The teachers who participated as speakers in the interview taught at schools for secondary education in the selected towns. These speakers were required to meet the following requirements. First, at the time of the interview, they all lived in one of the selected towns, or near that town in the dialectal region characteristic for that region. Second, they were born in the region or moved there before their eighth birthday. Third, they had to have lived in the region for at least eight years prior to their 18th birthday. This last requirement was formulated on the basis of research studies by Payne (1980) and by Scovel (1988). Payne stated that children younger than eight years old have no difficulty acquiring the phonological system of the place they have moved to. Scovel concluded that learners of a second language generally do not acquire near-native pronunciation of this language after puberty. This last requirement was furthermore used to make sure that the speakers had lived in the town/region from an age at which they had no difficulties in learning the language variety spoken in that region or town.

Table 5.1: *The selected towns per speech community per region and the names of the corresponding provinces.*

Speech community	Region	Name	Selected towns
Netherlands	Central (*N-R*)	Randstad	Alphen aan de Rijn, Gouda
	Intermediate (*N-M*)	South-Gelderland	Tiel, Veenendaal, Ede, Culemborg, Elst
	Peripheral 1 (*N-S*)	Limburg	Sittard, Geleen, Roermond
	Peripheral 2 (*N-N*)	Groningen	Assen, Veendam, Winschoten
Flanders	Central (*F-B*)	Brabant	Lier, Heist-op-den-berg
	Intermediate (*F-E*)	East-Flanders	Oudenaarde, Zottegem, Ronse, Brakel
	Peripheral 1 (*F-L*)	Limburg	Tongeren, Bilzen
	Peripheral 2 (*F-W*)	West-Flanders	Ieper, Poperinge

Gender

One of the aims of the larger project of which the present research forms a part, is to get more insight into the roles of both genders[36] in processes of language variation and change. Therefore, a division was made for speaker-sex in the design. It is hypothesized that men and women adopt different roles in the process of changes in the standard language. Women use more prestigious pronunciation varieties. Also, the idea that women act as pioneers in language changes with a high prestige is generally accepted (Trudgill, 1983; Cameron & Coates, 1988, Labov, 1990; Holes, 1997; Chambers, 2003).

Age

The speakers were divided into two age-groups, a younger and an older one. The speakers in the younger group were born between 1960-1978 and were therefore 22 to 44 years old at the time of the interview. The speakers in the older age group were born between 1940 and 1955 and were between 45-60 years old at the time of the interview.

[36]Nevertheless, in the present research I was concentrated on the acoustic effects of biological gender, or of speaker-sex, because the sociolinguistic variation in the acoustic representation of vowels related to cultural gender (cf. Chambers, (2003) for the terminology) is expected to be small. Variation related to cultural gender can be found to a larger extent in consonants, for instance in Van Hout & Van de Velde (2000).

Design

In Table 5.2, an overview is displayed of the distribution of the listeners over the factors region, speaker-sex, and age.

Table 5.2: *The design of the research project, with 160 listeners distributed over 32 cells.*

Speech community		Central	Intermediate	Peripheral I	Peripheral II
Netherlands		*N-R*	*N-M*	*N-S*	*N-N*
younger	male	5	5	5	5
	female	5	5	5	5
older	male	5	5	5	5
	female	5	5	5	5
Flanders		*F-B*	*F-E*	*F-L*	*F-W*
younger	male	5	5	5	5
	female	5	5	5	5
older	male	5	5	5	5
	female	5	5	5	5

5.3 Sociolinguistic interview

The target phonemes from all the 160 informants were elicited through a sociolinguistic interview. The interview consisted of a guided part and a spontaneous part. In the guided part, all vowels and consonants of Dutch were recorded:

- All vowels of Dutch (/ɛ, ɑ, ɔ, ɪ, ʏ, a, o, e, i, ø, y, u, œy, ɔu, ɛi/) in a neutral context and in several consonantal contexts,
- All consonants of Dutch (/p, b, t, d, k, g, f, s, v, z, ɣ, x, ʃ, ʒ, m, n, ŋ, l, r, w, j, h/) in word-initial, intervocalic and word-final position.

A complete description of the design and further specifics of the entire sociolinguistic interview can be found in Van Hout et al. (1999).

Neutral context sentences

The speech material used in the present research was taken from the ‘neutral context’ task in the interview. The goal of this task was to elicit all the phonemes of Dutch in a consonantal

context that would influence the target phoneme as little as possible. The vowel tokens in their neutral consonantal contexts were recorded in a carrier sentence.

Two different carrier sentences were constructed for recording the vowels: one for the short vowels and one for the long vowels of Dutch. Dutch vowels are traditionally categorized into short vowels (/ɑ, ɛ, ɪ, ɔ, ʏ), long vowels (/a, e, ø, i, o, u, y/), diphthongs (/au, ɛi, œy/), and schwa (/ə/) (Booij, 1995). Schwa was not included in the target sentences, because it does not occur in stressed syllables in Dutch. The sentences have the following generic structure for the short vowels ('V' indicates the target vowel):

'In sVs en in sVsse zit de V'
/ɪn sVs ɛn ɪn sVsə zɪt də V/
[*In sVs and in sVsse is the V*]

For the long vowels and the diphthongs, the sentences have the following structure:

'In sVs en in sVze zit de V'
/ɪn sVs ɛn ɪn sVzə zɪt də V/
[*In sVs and in sVze is the V*]

The distinctions between the short vowels on one hand and the long vowels and diphthongs on the other hand was made because in Dutch an /z/ never occurs after a short vowel in a CVCV[37] consonantal context. It was thus decided to use a postvocalic /z/ for long vowels and diphthongs, and a post-vocalic /s/.

Of the three different consonantal contexts in which the target vowel was recorded (CVC, CVCV, or V), the CVC contexts were selected for further processing throughout the present research, because the vowels in isolation (V) lacked a consonantal context. It may be possible that the vowels in isolation are not as 'natural' than vowel tokens in a consonantal context, because only the monophthongal vowel /y/ occurs in isolation in standard Dutch. Furthermore, the vowels in isolation were produced in sentence-final position, which may also affect the naturalness of the F_0 values. The vowels in the CVCV consonantal context were not selected, because they contain different postvocalic consonants depending on the target vowel in the neutral context task (for the short vowels, this was /s/, and for the long vowels and the diphthongs it was /z/)[38].

The vowel tokens in the syllabic structure CVC can be regarded as vowels in a neutral context for Dutch. No recordings were made of vowels pronounced in the 'traditional' neutral /hVt/ consonantal context (/hVd/ for English) because it could not be predicted how people from Flanders would pronounce the word-initial /h/. Almost all people interviewed in

[37] Except in a handful of loan words (e.g. in *mazzel*, which translates as 'luck').

[38] An additional argument not to use the CVCV contexts with postvocalic /z/ is that in the CVCV contexts for the short vowels postvocalic /s/ is ambi-syllabic, whereas a postvocalic /z/, as it occurs in the CVCV contexts for the long vowels, is not.

Flanders are originally dialect speakers. They usually speak Southern Standard Dutch as well as their (native) dialect. Some (West-)Flanders dialects do not have an /h/ in their phoneme inventory. When asked to produce a word-initial /h/, some speakers of West-Flanders dialects tend to replace /h/ by either a glottal stop, or by /χ/. A /sVs/ context was adopted to make sure that no additional sources of variation were added to the data.

During the interview, the carrier sentences were presented to the speaker on a computer screen, with a three-second interval between sentences. When the speaker made a mistake, the interviewer interrupted the computer program and went back at least two sentences and asked the speaker to read these sentences again, to make sure that all sentences were recorded correctly. The neutral sentences task was performed twice during the interview; each vowel token was therefore available twice in each syllabic structure.

A total of 4800 vowel tokens was recorded in the neutral context sentences task: two tokens of each of the 15 vowel categories of Dutch, produced by 160 speakers.

Recording conditions

The recordings were made on DAT with a TASCAM DA-P1 portable DAT-recorder, with an AKG C420 Headset condenser microphone. The recordings were digitized through a Lucid Technology PCI24 digital audio card, and stored at 48 kHz on a PowerMac 7500/100. The neutral context sentences used in this experiment were down-sampled to 16 kHz, after applying an anti-aliasing low-pass filter with a cut-off frequency just below 8 kHz.

Recording conditions were different for each of the speakers. Some were interviewed in an empty classroom while others were interviewed at home. Due to these differences in recording conditions, background noises were audible in some cases. Whenever this was the case, the speech segment was excluded from further processing[39].

[39] As is described when the vowel data is discussed in detail, in Chapters 6 and 7.

Chapter 6

Acoustic measurements

6.1 Introduction

This chapter describes the procedure through which the measurements of F_0, F_1, F_2, and F_3 were generated. These measurements were obtained from the vowel data from the 160 speakers (described in Chapter 5).

Section 6.2 describes how the acoustic measurements were generated. These measurements represent the raw acoustic description of the vowel data. This section describes how F_0 was measured using the program Praat. Next, a comparison of two programs for formant estimation, Praat, Boersma, (2001) and Nearey's program (described in Nearey, Assmann, & Hillenbrand, 2002), is given. This comparison is followed by a verification of the measurements that were obtained through the program that was found to the most suitable of the two. Section 6.3 provides a summary of this chapter.

6.2 Measurements

6.2.1 Segmentation of vowel tokens

As described in Chapter 5, recordings were made for a total of 4800 vowel tokens: 15 vowels of Dutch, pronounced twice by 160 speakers. Before the acoustic measurements were obtained, I segmented all 4800 vowel tokens by hand in the digitized speech wave. In doing so, I made sure that the surrounding speech sounds (/s/) were not audible in the remaining signal. Segment labels were placed at zero crossings. The labeled vowel segments were extracted from their carrier sentences automatically, using the program Praat, version 4.02 (Boersma, 2001).

6.2.2 Fundamental frequency

For each vowel token, F_0 was extracted automatically with the program Praat, using the autocorrelation method, Praat's default procedure, and which is evaluated as the best option in Boersma (1993)[40]. I set the range within which the algorithm was to estimate F_0 between 50 and 300 Hz for male speakers, and between 100 and 500 Hz for female speakers. F_0 was extracted at the vowel's temporal mid-point. Although I estimated F_0 for all 4800 vowel tokens in this fashion, I discuss only the measurements for the 2880 monophthongal vowel tokens (nine vowel categories $\times$ 160 speakers $\times$ two tokens).

As a first step, the data was checked for outliers (a case with a value between 1.5 and 3 times the interquartile range (IQR)), and extreme values (cases $>$ 3 IQR)[41]. Whenever I refer to outliers or extreme values, this definition is meant. The data of each speaker was investigated separately. Every outlier and extreme value was verified by hand.

In total, 74 outliers and extreme values were found in the 2880 cases. Of these 74 cases, 63 cases were remeasured and reinserted in the database and 11 cases were excluded from further analysis. Of the 63 cases, 26 were octave errors that were remeasured by hand and subsequently reinserted. The remaining 37 outliers were neither caused by octave errors, nor by anomalous behavior of the F_0 measurement algorithm, nor by idiosyncrasies of the speaker's voice. To explain these findings, I inspected the other vowels pronounced by the same speaker, and I concluded that each one of these 37 values must be attributed to variation in F_0 of that speaker. Therefore, these 37 values were included in the data set. The remaining 11 cases were excluded from further analysis, because the voice characteristics of the speaker (e.g hoarseness) did not allow F_0 to be measured reliably. These 11 cases were replaced by the mean F_0 for that speaker[42]. This was done because some of the procedures for vowel normalization – that are evaluated in Chapter 7 – use speaker-dependent scale factors (e.g., LOBANOV, GERSTMAN) and these factors could be biased if calculated using fewer vowels for that speaker.

6.2.3 Algorithms from Praat and Nearey

Two programs for obtaining formant measurements were compared, in order to obtain valid formant measurements. The first program, Praat, is widely-used for obtaining formant measurements. Using the algorithm embedded in this program, formant measurements can be obtained semi-automatically.

The second program, developed by Nearey (Nearey et al., 2002), uses a formant measuring algorithm comparable to the algorithm used in Praat, and formant frequencies can

[40] It was not feasible to compare Praat's estimations of F_0 with Nearey's, because Nearey's program did not provide estimations of F_0.

[41] Using the definition of the statistical package (SPSS).

[42] Which is a standard procedure for replacing missing values.

also be obtained semi-automatically. Nearey's program differs from Praat in two respects: it generates alternative sets of solutions (i.e., a different set of measurements) for each vowel token and it allows the user to manually verify and adjust the formant measurements.

In this section, the technical details of the two programs are described and subsequently a description is given of how both programs were used to generate formant measurements using the same set of vowel data. Next, a description is given of how the quality was assessed of the measurement resulting from the two programs.

Nearey's program

Nearey's program (described in Nearey et al., 2002) consists of two parts: a formant tracking algorithm and a user interface that allows the user to verify, and, where necessary, adjust the formant tracks generated by the tracking algorithm. The hypothesis is that Nearey's program has two advantages compared with Praat. The formant measurements obtained with this program should first be more reliable and accurate than measurements obtained with the other programs, as it goes through several cutoff frequencies and selects the best option for each vowel token separately when making the formant estimations. Other programs generally allow the user to select only one cutoff frequency that is identical for all vowel tokens to be measured. Second, the interface in Nearey's program allows for making adjustments to the output formant tracks: the course of the tracks proposed by the algorithm can be altered, or the user can select an alternative cutoff frequency than the one proposed by the tracking algorithm. This way, possible errors of the system can be corrected and the resulting formant tracks can be altered to obtain better measurements.

The setup of Nearey's program is as follows. The program's preprocessing consists of applying a cosine4 window with a time-step of two milliseconds[43]. Subsequently, three formant candidates are estimated by means of root extraction, using a version of Markel & Gray's (1976) "FORMNT" algorithm, followed by a five-point running median smoothing. The number of LPC-coefficients is fixed at nine. For each vowel token, the settings were identical.

Three parameters must be set by the user. The tracking program puts considerable emphasis on the frequency range within which the three formant tracks are to be estimated. This frequency range consists of two parts, a lower range and an upper range, the lower range is fixed between 0 and 3000 Hz, while the upper range is set between 3000 and a variable value constituting the highest cutoff frequency in Hz. The user determines the value of this highest cutoff frequency. Within this upper range (3000 to the highest frequency), the user must also provide the number of cutoff frequencies[44] to be evaluated. For instance, if the

[43] A cosine4 is a cosine window 4-cubed; interval from $-\frac{1}{2}\pi$ to $+\frac{1}{2}\pi$.

[44] These cutoff frequencies can be regarded as estimates of a frequency between F_3 and F_4: the cutoff frequency of the F_3.

user set the highest cutoff frequency to 4000 Hz, and the number of cutoff frequencies to six, the following six frequencies are evaluated: 3000 Hz, 3200 Hz, 3400 Hz, 3600 Hz, 3800 Hz, and 4000 Hz. The third parameter that must be set concerns the distribution of the F_3 cutoff frequencies across the total range, i.e., whether the distance between the cutoff frequencies is spaced logarithmically (Log) or linearly (Hz) across the upper range.

Subsequently, the program compares the three formant tracks associated with each of the cutoff frequencies to be evaluated. This comparison consists of tracking all three formants within the range specified by each cutoff frequency and by evaluating the results[45]. Finally, the program forwards one particular high cutoff frequency, together with the three formant tracks that were estimated using this cutoff frequency, as the best option.

The resulting formant tracks for the input vowel tokens are to be verified by hand in the user interface of the program. For this purpose, the tracks are plotted on the smoothed spectrogram. It is possible to alter the course of each track. For instance, if the F_1 is estimated at a frequency that is probably the frequency of F_0 (a 'formant jump'), which can be expected for speakers displaying a high F_0 for close front vowels, such as /i/, the track for F_1 can be altered to position it at the appropriate place in the spectrogram. In addition, the user can cut off the beginning and/or ending of the tracks, if the initial or final /s/ appears to have affected the course of the tracks. Finally, it is possible to select one of the five alternative tracks (using the other five cutoff frequencies), if another cutoff seems more appropriate when inspecting the smoothed spectrogram.

In order to find the optimal configuration of Nearey's program, I investigated the effect of three variables: the range within which the third formant was to be tracked (the highest cutoff frequency was set to 3900, 4200, or 4500 Hz), the number of cutoff frequencies in the upper range (5, 6, or 8 cutoff frequencies), and the spacing of these cutoff frequencies (log or Hz).

Table 6.1: *Combinations of input settings for calibrating Nearey's tracker.*

Combination	High cutoff	Nr. of cutoffs	Spacing
1	3900	5	Hz
2	4200	5	Hz
3	4200	5	log
4	4200	6	log
5	4500	5	Hz
6	4500	6	Hz
7	4500	8	Hz

[45] Using a complex system of calculating nine 'figures of merit' (Nearey, personal communication, 2002). The best cutoff frequency is calculated by multiplying the scores for all nine figures of merit. The program forwards the cutoff frequency with the largest compound score is forwarded as the best option.

When selecting combinations of these three variables, I used the smallest number of combinations possible. In total, I evaluated the seven combinations of settings, listed in Table 6.1. The rationale behind each of these seven combinations was the following. Combination 1 was selected to investigate the effect of a low F_3 cutoff frequency. Combination 2 was used to investigate the effect of a slightly higher cutoff frequency than the one used in combination 1. Combination 3 was used to test the effect of log spacing instead of linear (Hz) spacing. Combination 4 was used to assess whether a higher number of cutoff frequencies than in combination 2 in the same range would improve the quality of the measurements. Combinations 5 and 6 were used to establish whether the results found with combinations 2 and 3 would remain stable with wider spaced cutoff frequencies. Combination 7 was used to find out if using a considerably higher number of cutoff frequencies affects the quality of the measurements. Pairwise comparisons were made between combinations of settings, to be able to decide which one of two parameters yielded better results.

The seven combinations of input settings for the tracker were evaluated using the vowel data from 20 speakers from the N-R region in the sociolinguistically balanced data set described in section 5.3. For each of the N-R speakers, both occurrences of each monophthongal vowel (/ɑ/, /a/, /ɛ/, /ɪ/, /i/, /ɔ/, /u/, /ʏ/, /y/) were used. In total, 360 vowel tokens (nine vowel categories $\times$ 20 speakers $\times$ two tokens per vowel category) were selected for this evaluation. These 360 vowel tokens were passed through the tracker seven times.

The raw output data from the seven combinations was not verified by hand, or altered in any way, because checking the data by hand may obscure the differences between the combinations. For each combination, for each vowel token, the frequencies for F_1, F_2, and F_3 at each vowel token's temporal center point were extracted from the tracker's results.

The seven combinations were evaluated using linear discriminant analysis (LDA)[46]. Using LDA, I evaluated how well the formant data from each combination could be used to classify each vowel token into the corresponding intended vowel category. For each vowel token, the values of the first three formants were entered as predictors. For each procedure, the percentage of correctly classified vowel tokens was calculated on the basis of these three predictor variables. A higher percentage of successfully classified vowel tokens was thought to reflect fewer mistakes in the tracking process. For each combination of settings, the discriminant analyses were repeated three times, once for the 10 male speakers, once for the 10 female speakers, and once for all 20 speakers combined.

Before the discriminant analyses were performed, the correlations between the data resulting from the seven combinations were calculated for F_1, F_2, and F_3. The correlations between the combinations were overall high to very high. The correlation coefficients (Pearson's r) ranged between 0.97 and 1.00 for F_1, between 0.87 and 0.97 for F_2, and between

[46] A non-linear, or quadratic, discriminant analysis could also have been used here, because it can be expected that the covariance matrices between groups are unequal. However, it was decided to use linear discriminant analysis throughout this research, for reasons discussed in further detail in Chapter 7.

0.75 and 0.97 for F_3. The slightly lower correlation coefficients for F_3 were due to pairs involving combination 1; when combination 1 is not included, the coefficients range between 0.90 and 0.97. The results suggest that the value of 3900 for the highest cutoff frequency used in combination 1 may be too low for reliable calculation of F_3.

Table 6.2: *Percent correctly classified vowel tokens from the LDAs.*

Combination	Female speakers	Male speakers	All speakers
1	83	82	76
2	86	81	76
3	86	82	76
4	83	82	75
5	83	79	76
6	86	80	76
7	81	81	74

Table 6.2 shows the results of the discriminant analyses. All percentages are rounded of to the nearest whole number, as is done throughout this research. It can be observed that all percentages correctly classified vowel tokens are around 80%, for the male as well as for the female speakers, whereas the percentages for all speakers are slightly lower, around 76%. Therefore, it must be concluded that all seven combinations produced data that could be classified reasonably accurately, especially considering that this data was not verified manually and therefore contained some errors.

Table 6.2 shows furthermore that combinations 1, 2, 3, 5, and 6 yield the highest scores (76%) for all speakers. Of these five, combination 3 appears to be the best option, because it leads to the highest percentage of correctly classified vowel tokens for male and female speakers. Moreover, it can be observed that the log scale (combinations 3 and 4) produces slightly higher percentages than the Hz scale (combinations 1, 2, 5, 6, 7), a cutoff frequency of 4200 yields higher scores than higher and lower cutoff frequencies (3900 and 4500), and a smaller number of cutoff frequencies produces higher percentages correctly classified than a higher number.

Given the results of the evaluation described above, I decided that combination 3 was the best option. Therefore, combination 3 was used for all formant measurements described in the present research using Nearey's program.

Praat

As was done for Nearey's program, several combinations of settings were evaluated for Praat's formant measuring algorithm. It was decided to evaluate several settings of Praat's

default procedure. Praat's default procedure for formant estimation is the Burg-algorithm, which is implemented using algorithms from Press et al. (1992). For formant estimations, one signal is calculated every 10 ms using a Gaussian window with a length of 51.9 Hz, shifted every 10 ms. The number of LPC-coefficients is linearly dependent on the highest cutoff frequency: two coefficients are calculated per 1000 Hz, or per 1100Hz (for the female speakers). For female speakers, five formants are to be found in a range of 0-5500 Hz and five formants in the 0-5000 Hz range for male speakers. This means that for the females and for the males 10 LPC-coefficients were estimated. Although five formants were measured, I used only the values for the first three. I set the program to measure the formant frequencies at each vowel token's temporal midpoint. It should be noted that, because the formant frequencies are not passed through a tracking algorithm after their estimation by the LPC-procedure, Praat's procedure delivers formant candidates. However, this does not necessarily affect the quality of the formant frequencies. A tracking procedure can be seen as an algorithm that 'connects the dots' across all frames of the signal (where the 'dots' are the formant candidates in each frame). This implies that if the measurement algorithm is good, then tracking these formant candidates does not result in better measurements.

The height of the highest cutoff frequency and the shift of the Gaussian window were evaluated. Four combinations of settings were evaluated. Combination 1 was identical to the default settings: the shift was 10 ms and five formants were to be estimated in 5000 Hz for male speakers and in 5500 Hz for female speakers (only the values of the first three formants were used). In combination 2, all settings were identical to the default settings (and thus to combination 1), the only difference was that the shift was set at 2 milliseconds instead of 10. In combinations 3 and 4, three formants were to be estimated in 3000 Hz for male speakers and in 3300 Hz for female speakers. In combination 3 the window shift was set at 10 ms and the window shift was set at 2 ms. Table 6.3 lists the specifics of the four combinations.

Table 6.3: *Combinations of input settings for calibrating Nearey's tracker.*

Combination	Nr. Formants & Cutoff frequency	Window shift
1	5 formants in 5000/5500 Hz	10 ms
2	5 formants in 5000/5500 Hz	2 ms
3	3 formants in 3000/3300 Hz	10 ms
4	3 formants in 3000/3300 Hz	2 ms

The four combinations were used to generate four sets of measurements of F_1, F_2, and F_3, using the same 360 vowels as used to calibrate Nearey's program. These combinations were subsequently evaluated using linear discriminant analysis (LDA) to establish how well the data resulting from each combination could be used to classify each vowel token into the corresponding intended vowel category. For each vowel token, the values of the first three

Table 6.4: *Percent correctly classified vowel tokens from the LDAs for the four combinations for Praat.*

Combination	Female speakers	Male speakers	All speakers
1	76	79	72
2	73	77	71
3	64	74	66
4	58	75	65

formants were entered as predictors. For each combination, the percentage of correctly classified vowel tokens was calculated using the three predictor variables. A higher percentage successfully classified vowel tokens was thought to reflect fewer mistakes in the measuring process. For each combination, the discriminant analyses were repeated three times, once for the 10 male speakers, once for the 10 female speakers, and once for all 20 speakers combined. The formant data was not checked for outliers or extreme values.

Correlation coefficients were calculated between measurements resulting from the four combinations for F_1, F_2, and F_3. The correlation coefficients (Pearson's r) ranged between 0.94 and 0.99 for F_1, between 0.91 and 0.99 for F_2, and between 0.84 and 0.98 for F_3 and were thus overall high to very high.

Table 6.4 shows the resulting percentages correctly classified vowel tokens per combination for the female speakers, male speakers, and for all speakers. It can be observed that combination 1 yields the highest scores for all three sets of speakers (76% females, 79% for the males, and 72% for all speakers). Combinations 3 and 4 perform poorer than combinations 1 and 2, indicating that estimating five formants in 5000 or 5500 Hz leads to higher percentages correctly classified vowel tokens than estimating three formants in 30000 or 3300 Hz. In addition, combination 1 performed marginally better than combination 2 and combination 3 performed marginally better than combination 4, which indicates that a time shift of 10 ms is more suitable than a shift of 2 ms. Closer inspection of the measurements of combinations 3 and 4 revealed that their low scores can most likely be accounted for by the fact that F_3 could not be measured for over 45% of the vowel tokens for the female speakers and over 30% for the male speakers. It was found that this was the case for the vowels /i/, /ɪ/, /a/, and /ɑ/. A possible explanation is that the cutoff frequency of 3000 or 3300 Hz is too low for these vowels to allow F_3 to be measured reliably. Given the results in Table 6.4, it was decided to measure the formant frequencies for the comparison with Nearey's program using the settings of combination 1, i.e., Praat's default settings.

6.2.4 Comparing Nearey's and Praat's measurements

The 360 vowel tokens from the 20 speaker of the N-R region were measured twice, once using Nearey's program, with the settings of combination 3, and once with Praat's combination 1.

Table 6.5: *Parameters for Nearey's program and for the Praat program.*

Parameters	Nearey	Praat
Window shape	Cosine4	Gaussian
Window length	125 Hz	51.9 Hz
Time step	2 ms	10 ms
Estimating LPC-candidates	FORMNT (Markel & Grey, 1976	Burg (Press et al., 1992)
LPC-coefficients	9	10
Cutoff-frequency	Variable: 3000, 3263, 3550, 3861 or 4200 Hz	Not variable, males: 2500 Hz, females: 2750 Hz
Tracking	yes	no

The specific characteristics of each procedure are displayed in Table 6.5.

As a first step in the analysis of both data sets, the data was checked for outliers. All outliers were included in the data set and all extreme values were inspected visually, and where thought necessary, excluded from further analysis. This procedure was repeated for male and female speakers, for each of the nine vowel categories, and for the values of F_1, F_2, and F_3 separately. Four vowel tokens measured with Praat were remeasured. After remeasuring, they were included in the data set. The data from Nearey's program showed no extreme values, therefore no cases were excluded from the analysis.

To establish the correlation between the measurements from both programs, Pearson's r was calculated. For F_1 an r of 0.95 was found, an r of 0.94 for F_2, and r was 0.73 for F_3. No substantial differences were thus found between the measurements of the first two formants by the two programs, although the measurements for F_3 were found to differ more.

To get more insight in the precise nature of the differences between the two programs, three scatter plots are displayed in Figure 6.1. Figure 6.1 shows that the variation around the center line does not seem to be random. Overall, the measurements obtained using Praat are lower than those obtained using Nearey's program.

For F_1, the deviations from the center line seem to be caused mainly by the fact that Praat's estimations of /ɑ/ and /a/ are lower than Nearey's estimations. For some tokens for /ɑ/ in Figure 6.1 the estimations by Praat are between 300 and 400 Hz, while Nearey's program generated estimations between 600 and 900 Hz. A possible explanation may be that Praat estimated a frequency somewhere between the fundamental frequency and the first formant for these tokens. The same can be said about Praat's low values for /a/. For F_2, it seems that Praat's estimations of /i/, /ɛ/, /y/, and /ɪ/ were lower than Nearey's estimations. For instance, for /i/ some of Praat's estimations in Figure 6.1 are around 1500 Hz, while Nearey's estimations of those vowel tokens are around 2800 Hz. It seems that Praat missed

F_2 altogether for these tokens, and chose a value intermediate between F_1 and F_2 instead. For F_3, deviations from the center line can be observed for each vowel category, although it can be observed that the some of the tokens from vowels /i/, /y/, and /ɛ/ are lower for Praat's program, presumably due to estimations that were somewhere intermediate between F_2 and F_3.

Linear discriminant analyses (LDAs) were carried out on the formant data generated by both programs to establish if there were differences between Praat and Nearey's measurements. In each of these analyses, the values of the first three formants served as predictor variables, while the percentage vowel tokens that were classified into their corresponding vowel category was used as the dependent variable. It was assumed that higher percentages classified would reflect less random variation in the formant estimations and less overlap between the vowel categories. This analysis was repeated for the female and the males separately and for the pooled data of all 20 speakers.

Table 6.6: *Percent correctly classified vowel tokens resulting from the LDAs.*

Program	Female speakers	Male speakers	All speakers
Praat	79	81	72
Nearey	86	82	76

Table 6.6[47] shows the percentages correctly classified vowel tokens from the LDA on Nearey's and Praat's measurements. All percentages for the data estimated using Nearey's program are higher than those for Praat, 1% to 7 % higher. Overall, Nearey's program produced data that could be classified more easily into vowel categories. This indicates that Nearey's program represented the phonemic variation in the raw formant frequencies better than Praat. Therefore, Nearey's program was considered to be the best option for obtaining formant measurements of the two.

Given the analyses described in this section, i.e., the calibration of the ideal settings of Nearey's program and the comparison with Praat, it can be concluded that Nearey's program produced better formant measurements than Praat. Nearey's program was therefore used to measure the total set of vowels. All 4800 vowels were passed through Nearey's formant tracker, both the diphthongs (1920) and the monophthongal vowel tokens (2880). In the present research, only the results concerning the 2880 monophthongal vowel tokens are discussed in further detail.

Summarizing, the three formant frequencies for all vowel tokens described in the present research were measured using Nearey's program, using a highest cutoff frequency of 4200 Hz, with different five cutoff frequencies to be evaluated that were log-spaced throughout the higher range (3000-4200 Hz).

[47] The values for the Praat data are slightly higher than in Table 6.4, because here the data were checked for outliers and missing values, which was not the case for the data used in Table 6.4.

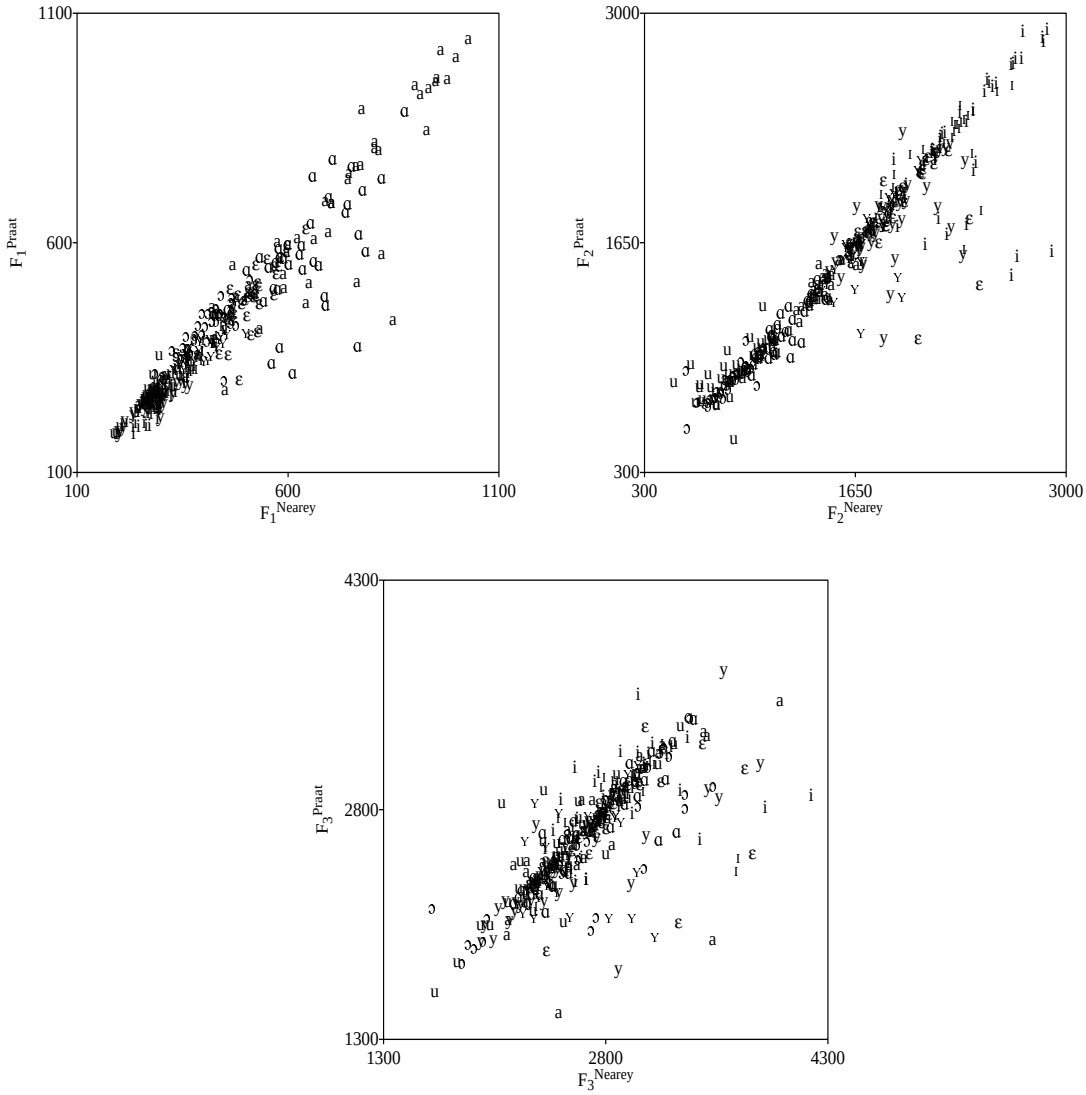

Figure 6.1: *Scatter plots of Praat's and Nearey's estimations of F_1, F_2, and F_3 in Hz.*

6.2.5 Verification of formant frequencies

All 2880 vowel tokens were verified with the user interface implemented in Nearey's program. The program presents the user with the smoothed LPC-spectrogram, with the three proposed tracks with their corresponding cutoff frequency printed on it.

All vowel tokens were verified as follows. First, the cutoff frequency proposed by the tracker was evaluated. In Nearey's program, it is possible to view the three formant tracks for each of the four alternative cutoff frequencies, and, if necessary, to select one of these alternatives (including the three corresponding formant tracks). Second, the formant tracks can be altered independently of the proposed cutoff frequency. For instance, if the second formant's track was set too high (e.g., in back vowels like /u/), the track can be put at the correct frequencies by hand. Third, the user can remove parts from the signal (for instance if noise from the surrounding /s/-sounds was still audible and visible). Finally, the user can choose to save the formant at the vowel's temporal midpoint or the user can choose to reject the entire set of measurements. Overall, the modifications that I made in the formant tracks were small, although in some cases (especially for /ɔ/, for which F_3 was often estimated at a too-low frequency) the position of F_3 was adjusted.

Table 6.7, which lists the percentage of altered and unaltered vowel tokens, shows that the data from the 40 female speakers from the Dutch community required the highest percentage of modifications[48], whereas the 40 Flemish males required the lowest percentage. The Flemish data required fewer modifications than the Dutch data.

After calculating the outliers and extreme values (for the male and female speakers and for each vowel separately), it was decided to include all outliers in the data set, while all extreme values were verified using the tracking program's graphical user interface again. This was done to ensure that the extreme value was not the result of a mistake that was made during the first round of manual verification with the interface. As a result, three vowel tokens were excluded from further analysis due to background noises: two from the Dutch data set and one from the Flemish data set. All three tokens were produced by male speakers. For these three tokens, the values for F_1, F_2, and F_3 were replaced by the mean values calculated for all remaining vowels for the corresponding speaker. This was done for the same reason as given for the measurements of F_0, described in section 6.2.2.

In addition, an LDA was carried out on unverified and verified data from the tracker, to ensure that manual verification did in fact improve the quality of the results. The analysis was carried out using only the monophthongal vowel data from all eight speaker groups (N-R, N-M, N-S, N-N, F-B, F-E, F-L, and F-W). Each speaker group consisted of 20 speakers. Both tokens of each monophthongal vowel were included in the data set. The total number of cases was 2880 for the unverified data and 2877 for the verified data. In each LDA, the values of the first three formants for each vowel token served as the predictor variables, while the percentage of correctly classified vowel tokens served as the dependent variable.

[48] The majority of the modifications were minor changes to the course of the formant tracks.

Table 6.7: *Percent unaltered and altered data for the 2880 monophthongal vowel tokens. The number of speakers per cell is 20.*

%	Netherlands		Flanders	
	Females	Males	Females	Males
Altered	54	49	47	45
Unaltered	47	51	53	56

Table 6.8: *Percent correctly classified vowel tokens from the linear discriminant analyses for the unaltered and altered data.*

%	Unverified data		Verified data	
Region	Female	Male	Female	Male
N-R	86	82	85	84
N-M	84	76	88	86
N-S	80	74	81	86
N-N	89	77	87	86
F-B	87	89	88	90
F-E	87	89	92	93
F-L	90	82	88	89
F-W	86	81	86	87
All regions	80	75	81	82

Table 6.8 displays the percentage of correctly classified vowel tokens from the LDA on the verified and the unverified data. Given these results, it can be concluded that verification of the data generally improved the classification of the male speakers; the percentages correctly classified vowel tokens are overall higher than the females' percentages. Both pooled percentages for the verified data are higher than the percentages for the unverified data. Some of the percentages in Table 6.8 are higher for the verified data, but, overall, those differences are minor, compared with the improvements found for, for instance, the N-S region for the male speakers (12% higher after verification). Given these results, it seemed justified to assume that manual verification of the data resulted in higher quality formant measurements.

6.3 Summary

This chapter presents an outline of how the measurements of the fundamental frequency and the first three formants of each vowel token were obtained. First, the measurements of F_0 are discussed, these measurements were generated using the program Praat. For each vowel token, the fundamental frequency was extracted automatically and subsequently verified by hand. Second, process is described through which the measurements of F_1, F_2, and F_3 were obtained. To obtain valid measurements, I compared two programs for semi-automatically measuring formant frequencies: the program Praat and a program developed by Nearey. Before carrying out formant measurements using Nearey's program, its optimal settings were estimated, using a subset of the sociolinguistically balanced database, described in Chapter 5. This procedure was repeated for Praat; several settings of its default procedure were evaluated using the same data as used for the calibration of the settings of Nearey's program. After the optimal settings were established for both programs, a comparison of the formant frequencies measured with the two programs was carried out using Linear Discriminant Analysis. I concluded that the measurements made with Nearey's program were better than Praat's. Third, it is described how all 4800 vowel tokens of the sociolinguistically balanced data set (described in Chapter 5) were measured using Nearey's program and verified by hand.

Chapter 7

Acoustic comparisons

7.1 Introduction

This chapter describes the comparison of the 12 normalization procedures in the acoustic domain. The procedures are evaluated on how well they meet the criterion proposed in Chapter 1. A procedure is considered to meet this criterion when the application of the procedure results in better representation of the phonemic variation compared with the baseline (no normalization). This improved performance should be accompanied by a reduction of the anatomical/physiological variation, and by the preservation of the sociolinguistic variation. If a procedure performs poorly at one (or more) of these tasks, it is decided that it does not meet the criterion in the acoustic domain.

The acoustic comparisons were carried out as follows. As a first step, the normalization procedures were applied to the raw acoustic measurements of F_0, F_1, F_2, and F_3 (described in Chapter 6). These measurements were obtained using the vowel data from all 160 speakers from the sociolinguistically balanced data set (described in Chapter 5). Thus, 12 normalized sets of measurements were obtained.

Section 7.2 describes to what degree the 12 normalization procedures meet the criterion. In section 7.3, the three procedures that meet the criterion best are discussed in more detail as to how they deal with specific sources of sociolinguistic variation. Section 7.4 presents the conclusions of this chapter. Parts of the research that is described in this chapter are also described in Adank (1999).

7.2 General comparisons

7.2.1 Applying normalization procedures

In Chapter 2, a detailed description of the 12 normalization procedures was given. In addition, it was described how the procedures are classified as vowel-intrinsic/formant-intrinsic, vowel-intrinsic/formant-extrinsic, vowel-extrinsic/formant-intrinsic, or as vowel-extrinsic/formant-extrinsic procedures. All procedures were applied to the raw measurements of F_0, F_1, F_2, and F_3 for each of the 2880 vowel tokens. However, because the implementation of NORDSTRÖM & LINDBLOM and MILLER involves the calculation of scale factors that depend on the specifics of the speaker groups that are compared, a short description of these procedures is given here.

First, my interpretation of NORDSTRÖM & LINDBLOM (formula (2.12)) entails multiplying each frequency of F_0, F_1, F_2 and F_3 in Hz of the female speakers with a scale factor k. k was calculated using equation (2.13), as described in section 2.4.4, and was set at 0.87 for the population of 160 speakers.

Second, MILLER's three dimensions were calculated using equations (2.15), (2.16), and (2.17). MILLER uses a sensory reference SR (equation (2.14), which incorporates a scale factor k. Here, k is the geometric mean of μF_0 (mean F_0) for the female speakers and μF_0 for the male speakers. Because μF_0 for the 80 female speakers was 233.9 Hz, and μF_0 for the 80 male speakers was 147.7 Hz, the geometric mean (k) was set to 185.9 Hz.

7.2.2 Preserving phonemic variation

A series of discriminant analyses was carried out to establish how well the 12 normalization procedures preserved phonemic variation in the transformed version of each vowel token. The purpose of these analyses was to establish how well the vowel data produced by younger and older male and female speakers – containing considerable anatomical/physiological speaker-related variation – can be categorized. The transformed acoustic variables (D_0, D_1, D_2, D_3) for each procedure served as predictor variables, while the intended vowel category, having nine possible values, was the dependent variable. The percentage of correctly classified vowel tokens was used as the measure to express how well the 12 procedures preserve phonemic variation.

Linear discriminant analysis (LDA) is a standard pattern recognition technique that assumes that the covariance matrices are equal across categories. The expectation is that the data cannot meet this assumption. In cases like these, quadratic discriminant analysis (QDA) is the appropriate analysis. A disadvantage of QDA is that it requires much larger numbers of parameters to be estimated. Therefore, for the first analysis LDA as well as QDA were carried out, to establish whether the percentages correctly classified vowel tokens by the two analyses were substantially different.

Table 7.1: *Results for LDA 1 and for QDA 1: percent correctly classified vowel tokens from the linear and quadratic discriminant analyses on the pooled data from 160 speakers. For LDA 1, all percentages higher than 81% or lower than 77% (all percentages are rounded off to the nearest whole number) are significantly different from the baseline condition (*HZ*). For QDA 1, these percentages are 79% and 83%, respectively. The four classes of procedures (vowel-intrinsic/formant-intrinsic, vowel-intrinsic/formant-extrinsic, vowel-extrinsic/formant-intrinsic, and vowel-extrinsic/formant-extrinsic) are separated by horizontal lines.*

%	LDA 1	QDA 1
HZ	79	81
LOG	80	81
BARK	80	82
ERB	80	82
MEL	80	82
SYRDAL & GOPAL	69	70
LOBANOV	92	93
GERSTMAN	84	86
CLIH$_{i4}$	90	91
CLIH$_{s4}$	82	83
NORDSTRÖM & LINDBLOM	82	84
MILLER	76	77

Table 7.1 shows the following results. For LDA 1, five out of 11 procedures performed significantly better than HZ, two performed significantly poorer and four show no difference in performance. Of the five procedures that performed above the baseline, LOBANOV's procedure and Nearey's CLIH$_{i4}$ procedure show the highest percentages (92% and 90% of the vowel tokens was categorized correctly, respectively), followed by GERSTMAN (84%). SYRDAL & GOPAL (69%) and MILLER (76%) show a deterioration compared with the baseline. No differences in performance are observed between the four scale transformations and the baseline.

The results for QDA 1 show that four out of 11 procedures performed significantly better than HZ, two performed significantly poorer and five show no difference in performance. Furthermore, the percentages for QDA are only 1-2% higher than the results for LDA 1. Therefore, for the analyses described in the present chapter, LDA was used instead of QDA.

The next issue to be addressed is the precise nature of the classification errors. The proportion of misclassified vowel tokens per normalization procedure is displayed in Table 7.2, to

Table 7.2: *Percent errors per vowel category for each of the 12 procedures, based on LDA 1. Percent per vowel type is displayed in the bottom row.* S & G *refers to* SYRDAL & GOPAL *and* N & L *refers to* NORDSTRÖM & LINDBLOM.

%	/ɑ/	/a/	/ɛ/	/ɪ/	/i/	/ɔ/	/u/	/ʏ/	/y/	mean
HZ	26	14	31	40	14	13	12	19	15	21
LOG	23	10	29	32	16	15	14	23	17	20
BARK	26	11	29	34	14	15	13	23	14	20
ERB	24	10	28	34	15	14	13	23	14	20
MEL	25	11	29	36	14	15	13	22	14	20
S & G	26	16	31	40	39	16	18	36	55	31
LOBANOV	11	7	17	9	2	6	7	4	7	8
GERSTMAN	24	11	32	28	2	14	8	16	14	17
CLIH$_{i4}$	14	7	17	13	4	9	10	9	7	10
CLIH$_{s4}$	23	11	28	24	11	14	14	13	13	17
N & L	23	14	31	31	8	17	12	15	13	18
MILLER	27	14	13	41	17	16	17	35	20	22
Mean	23	11	26	30	13	14	13	20	17	20

establish whether the errors show a systematic pattern, or whether merely random variation is present. The percentages in Table 7.2 were obtained from confusion matrices that were calculated for LDA 1 in Table 7.1.

Table 7.2 shows that vowel tokens in the vowel categories /ɪ/ and /ɛ/ were misclassified most often (30% and 26%, respectively), followed by /ɑ/ (23%) and /ʏ/ (20%). The lowest percentages were found for the point vowel categories /a/ (11%), /i/ (13%), and /u/ (13%). This pattern can be observed for all 12 procedures, except for SYRDAL & GOPAL. Apparently, SYRDAL & GOPAL's low LDA 1 scores in Table 7.1 can be explained by the high error percentages for the vowels /y/ (55%), /ɪ/ (40%), /i/ (39%) or /ʏ/ (36%). Finally, LOBANOV (and, to a lesser extent CLIH$_{i4}$) shows a considerable lower error percentage, compared with the other procedures, especially for /i/ (2%), /ʏ/ (4%), /ɪ/ (9%), and /y/ (7%).

Comparison with results from earlier studies

Chapter 2 discusses a variety of studies that evaluate how well the normalization procedures represent phonemic variation in the transformed vowel tokens. Here, my results are compared with the results of two of these studies, Syrdal (1984) and Deterding (1990).

Syrdal (1984) evaluated eight normalization procedures, the baseline procedure (raw data in Hz), the log-transformation, the bark-transformation, Syrdal's bark-difference model

(1984), two versions of Miller's log-ratio model (1980), two versions of Nearey's (1978) model (CLIH_{s4} and CLIH_{i4}), and Gerstman's range normalization (1968). She applied these procedures to Peterson & Barney's (1952) data set and used the percentage of correctly classified vowel tokens from an LDA to evaluate the success of the procedures.

Overall, my results show a pattern similar to Syrdal's. Syrdal found that CLIH_{i4} is the best of the eight procedures that she compared with her baseline condition (Hz), while I found that CLIH_{i4} is the second best procedure, after LOBANOV (not evaluated by Syrdal). One major difference between Syrdal's results and the results presented in the present research lies in the fact that Syrdal found that the 'bark-difference' procedure (85.9%) – nearly identical to SYRDAL & GOPAL discussed here – performs better than her baseline condition (82.3%), while I found that SYRDAL & GOPAL clearly performed below the baseline. In LDA 1, SYRDAL & GOPAL scored 69.8% correct (nine response categories). LDA 1 can be seen as a test similar to Syrdal's test (her test with all speakers pooled) for which she reported a percentage of 85.9% (for 10 response categories) for her bark-difference procedure. A small part of the difference between my results and Syrdal's could be accounted for by differences in implementation: a different formula was used to calculate the bark-transformation. Syrdal used Zwicker & Terhardt's (1980) formula and I used Traunmüller's (1990) formula. A second explanation for this difference in performance can possibly be attributed to the speech material to which the procedures were applied. I used a database of speakers of Dutch and Syrdal's speakers spoke American English. It could be the case that Dutch is one of the languages that cannot be described as adequately by SYRDAL & GOPAL's second dimension as American English, as was suggested in Syrdal & Gopal (1986), described in detail in Chapter 2.

Deterding (1990) evaluated eight formant-based procedures. He evaluated a baseline procedure (raw data in Hz), two versions of Gerstman (1968), Lobanov (1971), Nearey's CLIH_{i2} (1978), four scale transformations (log, mel, bark, ERB). He applied the procedures to a small data set consisting of female speakers, male speakers, and children. He evaluated the procedures using his own pattern classification algorithm. Deterding found that all eight procedures perform better than the baseline. In addition, he found that the performances of the procedures LOBANOV and CLIH_{i2} are superior to all other procedures. Also, he did not find considerable differences between the four scale transformations (log, mel, bark, ERB). My results show a pattern similar to Deterding's.

In summary[49], my results are comparable with most of the results reported in the literature (Syrdal, 1984; Deterding 1990), except for the results found for SYRDAL & GOPAL, as reported in Syrdal (1984).

[49] Pols, Tromp, & Plomp (1973) used their vowel normalization procedure (identical to CLIH_{i4}) to obtain higher percentages correctly classified vowel tokens using Dutch vowels that were produced by 50 male speakers. They found that the percentages increased considerably (e.g. from 86.7% for the raw data in Hz to 97.2% for the normalized data (cf. page 1097, Pols et al., 1973).

7.2.3 Minimizing anatomical/physiological variation

To establish how well the 12 procedures minimize anatomical/physiological speaker-related variation in the acoustic data, four series of linear discriminant analyses were carried out: LDA 2-5. In these analyses, combinations of the four acoustic variables, F_0, F_1, F_2, and F_3, transformed through each of the 12 normalization procedures, were entered as predictors.

In LDA 2, the normalization procedures were evaluated on how well they produce output that can be classified as produced by a male or by a female speaker. In this LDA, speaker-sex was the dependent variable, having two levels; (transformed versions of) F_0, F_1, F_2, and F_3 were entered as predictors. For LDA 2, it was expected that F_0 would be the dominant predictor, especially for the baseline transformation (HZ). Large differences in values of F_0 between both sexes were expected.

To investigate whether differences between the procedures found for LDA 2 can be attributed to differences in F_0, or that they are due to differences in F_1, F_2, and F_3, two additional LDAs were carried out. In both analyses, speaker-sex served as the dependent variable. In LDA 3, F_0 was entered as the sole predictor and in LDA 4, F_1, F_2, and F_3 were entered as predictors.

In LDA 5, F_0, F_1, F_2, and F_3 were entered as predictors. Here, it was evaluated how well the procedures classify vowel tokens as being produced by a younger or older speaker. The dependent variable was the factor age, also having two levels. Any procedure that effectively eliminates anatomical/physiological variation related to the speaker's sex or age must not perform above chance level in the LDA. If a normalization procedure is performing at chance level in classifying tokens as male of female, it has eliminated all anatomical/physiological variation.

Table 7.3 presents the results for all four LDAs, i.e., the percentage of vowel tokens successfully classified as one of the two sexes (LDA 2, LDA 3, LDA 4) or as one of the two age-groups (LDA 5). For SYRDAL & GOPAL and MILLER, the analyses could not be carried out for LDA 3 and LDA 4, because these two procedures do not use F_0, or F_1, F_2, or F_3 in the same way as the other procedures (cf. Chapter 2). For instance, SYRDAL & GOPAL use $D_1^{\mathrm{B}} - D_0^{\mathrm{B}}$ as their first dimension (see equation (2.6)).

Table 7.3 shows that, for LDA 2, 93% of the vowel tokens were categorized correctly on speaker-sex for HZ. This can be interpreted as that most of the anatomical/physiological was preserved in the raw data. If all speaker-sex related variation would have been eliminated, the percentage of correctly classified vowel tokens would have been 50%. LOBANOV (50%) and CLIH$_{i4}$ (50%) performed best, they removed all variation related to the speaker's sex. GERSTMAN (53%) and SYRDAL & GOPAL (53%) removed nearly all sex-related variation, followed by MILLER (79%), CLIH$_{s4}$ (81%), NORDSTRÖM & LINDBLOM (83%). The scale transformations LOG, BARK, MEL, and ERB did not eliminate any anatomical/physiological variation related to the speaker's sex. Only three procedures perform at chance level for

Table 7.3: *Percent correctly classified vowel tokens for LDA 2-5. For all four LDAs, the chance level was 50%. For LDA 2, all percentages lower than 92% are significantly different from the baseline (*HZ*). For LDA 3, this is 87%, and for LDA 4, this is 78%. For all LDAs: all percentages are higher than 53% are significantly higher than chance level. All percentages were rounded off to the nearest whole number.*

%	LDA 2	LDA 3	LDA 4	LDA 5
Dependent variable	Speaker-sex	Speaker-sex	Speaker-sex	Speaker-age
Predictor variables	F_0, F_1, F_2, F_3	F_1, F_2, F_3	F_0	F_1, F_2, F_3
HZ	93	89	80	57
LOG	93	89	80	57
BARK	93	89	80	58
ERB	93	89	80	57
MEL	92	89	80	58
SYRDAL & GOPAL	53	-	-	51
LOBANOV	50	51	51	52
GERSTMAN	53	53	51	52
CLIH$_{i4}$	50	51	49	50
CLIH$_{s4}$	81	78	69	57
NORDSTRÖM & LINDBLOM	83	82	52	57
MILLER	79	-	-	51

LDA 2: LOBANOV, CLIH$_{i4}$, and GERSTMAN. All other procedures do not eliminate variation related to speaker-sex from the vowel data effectively enough.

The results for LDA 3 and LDA 4 show that F_0 contains considerable anatomical/physiological variation. This variation can be attributed to differences between male and female speakers, because 89% of the vowel tokens could be correctly classified in LDA 3 (in which F_0 was entered as the sole predictor). The variation in F_0 stems most likely from differences in the anatomy and physiology of the larynx of males and females. However, the three formant frequencies display anatomical/physiological sex-related variation as well, although less than F_0. The variation in the formant frequencies is caused by differences in vocal-tract-length between males and females. The results for LDA 4 show that NORDSTRÖM & LINDBLOM, a procedure designed to account for vocal-tract-length differences, succeeded in eliminating these differences. The results for NORDSTRÖM & LINDBLOM in LDA 2, LDA 3, and LDA 4 indicate that this procedure dealt effectively with (vocal-tract-related)

anatomical/physiological variation in the formants, but that it did not succeed in eliminating the (larynx-related) anatomical/physiological variation in the fundamental frequency.

The results for LDA 5 in Table 7.3 show that less anatomical/physiological age-related than sex-related variation was present; the percentages of correctly classified vowel tokens according to speaker-age are considerably lower than the percentages for speaker-sex. Five procedures perform at chance level for LDA 5: SYRDAL & GOPAL, LOBANOV, GERSTMAN, CLIH$_{i4}$, and MILLER. All other procedures perform (slightly) above chance level and did not eliminate all age-related anatomical/physiological variation from the vowel tokens[50].

To sum up, the results in Table 7.3 show that the acoustic consequences of the anatomical/physiological related to speaker-age were overall considerably smaller than the acoustic consequences of the speaker-sex. For age, the percentages across all procedures are overall just above chance level. Most of the variation in the acoustic signal seems to be related to the anatomical/physiological differences in the vocal tract and larynx of female and male speakers, whereas the differences related to speaker-age could not be attributed univocally to specific anatomical or physiological differences between younger and older speakers. Overall, LOBANOV, CLIH$_{i4}$, and GERSTMAN eliminated anatomical/physiological variation from the acoustic measurements best of all 12 procedures.

7.2.4 Preserving sociolinguistic variation

The 160 speakers were stratified for the three sociolinguistic variables speaker-sex (male or female), speaker-age (young or old), and regional background (regions 1-8). LDA 6 was carried out to establish to what extent regional variation was preserved in the transformed acoustic representations of the vowel data. The acoustic variables F_0, F_1, F_2, and F_3, transformed through each of the 12 normalization procedures, were entered as predictors. Region served as the dependent variable. The analysis was carried out for each of the nine vowel categories separately, to eliminate the effect of the vowel token's category and to evaluate whether some vowels displayed more regional variation than others. If all percentages for a certain procedure are at or near chance level (12.5%), it must be concluded that the procedure eliminates all systematic sociolinguistic variation related to the speaker's regional background. The results are displayed in Table 7.4.

Table 7.4 shows that the percentages correctly classified vowel tokens are above chance level across all procedures for all vowels. This indicates, first, that regional variation was present in the data of the 160 speakers, and second, that none of the investigated procedures eliminated all regional variation from the data. Third, vowel tokens intended to belong to the vowel categories /ɛ/, /ɪ/, and /ʏ/ show most regional variation. The mean percentages for

[50]The analyses for LDA 2 and LDA 5 were repeated per normalization procedure and per vowel category, to ensure that the effects described in Table 7.3 were not due to one specific vowel, or due to a subset of the vowels. Overall, percentages of the same magnitude were found as found for LDA 2 and LDA 5 across all nine vowels.

Table 7.4: *Results for LDA 6: percent vowel tokens that were classified into the corresponding region, for each vowel category for each normalization procedure. The number of tokens per vowel category is 320. Percentages higher than 18% (rounded off) are significantly different from chance level (12.5%).* SYRDAL & GOPAL *is referred to as* S & G, *and* NORDSTRÖM & LINDBLOM *as* N & L.

%	/ɑ/	/a/	/ɛ/	/ɪ/	/i̇/	/ɔ/	/u/	/ʏ/	/y/	mean
HZ	27	23	36	35	29	29	33	38	26	31
LOG	26	20	37	33	26	31	33	36	26	30
BARK	27	22	35	34	26	29	33	37	27	30
ERB	26	22	35	34	26	30	33	37	27	30
MEL	27	22	35	33	26	29	33	37	25	30
S & G	22	19	32	30	20	25	25	28	22	25
LOBANOV	26	18	35	31	28	27	32	25	31	28
GERSTMAN	25	22	36	34	19	26	25	31	26	27
CLIH$_{i4}$	23	19	34	31	29	29	33	31	28	28
CLIH$_{s4}$	28	20	37	35	31	31	30	32	25	30
N & L	27	21	37	33	29	30	33	34	27	30
MILLER	23	17	35	31	31	25	29	32	23	27
Mean	26	20	35	33	26	28	31	33	26	29

these vowels are higher than for the other six vowels (35%, 33% and 33%, respectively). Of the point vowels, /a/ and /i/ show little regional variation (20% and 26%, respectively), /u/, on the other hand, shows more variation (31%).

In Table 7.4 some differences between procedures can be observed. SYRDAL & GOPAL reduced more sociolinguistic variation than the other procedures, followed by GERSTMAN and MILLER, LOBANOV, and CLIH$_{i4}$. Given the results for GERSTMAN, MILLER, LOBANOV, and CLIH$_{i4}$, it can be concluded that some of the procedures that were performed best at reducing anatomical/physiological variation show a small reduction of the sociolinguistic variation, if that reduction is substantial at all.

7.2.5 Discussion

In sections 7.2.2, 7.2.3, and 7.2.4, the procedures were evaluated on how well they meet the criterion described in section 7.1 (as well as in Chapter 1). A procedure is considered to meet this criterion when the application of the procedure resulted in better representation of the phonemic variation compared with the baseline. This improved performance should be accompanied by a reduction of the anatomical/physiological variation, and by the preservation

of the sociolinguistic variation. If the procedure performs poorly at one (or more) of these tasks, then the criterion is not met. In the present section, the classification scheme described in Chapter 2 was used to discuss the performance of the normalization procedures.

Table 7.5: *Rank scores for the acoustic comparisons for each (class of) normalization procedures. The results from LDA 1 were used for the column "Preserve phonemic", results from LDA 2 and 5 were used for the column 'Reduce anatomical/physiological', and the mean results from LDA 6 were used for the column "Preserve sociolinguistic".*

Procedure	Preserve phonemic	Reduce anatomical/physiological	Preserve sociolinguistic
HZ	10	10.5	1
LOG	7.5	10.5	4.5
BARK	7.5	10.5	4.5
ERB	7.5	10.5	4.5
MEL	7.5	8	4.5
S & G	12	3	12
LOBANOV	1	1.5	8.5
GERSTMAN	3	4	10.5
CLIH$_{i4}$	2	1.5	8.5
CLIH$_{s4}$	4.5	6	4.5
N & L	4.5	7	4.5
MILLER	11	5	10.5

Table 7.5 summarizes the results for all normalization procedures at the tasks of preserving phonemic variation, reducing anatomical/physiological variation, and preserving sociolinguistic variation. The results for the vowel-intrinsic/formant-intrinsic procedures, HZ (baseline) and the four scale transformations LOG, BARK, ERB, and MEL, are as follows. The results for the scale transformations show that applying the four scale transformations to the measurements did not result in an improvement (nor deterioration) in performance compared with the baseline. The performance at the first two tasks (preserve phonemic and reduce anatomical/physiological variation) is identical across all five procedures. For the third task, preserving sociolinguistic variation, it was found that the four scale transformations performed slightly poorer than the baseline (rank score 7.5 vs. 1, respectively, in Table 7.5).

SYRDAL & GOPAL, the vowel-intrinsic/formant-extrinsic procedure, performed poorer than the baseline at representing phonemic variation (rank score 12 in Table 7.5). Nevertheless, SYRDAL & GOPAL performed considerably better than the baseline at minimizing anatomical/physiological variation (rank score 3). Finally, this procedure was found to eliminate more sociolinguistic variation from the data than all other procedures (rank score 12).
Overall, the class of vowel-extrinsic/formant-intrinsic procedures, LOBANOV, CLIH$_{i4}$, and

GERSTMAN, performed better than the other three classes. The three vowel-extrinsic/formant-intrinsic procedures performed best at preserving phonemic variation (rank scores 1, 2, and 3 for LOBANOV, CLIH$_{i4}$, and GERSTMAN, respectively), and minimized the anatomical/physiological variation best (rank scores 1.5, 1.5, and 4, respectively). However, these procedures were found to minimize some of the sociolinguistic variation from the data (rank scores 8.5, 8.5, and 10.5 respectively). LOBANOV and CLIH$_{i4}$ thus performed best of all procedures in two of the three tasks.

The performance of the class of vowel-extrinsic/formant-extrinsic procedures, CLIH$_{s4}$, NORDSTRÖM & LINDBLOM, and MILLER was poorer than the vowel-extrinsic/formant-intrinsic procedures. Overall, of the three procedures, NORDSTRÖM & LINDBLOM ('N & L') and CLIH$_{s4}$ performed better than MILLER at preserving phonemic variation (rank score 3.5 vs. 11, respectively). MILLER performed better than NORDSTRÖM & LINDBLOM and CLIH$_{s4}$ at minimizing anatomical/physiological variation (rank score 5 vs. 7 and 6, respectively). NORDSTRÖM & LINDBLOM ('N & L') and CLIH$_{s4}$ performed better than MILLER at preserving sociolinguistic variation (rank score 4.5 vs. 10.5, respectively).

The results for all four classes of procedures show that the vowel-extrinsic/formant-intrinsic procedures performed best at transforming the acoustic measurements in such a way that the phonemic variation was represented better than in the baseline procedure. These procedures further performed best at the task of eliminating the anatomical/physiological variation from the data. Finally, these procedures performed only slightly poorer than most other procedures at the task of preserving sociolinguistic variation; it is by no means the case that all variation was normalized away. In sum, it seems justified to conclude that the vowel-extrinsic/formant-intrinsic procedures met the criterion best of all procedures in the acoustic comparison of the procedures described in this section, because they performed best at two of the three tasks that were evaluated.

7.3 Performance of the three best procedures.

7.3.1 Multivariate analysis

This section aims to get more insight into how the procedures deal with the acoustic consequences of specific (sociolinguistic) variation sources in the acoustic measurements. In section 7.2, linear discriminant analysis was used to evaluate the performance of the procedures, whereas in the present section multivariate analysis of variance (manova) is used. Manova provides a different perspective on the performance of the procedures. Linear discriminant analysis can be used to establish which combination of predictors allows the vowel tokens to be categorized best (e.g., into the corresponding vowel category, region or speaker-sex) and manova can be used to establish the relative proportion of variance in the data corresponding to specific variation sources. For instance, LDA 2 (section 7.2) used the four acoustic

variables as predictors, and speaker-sex as the dependent variable. When manova is used, the predictor and the dependent variables are switched; this way, it can be assessed which proportion of the variance can be attributed to anatomical/physiological differences between speakers and to regional differences between speakers. Furthermore, it can be established how much this variance decreases when the acoustic variables are transformed using the normalization procedures. Finally, it can be established which interactions exist between predictor variables.

In this section, I describe various manovas in which the speaker-sex, age, regional background and the vowel token's category were used to predict the variation in the four (transformed) acoustic variables. The four acoustic variables were the dependent variables. To limit the present evaluation of the normalization procedures, only the baseline and the three vowel-extrinsic/formant-intrinsic procedures LOBANOV, CLIH$_{i4}$, and GERSTMAN were investigated. Section 7.2 shows that the vowel-extrinsic/formant-intrinsic procedures performed best at two of the three tasks described in the previous section. The baseline was included to allow for comparisons.

Variance reduction

Four manovas were carried out to get a more detailed picture of the differences between the baseline data and the three vowel-extrinsic/formant-intrinsic procedures. One manova was run for HZ, one for LOBANOV, one for GERSTMAN, and one for CLIH$_{i4}$. For each analysis, the values of (subsets of) the raw or transformed fundamental frequency and the first three formant frequencies served as dependent variables. Vowel, Region, Speaker-sex and Speaker-age were entered as fixed factors. These manovas were repeated three times, once with F_0, F_1, F_2, and F_3 as dependent variables, once with F_1, F_2, and F_3 as dependent variables, and once with F_1 and F_2 as dependent variables. This was done to get more insight into the role of F_0 and F_3 in the reduction of variance, and to establish whether a combination consisting of only F_1 and F_2 would explain the variation equally well.

The multivariate measure of effect size for each set of factors and interaction terms was η^2. η^2 reveals the proportion of the total variation in the dependent variable that is accounted for by the variation in the independent variable. The significance level of this measure was estimated using Pillai's trace[51]. Table 7.6 displays the significant results ($p < 0.001$). Here, first the differences between the four procedures are discussed and second the differences between the three different sets of dependent variables are discussed.

Table 7.6 shows that the values of the dependent variables varied primarily depending on the factor Vowel: the values of η^2 are highest for the factor Vowel across all procedures. For HZ,

[51] One of the tests available in multivariate analysis of variance, used for reflecting the proportion of the variance in the dependent variable that can be accounted for, given variation in the independent variable(s). See Rietveld & Van Hout (1993) for more information on Pillai's trace.

the largest variation in the dependent variables could be accounted for by the factor Speaker-sex (for F_0, F_1, F_1, F_3, Speaker-sex shows a larger effect than Vowel). In contrast, there is no effect of Speaker-sex for LOBANOV and for CLIH$_{i4}$, and only a minor effect of Speaker-sex for GERSTMAN. This corroborates the earlier finding that these three procedures effectively removed anatomical/physiological variation from the acoustic measurements. Furthermore, for all four procedures a relatively large effect was found for the interaction between vowel and region. This indicates that other vowels showed region-dependent variation. Finally, the factor Speaker-age showed small effects for all four procedures (for HZ and GERSTMAN), or no effects (LOBANOV and CLIH$_{i4}$). Small effects were also found for the majority of the other interaction terms.

The results for the three sets of dependent variables show the following pattern for the analyses that first excluded F_0, and F_3 in a second step. For HZ, excluding F_0 had the following four consequences. First, for F_1, F_2, and F_3, the value of η^2 for the factor Vowel increases compared with the results for F_0, F_1, F_2, and F_3. The factor Vowel explains the variation in the formant measurements better than the variation in F_0. Second, the value of η^2 for Speaker-sex decreases. F_0 was therefore affected more by the factor Speaker-sex than the other acoustic variables. Third, the effect of Speaker-age disappeared altogether, meaning that this factor only affected F_0. Fourth, the interaction term Vowel$\times$Region increases, as well as the interaction term Vowel$\times$Speaker-sex. This indicates that some vowels show more systematic acoustic variation across different regions after exclusion of F_0. Excluding F_0 resulted in a clearer perspective on variation patterns in the data.

For the analysis with F_1 and F_2 as dependent variables (in which F_3 and F_0 were excluded) the following pattern was observed in the results. Compared with results for F_1, F_2, and F_3, the value for η^2 increased further for the factor Vowel as well as for Vowel$\times$Region. For the three procedures LOBANOV, CLIH$_{i4}$, and GERSTMAN, a pattern identical to the one for HZ is found for Vowel and for Vowel$\times$Region. The values of η^2 in Table 7.6 further show that the proportion of variation in the measurements is attributable to different sources of variation. The anatomical/physiological variation is most prominent in the raw data (Speaker-sex and Speaker-age), followed by the phonemic variation (Vowel), and the sociolinguistic variation (Region). When the data is normalized following the successful normalization procedures, a different pattern emerges; then the phonemic variation is most prominent, followed by the sociolinguistic variation, while the anatomical/physiological variation does not play a role of any importance[52].

[52] Although Pols, Tromp & Plomp (1973) describe a similar division of the variation sources in their data set (cf. Chapter 1 of this research), I did not compare my results with theirs. No systematic sociolinguistic variation could have been present their data; they did not include information about the speaker's regional background.

Table 7.6: *Results for the four multivariate analyses of variance: η^2 for each significant factor, for each of the four procedures ($p<0.001$). For each procedure, the analysis is repeated for three different sets of dependent variables. The number of tokens per analysis is 2880.*

η^2	HZ			LOBANOV			GERSTMAN			CLIH$_{i4}$		
	F_0, F_1, F_1, F_3	F_1, F_2, F_3	F_1, F_2	D_0, D_1, D_2, D_3	D_1, D_2, D_3	D_1, D_2	D_0, D_1, D_2, D_3	D_1, D_2, D_3	D_1, D_2	D_0, D_1, D_2, D_3	D_1, D_2, D_3	D_1, D_2
Vowel	0.527	0.695	0.893	0.579	0.760	0.932	0.556	0.731	0.914	0.568	0.743	0.917
Region	0.075	0.080	0.063	-	-	-	0.041	0.051	0.067	-	-	-
Speaker-sex	0.770	0.656	0.537	-	-	-	0.018	0.014	0.014	-	-	-
Speaker-age	0.075	-	-	-	-	-	-	-	-	-	-	-
Vowel×Region	0.120	0.151	0.183	0.150	0.190	0.236	0.126	0.159	0.200	0.139	0.173	0.207
Vowel×Speaker-sex	0.064	0.079	0.108	0.014	0.017	0.019	0.011	0.014	0.016	0.019	0.024	0.025
Region×Speaker-sex	0.017	0.010	0.011	-	-	-	0.016	0.016	0.019	-	-	-
Vowel×Region×Speaker-sex	0.031	0.036	0.036	0.030	0.032	-	-	-	-	0.039	0.043	0.039
Vowel×Speaker-age	0.005	-	-	0.008	0.008	0.010	-	-	-	0.008	0.008	0.010
Region×Speaker-age	0.029	0.027	0.029	-	-	-	-	-	-	-	-	-
Vowel×Region×Speaker-age	-	-	-	0.030	0.033	0.033	-	-	-	0.029	0.033	0.032
Speaker-sex×Speaker-age	0.009	0.009	-	-	-	-	0.016	0.015	0.014	-	-	-
Vowel×Speaker-sex×Speaker-age	-	-	-	-	-	-	-	-	-	0.007	0.008	-
Region×Speaker-sex×Speaker-age	0.020	0.019	0.017	-	-	-	0.013	0.014	0.018	-	-	-
Vowel×Region×Speaker-sex×Age	-	0.030	-	0.032	0.038	0.033	0.030	0.035	0.032	0.039	0.043	0.037

7.3.2 Analyses with two formants

Given the results for the three sets of dependent variables, it seems justified to exclude D_0 and D_3 from further analysis in this section and to use only D_1 and D_2. Furthermore, as can be observed in Table 7.7, when LDA 1 (displayed in Table 7.1) is repeated with only D_1 and D_2 as predictors for HZ, LOBANOV, CLIH$_{i4}$, and GERSTMAN. The decrease in performance is highest for HZ (7%) and lowest for LOBANOV and GERSTMAN (both 1%). All percentages are only 1-7% lower than when all four acoustic variables were added.

Table 7.7: *Results for the LDA with F_1, F_2 as predictors, and from the LDA with F_0, F_1, F_2, and F_3 as predictors (i.e., LDA 1). All percentages higher than 75% or lower than 69% (all percentages rounded off to the nearest integer) are significantly different from the baseline condition (*HZ*).*

%	F_1, F_2	F_0, F_1, F_2, F_3 (LDA 1)
HZ	72	79
LOBANOV	91	92
GERSTMAN	83	84
CLIH$_{i4}$	87	90

The next step in the analysis of the four procedures was to focus on the (sociolinguistic) variation within the vowel categories. To this end, the same four manovas were carried out as displayed in Table 7.6, the only difference was that this time the analyses were repeated per vowel. The dependent variables were F_1 or D_1 and F_2 or D_2, and the independent variables were Region, Speaker-age and Speaker-sex. In Table 7.8 the significant results ($p<0.001$) are displayed. Given the results in Table 7.6, it was expected that different vowel categories would show different values of η^2. This seemed plausible, because effects across all four procedures were found for the interaction term Vowel$\times$Region, and no, or very small, effects for Region. This indicates that my data displays no differences between the whole set of vowels, but rather that regional differences exist within individual vowels.

Table 7.8 shows that the values for η^2 for the factor Speaker-sex are much lower, or nonexistent, for the three normalization procedures. The same pattern can be observed in the results for factor Speaker-age.

When comparing the results in Table 7.8 with those in Table 7.6, it can be observed that the values for η^2 for the factor Region are considerably higher for all individual vowels than when the data was not split up per vowel category. Little or no differences in the shape and or size of the entire vowel system seem to be present between the eight regional varieties, instead, the differences could be found for individual vowels. This could be expected given the significance of the interaction term Vowel$\times$Region in Table 7.6. For the vowels /ɛ/, /ɪ/,

Table 7.8: *Results for the four multivariate analyses of variance: η^2 for each significant factor, per procedure and vowel category ($p<0.001$). The dependent variables are F_1/D_1 and F_2/D_2. The number of tokens per vowel category is 320.*

η^2	Factor	HZ	LOBA-NOV	GERST-MAN	CLIH$_{i4}$
/ɑ/	Region	0.155	0.149	0.205	0.139
	Speaker-sex	0.446	-	-	0.081
	Region×Speaker-age	0.083	-	-	
	Region×Speaker-sex ×age	0.069	-	0.071	0.060
/a/	Region	0.062	-	0.061	-
	Speaker-sex	0.564	-	-	-
	Speaker-sex×Speaker-age			0.062	-
/ɛ/	Region	0.342	0.366	0.399	0.368
	Speaker-sex	0.622	-	-	
	Region×Speaker-sex	0.064	-	-	-
	Speaker-sex×Speaker-age	-	-	0.048	-
/ɪ/	Region	0.282	0.295	0.333	0.274
	Speaker-sex	0.716	0.059	-	0.059
	Speaker-age	0.052	-	-	-
	Region×Speaker-sex ×Speaker-age	-	-	-	0.069
/i/	Region	0.076	0.137	-	0.097
	Speaker-sex	0.645	0.053	0.146	0.157
	Speaker-age	0.058	-	-	-
/ɔ/	Region	0.170	0.150	0.144	0.178
	Speaker-sex	0.318	0.066	0.063	
	Region×Speaker-sex	-	-	0.063	-
/u/	Region	0.205	0.220	0.133	0.221
	Speaker-sex	0.330	-	-	-
	Region×Speaker-sex	-	-	0.065	-
/ʏ/	Region	0.210	0.171	0.247	0.204
	Speaker-sex	0.731	0.105	-	0.099
	Region×Speaker-sex	-	-	0.061	
/y/	Region	0.194	0.236	0.109	0.169
	Speaker-sex	0.564	0.101	0.095	-
	Region×Speaker-sex	-	-	0.063	-

and /ʏ/, the values of η^2 for the factor region are highest for all for the three normalization procedures (but not for HZ).

Given the results in Tables 7.6 and 7.8, the effects for the four procedures can be summarized as follows. The measurements for the baseline procedure, HZ, displayed considerable variation related to the speaker-sex and age. Furthermore, for the three procedures, it seems that the transformed data displayed no (LOBANOV and CLIH_{i4}) or very little (GERSTMAN) anatomical/physiological sex- or age-dependent variation. Instead, the transformed data showed greater clustering according to the vowel token's category. For these three procedures, the data was found to vary depending of the region for a subset of the vowels: the values of η^2 were overall highest for the vowel categories /ɛ/ and /i/, the vowel categories that show the highest mean scores in Table 7.8.

To sum up, the results of the comparisons carried out throughout this section indicate that the sociolinguistic variation in the sociolinguistically database was predominantly related to differences in the regional background of the 160 speakers, while the speaker's sex and age played a minor role after normalization.

7.3.3 Specific differences between regional varieties

Randstad Dutch versus Valkenburg Dutch

This section evaluates how well sociolinguistic differences were preserved in the acoustic measurements after normalization through LOBANOV, CLIH_{i4}, and GERSTMAN. The sociolinguistic differences are those reported by Adank, Van Heuven & Van Hout (1999), a pilot study to the present research. Adank et al. report sociolinguistic differences for a subset of the Dutch vowels, i.e., for /ɛ/, /ɪ/, and /ʏ/. They aimed to uncover sociolinguistic differences in the vowels of different regional varieties of standard Dutch (SD). To this end, they used data sets from two groups of female speakers (speakers from the Randstad and from Valkenburg); a different set of data than the set described in Chapter 5; for further details about this data set, see Van Rie, Van Bezooijen & Vieregge (1995)[53]. These speakers were thought to differ minimally in their anatomical/physiological characteristics and maximally in their socioeconomic characteristics. Because these latter differences were maximized, it was expected that their acoustic consequences would be considerable.

The speech material used by Adank et al. was set up as follows. The speakers were divided into two groups of 15 female speakers,. Of the se 15 five were between 20 and 30 years of age, five were between 30 and 40 years, and five were between 40 and 50 years. They had comparable body size and body height. The Randstad speakers were highly educated and the speakers from Valkenburg had a lower than average education level. The speech material was obtained through a sociolinguistic interview. One of the tasks in this interview was to

[53] I wish to thank Renée van Bezooijen for supplying this speech material.

read the 15 vowels of Dutch in isolation.

Two phonetically-trained listeners reported that the vowels /ɛ/, /ɪ/, and (to a lesser extent) /ʏ/ were pronounced with minimal (for /ʏ/) to considerable (for /ɛ/ and /ɪ/) lower height (more open), while /ɛ/ was pronounced with less advancement (more back), when pronounced by the speakers from Valkenburg than when pronounced by the Randstad speakers.

The sociolinguistic differences reported by Adank et al. should also be observable in the data set of the 160 speakers that is used in the present research. However, the differences of the 160 speakers were expected to be smaller than those reported on the 30 speakers used in Adank et al.'s study. Nevertheless, the same differences should be found comparing N-R (speech from towns in the Randstad in the present research, cf. Chapter 5) with N-S (speakers from towns in Limburg) speaker groups. To make an overview of the differences found for the N-R and N-S regions visible, the mean values of D_1 and D_2 across all 20 speakers for the region N-R and the mean values across the 20 speaker for region N-S for all four procedures are displayed according to their values for the first two formants in Figure 7.1. Figures displaying the mean values for all eight regions for all four procedures are presented in Appendix A.

In Figure 7.1, differences between the mean values for N-R and N-S region can be observed for nearly every vowel category, for all four conditions. In addition, the differences in height for the vowels /ɛ/ and /ɪ/ are relatively large across all four conditions, while for /ʏ/ and /u/, an advancement difference was found. In order to establish whether these differences between the two regions are significant, a manova was carried out. In this manova, F_1 and F_2 (or D_1 and D_2) served as the dependent variables, while the factor region (N-R or N-S) was entered as a predictor variable. The manova was repeated for each vowel category (to see if other differences in Figure 7.1 were significant too) and for each of the four normalization procedures.

Table 7.9 shows the results for the manovas for the differences between the two regional varieties. For HZ, effects were found for three vowels: /ɛ/, /ɪ/, and /ʏ/, plus an extra effect for /u/. If it is assumed that perceived height correlates primarily with F_1 and that perceived advancement correlates primarily with F_2 (as was reported by Assmann, 1979), then the effects are, generally, in the direction predicted by Adank et al. For /ɛ/ and /ɪ/, the effects are largest for F_1 (0.429 and 0.252, respectively), and smaller for F_2 (0.190 and 0.154, respectively). For /ʏ/ no effect was found for F_1, but an effect was found for F_2 (0.167). Finally, the results for HZ show an effect of F_2 (0.329) for /u/, indicating that an Advancement difference exists between the N-R and the N-S data.

For LOBANOV the same pattern in the results can generally be observed as was found for HZ. However, there are some differences. First, all values of η^2 are higher for LOBANOV, except for D_1 for /ɪ/. Second, three significant effects were found for LOBANOV that were not found for HZ: effects were found for D_1 and D_2 for /i/ (0.157 and 0.296, respectively), and D_2 for /ɑ/ (0.187). GERSTMAN's results show the same pattern as HZ. One difference

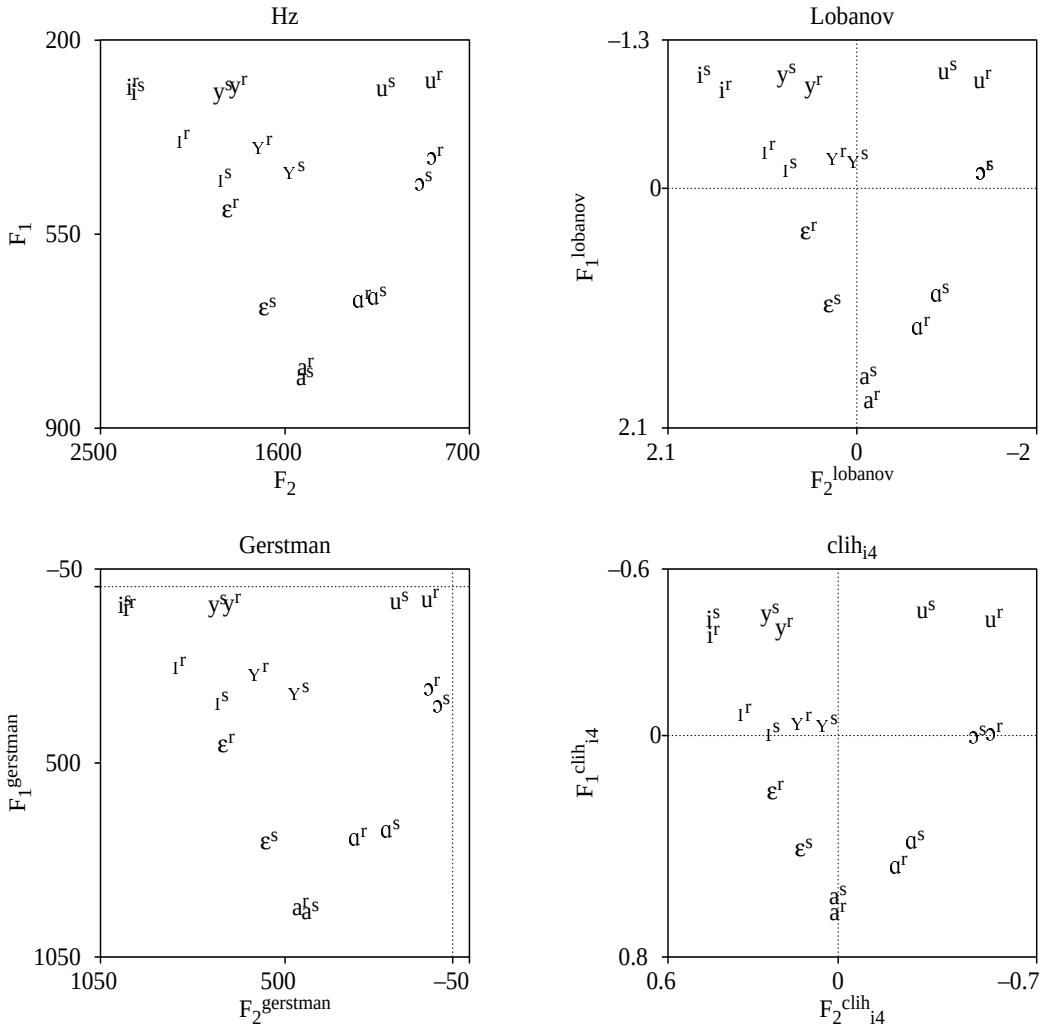

Figure 7.1: *Mean values per vowel across the 20 speakers in the N-R region,* ^{r}*, and the N-S region,* ^{s}*, for* HZ, LOBANOV, CLIH$_{i4}$, *and* GERSTMAN. *Each mean frequency represents the measurements of 40 tokens. Some mean values are not (clearly) visible because they overlap with other mean values.*

Table 7.9: *Significant results ($p<0.001$) for the multivariate analyses of variance for* HZ, LOBANOV, CLIH$_{i4}$, *and* GERSTMAN*: η^2 per vowel for the differences between the regions N-R and N-S. N per vowel category is 40.*

η^2	Hz		LOBANOV		CLIH$_{i4}$		GERSTMAN	
Vowel	F_1	F_2	D_1	D_2	D_1	D_2	D_1	D_2
/ɑ/	-	-	-	0.187	-	-	-	0.303
/a/	-	-	-	-	-	-	-	-
/ɛ/	0.429	0.190	0.544	0.360	0.565	0.444	0.497	0.479
/ɪ/	0.252	0.154	0.149	0.336	0.168	0.472	0.210	0.465
/i̇/	-	-	0.157	0.296	-	-	-	-
/ɔ/	-	-	-	-	-	-	-	-
/ʉ/	-	0.329	-	0.336	-	0.443	-	0.171
/ʏ/	-	0.167	-	0.473	-	0.547	-	0.555
/y/	-	-	-	-	-	-	-	-

with HZ is that an effect was found for D_2 for /ɑ/ (0.303), an effect also found in LOBANOV's results (0.187). The results for CLIH$_{i4}$ show the same pattern as HZ, the only difference is that all values of η^2 are higher for CLIH$_{i4}$. This indicates that more of the variation in the values of D_1 and D_2 can be attributed to the two different regions.

In conclusion, the results show that similar effects were found for the three vowels reported on by Adank et al. It must be concluded that the three procedures transformed the data in such a way that sociolinguistic differences in the data were preserved. However, two out of three procedures also found effects that are not predicted by Adank et al.: LOBANOV shows three such effects and GERSTMAN shows one effect. It is not clear whether these results would also be reported by phonetically-trained listeners when listening to these data.

Regional variation in /ɛ/

To get a better view of how the four procedures preserved sociolinguistic variation, the mean values for all eight regions for the vowel /ɛ/ are shown in Figure 7.2. I selected the vowel /ɛ/, because it appears from Figure 7.1 that the difference between the two regions (N-R and N-S) are largest for this vowel. It can furthermore be observed in Figure 7.2 that the general structure of the variation between the eight region is as follows. The three regions N-N, N-M, N-R are clustered together, F-W and F-L are clustered together, F-B is positioned between these two clusters. N-S is the most deviant of all. F-E is positioned between the cluster F-W and F-L and N-S. Large differences between regions appear to have been preserved in the data normalized following the three procedures, some smaller differences between regions

Table 7.10: *Results for the analyses of variance for the baseline data (*HZ*) for the 28 combinations of the eight regional varieties for /ɛ/. '+': significant (*$p < 0.001$*).*

HZ	Variable	N-R	N-M	N-S	N-N	F-B	F-E	F-L	F-W
N-R	F_1	0	-	+	-	-	+	+	+
	F_2	0	-	+	-	-	+	+	+
N-M	F_1		0	+	-	-	+	+	+
	F_2		0	+	-	-	+	+	+
N-S	F_1			0	+	+	-	+	+
	F_2			0	+	-	-	-	-
N-N	F_1				0	-	+	-	-
	F_2				0	-	+	+	+
F-B	F_1					0	+	-	-
	F_2					0	-	-	-
F-E	F_1						0	-	-
	F_2						0	-	-
F-L	F_1							0	-
	F_2							0	-
F-W	F_1								0
	F_2								0

were inverted (horizontally as well as vertically). For instance, in the data following the three normalization procedures the mean values for F-L and F-W are inverted compared with the mean data in HZ.

It would be interesting to establish whether the procedures eliminated certain differences that were present in the HZ data, or whether the procedures created differences that are not present in the HZ data. To this end, a series of pairwise comparisons of the eight regional varieties was carried out. For each pair of combinations (28 in total) an analysis of variance was carried out on the values of F_1 and F_2 for /ɛ/.

Table 7.10 shows the results for the 56 anovas (28 combinations for F_1 and for F_2), carried out on pairs of language varieties for HZ. The results in this table indicate that significant differences were found for mean values that differ considerably in Figure 7.2 (for HZ), such as the mean values for N-N and N-S, or for N-M and F-L. Mean frequencies that are closer together, such as N-M and N-R, show no effects.

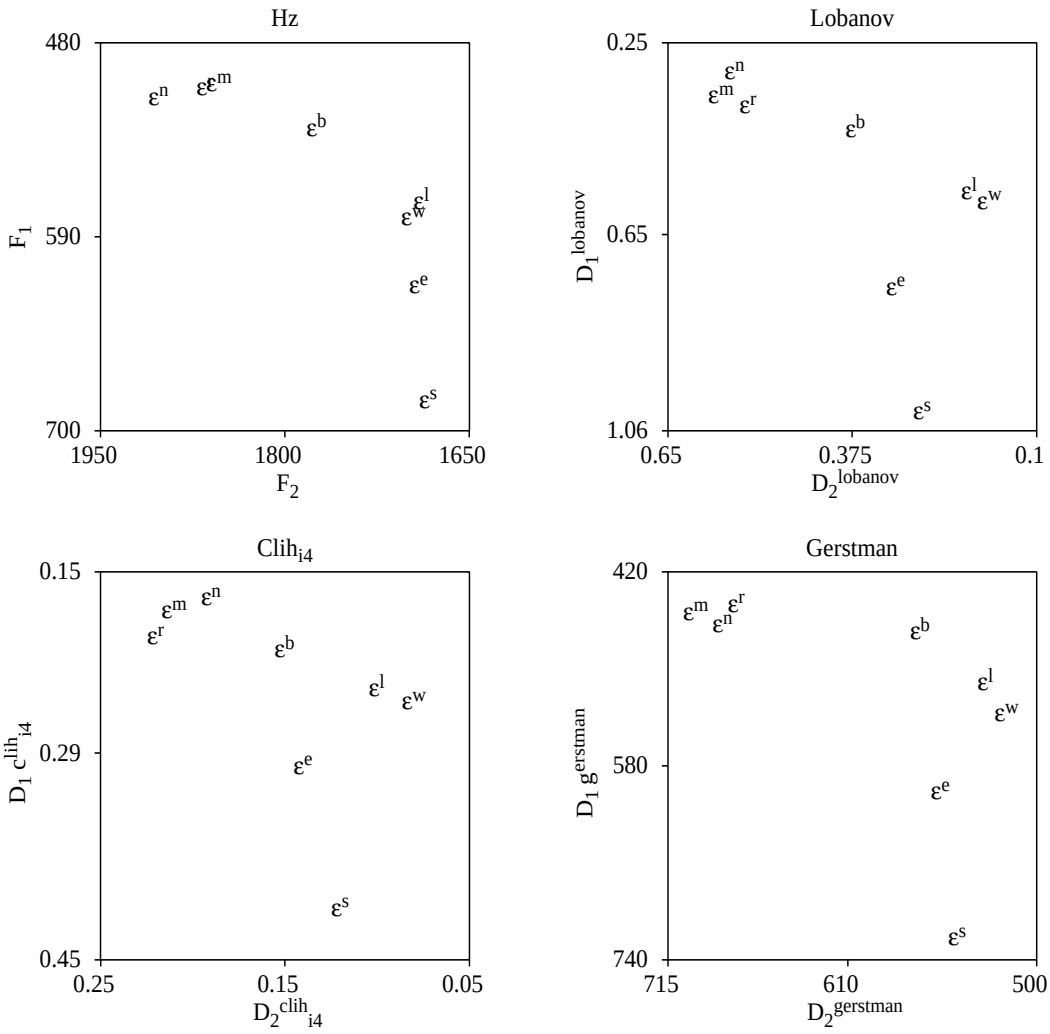

Figure 7.2: *Mean values for each of the eight regions,* ^{r}*: N-R,* ^{m}*: N-M,* ^{s}*: N-S,* ^{n}*: N-N,* ^{b}*: F-B,* ^{e}*: F-E,* ^{l}*: F-L, and* ^{w}*: F-W, for /ɛ/, for the four procedures for* HZ, LOBANOV, CLIH$_{i4}$, *and* GERSTMAN. *Each mean frequency represents the measurements of 40 tokens.*

Table 7.11: *Number of 'Extra' effects (effects not found for* HZ*), and 'Missed' effects (effects found for* HZ*, but not for these procedures) for the analyses of variance carried out on pairs of regional varieties on F_1 and F_2 for /ɛ/ ($p < 0.001$) for the three vowel-extrinsic/formant-intrinsic procedures.*

Effects	LOBANOV	GERSTMAN	CLIH$_{i4}$	Total
Extra F_1	2	1	2	5
Missed F_1	4	3	4	11
Extra F_2	5	1	3	9
Missed F_2	0	3	0	3
Total	11	8	9	28

The analyses reported on in Table 7.10 were repeated for the data transformed following LOBANOV, CLIH$_{i4}$, and GERSTMAN. For each of these three procedures, each significant and non-significant result was compared with the results found for HZ. For each procedure, the number of additional, or 'extra', differences was calculated, i.e., the number of significant differences that was present in the transformed data but not in the raw data. In addition, the number of 'missed' differences was calculated, these are the differences between the mean frequency values that were significantly different in the raw data, but not in the transformed data.

Table 7.11 shows the number of extra and missing significant effects compared with HZ. It can be observed that the total number of differences for LOBANOV is highest (11), followed by that for CLIH$_{i4}$ (9), and GERSTMAN (8). It can further be seen that the number of differences is higher for F_1 (total 16) than for F_2 (total 12). Finally, the number of extra differences is equal to the number of missing differences (both 14).

Table 7.12 shows differences between the three normalization procedures between pairs of the eight regional varieties. These pairs are associated to the differences in Table 7.11. Table 7.12 shows that a considerable number of differences was found for all three procedures (e.g. N-S$\times$F-E). This was taken to indicate that the discrepancies in the significant results between HZ and the three vowel-extrinsic/formant-intrinsic procedures were not random. At present, it is not possible to establish how large a difference between two mean values must be to constitute a difference that would also be reported by phonetically-trained listeners.

Estimating scale factors using three vowels

In section 7.2, it was concluded that the vowel-extrinsic/formant-intrinsic procedures performed best in the acoustic domain. However, one of the disadvantages of these procedures is that they require information across all (monophthongal) vowels of a single speaker to

Table 7.12: *Pairs of regional varieties for which of 'Extra' effects (effects not found for* HZ*), and 'Missed' effects (effects found for* HZ*, but not for these procedures) were reported. Italics indicate that the effects were found for all three procedures, for /ɛ/ (*$p < 0.001$*).*

Effects	LOBANOV	GERSTMAN	CLIH$_{i4}$
Extra F_1	*N-S×F-E*	*N-S×F-E*	*N-S×F-E*
	N-N×F-L	-	F-B×F-W
Missed F_1	*N-R×F-L*	*N-R×F-L*	*N-R×F-L*
	N-R×F-W	*N-R×F-W*	*N-R×F-W*
	N-M×F-L	-	N-M×F-L
	N-M×F-W	*N-M×F-W*	*N-M×F-W*
Extra F_2	*N-R×F-B*	*N-R×F-B*	*N-R×F-B*
	N-M×F-B	*N-M×F-B*	*N-M×F-B*
	N-N×F-B	N-N×F-B	-
	F-B×F-L	-	-
	F-B×F-W	-	F-B×F-W
Missed F_2	-	N-M×F-L	-

estimate the scale factors necessary to normalize the raw measurements. For LOBANOV, this scale factor is the standard deviation per formant, for CLIH$_{i4}$ this is the log-mean per formant and for GERSTMAN these are the minimum and maximum frequency for each formant. In the present research, I calculated these values using the nine monophthongal vowels per speaker[54].

However, measuring all vowels can be a laborious task, especially when a lot of speakers need to be investigated, which is quite common in sociolinguistic research. In addition, one can imagine that a subset of vowels is sufficient to estimate the overall mean frequency and the size of a vowel space. The point vowels seem plausible candidates for such a subset. Deterding (1990) described how he calculated Gerstman's (1968) scale factors using only two vowels per speaker, two of the cardinal vowels for English /ɑ/ and /i/, with good results.

To find out if the scale factors for the three procedures can be estimated with only the three point vowels per speaker, I estimated these scale factors for LOBANOV, CLIH$_{i4}$, and

[54]For the present research, a total of 15 vowels per speaker was available (each pronounces twice). However, six of these were (semi)-diphthongs (/o, e, ø, œy, ɔu, ɛi/). I decided against estimating the scale factors using all 15 vowels for two reasons. First, because I excluded the diphthongal vowels due to the dynamic character of their formant frequencies (cf. Chapter 4). Second, I did not expect that the diphthongal vowels would provide additional information about the speaker's vowel system. It is reported for these vowels that their formant frequencies measured at the begin and end points of the Dutch diphthongs coincide with the values of the formant frequencies of one of the nine monophthongal vowels (Pols, Tromp & Plomp, 1973).

GERSTMAN using the fundamental and formant frequencies of the three Dutch point vowels (/a/, /i/, and /u/). Three vowels were used, because LOBANOV – which uses a standard deviation – was included as well, and to include both a maximum and a minimum frequency for F_1 and F_2. Subsequently, using these scale factors, the normalization procedures were applied to the raw data of F_0, F_1, F_2, and F_3. The correlations were calculated between the measurements obtained with all nine vowels, and the measurements obtained based on the three vowels per speaker.

Table 7.13: *Correlation coefficients (Pearson's r) for the four acoustic variables, transformed following the three normalization procedures, for the measurements normalized using a scale factor obtained using nine vowels, and normalized using a scale factor obtained using three vowels per speaker.*

r	LOBANOV	GERSTMAN	CLIH$_{i4}$
D_0	0.90	0.86	0.95
D_1	0.97	0.99	0.99
D_2	0.98	0.97	0.99
D_3	0.87	0.87	0.96

Table 7.13 shows the correlations between the two versions of each normalization procedure. It can be observed that all correlations are high (between 0.86 and 0.99). All measurements that were obtained with scale factors estimated based on the three point vowels per speaker correlate strongly with measurements obtained using scale factors estimated using all nine vowels per speaker. Furthermore, the correlations for D_0 and D_3 are lower than those for D_1 and D_2, for all three procedures. As a final step, I carried out an analysis identical to LDA 1 (described in Table 7.1) using both sets (nine vowels and three vowels), to evaluate whether the percentages correctly classified vowel tokens would drop when the measurements were normalized using scale factors that were estimated using only three vowels per speaker.

Table 7.14 shows that the percentage of correctly classified vowel tokens for all three procedures is lower for the scale factors estimated using three vowels. For LOBANOV and for CLIH$_{i4}$, the percentages decrease (8% and 6%, respectively), for GERSTMAN, this decrease is less dramatic (2%). This result shows that the three procedures represent the phonemic variation in the transformed measurements less well when only three vowels were used. As the correlations were lower than 1.00 and the percentages correctly classified vowel tokens were lower for three vowels than for nine vowels, it can be concluded that the normalization based on three vowels is of a lesser quality than a normalization based on nine vowels. Apparently, some relevant information about the speaker's vowel system was excluded when only three vowels were used to estimate the scale factors.

The results in Table 7.13 and Table 7.14 thus indicate that it is better to obtain recordings

Table 7.14: *Results for the LDA with D_0, D_1, D_2, and D_3 as predictors for the three procedures. The set of predictors was calculated once using scale factors estimated with three vowels and once using all nine vowels from the LDA with F_0, F_1, F_2, and F_3 as predictors.*

%	Three vowels	Nine vowels)
LOBANOV	84	92
GERSTMAN	82	84
CLIH$_{i4}$	84	90

and measurements of the three point vowels and to normalize the vowels using the corresponding scale factors than to use no normalization at all. The results for data normalized using these scale factors are still better than carrying out no normalization (the percentage of correctly classified vowel tokens for HZ in Table 7.1 was 79%). However, this matter needs to be investigated in further detail before more conclusive remarks can be made.

7.4 Conclusions

In this chapter, it was first established that the three vowel-extrinsic/formant-intrinsic procedures, LOBANOV, GERSTMAN, and CLIH$_{i4}$, meet the criterion best. These three procedures were subsequently evaluated using multivariate analysis of variance on how well they preserved sociolinguistic variation from different sources, age-related, sex-related, and regional variation. Second, it was found that when my acoustic data was represented using only F_1 and F_2, the phonemic and the sociolinguistic variation could be represented adequately. F_1 or D_1 and F_2 or D_2 were related primarily to the vowel and D_0 and D_3 appeared to be related primarily to the speaker's sex, and to a lesser extent, to the speaker's age. Third, the evaluations carried out on the raw data and on the three vowel-extrinsic/formant-intrinsic procedures revealed that the sociolinguistic variation related to the speaker's age and speaker's sex was considerably smaller than the variation related to the speaker's regional background. The vowel-extrinsic/formant-intrinsic procedures preserved this regional variation. It was further found that some vowels showed considerable variation related to the speaker's regional background (such as /ɛ/ and /ɪ/), while other vowels showed very little regional variation (/a/ and /i/). Furthermore, it was found that the three procedures preserved previously reported sociolinguistic differences, although two of the procedures, LOBANOV and GERSTMAN reported differences that cannot be observed in the literature. However, this matter requires further investigation. Fourth, the scale factors used by the three vowel-extrinsic/formant-intrinsic procedures could be estimated reasonably well using three vowels per speaker, although the resulting acoustic data was less good than when the scale factors were estimated with all nine vowels.

Chapter 8

Perceptual comparisons

8.1 Introduction

This chapter explores the perceptual side of the normalization problem. In Chapters 1 and 4, it was argued that the output of the 12 acoustic normalization procedures must be compared to a human benchmark, i.e., the articulatory perceptual representation. The primary purpose of the present chapter is to describe how this perceptual representation was obtained. The articulatory perceptual representation consists of phonetically-trained listeners' judgments of the perceived Height, Advancement, and Rounding of a set of vowel tokens. These judgments were obtained through a listening experiment. In this experiment, a category judgment was also obtained for the set of vowel tokens. The secondary purpose of the present chapter is to evaluate the reliability of the articulatory judgments. In Chapter 3, it was concluded that it is unclear whether phonetically-trained listeners provide reliable articulatory judgments and that it is unclear how these judgments are affected by the availability of different vowel-intrinsic and vowel-extrinsic sources of information. For this reason, it was decided in Chapter 4 that the reliability of category and articulatory judgments of phonetically-trained experts had to be established before comparing these judgments to the normalized acoustic data. Furthermore, it was decided in Chapter 4 that it is necessary to establish how the articulatory judgments are affected by the availability of information about the speaker and/or the vowel token's intended category label.

Section 8.2 describes the design of the listening experiment and section 8.3 describes its results. The selection of the articulatory perceptual data used in Chapter 9 for the comparison with the acoustic data is described in section 8.4. Section 8.5 discusses the results.

8.2 Method

8.2.1 Stimulus material

The speech material used in the experiment consists of read vowels in a neutral context. The stimuli were taken from the 'neutral sentences' task from the VNC-database, described in Chapter 5. In this task, the speaker had to read aloud sentences containing the target vowel in three syllabic positions: in a closed syllable (CVC), in an open syllable (CVCV), and in isolation (V). The vowels in the closed syllables are used throughout the present research (cf. Chapter 5).

A subset of the VNC-database was used as the stimulus material in the experiment: i.e., the vowels from the 20 speakers from the N-R region. As argued in Chapter 4, this region was selected because the sociolinguistic variation was expected to be relatively moderate, while at the same time a considerable amount of anatomical/physiological variation was expected to be present. The sociolinguistic variation must be moderate, as I expected in Chapter 4 that presenting listeners with stimuli that reflects subtle sociolinguistic differences may induce them to use their perceptual scale optimally. If the sociolinguistic differences are too large, it may be impossible to establish whether the listeners can perceive (and reliably record) subtle sociolinguistic differences.

For each speaker in the N-R region, two tokens were available in a /sVs/-context for each of the nine monophthongal vowels (two tokens $\times$ 20 speakers $\times$ nine vowel categories = 360). To limit the number of experimental trials, only the second token for each speaker was selected to serve as a stimulus, thus selecting a total of 180 stimuli (one token for each of the nine vowel categories for each of the 20 speakers). However, on five occasions it was decided to select the first token. This was done whenever something was wrong with the second token: for instance, if the token displayed a deviant F_0-pattern, or when background noises were clearly audible during the realization of the vowel token.

The stimulus words were extracted semi-automatically from their carrier sentences. When extracting the syllable, I ensured that no part of the surrounding sounds was audible in the final stimulus. Markers used for cutting the stimulus word out of the carrier sentence were placed at zero crossings, to avoid possible 'clicks' in the stimuli.

8.2.2 Listeners

Eleven phonetically-trained experts participated as listeners in the experiment. One of them participated in a pilot version (his data was excluded from further analysis), while the other 10 participated in the actual experiment. The participating listeners were selected on the basis of their extensive experience with narrow phonetic transcription of speech sounds. Table 8.1 presents information about the 10 listeners who participated in the experiment.

Information about the listeners' backgrounds was collected through a questionnaire (see Appendix B for a translation of this questionnaire from Dutch). Note that whenever the results for individual listeners are presented in section 8.3, their listing is not identical to the listing in Table 8.1, thus preventing the listeners' identity to be revealed.

Table 8.1: *Information about the 10 phonetically-trained listeners participating in the experiment.*

Listener	Age	Gender	Native tongue	Years of experience	Transcription system
1	51-65	male	German	>25	IPA
2	51-65	male	Dutch	>30	IPA
3	21-35	female	Dutch	2	IPA
4	51-65	male	Dutch	>10	DJCVS, IPA
5	51-65	male	Dutch	>20	DJCVS, IPA
6	36-50	male	Dutch	11	IPA
7	36-50	male	Dutch	20	IPA
8	21-35	female	Dutch	4	IPA
9	51-65	male	Dutch	35	IPA
10	21-35	female	Dutch	2	IPA

8.2.3 Procedure

Experimental interface

The experiment was designed with NESU ("Nijmegen Experiment Set Up"), a software package designed at the Max Planck Institute for Psycholinguistics in Nijmegen for carrying out psychological experiments. The experimental response screen is displayed in Figure 8.1. This response screen contains 11 functional objects: nine vowel buttons, a copy of the IPA vowel quadrilateral for judging Height and Advancement, and a rectangular field for judging Rounding. All these objects were presented on a computer monitor with a resolution of 640 $\times$ 480 pixels. The size of the vowel buttons was 40 $\times$ 32 pixels, the rounding scale was 352 $\times$ 48 pixels and the vowel quadrilateral was a rectangle, 400 $\times$ 304 pixels in size.

The nine vowel buttons on the left-hand side of the response screen were used for the category judgments. Each button contains a phonetic symbol corresponding to one of the nine monophthongal vowels of Dutch. The placement of the individual vowel buttons on the screen from top to bottom is (roughly) alphabetical.

A copy of the IPA vowel quadrilateral (the 1996 corrected version) was used as the response area for the articulatory judgments in this experiment, as I expected that phonetically-trained listeners know it well. In addition, it is in widespread use in teaching phonetics and phonetic transcription courses. All listeners were asked whether they were familiar with the 1996 IPA-chart as described in "The Handbook of the International Phonetic Association" (1999). They all replied affirmatively.

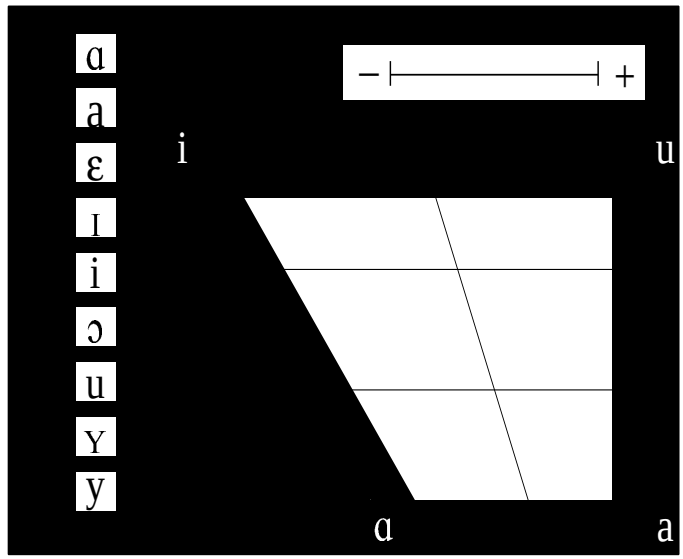

Figure 8.1: *The experimental interface used in the experiment.*

The quadrilateral is empty, except for the horizontal and vertical slanted lines (also present in the original version). The listeners were instructed to regard the four corners as theoretical end points of the quadrilateral. In addition, they were instructed to regard the horizontal axis as the axis along which the articulatory dimension 'place of constriction' (from left to right representing front to back) was displayed and that the vertical axis represented the articulatory dimension 'vowel height' (from bottom to top representing low to high).

Rounding was judged independently of Height and Advancement. This was necessary to obtain explicit lip rounding judgments. If Rounding was to be judged inside the vowel quadrilateral, the listeners may use two different strategies for judging vowel rounding. The first strategy entails using the quadrilateral as a two-dimensional chart *and* using the phoneme

category as an explicit indicator for the perceived lip rounding[55]. When using this strategy, the listener follows the IPA description that rounding is superimposed on the quadrilateral and that differences in location on the chart refer only to perceived differences in vowel height and place of constriction.

The second strategy entails using the left-right dimension of the IPA-chart for the judgments of place and vowel rounding. When this strategy was used, judgments for Rounding would be confounded with Advancement judgments. For instance, if the stimulus was perceived as high and advanced, the listener would place the mouse in the top half of the chart and click in the left hand corner if the vowel was perceived as spread, and more to the right if the vowel was perceived as rounded. The listener would then be expected to categorize this vowel as either /i/ or /y/, and click the mouse more to the left when the vowel was categorized /i/ than when the vowel was categorized /y/. Because of the possibility of obtaining results in which the Advancement and Rounding judgments are confounded, it was not desirable that listeners would use this strategy.

Thus, to encourage the listeners to use the first strategy, Rounding was judged outside the quadrilateral: in the rectangular field displayed in the top right-hand side of the computer screen in Figure 8.1. Inside the rectangle a horizontal scale is shown, flanked by a minus sign on the left and a plus sign on the right; the left side ($-$) of the scale represented a maximally spread vowel and the right side ($+$) represented a maximally rounded vowel. If the listener clicks on the scale, the horizontal coordinate corresponding to that pixel on the screen is recorded.

The listeners were informed that they should regard the end points of the axis as unrelated to the specific vowel category of the stimulus vowel to be judged. Instead, this axis represents the entire phonetic space for all vowels. For instance, when they perceived an /a/ as very spread for an /a/, then they should place the mouse only slightly to the left of in the middle of the axis, and when they perceive an /i/ as extremely spread, they should place the mouse more towards the extreme left side of the axis.

Experimental cycle

The sequence of events in one experimental trial (in which one stimulus vowel was judged) was as follows. First, a signal tone was played (a sine wave of 1000 Hz and a duration of 200 ms). Next, the stimulus was repeated 10 times with 1.5-second intervals. While the stimulus was repeated, the listener had to perform three tasks: first, a category judgments was made by clicking on one of the vowel buttons[56], second, the articulatory judgment was made. This

[55] The 1996 IPA-chart is essentially a three-dimensional space: for place of constriction, vowel height, and vowel rounding. In the IPA chart, if two vowel symbols are printed directly next to each other, the right-hand one of the two represents the rounded version.

[56] As explained in section 8.2.4, there were three experimental conditions, in one of these conditions, ("sub-experiment 2") it was not mandatory for the listeners to identify the stimulus by selecting one of the vowel buttons.

was done by clicking on the axis in the rounding scale to judge the vowel's Rounding, and by clicking in the vowel quadrilateral to judge the vowel token's Height and Advancement. After the listener had performed these three tasks, the experimental program would proceed to the next experimental trial, even if all 10 repetitions had not yet been completed. If the stimulus had been presented 10 times, the program waited for the listener to click in the vowel quadrilateral before it moved on to the next trial. If the listener clicked on the vowel quadrilateral before the rounding scale was clicked on, the experimental program proceeded as well. This results in a missing value for the judgment of rounding. It was attempted to prevent this from happening by explicitly warning the listener about this characteristic of the experimental program.

The experiment was divided into a number of experimental blocks. After each experimental block, a screen appeared with the word 'pauze' (pause). Whenever this screen appeared, the listener could temporarily interrupt the experiment to take a break or to ask a question, but the listener could also proceed with the next block without pausing.

The entire experiment consisted of three parts in which the stimuli were judged under three different conditions (sub-experiments 1, 2, and 3). The setup of these sub-experiments is explained in detail in section 8.2.4. Each sub-experiment was preceded by several familiarization stimuli (five in experiments 1 and 2 and nine in sub-experiment 3). These familiarization trials were used to explain the course of events and to allow users to get used to the tasks in the experiment. The five familiarization trials were taken from the VNC-database, from the N-M region. The familiarization trials were identical for each listener.

The vowels of three out of the 20 speakers were presented twice in each sub-experiment. The judgments of these three speakers were used to establish the consistency of each listener within each sub-experiment. In all, 207 stimuli were judged ((20 + 3 speakers) $\times$ 9 vowels = 207 stimuli). All listeners judged the data from speaker 17 and from two other speakers twice. These two other speakers were different across listeners. It was decided to use only one common speaker, to prevent possible effects related to individual idiosyncrasies of the speakers. Speaker 17 (female, young) was randomly chosen as the common speaker. The distribution of speakers who were judged twice per listener, is displayed in Table 8.2.

8.2.4 Three sub-experiments

When phonetically-trained listeners are asked to locate the perceived articulatory characteristics of a vowel token as a point in a three-dimensional space, one would expect that the availability of information about the speaker or about the vowel category influences that localization. For instance, if one of these sources of information is not available, the listener may be less sure about the precise placement of the point and the variation of the localizations may increase as a consequence.

To find out if this is the case, three sub-experiments were designed, in which two features

Table 8.2: *Speakers that were judged by each listener across the three sub-experiments.*

Listener	Speakers judged twice
1	17, 1, 2
2	17, 3, 4
3	17, 5, 6
4	17, 7, 8
5	17, 9, 10
6	17, 11, 12
7	17, 13, 14
8	17, 15, 16
9	17, 18, 19
10	17, 20, 1

were systematically varied: the presence of the vowel token's category label and the presence of systematic information about other vowel tokens from a speaker. The stimuli were presented either speaker-mixed or speaker-blocked, and the vowel category label was made available or not. The blocked and mixed presentation were varied to evaluate the role of the availability of speaker-specific information. In the speaker-blocked presentation, vowel-extrinsic information as well as vowel-intrinsic information is available. In this condition – after being presented with a few stimuli – the listener is expected to be able to make a more reasonable estimation of the speaker's vowel space than in a speaker-mixed condition. Consequently, the listener may locate the stimulus more reliably in a speaker-blocked condition than in a speaker-mixed condition. This expectation is based on findings that were reported by Verbrugge et al. (1976), Strange et al. (1976), Macchi (1980), Assmann, Nearey & Hogan (1982), and Mullennix, Pisoni & Martin (1989) on vowel-categorization experiments involving speaker-blocked and speaker-mixed presentation of vowel stimuli, as discussed in Chapter 3.

Information about the vowel token's label was made available to the listener or not. In two of the three sub-experiments, the listeners had to provide a category label and in one sub-experiment they did not have to provide a label. This was done as the expectation was that listeners maximize their perceived differences in the vowel tokens when making articulatory judgments in a vowel-blocked condition.

By combining a random presentation with two blocked presentations (for speaker and for vowel), three experimental conditions were obtained. In the first condition, referred to as sub-experiment 1, all stimuli were presented in a random order and the listeners were required to provide category judgments. In the second condition, sub-experiment 2, the stimuli were presented blocked per intended vowel category, thus providing the listeners with the intended

Table 8.3: *The presentation of the stimulus vowels and the availability of the vowel's category label per sub-experiment.*

Sub-experiment	Blocking	Vowel labels
1	no	no
2	vowel	yes
3	speaker	no

category label for each stimulus within a block. In the third condition, sub-experiment 3, all stimuli were presented blocked per speaker and the listeners were required to provide category judgments[57]. In all three sub-experiments, the same 207 stimuli were used (180 stimuli plus the 27 stimuli that were judged twice). Table 8.3 presents an overview of the three experimental conditions.

It was expected that judging stimuli in a speaker-blocked condition provides the listeners with extrinsic information that may allow them to build a 'mental image' of the vowel constellation for that speaker. Therefore, the speaker-blocked condition (sub-experiment 3), which was expected to provide more extrinsic information than sub-experiments 1 and 2, was run last. The fully random condition (sub-experiment 1) was run first, because in this condition less extrinsic information was expected to be available than in sub-experiments 2 and 3.

Sub-experiment 1: random condition

In sub-experiment 1, all stimuli were presented in a random order. Because the listener was not provided with information about the vowel category or with systematic information about other vowels by the same speaker, this sub-experiment was expected to be the most difficult one of the three. It was hypothesized that the listener has to rely solely on the vowel-intrinsic information within one stimulus and has to judge its articulatory characteristics in the absence of extrinsic information about the speaker's other vowels[58].

Sub-experiment 1 was divided into 11 blocks: the first one consisted of the five familiarization trials, and – because a total of 207 stimuli was to be judged – it was decided to create 10 experimental blocks: one block consisting of 27 trials and nine blocks consisting of 20 trials each. For each listener, a different randomized stimulus list was created. The stimuli were randomized with two restrictions: no two vowels from the same speaker, or three vowels belonging to the same intended vowel category were allowed to be presented in succession.

[57] The fourth possible condition, blocking by speaker and by vowel would result in blocks consisting of a single trial and was not implemented.

[58] Strictly speaking, this is only true for the very first time a stimulus from a speaker is presented, but in this sub-experiment the extrinsic (speaker-specific) information is minimal compared with sub-experiment 3.

The five familiarization trials were identical for each listener and were selected from the N-M region in the VNC-database. This was token 2 from 'saas' (/sas/) from speaker 19, token 2 from 'suus' (/sys/) from speaker 47, token 2 from 'ses' (/sɛs/) from speaker 68, token 2 from 'sies' (/sis/) from speaker 69, and token 2 from 'soes' (/sus/) from speaker 73. When selecting the tokens and speakers, an attempt was made to select tokens and speakers that displayed no or very faint regional accents and no prominent voice characteristics.

Sub-experiment 2: vowel-blocked condition

In sub-experiment 2, all stimuli were presented blocked by vowel category. It was hypothesized that this would induce listeners to judge the vowels without systematic information about the other vowels of the speaker. Therefore, the listeners were expected to rely solely on vowel-intrinsic information. Second, presenting the stimuli blocked by vowel may encourage the listeners to use the vowel quadrilateral in a different way than in sub-experiments 1 and 3. In particular, it was hypothesized that listeners may 'enlarge', or maximize, the perceived differences between vowel tokens within the same category. In experiments 1 and 3 it could be the case that listeners merely categorize the vowel tokens and pay less attention to within-vowel differences between vowel tokens. However, in this sub-experiment, it was hypothesized that listeners pay more attention to these differences.

The course of an experimental cycle in sub-experiment 2 was different from sub-experiments 1 and 3. In sub-experiment 2, the CVC-syllable that the speaker was supposed to read out aloud during the interview was printed in the lower mid-section of the experiment screen. The listeners were told that they were not required to provide a category judgment; they could start with the articulatory judgments by clicking on the rounding scale and subsequently in the vowel quadrilateral. They were given the option to indicate that they disagreed with the vowel token's label. If for instance they perceived a stimulus as /ɛ/, while the vowel was supposed to be /ɪ/ (and the word 'sis' was printed on the screen) they could indicate this by clicking on the button for /ɛ/, before judging its height, tongue advancement and rounding. They were asked to use the same procedure as in sub-experiment 1 and to click on the vowel button before they clicked on the rounding scale and the vowel quadrilateral. The familiarization trials in this sub-experiment were the second tokens from the CVC-syllable 'saas' (/sas/) from five speakers from the N-M region of the data base described in Chapter 5.

Sub-experiment 3: speaker-blocked condition

In sub-experiment 3, all stimuli were presented blocked by speaker. Sub-experiment 3 consisted of 24 blocks: the first one with nine familiarization trials (all nine monophthongal vowel tokens of speaker 19) and 23 blocks for the 20 speakers and the three repeated speakers. The 23 blocks each consisted of nine stimuli (of the nine monophthongal vowels). Again, a stimulus list was created. The nine vowel tokens within each block were randomized, as well

as the sequence in which the 23 blocks were presented. When randomizing these blocks, it was ensured that no two blocks of the same speaker were presented in succession (at least two other speakers were judged between two blocks of the same speaker). The nine familiarization trials were identical for each listener.

8.2.5 Expectations

Table 8.4 presents an overview of the information hypothesized to be available in each experimental condition. Given the information hypothesized to be present in each condition, I formulated the following expectations about the results for the three conditions on the category judgment task and the articulatory judgment task. For the category judgment task, a different number of misclassified vowel tokens (i.e., the listener categorized a vowel token into another category than the intended vowel category) was expected across the three conditions. Because the listener was expected to have less vowel-extrinsic information in experiment 1, the number of misclassifications (confusions) was expected to be higher than for sub-experiments 2 and 3. In sub-experiment 2 the listener had more information than in sub-experiment 3 (i.e., the vowel token's category label) therefore the number of misclassification was expected to be lower for sub-experiment 2 than for 3.

Table 8.4: *Information hypothesized to be available in each experimental condition.*

Sub-experiment	Information
1 (random)	Intrinsic
2 (vowel-blocked)	Intrinsic + vowel category
3 (speaker-blocked)	Intrinsic + extrinsic

For the articulatory judgment task, it was expected that the variance within vowel categories was different under each of the three experimental conditions. These articulatory judgments may be affected by the availability of information about other vowels produced by the same speaker, or by the vowel token's category label. It is first expected that the variance is highest in sub-experiment 2. Listeners have more information about the vowel token's intended category. This may induce them to enlarge the area for their judgments. Furthermore, I expected that the variance within vowels would be higher in sub-experiment 1 than in sub-experiment 3. It can be hypothesized that the listener may be less certain where to locate the stimulus in the vowel quadrilateral or rounding rectangle, if less information about the speaker or vowel category label is available. The expected differences per sub-experiment are listed in Table 8.5.

It is not unimaginable that the three conditions affect the overlap between vowels and the dispersion of the mean values per vowel per condition. However, at this point in the

research it is not feasible to formulate expectations about how the availability of (additional) information may affect the articulatory judgments.

Table 8.5: *Expected pattern in the results for the two tasks in the three sub-experiments.*

Sub-experiment	Category judgments: number of confusions	Articulatory judgments: variance within vowels
1 (random)	High	Intermediate
2 (vowel-blocked)	Low	High
3 (speaker-blocked)	Intermediate	Low

8.3 Results

8.3.1 Raw data

Four dependent variables were evaluated in the experiment. The first was a discrete variable: the response vowel category from the categorization task. The other three were the articulatory variables: Height (the corresponding y-coordinate of the vowel quadrilateral), Advancement (i.e., the corresponding x-coordinate of the vowel quadrilateral), and Rounding (the x-coordinate of the rounding scale). Before further processing, the values were transformed to a scale between 0 and 100.

High values for Height correspond to perceived lower openness (e.g., for /i/) and lower values to perceived higher openness (e.g., for /a/). Low values for Advancement correspond to perceived fronting (e.g., for /i/) and high values with perceived backing of the vowels (e.g., for /u/). Finally, low values for Rounding correspond to perceived spreading (e.g., for /i/) and high values to rounding (e.g., /u/). Table 8.6 lists the four dependent variables.

Table 8.6: *The dependent variables in the experiment.*

Dependent variable	Experimental origin	Level of measurement
Response vowel category	identification: one of nine vowel buttons	nominal
Height	y-coordinate of quadrilateral	interval
Advancement	x-coordinate of quadrilateral	interval
Rounding	x-coordinate of rounding scale	interval

Outliers and extreme values

The outliers and the extreme values were explored for the three variables Height, Advancement, and Rounding separately, to detect possible mistakes made by the listeners. For each listener, the data for the three variables was sorted into the nine vowel categories using the category judgments. An outlier was defined as a case with a value between 1.5 and three times the interquartile range (IQR). An extreme value was defined as a case with a value larger than three times the IQR.

All extreme values were removed, except when it was obvious that the extreme value was not a mistake, but a 'genuine' judgment. This was taken to be the case when the extreme value belonged to a larger series of outliers and extreme values (i.e., when the listener expressed larger differences between vowel tokens in this manner more often), or when the listener had judged the same vowel token twice within the same sub-experiment and both judgments were extreme values or outliers.

All outliers were in principle included in the data set, because it was not always possible to decide whether they were mistakes or formed a part of the listeners's judging strategy. An exception was made for cases with a value close to three IQR (but lower than three IQR) and for cases that were the sole outlier for a vowel for a listener.

If one of the responses was taken to be the result of a mistake, the entire case was removed (including the response category, Height, Advancement, and Rounding). I removed a total of 33 cases from the data set. In addition, a total of 36 cases was missing due to mistakes made by the listeners when performing the categorization task[59]. The total number of missing cases is therefore 33 + 36 = 69 (i.e., only 1.2% of the total number of cases, which was 5589).

After cleaning the data by removing mistakes and after the verification of the outliers, I carried out a series of diagnostic tests. This was done to establish if the 10 listeners were equally consistent, equally reliable, and whether they produced results that were overall comparable to each other. After having performed these analyses (described in section 8.3.2 and 8.3.3), I decided to exclude one of the listeners' results from further analyses, because this listener showed the lowest intra-rater consistency values for the labeling task (Cohen's κ values of 0.833 for sub-experiment 1, 0.958 for sub-experiment 2, and 0.875 for sub-experiment 3), the highest number of misclassified vowel tokens (a value of 48; Table 8.9), and the lowest intra-rater consistency values for the judgment task (Cochran's α values of 0.997 for Height, 0.981 for Advancement, and 0.924 for Rounding; Table 8.12). In addition, the misclassifications of this listener were generally in a different 'direction' than those of the other listeners, for instance, where the majority of the listeners categorized a vowel token as /ʏ/, while the vowel token was intended to be /u/, this listener categorized this vowel token as /y/[60]. Therefore, in the description of the results in this section, only the results from the remaining nine listeners are described.

Table 8.7 shows the number of valid cases per dependent variable. The number of cases

[59]For instance, when they forgot to provide a category label before clicking in the rounding rectangle and the

Table 8.7: *Number of cases (N) for the three variables Height, Advancement, and Rounding. N is corrected for mistakes, missing values and deleted cases.*

Variable	Number of cases (N)
Height	5520
Advancement	5520
Rounding	3657

(N) of 5520 was composed as follows: 3 experiments $\times$ 9 listeners $\times$ (20 speakers + 3 speakers) $\times$ 9 vowel categories = 5589. The value 5589 was corrected for the number of missing values (36) and excluded extreme values (33), the final number of cases for Height and Advancement is therefore 5520. Rounding shows a smaller number of cases (3657), because the responses for Rounding were not recorded in sub-experiment 2, due to an error[61] in the experimental program (3657 = $\frac{2}{3}$ $\times$ 5589 $-$ 69 missing cases and removed extreme cases).

8.3.2 Category judgments

Consistency

To establish whether the listeners were consistent in their judgments, the intra-rater consistency was calculated. When a listener is consistent with him- or herself, he or she gave a token the same category label whenever this token was judged, across and within sub-experiments. The intra-rater consistency is examined using the categorizations of the tokens from the three speakers that were judged twice within each experiment (see Table 8.2). Cohen's κ (Cohen, 1960), a measure for agreement suitable for nominal variables, was calculated for each listener separately.

Table 8.8 shows the values for Cohen's κ for each listener for each sub-experiment. A value close to 1 implies perfect consistency and a value close to 0 implies randomness. An analysis of variance (rm-anova) was carried out on the κ values in Table 8.8 for the nine listeners, with the sub-experiment as the independent variable. This analysis revealed no significant differences.

Table 8.8 shows further that all listeners categorized the vowel tokens with high consistency; two listeners (3 and 8) even reached a perfect score. This means that these two listener assigned the same category label to the stimulus on both presentations within each sub-experiment. Of course, it may well be the case that the listeners assigned the vowel

quadrilateral.

[60] These misclassifications were not included in Table 8.10.

[61] Which unfortunately remained undisclosed during all trial runs and pilot tests of the experimental program.

Table 8.8: *Intra-listener consistency: Cohen's κ for the 27 tokens that were categorized twice per sub-experiment. A value of 1 indicates that all the 27 tokens were assigned to the same category within a sub-experiment.*

Listener	Sub-experiment 1 (random)	Sub-experiment 2 (vowel-blocked)	Sub-experiment 3 (speaker-blocked)	Mean
1	0.917	1	0.958	0.958
2	0.958	1	0.958	0.972
3	1	1	1	1
4	0.916	1	1	0.972
5	0.958	1	1	0.986
6	0.958	0.958	0.917	0.944
7	0.958	0.958	0.873	0.930
8	1	1	1	1
9	1	1	0.958	0.972
mean	0.963	0.991	0.963	0.972

tokens to different categories across experiments, e.g., twice as /ɔ/ in sub-experiment 1 and the same stimulus twice as /ʏ/ in sub-experiment 3. Other listeners performed less well: the scores for listener 7 are less than perfect, meaning that this listener assigned the same vowel token to different categories on several occasions. However, the intra-rater consistency is overall high.

Vowel confusions

A vowel confusion occurred whenever a listener chose a category label other than the intended category label (e.g., when the vowel token in the stimulus 'sas' is categorized as /ɛ/). To find out whether the number of confusions differed across the three experimental conditions, this number was calculated for the three sub-experiments per listener. An overview of the number of confusions per sub-experiment per listener is listed in Table 8.9.

A repeated measures analysis of variance was carried out on the confusions data from the individual nine listeners displayed in Table 8.9 to establish the significance of the differences in confusions per sub-experiment. An effect was found for the factor 'sub-experiment' ($F(1.818, 8)=17.6$, $p< 0.01$, Huynh-Feldt correction). All pairwise comparisons showed an effect for 'sub-experiment' as well, indicating that all pairs of combinations of sub-experiments differed significantly from each other.

An important observation from Table 8.9 is that few confusions were made: between 1.2% and 5.6% of the stimuli received a label other than the intended label. In addition, the expectations about the differences between the sub-experiments described in Table 8.5 were

Table 8.9: *Confusions per experiment per listener. The number of confusions is displayed in every left-hand column per experiment, the corresponding percentage is displayed in every right-hand column.*

	Sub-experiment 1		Sub-experiment 2		Sub-experiment 3		
Listener	Confusions	%	Confusions	%	Confusions	%	Total
1	15	8.3	0	0	9	5.0	24
2	11	6.1	2	1.1	3	1.7	16
3	5	2.8	0	0	3	1.7	8
4	11	6.1	0	0	5	2.8	16
5	7	3.9	0	0	2	1.1	9
6	20	11.1	5	2.8	6	3.3	31
7	12	6.7	10	5.6	10	5.6	32
8	4	2.2	2	1.1	3	1.7	9
9	6	3.3	0	0	5	2.8	11
Total	91	5.6	19	1.2	46	2.8	156

confirmed; fewer stimuli were confused in sub-experiment 2 (the vowel-blocked condition) than in sub-experiments 1 (the random condition) and 3 (the speaker-blocked condition), and more confusions were made in sub-experiment 1 than in 3. This pattern in the results can be observed for all listeners. This finding can be interpreted as that the listeners accepted more variation in the vowel tokens when the labels were supplied beforehand, even though they were allowed to indicate disagreement with the provided label.

Because so few vowel confusions were found for the categorization task, it seems justified to pool the results of the three sub-experiments. In Table 8.10 an overview is displayed for all the vowel confusions per intended vowel pooled for the three sub-experiments and for the nine listeners. Only category judgments that were obtained the first time each vowel tokens was presented were included (thus excluding the categorizations of the stimuli from the three speakers that were presented twice). The total number of cases was therefore 4860 (3 experiments $\times$ 9 listeners $\times$ 9 categories $\times$ 20 speakers).

Table 8.10 shows that mid vowels /ɪ/, /ɛ/ and /ʏ/ generated the highest percentages of confusions (8.5%, 6.7%, and 4.7%, respectively), while the point vowels /ʉ/ (0.2%), /a/ (0.8%) and /i/ (0.7%) were rarely confused. Another observation from Table 8.10 is that the confusions are asymmetric: /ɛ/ was categorized 30 times as /ɪ/, whereas /ɪ/ was categorized as /ɛ/ only on 5 occasions. The same holds for /ɪ/ - /i/ (28, versus 3 for /i/-/ɪ/), /ʏ/-/y/ (18 versus 4 for /y/-/ʏ/), and /ɔ/-/ɑ/ (15 versus 8 for /ɑ/-/ɔ/).

To verify that the patterns in Table 8.10 were not the result of idiosyncrasies in the

Table 8.10: *Confusion matrix for the intended and response vowel category. The data was pooled for sub-experiment 1-3, and for the nine listeners. The rows represent the intended vowel category and the columns represent the response vowel category. The total number of cases is lower than 4860, because some cases were missing due to mistakes. The last two columns represent the number of confusions for each intended vowel category and the corresponding percentage, respectively.*

Vowel	/ɑ/	/a/	/ɛ/	/ɪ/	/i̇/	/ɔ/	/ɯ/	/ʏ/	/y/	To-tal	#	%
/ɑ/	524	2	2			8				536	12	2.2
/a/	4	528								532	4	0.8
/ɛ/			501	30				6		537	36	6.7
/ɪ/			5	487	28			8	4	532	45	8.5
/i̇/			1	3	536					540	4	0.7
/ɔ/	5					524				539	15	2.8
/ɯ/							536	1		537	1	0.2
/ʏ/			2	5				508	18	533	25	4.7
/y/					5		5	4	521	535	14	2.6
total	543	530	511	525	569	532	541	527	543	4821	156	3.3

stimulus material, the responses to individual stimuli that elicited four or more confusions (i.e., four or more times into the same vowel category) are listed in Table 8.11.

The six stimuli (a total of 49 confusions) displayed in Table 8.11 apparently accounted for 31.4% of the total percentage of confusions for the three sub-experiments (the total number of confusions of the entire data set is 156). The results shown in Table 8.10 could therefore be accounted for in part by responses to individual stimuli, but even if the responses to these six stimuli would be removed, the same trends are found.

Summary category judgments

The following results were found for the category judgment task. First, a significant difference was found in the performance of the listeners under the three experimental conditions. The performance was highest for sub-experiment 2, followed by the sub-experiment 3, while the performance was lowest for sub-experiment 1. Overall, the number of vowel confusions was low, this points to a possible ceiling effect operating on the results. Second, the number of confusions between the intended vowel category and the response vowel category was relatively high for vowels in the middle of the vowel space, such as/ɛ/, /ɪ/, and /ɑ/, while the number of confusions was lowest for point vowels such as /i/, /u/, and /a/. Third, the

Table 8.11: *Overview of the stimuli that elicited four or more confusions. The data was pooled for sub-experiments 1-3.*

Speaker	Intended vowel	Response vowel	Nr.
5	/ɔ/	/ɑ/	7
6	/ɛ/	/ɪ/	14
6	/ʏ/	/y/	7
13	/ʏ/	/y/	8
17	/ɪ/	/i/	4
18	/ɪ/	/i/	9

confusions showed an a-symmetric pattern, for instance, tokens intended to belong to the vowel category /ɛ/ were categorized more often as /ɪ/ than tokens intended to belong to the vowel category /ɪ/ were categorized as /ɛ/.

8.3.3 Articulatory judgments

Reliability

Per listener, 27 vowel tokens were judged six times in total (twice in each sub-experiment). The intra-listener reliability was established using Cochran's α, a measure of reliability suitable for variables at the interval level of measurement (Rietveld & Van Hout, (1993)). In Table 8.12, the values for Cochran's α are listed per dependent variable, for each listener. The values of α in this Table are the mean values of the αs for the three sub-experiments. Values close to 1 indicate high reliability, values close to 0 indicate poor reliability.

In Table 8.12, it can be seen that all listeners show high intra-listener reliability values (higher than 0.9). Overall, the average reliability is highest for Height, slightly lower for Advancement and lowest for Rounding. An analysis of variance was carried out on these data. The differences between the three variables appeared to be significant ($F(2, 24)=10.148$, $p<0.01$), a post-hoc comparison pointed out that this was due to Rounding: Rounding differed significantly from Height as well as from Advancement (Bonferroni, $p<0.01$), while Advancement and Height were not found to differ significantly.

Second, to assess the inter-rater reliability, Cochran's α was calculated for the three variables for each sub-experiment. For each of the three variables Height, Advancement, and Rounding, the judgments of the nine listeners were compared with each other. Table 8.13 lists the results. The values of Cochran's α represent the mean value across the three sub-experiments. The results show very high values of Cochran's α. These results indicate the listeners' behavior was very similar, and that it can thus be concluded that the listeners behaved reliably.

Table 8.12: *Intra-listener reliability,: Cochran's α calculated per listener for each dependent variable, for the six times a token was judged during the course of the entire experiment. The α values were first calculated for sub-experiments 1-3 separately, and subsequently the mean values were calculated over the three αs for the three sub-experiments. The latter values are displayed here.*

Listener	Height	Advancement	Rounding	Mean
1	0.997	0.993	0.994	0.996
2	0.990	0.988	0.978	0.985
3	0.992	0.985	0.980	0.986
4	0.987	0.992	0.917	0.965
5	0.986	0.984	0.952	0.974
6	0.993	0.992	0.941	0.975
7	0.993	0.984	0.963	0.980
8	0.995	0.991	0.987	0.991
9	0.995	0.992	0.965	0.984
Mean	0.992	0.989	0.964	0.982

Table 8.13: *Inter-listener reliability: Cochran's α calculated per dependent variable for the 180 stimuli per experiment.*

Experiment	Height	Advancement	Rounding
Sub-experiment 1	0.993	0.992	0.978
Sub-experiment 2	0.994	0.990	*not recorded*
Sub-experiment 3	0.994	0.992	0.981
Mean	0.994	0.992	0.978

Sub-experiments

The next step in the analysis of the results was assessing whether the listeners behaved differently in the three sub-experiments when judging Height, Advancement, and Rounding. This was done to test the predictions from Table 8.5 regarding the use of the area of the vowel quadrilateral and the rounding scale.

Three possible patterns in the results are discussed here. First, the conditions could differ in the locations of the mean values per vowel category for the three judgment variables. Second, differences could be found between the standard deviations around those means per sub-experiment. Third, it is possible that the covariance between the three dependent variables varies across the sub-experiments. Each of these possible patterns could be found independently of the other two patterns. For instance, it could be the case that while differ-

Table 8.14: *Significant (+) results ($p<0.01$) from the rm-anovas for the intended and response vowel categories. Results are shown for the Intended (I) and Response vowel category (R), for the three variables Height (H), Advancement (A), and Rounding (R). μ refers to the mean value, σ refers to the standard deviation, and r refers to the correlation coefficients.*

	Variable	Sub-experiment		Vowel		Vowel×Sub-experiment	
Effect		I	R	I	R	I	R
μ	H	-	-	+	+	-	-
	A	-	-	+	+	-	-
	R	-	-	+	+	-	-
σ	H	+	-	+	+	-	-
	A	+	-	-	+	-	-
	R	+	+	+	-	-	-
r	H×A	-	-	-	+	-	-
	H×R	-	-	-	+	-	-
	A×R	-	-	-	+	-	-

ences were found between the mean values, no differences between the standard deviations per vowel category were found. And in the case where there are no differences between the means and the vowel categories' standard deviations, the three dependent variables could co-vary differently across the three sub-experiments.

To get more insight into possible patterns of differences in the results, repeated measures analyses of variance (rm-anova) were carried out on the mean values per vowel category per listener, the standard deviations per vowel category per listener and the correlation coefficients (Height with Advancement, Height with Rounding, and Advancement with Rounding) as a measure for the covariance for the three sub-experiments. These analyses were carried out twice, once with the data grouped according to the intended vowel category and once with the data grouped according to the response vowel category. The significant results are listed in Table 8.14.

The nine rm-anovas on the data grouped according to the intended vowel category were carried out to assess how the listeners judged the same stimulus vowel tokens in the three sub-experiments. In the first three rm-anovas, the dependent variables were the mean (μ) judgments per listener per vowel category, in the next three rm-anovas, the dependent variables were the standard deviations (σ) of the judgments per listener per vowel category, and in the final three rm-anovas, the dependent variables were the correlation coefficients (r) between the three variables per listener per vowel category. The independent variables were identical across all nine rm-anovas. i.e., the three sub-experiments (Sub-experiment) and the nine vowel categories (Vowel). All calculations that involved the Rounding judgments were carried out on the data for sub-experiments 1 and 3 only. All degrees of freedom were

subjected to the Huynh-Feldt correction for possible non-sphericity.

Table 8.14 shows that the factor Vowel had a significant effect of the intended vowel category on the mean Height judgments ($F(3.989, 31.91)=273.225$, $p<0.01$). No significant effects were found for Height for the factor Sub-experiment, nor for the interaction between vowel category and sub-experiment (Vowel$\times$Sub-experiment). The results for Advancement show a similar pattern: a significant effect for Vowel ($F(2.833, 22.661)=230.093$, $p<0.01$) was found, while no significant effects for Sub-experiment nor for Vowel$\times$Sub-experiment were found. The mean Rounding judgments were significantly affected by the factor Vowel ($F(3.066, 24.526)=75.926$, $p=<0.01$). No significant effects were found for Sub-experiment nor for Vowel$\times$Sub-experiment. This indicates that the mean positions per vowel category do not vary across the three sub-experiments for Height, Advancement, or Rounding.

The results for the standard deviations per listener revealed a different pattern. For the standard deviations for Height, a significant effect was found for Sub-experiment ($F(1.95, 15.598)=6.204$, $p<0.05$) and for Vowel ($F(5.546, 44.366)= 12.979$, $p<0.01$), while Vowel$\times$-Sub-experiment was not significant. For Advancement, again a significant effect was found for Sub-experiment ($F(2, 16)= 4.369$, $p<0.05$). No significant effects were found for Vowel or Vowel$\times$Sub-experiment. For Rounding, a significant effect was, first, found for Sub-experiment ($F(1, 8)= 11.193$, $p<0.01$). A second significant effect was found for Vowel ($F(3.785, 30.278)=2.85$, $p<0.05$) and Vowel$\times$Sub-experiment was not significant. These results indicate that the judgments showed variation depending on the presentation of the stimuli in the three sub-experiments for Height, Advancement, and Rounding.

The correlation coefficients show no significant effects. This indicates that the covariance per vowel category did not vary across the three sub-experiments.

In summary, the main findings from these nine rm-anovas are that the vowel category has a large effect on the judgments; vowel category affected the mean values and the standard deviations. Second, the results show that the standard deviations per vowel category were affected by the division in sub-experiments.

It seems possible that the listeners relied entirely on the vowel category and did not reliably judge the variation within vowel categories. If this is the case, then the effect of the sub-experiment on the standard deviations may be an artifact of the categorization task. In the current series of rm-anovas, the response vowel category was not yet taken into account. The expectation is that the effect found for sub-experiment disappears, when the data is sorted according to the response vowel category. This seems plausible, because it was found that the listeners categorized the vowel stimuli differently under the three experimental conditions (as can be observed in Table 8.9). Thus, if the listeners' articulatory judgments are affected predominantly by their response vowel category and not by the experimental condition, then the effect for sub-experiments should disappear when the rm-anovas are carried out again, this time with the stimuli sorted according to the response vowel category. If the effect for the standard deviations disappears, then this effect was an artifact of the grouping of the

responses according to the intended vowel categories. However, if no effect is found, it must be concluded that the listeners were primarily influenced by their response vowel category label, and not by the experimental condition.

Therefore, to establish whether the response vowel category affected the means, standard deviations, and correlation coefficients for the articulatory judgments, nine rm-anovas were carried out. The significant effects for these analyses are listed in Table 8.14.

The first three rm-anovas were carried out on the aggregated mean values (μ) per response vowel category per listener. The results for the means reveal a significant effect of the factor Vowel on the Height judgments ($F(3.933, 31.465)=273.939$, $p<0.05$). No significant effects were found for the Height judgments for Sub-experiment, nor for Vowel$\times$Sub-experiment. The results for Advancement show a significant effect of vowel category on the mean judgments ($F(2.651, 21.211)=222.663$, $p<0.05$) and no significant effects for Sub-experiment, nor for Vowel$\times$Sub-experiment. The results for the mean Rounding judgments show a significant effect for vowel category ($F(3.025, 24.199)=80.165$, $p=<0.05$). No significant effects were found for the factor Sub-experiment, nor for the Vowel$\times$Sub-experiment interaction term.

The results for the standard deviations (σ) per listener per vowel category reveal no significant effects of sub-experiment on the standard deviations of the Height judgments. A significant effect was found for the vowel category ($F(3.663, 29.306)=9.217$, $p<0.05$), while the factors Sub-experiment and Vowel$\times$Sub-experiment interaction were not significant. For Advancement, a significant effect was found for vowel category ($F(6.989, 55.913)=3.617$, $p<0.05$). No significant effects were found for Sub-experiment, nor for Vowel$\times$Sub-experiment. For Rounding, a significant effect was found for sub-experiment ($F(1, 8)=8.865$, $p<0.05$). Neither the factor Vowel, nor Vowel$\times$Sub-experiment was found to significantly affect the standard deviations of the Rounding judgments.

The rm-anova on the correlation coefficients (r) of Height $\times$ Advancement show a significant effect for vowel category ($F(7.638, 61.105)=11.367$, $p<0.05$). The factors Sub-experiment and the Vowel$\times$Sub-experiment were not significant. Vowel category shows a significant effect on the correlation coefficients for Height $\times$ Rounding ($F(8, 64)=4.012$, $p<0.05$). No significant effects were found for the factors Sub-experiment and Vowel$\times$-Sub-experiment. The vowel category significantly affects the correlation coefficients for Advancement$\times$Rounding ($F(7.842, 62.738)=6.805$, $p<0.05$). The factor sub-experiment and the interaction between vowel category and sub-experiment show no significant effects.

The results of the nine rm-anovas on the response vowel categories can be summarized as follows. First, there appear to be no systematic differences between the judgments in the three sub-experiments, neither in the mean values, nor in the standard deviations, nor in the correlation coefficients. Only one significant effect was found for Sub-experiment: for the standard deviations for Rounding. Second, the judgments are shown to be systematically affected by the response vowel category, for Height, Advancement, and Rounding, for the mean values, standard deviations and correlation coefficients, except for the standard deviations for the Rounding judgments. This last result was in the direction predicted in Table 8.5,

the standard deviations were higher for sub-experiment 1 (the mean value for σ across all nine response vowel categories was 13.13) than for sub-experiment 3 (mean σ was 12.15). Nevertheless, this result is not of much importance; the difference between the mean values is small and the results for Height and Advancement did not show similar significant results for the standard deviations.

From the results in Table 8.14, it appears that the differences in the standard deviations across the three sub-experiments in the judgments disappeared for the largest part when the analyses were performed on means, standard deviations, and correlation coefficients grouped according to the response vowel category. It must therefore be concluded that the difference in standard deviations found for the analyses for the judgments grouped according to the intended vowel category was an artifact of the analysis.

Sociolinguistic variation

The next step in the analysis of the articulatory judgments was establishing how reliably listeners judged the vowel tokens within each vowel category. In other words, establishing how reliably the listeners judged the sociolinguistic variation. To this end, Cochran's α was calculated for each vowel category for each listener for the three articulatory variables. If the values of Cochran's α are equal to zero, this indicates that the listeners did not reliably judge sociolinguistic variation.

Table 8.15: *Mean Cochran's α per vowel category across sub-experiments 1, 2, and 3 for Height, Advancement, and Rounding (for Rounding the values were calculated using only sub-experiments 1 and 3); mean values calculated using the values for the nine listeners.*

Vowel	Height	Advancement	Rounding	Mean
/ɑ/	0.235	0.433	0.233	0.299
/a/	0.378	0.654	0.237	0.423
/ɛ/	0.498	0.452	0.108	0.353
/ɪ/	0.378	0.283	0.150	0.270
/i̇/	0.230	0.343	0.052	0.208
/ɔ/	0.367	0.598	0.206	0.390
/ʊ/	0.299	0.311	0.176	0.262
/ʏ/	0.452	0.336	0.211	0.333
/y/	0.110	0.381	0.106	0.199
Mean	0.327	0.421	0.164	0.304

Table 8.15 lists the mean values per vowel category per dependent variable. The values were first calculated for each listener separately and subsequently the mean of these values was calculated. The values for each individual listener were calculated for the intended vowel

Table 8.16: *Cochran's α per vowel category (mean value calculated across the sub-experiments 1, 2 and 3) for Height, Advancement, and Rounding (for Rounding the values were calculated only for sub-experiments 1 and 3); mean values calculated using the values for all listeners, except for listeners 4 and 8.*

Vowel	Height	Advancement	Rounding	Mean
/ɑ/	0.442	0.669	0.242	0.451
/a/	0.485	0.645	0.334	0.488
/ɛ/	0.674	0.455	0.252	0.460
/ɪ/	0.438	0.250	0.145	0.278
/i̇/	0.178	0.333	0.129	0.213
/ɔ/	0.501	0.650	0.281	0.477
/u/	0.467	0.470	0.124	0.354
/ʏ/	0.572	0.394	0.304	0.423
/y/	0.211	0.399	0.323	0.311
Mean	0.441	0.474	0.237	0.355

categories[62] and are displayed in Appendix D. Table 8.15 shows that the reliability scores are considerably lower within vowel categories than between vowel categories (displayed in Table 8.12). This indicates that the listeners appear to be able to judge phonemic variation (between vowel categories) more reliably than sociolinguistic variation (within vowel categories). Some listeners are less reliable than others: Appendix D shows that listeners 4 and 8 (see Table D.1) show the lowest values of all listeners for all three dependent variables. Therefore, to investigate the effect of listeners 4 and 8 on the mean reliability scores, the mean values of Cochran's α were calculated again, this time excluding the values from listeners 4 and 8.

In Table 8.16, it can be seen that, overall, the values for Cochran's α are higher when the scores for listeners 4 and 8 are excluded. In addition, the following pattern can be observed in Table 8.16 more clearly than in Table 8.15. The reliability scores for Height and Advancement are moderately high for some vowel categories, i.e., for /ɛ/, /ɔ/, and /ʏ/. For some other categories, low values were found, e.g., for /ɪ/, /i̇/, and /y/, indicating almost random judgment behavior within these vowel categories. Finally, it can be observed that all vowel categories show very low values for Rounding; these values depressed the overall mean values considerably.

Given the results in Tables 8.15 and 8.16, it can be concluded that the hypothesis, that

[62] Using the response vowel categories would result in missing values.

listeners did not reliably judge sociolinguistic variation, is not valid: the listeners judged sociolinguistic variation in Height and Advancement reliably and systematically for vowel tokens in all (intended) categories except for /ɪ/, /i/, and /y/.

Anatomical/physiological variation

The last issue to be addressed is whether the articulatory judgments showed variation depending on the anatomical/physiological characteristics of the speaker. Both older and younger male and female speakers were included in the design of the experiment, so the anatomical/physiological variation was expected to be considerable in the stimulus material.

To establish if the listeners were influenced by the anatomical/physiological characteristics of the speakers of the vowel stimuli, various analyses of variance were performed, for the three articulatory variables, for each of the three sub-experiments, and for each of the nine listeners. Speaker-age and speaker-sex served as independent variables, and Height, Advancement, and Rounding served as dependent variables in each analysis. No significant effects were found for the factors gender or age in any of these analyses. It was concluded that the articulatory judgments did not vary systematically depending on the anatomical/physiological characteristics of the speaker.

Summary articulatory judgments

The findings of the articulatory judgments task can be summarized as follows. First, the variation in the judgments was related primarily to the stimulus token's response category label. Second, overall, no systematic effects could be found for the presentation of the stimuli in three sub-experiments except for Rounding (the variation in the Rounding judgments was higher in sub-experiment 1 than in sub-experiment 3 for the data grouped into the response vowel categories). Third, the listeners appeared to be able to judge phonemic variation (between vowel categories) more reliably than sociolinguistic variation (within vowel categories). Fourth, it was found that the sociolinguistic variation in the stimulus material could be judged with relative reliability for all vowels except for /ɪ/, /i/, and /y/, the vowel tokens for the remaining six vowel categories showed higher reliability scores. Fifth, it was found that listeners' judgments were not affected by the anatomical/physiological characteristics of the speaker.

8.4 Selection of articulatory perceptual data

As announced in section 8.1, the purpose of the experiment was not only to get more insight into the judgment strategies and use of information sources by the listeners, but also to generate an articulatory perceptual description of a set of vowel data used for the comparison (described in the next chapter) with acoustic data (measurements of F_0, and F_1, F_2, and F_3,

transformed according to the procedures for vowel normalization that are evaluated in the present research). In this section, the selection criteria for the articulatory perceptual data that were used for this comparison are described.

8.4.1 Listeners

The first step that is to be taken in the selection of the articulatory perceptual data, is to decide which listeners to include. In section 8.3.3, it was concluded that listeners 4 and 8 had the lowest within-vowel reliability. Furthermore, the results in Table 8.16 show that the values for Cochran's α are higher when listeners 4 and 8 are excluded. The reliability values should be as high as possible, to allow the comparison in Chapter 9 between the articulatory perceptual data and the acoustic data to be successful. It seems therefore reasonable to exclude listeners 4 and 8 from further processing. However, there is one aspect that should be taken into account, which relates to the reliability of the listeners as a group. Because the goal of the comparison of the articulatory perceptual and acoustic data is to model not only the judgments of the individual listeners but also of the listeners as one instrument, it is important that the reliability of the group is not decreased by including the judgments of a few less reliable listeners. To assess how the judgments of the two relatively unreliable listeners 4 and 8 influence the reliability of the whole listener population, Cochran's α was calculated per vowel category on the data across the three sub-experiments, once on the pooled data from all nine listeners, and once on the data from the seven listeners (excluding listeners 4 and 8).

Table 8.17 shows the results of the two sets of reliability analyses. It turns out that the exclusion of the data from listeners 4 and 8 did not substantially influence the results. Excluding the data from listeners 4 and 8 appears improve the reliability scores for the closed rounded vowel categories: /ɔ/, /u/, /ʏ/, and /y/). The scores for the other categories are slightly lower, when listeners 4 and 8 are excluded. In general, when reliability analysis are performed on a smaller number of raters, the overall reliability scores decrease, even if the excluded raters' data contains a lot of noise. However, that does not appear to be the case here. Removing the two listeners did not considerably lower the reliability scores of the group. For some vowel categories, especially for /u/ and /ɔ/ considerable higher scores were obtained.

Finally, in Table 8.17, it can be seen that, when the data of all nine listeners are pooled, the reliability values are considerably higher for Rounding than in Table 8.15 and 8.16. This can partially be explained by the fact that this analysis is performed on more data; the reliability was calculated for a group of judges and not for a single judge. Given the results in Table 8.17, it was decided to exclude listeners 4 and 8's data from further analysis. The data from the remaining seven listeners was used in the comparison with the acoustic data in Chapter 9.

Table 8.17: *Cochran's α per vowel category across the sub-experiments 1, 2, and 3 for Height and Advancement, and across sub-experiments 1 and 3 for Rounding. The values were calculated on the data for 7 listeners (listeners 4 and 8 excluded) and for 9 nine listeners, 'list.': listeners.*

Vowel	Height		Advancement		Rounding		Mean	
	9 list.	7 list.	9 list.	7 list.	9 list.	7 list.	9 list.	7 list.
/ɑ/	0.724	0.713	0.783	0.781	0.908	0.898	0.805	0.797
/a/	0.702	0.690	0.831	0.840	0.910	0.881	0.814	0.803
/ɛ/	0.877	0.888	0.690	0.657	0.825	0.801	0.797	0.782
/ɪ/	0.858	0.857	0.804	0.820	0.553	0.519	0.738	0.732
/i̇/	0.773	0.759	0.822	0.852	0.755	0.734	0.783	0.781
/ɔ/	0.853	0.886	0.808	0.818	0.452	0.739	0.704	0.814
/u/	0.736	0.756	0.660	0.694	0.658	0.761	0.685	0.737
/ʏ/	0.902	0.859	0.924	0.911	0.586	0.704	0.804	0.824
/y/	0.792	0.773	0.765	0.798	0.678	0.698	0.745	0.756
Mean	0.802	0.798	0.787	0.797	0.703	0.748	0.764	0.781

8.4.2 Sub-experiments

The articulatory perceptual data that was used for the comparison with the acoustic data in the next chapter could be selected in two possible ways. An option would be to pool over the data from the three sub-experiments. However, as a small effect was found for Rounding, this was not an option. Another option would be to use the articulatory data from either sub-experiment 1, 2, or 3. However, sub-experiment 2 is excluded, for reasons mentioned earlier. Although the data from sub-experiment 1 and sub-experiment 3 seem both equally suitable (because only minimal differences in performance between the two sub-experiment were found), I decided that the data from sub-experiment 1 was most suitable. The reason for this is that the judgments obtained in sub-experiment 1 more closely resemble the way the acoustic data was obtained than the way the judgments were obtained in sub-experiment 3. In sub-experiment 1, no information about other vowel tokens from the same speaker was made available. Analogously, the measurement program that extracted the F_0 and the formant frequencies, could not make use of information about other vowel tokens from the same speaker either. Consequently, I decided to use only the data from sub-experiment 1 for the comparison with acoustic data in the next chapter. For the stimuli that were judged

twice, the first occurrence was selected, thus selecting 180 articulatory judgments from the seven selected listeners. The precise nature of the setup of the articulatory perceptual data is discussed in further detail in Chapter 9.

To get a clearer picture of the data of the seven listeners for the three variables from sub-experiment 1, the data was plotted in Figure 8.2. This figure shows all judgments for the seven listeners, for Height, Advancement, and Rounding, labeled according to the response vowel. It can be observed that the data appear to be (more or less) grouped per response vowel in all three plots.

8.5 Discussion

8.5.1 Main findings

The main findings of the experiment were as follows. First, although the presentation type (sub-experiments 1, 2, and 3) affected the category judgments, no systematic effect of presentation type could be found for the articulatory judgments. Second, the reliability of the judgment of phonemic variation was high; high scores were found for the category judgments and for the articulatory judgments between vowel categories. Third, the overall reliability of the listeners was considerably lower for the articulatory judgments of sociolinguistic variation. However, all vowel judgments were reliable (all values of Cochran's α were significantly different from 0, as can be seen in Tables 8.15 and 8.16), and some vowels (/ɛ/, /ɔ/, /ʏ/, and, to a lesser extent, /ɑ/ and /a/) were judged with higher reliability than others. These findings are discussed in further detail below.

Category judgments

It was found that the overall percentage of confusions was low, between 1.2% and 5.6% of the vowel tokens was misclassified. I concluded that a ceiling effect seemed to operate on the results for the three sub-experiments. Furthermore, it was found that the differences between sub-experiments went in the predicted direction (cf. section 8.2.4, Table 8.5): when only vowel-intrinsic information was available to listeners (sub-experiment 1), the results showed a higher number of misclassifications, than when extrinsic as well as intrinsic information was made available (sub-experiment 3), and the lowest number of misclassifications was found when the vowel token's intended category label was made available.

Articulatory judgments

The results for the articulatory judgment task show a small but significant difference between the judgments in the three sub-experiments when the data was sorted according to the stimulus token's intended vowel category. However, this difference disappeared when the data was sorted according to the listeners' response vowel category. Given this latter result,

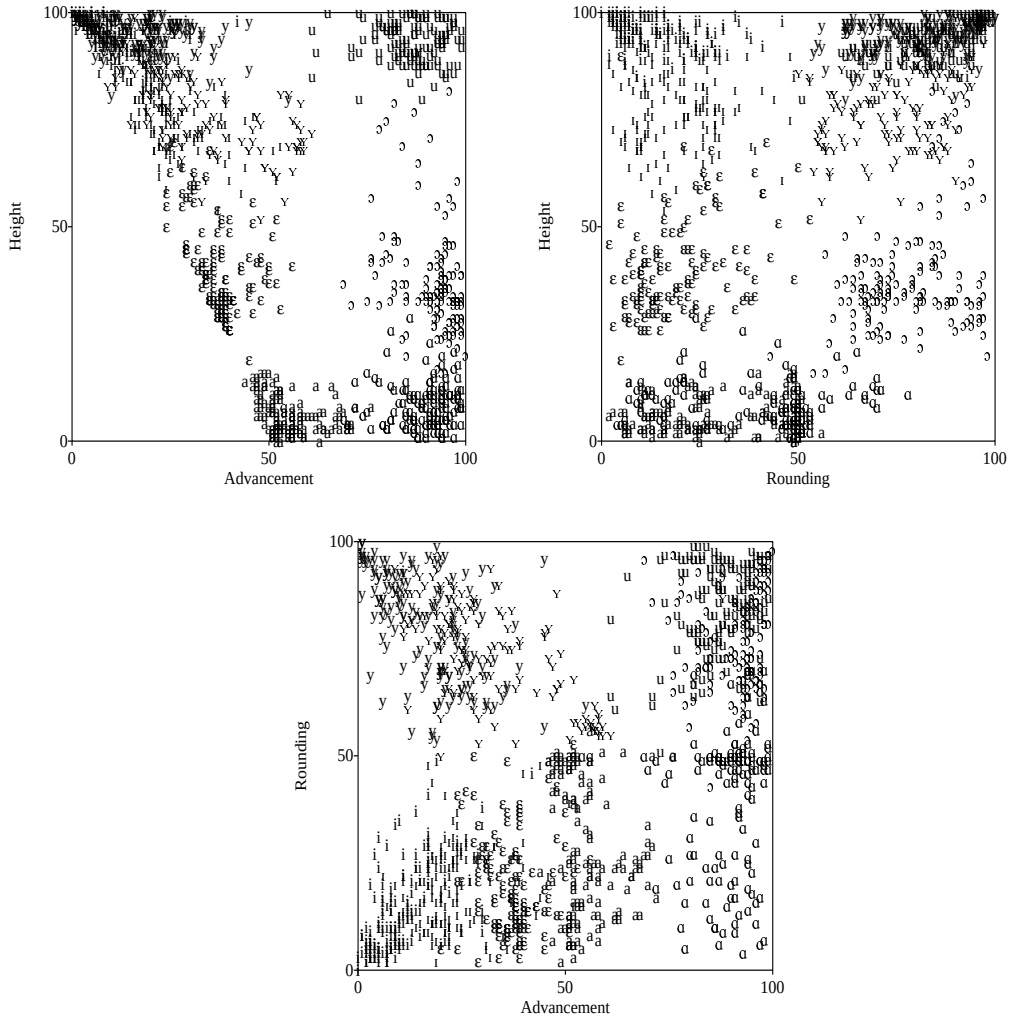

Figure 8.2: *Scatter plots of the Height, Advancement, and Rounding judgments for the 180 vowel tokens for the seven listeners for sub-experiment 1. Each phonetic symbol represents a single judgment. The judgments are expressed on a scale from 0-100 (original pixel values from the experiment transformed to 0-100).*

I concluded that there were no systematic differences between the judgments for Height, Advancement, and Rounding, regardless of the information hypothesized to be available to the listener in each sub-experiment. Only for the standard deviations for the Rounding judgments a difference was found for presentation type. However, the importance of this difference should not be overestimated, because the reliability scores were overall lower for Rounding for individual listeners than for Height and Advancement (see the results in Tables 8.12, 8.13, and 8.17).

It can thus be concluded that listeners were not influenced systematically by the availability of extra information (the vowel token's category label or information about other vowels from the same speaker) when performing the judgment task. Consequently, because no systematic differences in judgment strategies were found across the three sub-experiments, none of the predictions about the means, correlation coefficients and variance in Table 8.5 were confirmed.

It is unclear why the listeners were influenced by the three presentation types when making category judgments but not when making articulatory judgments. It seems possible that the listeners did not judge sociolinguistic variation in the stimulus vowel tokens, and focused only on the phonemic variation when making articulatory judgments. However, this explanation does not seem plausible: if listeners had indeed randomly clicked in the vowel token's category area, low reliability values should have been found for each listener for each vowel, given the results in Appendix D and in Table 8.17. It must therefore be concluded that the listeners judged the sociolinguistic variation within certain vowels more reliably than within other vowels.

The finding that listeners judged a subset of the vowels relatively reliably, corroborates an alternative explanation: the listeners judged the sociolinguistic variation in the vowel tokens reliably whenever enough variation was present. If it was indeed the case that, for instance, less linguistic within-vowel variation was present in the vowel category /ɪ/ than in the vowel category /ɛ/ in the N-R variety of standard Dutch, then this explains why listeners apparently judged vowel tokens that they had categorized as /ɪ/ less reliably than vowel tokens they categorized as /ɛ/. However, it is not possible to test this hypothesis until the comparisons with the acoustic representations are presented (in Chapter 9).

As mentioned before, the results show that the articulatory judgments were not affected by the vowel-specific or speaker-specific information made available to the listeners across sub-experiments. This indicates that listeners were primarily influenced by information contained within the stimulus token itself. If a single token contains enough information about the speaker, it can be hypothesized that the listener uses the speaker-specific information present in that vowel token to estimate a speaker-related frame of reference, against which the perceived articulatory characteristics of the single vowel tokens can be judged. If this is the case, then providing additional information about the speaker or the vowel does not considerably influence these judgments[63].

[63]If this is true, then the fact that listeners could listen up to 10 times may have improved the reliability of the

8.5.2 Comparison with previous studies

Category judgments

Chapter 3 discusses various studies investigating the effect of speaker-blocked versus speaker-mixed presentation of vowel tokens. The results of these studies were summarized in Table 3.1. The results from my experiment (2.8% and 5.6% error rates in the speaker-blocked and speaker-mixed conditions, respectively) are in agreement with those found by Verbrugge et al. (1976), Macchi, (1980), Assmann et al. (1982), and Mullennix et al. (1989). These authors found lower percentages for speaker-blocked presentations than for speaker-mixed presentations. Furthermore, my results also indicate a ceiling effect, as reported by all these studies, except by Mullennix et al.[64]. Finally, I found vowel-specific differences in the percentages of confusions. The percentages were highest for the (intended) vowel categories in the middle part of the (monophthongal) vowel system of Dutch, such as /ɪ/, /ʏ/, and /ɛ/. The percentages were lowest for vowels at the corners of the vowel system, such as /i/, /a/, and /u/. Other authors, such as Verbrugge et al. (1976) and Macchi (1980), reported a similar pattern in the error percentages for central and point vowels.

Articulatory judgments

In Chapter 3, I discussed Assmann (1979). I interpreted his results as that phonetically-trained listeners are strongly influenced by the perceived vowel category when judging the height and advancement of vowel tokens. My results showed the same effect as found be Assmann. The articulatory judgments in my experiment appeared to vary predominantly depending on the response category label of the stimulus token.

Both Ladefoged (1960) and Laver (1965) compared the reliability of articulatory judgments made by phonetically-trained listeners. They both reported that lip rounding was judged considerably less reliable than tongue height and tongue advancement. In my experiment, I found a similar pattern in the results.

---

articulatory judgments, or have decreased the variance around the means per vowel of these judgments. However, it could not be tested whether presenting a stimulus token only once led to a deterioration in the reliability scores (as displayed in Table 8.13), or an increase in the variance per vowel, because the experiment had to resemble a situation in which a listener makes a narrow phonetic transcription of a vowel token. Whenever a listener does this in a 'real-life' listening situation, he or she usually has the opportunity to listen to the vowel token to be transcribed more than once.

[64] It was not possible to attempt to avoid a ceiling effect by presenting the stimuli in noise, like was done by Mullennix et al., because the vowel stimuli were used for the articulatory judgment tasks, a task that required a good sound quality.

8.5.3 Conclusions

The purpose of the listening experiment was twofold. First, it was carried out to evaluate the effect of the availability of different vowel-intrinsic and vowel-extrinsic sources of information on category and articulatory judgments of vowel tokens by phonetically-trained listeners. The results of the experiment show that the category judgments varied depending on the presentation type. Furthermore, the articulatory judgments varies primarily depending on the response vowel category. The results further indicate that phonetically-trained listeners reliably judged phonemic information present in vowel tokens. Finally, the results indicate that a single vowel token may contain enough information to allow phonetically-trained listeners to reliably judge the articulatory characteristics of that vowel token.

The second purpose was to collect articulatory perceptual data necessary for a comparison with acoustic data in the following chapter. I decided to use the mean judgment values (of seven individual listeners) from sub-experiment 1 (the random condition) for the comparison with the (transformed) acoustic vowel measurements in the next chapter.

Chapter 9

Perceptual-acoustic comparisons

9.1 Introduction

This chapter describes the comparison between the perceptual representation and the 12 acoustic representations of the set of 180 vowel tokens from the N-R region of the sociolinguistically balanced database (described in Chapter 5). The perceptual description consists of judgments of each vowel token's perceived articulatory characteristics. These judgments were elicited by means of the experiment described in Chapter 8. The acoustic description consists of measurements of each vowel token's fundamental frequency and its first three formants, which were transformed through 12 vowel normalization procedures, as described in Chapters 2 and 7. This chapter aims to establish which procedure for vowel normalization produces acoustic data that allows the articulatory perceptual data to be modeled best.

The setup of this chapter is as follows. Section 9.2 describes the articulatory perceptual data and the acoustic data. Section 9.3 describes how the phonemic variation in the perceptual data can be modeled using (transformed versions of) the acoustic data. Section 9.4 describes how the comparisons between the articulatory perceptual and acoustic data were carried for each individual vowel, to establish how well the sociolinguistic variation could be modeled. Section 9.5 elaborates on the results found in Section 9.4. Section 9.6 lists the conclusions of this chapter. Parts of the research in this chapter are also described in Adank (2000) and in Adank, Van Hout & Smits (2001).

9.2 The acoustic and perceptual representation

Acoustic representation

The acoustic data consist of measurements of F_0, F_1, F_2, and F_3, taken from the vowel token's temporal midpoint, as described in Chapter 6. The measurements of tokens from the nine monophthongal vowel categories (/ɑ/, /a/, /ɛ/, /ɪ/, /i/, /ɔ/, /u/, /ʏ/, /y/) from the 20 speakers in the N-R speaker group (10 female and 10 male speakers) were used as the raw acoustic representation to be compared to the perceptual representation. Each speaker produced two tokens per vowel category during the interview; in general, the second token was used (on five occasions, the first token was used instead).

Figure 9.1 shows the raw acoustic data. It can be observed that the measurements for the four acoustic variables are scattered across the acoustic $F_2 \times F_1$ and $F_3 \times F_0$ spaces. The measurements are not tightly clustered, which means that the vowels show considerable overlap.

Perceptual representation

As described in Chapter 8, section 8.4, the perceptual representations consists of the values of Height, Advancement, and Rounding. The perceptual data was generated using the same 180 vowel tokens used to generate the acoustic data: from each of the 20 speakers from the N-R region, the second token of each realization of each of the nine monophthongal was selected. For each vowel token, the mean across the seven listeners was calculated for Height, Advancement, and Rounding. These mean values were calculated using the results from sub-experiment 1. In this sub-experiment, the data was presented to the listeners fully mixed.

Figure 9.2 shows the mean judgments of the 180 vowel tokens of the three dimensions Height, Advancement, and Rounding from the judgments made in sub-experiment 1. As mentioned earlier in section 8.3, all judgment values were transformed to a scale between 0 and 100. Two differences can be observed between Figure 9.2 and Figure 9.1 in the dispersion of the measurements. First, the perceptual data is concentrated mainly across the edges of the auditory spaces, while the acoustic measurements are scattered across the entire acoustic space. Second, the perceptual data are clustered more tightly per vowel category than the acoustic data.

Correlations

Three sets of correlation coefficients (Pearson's r) were calculated to obtain a preliminary idea of how the raw acoustic descriptions and the perceptual description of the vowel data set relate to each other. The first set was calculated for the four acoustic dimensions separately

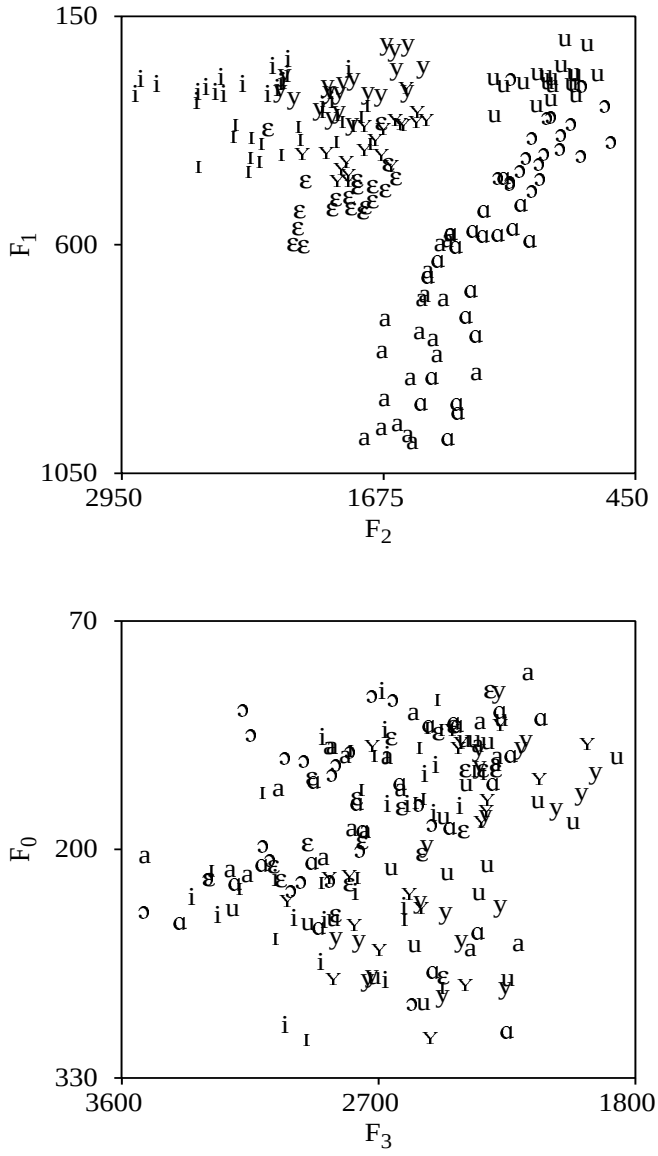

Figure 9.1: *Scatter plots of $F_2 \times F_1$ and $F_3 \times F_0$ for the 180 vowel tokens. Each phonetic symbol represents one vowel token. All values are in Hz.*

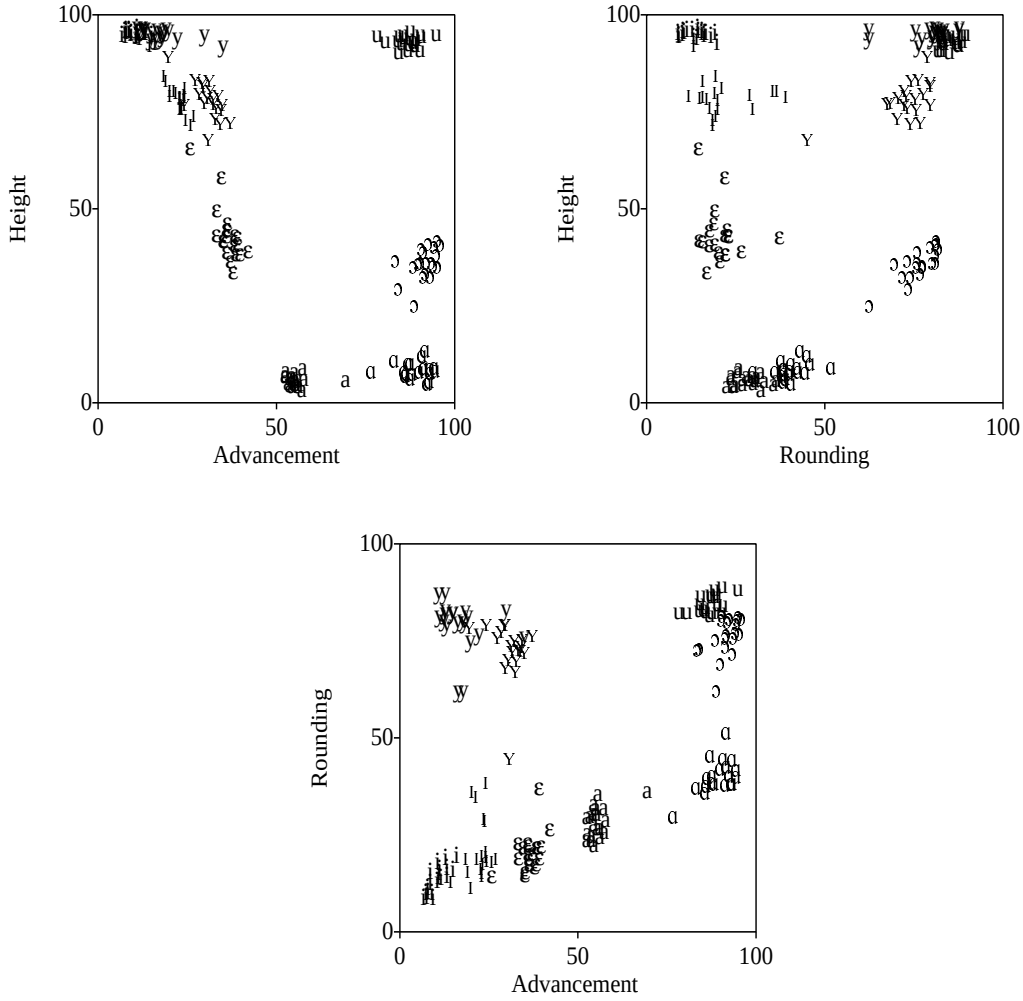

Figure 9.2: *Scatter plots of the Height, Advancement, and Rounding judgments for the 180 vowel tokens. Each phonetic symbol represents the mean across the judgments of seven listeners. The judgments are expressed on a scale from 0-100.*

(F_0, F_1, F_2, and F_3), and is listed in Table 9.1. Second, a set of correlation coefficients was calculated between the three perceptual dimensions (Height, Advancement, and Rounding), these are given in Table 9.2. The third set was calculated between the acoustic and perceptual dimensions (between F_0, F_1, F_2, and F_3, and Height, Advancement, and Rounding). The results are displayed in Table 9.3.

Table 9.1 shows that F_0 correlates positively with F_3 (0.29), and with F_2 (0.19). F_1 correlates positively with F_3 (0.24).

The results in Table 9.2 show a high negative correlation between Height and Advancement (-0.55). This means that high values for Height correspond to low Advancement values, and vice versa. In addition, a positive correlation was found between Advancement and Rounding (0.40), indicating that vowel tokens with high Advancement scores yield high Rounding scores. The correlation between Height and Rounding was significant as well (0.27), indicating a positive correlation between Height and Rounding (high Height values correspond to high Rounding values).

Table 9.1: *Correlation coefficients (Pearson's r) for the acoustic dimensions (F_0, F_1, F_2, and F_3). Based on the mean values for the 180 vowel tokens. * p<0.013, which is p<0.05 after Bonferroni correction for the number of correlations tested.*

r	F_0	F_1	F_2	F_3
F_0	-	-0.02	0.19^*	0.29^*
F_1		-	-0.09	0.24^*
F_2			-	0.18
F_3				-

Table 9.2: *Correlation coefficients (Pearson's r) for the perceptual variables (Height, Advancement, and Rounding). Based on the mean values for the 180 vowel tokens. * p<0.017, which is p<0.05 after Bonferroni correction for the number of correlations tested.*

r	Height	Advancement	Rounding
Height	-	-0.55^*	0.27^*
Advancement		-	0.40^*
Rounding			-

Table 9.3[65] shows that F_1 correlates negatively with Height (-0.85), indicating that high values for Height correspond to low values for F_1 and vice versa. Height further correlates

[65] A similar series of correlations was carried out for log-transformed acoustic data (log-transformed) and the

positively with F_2 (0.35), as well as with F_0 (0.21), indicating that low values for Height were found to correspond to low F_2 and low F_0 values. F_3 correlates negatively with Height (-0.19).

Table 9.3: *Correlation coefficients (Pearson's r) between the perceptual (Height, Advancement, and Rounding) and acoustic variables (F_0, F_1, F_2, and F_3). Based on the mean values for the 180 vowel tokens. * $p<0.004$, which is $p<0.05$ after Bonferroni correction for the number of correlations tested.*

r	F_0	F_1	F_2	F_3
Height	0.21^*	-0.85^*	0.35^*	-0.19^*
Advancement	-0.09	0.31^*	-0.89^*	0.08
Rounding	0.07	-0.36^*	-0.63^*	-0.25^*

For Advancement, the correlation coefficients are largest for F_2 (-0.89). The minus sign indicates that high values for Advancement (i.e., to the right in the vowel quadrilateral, indicating back articulation) were found to correspond to low values for F_2, and vice versa. It was further found that F_1 (0.31) correlates positively with Advancement (low values for F_1 correspond to higher values for Advancement and vice versa).

For Rounding, three significant correlations were found. The first, and largest, was found for F_2, -0.63, indicating that high Rounding values correspond to low values for F_2. F_1 correlates negatively with Rounding (-0.36), indicating that high Rounding values correspond to low values for F_1. F_3 correlates negatively with Rounding (-0.25), indicating that high Rounding values correspond to low F_3 values. In sum, the relations between the three perceptual variables and the four acoustic variables appear to be relatively straightforward, Height's most important correlate is $-F_1$, the most important correlate for Advancement is $-F_2$, while Rounding's most important correlate is $-F_2$.

9.3 Modeling perceived phonemic variation

9.3.1 Comparison of categorization performance

This section aims to establish to what extent the variation in the articulatory perceptual data (Height, Advancement, and Rounding) and in the transformed acoustic data (D_0, D_1, D_2, and D_3) can be used to categorize the vowel tokens into the corresponding intended vowel categories. To this end, 13 linear discriminant analyses (LDAs) were carried out on the 180

(linear) perceptual data, because it was suspected that the relationship between the perceptual data and the acoustic data was not completely linear. The correlation coefficient for Height for log-transformed F_1 was slightly higher (it increased from -0.85 to -0.88), and that the correlation coefficient for Rounding for F_3 was slightly higher (from -0.25 to -0.27), but the other correlations were unchanged as compared with Table 9.3.

vowel tokens. In the first LDA, the values of Height, Advancement, and Rounding for each vowel token were entered as predictors and the nine vowel categories served as the dependent variable. For the 12 remaining LDAs, the transformed values of D_0 through D_3 were entered as predictors, and the nine vowel categories were the dependent variable. For each of these LDAs, the percentage of correctly vowel tokens classified (into the corresponding vowel category), and the corresponding number of misclassified vowel tokens were calculated.

Table 9.4: *Percent correctly classified vowel tokens plus the absolute number of misclassified vowel tokens for the perceptual data (Height, Advancement, and Rounding), the raw acoustic data (F_0, F_1, F_2, and F_3 in* HZ*), and the acoustic data transformed according to the 11 normalization procedures. Height is referred to a H, Advancement as A, and Rounding as R. The number of vowel tokens is 180.*

Procedure	% Correct	# Misclassified
Perceptual (H, A, R)	98	4
HZ	79	37
LOG	83	27
BARK	81	31
ERB	83	30
MEL	82	31
SYRDAL & GOPAL	73	40
LOBANOV	95	9
GERSTMAN	88	22
CLIH$_{i4}$	95	10
CLIH$_{s4}$	86	24
NORDSTRÖM & LINDBLOM	82	33
MILLER	79	38

Table 9.4 shows the results for the LDAs. It can be seen that the perceptual variables Height, Advancement, and Rounding have very high scores; only four vowel tokens were put into another vowel category than the intended vowel category (/ɪ/ once as /ɛ/, /ɪ/ once as /i/, /ʏ/ once as /y/, and /y/ once as /ʏ/).

Overall, a similar patter in the performance of the procedures was found here (with the 180 vowels from the N-R region) as was found when the data from all eight regions was used, as was the case in Chapter 7. The vowel-extrinsic/formant-intrinsic procedures show the highest scores: LOBANOV and CLIH$_{i4}$ obtain equally high scores (95% correctly

classified) and both procedures obtain a roughly equal corresponding number of misclassifications (9 for LOBANOV and 10 for CLIH$_{i4}$); GERSTMAN (88% correct and 22 misclassifications) and CLIH$_{i4}$ (86% correct and 24 misclassifications) performed slightly poorer. The vowel-intrinsic/formant-intrinsic procedures performed second best: ERB (83% correct and 30 misclassifications) and LOG (83% and 27 confusions) performed best, followed by MEL (82% correct and 31 misclassifications) and, finally, BARK 81% correct and 31 misclassifications). The vowel-extrinsic/formant-extrinsic procedures performed slightly poorer than the vowel-intrinsic/formant-intrinsic procedures, MILLER performed rather poorly (79% and 38 misclassifications), while NORDSTRÖM & LINDBLOM's performance was somewhat better (82% and 33 misclassifications). The vowel-intrinsic/formant-extrinsic procedure, SYRDAL & GOPAL performed poorest of all procedures (73% and 40 misclassifications), even poorer than HZ.

LOBANOV and CLIH$_{i4}$ produced data in which the phonemic variation was strengthened. This can be assumed, because the overlap between vowel categories was reduced considerably. The other procedures represent the phonemic variation less effectively; the overlap between vowel tokens was not reduced as much as was the case for LOBANOV and CLIH$_{i4}$.

Because so few vowel tokens were misclassified when entering the three perceptual variables and when entering normalized sets of acoustic variables into the analysis, it seemed rather trivial to analyze the precise nature of the direction of the misclassified vowel tokens. It suffices to conclude that the vowel-extrinsic/formant-intrinsic procedures represented phonemic variation in agreement with the perceptual data.

9.3.2 Models for baseline data

This section aims to establish how well the phonemic variation in the articulatory perceptual representations can be modeled using each of the 12 acoustic representations. The performance of the 12 articulatory representations (i.e., the 12 normalization procedures) in this task was tested by carrying out 12 linear regression analyses (LRA). Using the results from the analyses on the raw measurements of F_0 through F_3, models for raw acoustic data were drawn up. These models served as the baseline models to which all other models were compared.

The articulatory baseline models were obtained as follows. Three linear regression analyses (LRAs) were carried out. In these LRAs, the four acoustic variables (F_0, F_1, F_2, and F_3) were stepwise entered as predictor variables. In the first LRA, the 180 mean values for Height served as the criterion variable. In the second and third LRA, the mean values of Advancement and Rounding served as criterion variables. Only the significant ($p < 0.001$)[66] predictor variables were used in the models. The resulting values for R^2 for the models for Height, Advancement, and Rounding are displayed in Table 9.5. The predictor variables in the equations are βs, i.e., standardized regression coefficients. These parameters were set in

[66] Because so many analyses were ran, a low significance level was adhered to.

this fashion for all the analyses that are described in this chapter. All R^2 values in this chapter are multiplied by 100, so that percentages explained variance are obtained.

Table 9.5: $R^2 \times 100$ *for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the acoustic variables* F_0, F_1, F_2, *and* F_3 *as predictor variables.*

$R^2 \times 100$	Height	Advancement	Rounding
HZ	82	88	61

In Table 9.5, it can be seen that the values for R^2 are generally high (above 0.61). The percentage for Rounding (61%) is substantially lower than for Height (83%) and Advancement (88%). The corresponding regression models for Height, Advancement, and Rounding with their respective β coefficients for the predictor variables are displayed in the equations in (9.1).

$$\begin{array}{lcrrrr} \text{Height} & = & 0.18F_0 & -0.80F_1 & +0.26F_2 & -0.10F_3 \\ \text{Advancement} & = & & 0.18F_1 & -0.92F_2 & +0.18F_3 \\ \text{Rounding} & = & 0.22F_0 & -0.40F_1 & -0.69F_2 & \end{array} \tag{9.1}$$

Equation (9.1) shows first, that, for Height all four predictor variables are significant. Second, F_1 is the highest contributing predictor and F_2 has the second highest β coefficient (0.26). F_0 and F_2 contribute relatively little (0.18 and 0.10, respectively). For Advancement, it can be observed that three out of four predictor variables are significant; F_2 is the most relevant (0.92), F_1 and F_3 have equal β coefficients (0.18), while F_0 did not contribute significantly. For Rounding, three out of four predictor variables contributed significantly. F_2 is the most relevant predictor variable (0.69), followed by F_1 (0.40) and F_0 (0.22). F_3 did not contribute significantly to the model for Rounding.

In summary, the results for the models for HZ show a pattern that is, overall, similar to the pattern in the results found for the correlation coefficients in section 9.2. For Height and Advancement, it was found that the acoustic predictors that are traditionally associated with articulatory vowel height (F_1), and articulatory tongue advancement (F_2) were indeed most relevant for the models for Height and Advancement (see equation (9.1)). The variance for Height and Advancement can for a large part be predicted using the raw acoustic correlates F_1 and F_2. Most of the variance in the scores for Rounding, on the other hand, can be explained using F_2 and F_1.

9.3.3 Models for normalized data

The remaining 11 normalization procedures (12 minus HZ, the baseline procedure) were carried out on the raw acoustic data. This process was identical to the process described in Chapter 7, the only difference is that this time the procedures were applied to a subset of the database used in Chapter 7.

Because the procedures were applied to different sets of vowel data than was the case in Chapter 7, some of the scale factors that were used by the procedures were re-calculated. The scale factors used in the two vowel-extrinsic/formant-extrinsic procedures, NORDSTRÖM & LINDBLOM and MILLER were set as follows. The scale factor k in equation (2.13) was calculated as described in section (2.4.4) and set to 0.91 based on the values for the 20 speakers. MILLER scale factor k in equation (2.14) was set to 185.66 Hz, the geometric mean of μF_0 for the 10 female speakers (230.30 Hz) and for the 10 male speakers (149.68 Hz).

For each of the 11 procedures, three LRAs were carried out. In each of these LRAs, the four transformed acoustic variables (D_0, D_1, D_2, and D_3) were stepwise entered as predictor variables, with either the mean Height values, Advancement values, or the Rounding values as criterion. Again, only the significant ($p < 0.01$) predictor variables are used in the final models.

Vowel-intrinsic/formant-intrinsic procedures

Table 9.6 shows $R^2 \times 100$ from the LRAs on the scale transformations. It can be observed that the portions of explained variance ($R^2 \times 100$) do not differ considerably from the baseline. The highest portion explained variance for all three criterion variables is observed for ERB. Compared with the baseline, the values are 4% higher for Height, 3% higher for Advancement, while for Rounding a deterioration can be observed: 2% lower than the baseline. In addition, no great differences can be observed between the four vowel-intrinsic/formant-intrinsic procedures.

The regression models for Height, Advancement, and Rounding for the LOG-transformed data are displayed in the equations in (9.2).

$$\begin{array}{lcrrrr} \text{Height} & = & & -0.88D_1^{\text{L}} & +0.26D_2^{\text{L}} & -0.14D_3^{\text{L}} \\ \text{Advancement} & = & & 0.29D_1^{\text{L}} & -0.91D_2^{\text{L}} & +0.11D_3^{\text{L}} \\ \text{Rounding} & = & 0.21D_0^{\text{L}} & -0.33D_1^{\text{L}} & -0.64D_2^{\text{L}} & -0.16D_3^{\text{L}} \end{array} \tag{9.2}$$

The regression models for acoustic data transformed to BARK are displayed in the equations in (9.3).

Table 9.6: $R^2 \times 100$ *for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the acoustic variables* D_0, D_1, D_2, *and* D_3 *as predictor variables, normalized to* LOG, BARK, ERB, *and* MEL, *respectively.*

$R^2 \times 100$	Height	Advancement	Rounding
HZ	82	88	61
LOG	86	90	58
BARK	85	91	59
ERB	86	91	59
MEL	85	90	59

$$\begin{array}{lcccccc} \text{Height} & = & 0.17D_0^{\text{B}} & -0.84D_1^{\text{B}} & +0.25D_2^{\text{B}} & -0.065D_3^{\text{B}} \\ \text{Advancement} & = & & 0.27D_1^{\text{B}} & -0.91D_2^{\text{B}} & +0.12D_3^{\text{B}} \\ \text{Rounding} & = & 0.22D_0^{\text{B}} & -0.34D_1^{\text{B}} & -0.65D_2^{\text{B}} & -0.16D_3^{\text{B}} \end{array} \tag{9.3}$$

The regression models for acoustic data transformed to ERB are displayed in the equations in (9.4).

$$\begin{array}{lcccccc} \text{Height} & = & 0.14D_0^{\text{E}} & -0.88D_1^{\text{E}} & +0.25D_2^{\text{E}} & \\ \text{Advancement} & = & & 0.27D_1^{\text{E}} & -0.91D_2^{\text{E}} & +0.12D_3^{\text{E}} \\ \text{Rounding} & = & 0.21D_0^{\text{E}} & -0.34D_1^{\text{E}} & -0.65D_2^{\text{E}} & -0.15D_3^{\text{E}} \end{array} \tag{9.4}$$

The regression models for acoustic data transformed to MEL are displayed in the equations in (9.5).

$$\begin{array}{lcccccc} \text{Height} & = & 0.15D_0^{\text{M}} & -0.86D_1^{\text{M}} & +0.25D_2^{\text{M}} & \\ \text{Advancement} & = & & 0.25D_1^{\text{M}} & -0.91D_2^{\text{M}} & +0.14D_3^{\text{M}} \\ \text{Rounding} & = & 0.22D_0^{\text{M}} & -0.35D_1^{\text{M}} & -0.66D_2^{\text{M}} & -0.15D_3^{\text{M}} \end{array} \tag{9.5}$$

The following general observations can be made about these sets of regression models. First, the models show patterns similar to those found for the baseline data: Height's primary predictor is D_1, Advancement can be modeled primarily by D_2, and D_2 served as the most relevant predictor for Rounding. However, these models differ from the baseline models in that D_3 plays a (small) role in the explained variance for Rounding; for all four procedures D_3 was significant for Rounding, whereas F_3 was not.

Vowel-intrinsic/formant-extrinsic procedures

The resulting values of $R^2 \times 100$ from the LRAs on the data transformed following SYRDAL & GOPAL are displayed in Table 9.7. Here, it can be observed that the value of $R^2 \times 100$ for SYRDAL & GOPAL is higher than the baseline values for Height (2% higher than the raw data). However, the values show a deterioration compared to the baseline: 1% lower for Advancement and 15% lower for Rounding. The corresponding regression models for acoustic data transformed with SYRDAL & GOPAL are displayed in the equations in (9.6).

Table 9.7: *$R^2 \times 100$ for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the acoustic variables D_0, D_1, D_2, and D_3 as predictor variables, normalized following* SYRDAL & GOPAL.

$R^2 \times 100$	Height	Advancement	Rounding
HZ	82	88	61
SYRDAL & GOPAL	84	87	46

$$\begin{aligned} \text{Height} &= -0.85D_1^{\text{s\&g}} - 0.25D_2^{\text{s\&g}} \\ \text{Advancement} &= -0.22D_1^{\text{s\&g}} + 0.88D_2^{\text{s\&g}} \\ \text{Rounding} &= -0.44D_1^{\text{s\&g}} + 0.58D_2^{\text{s\&g}} \end{aligned} \tag{9.6}$$

In the equations in (9.6), it can be observed that both dimensions are included in the models for Height, Advancement, and Rounding. $D_1^{\text{s\&g}}$ was found to correspond primarily to Height, while $D_2^{\text{s\&g}}$ could be used to model Advancement. For Rounding, the difference between the β coefficients is smaller than for Height and Advancement, $D_1^{\text{s\&g}}$ shows a larger coefficient than $D_2^{\text{s\&g}}$.

Vowel-extrinsic/formant-intrinsic procedures

A total of nine LRAs was carried out to obtain the results for the three vowel-extrinsic/formant-intrinsic procedures. Table 9.8 shows the results for GERSTMAN, LOBANOV, and CLIH-$_{i4}$. In this table, it can be seen that the explained variance is considerably higher for the vowel-extrinsic/formant-intrinsic procedures than for the baseline. The highest portion explained variance for all three criterion variables can be observed for LOBANOV. Compared with Table 9.5, the values are 7% higher for Height, 3% higher for Advancement, and the largest improvement, 11% was obtained for Rounding.

The regression models for acoustic data transformed using LOBANOV are displayed in the equations in (9.7).

Table 9.8: R^2 *for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the acoustic variables* D_0, D_1, D_2, *and* D_3 *as predictor variables, normalized following* GERSTMAN, LOBANOV, *and* CLIH_{s4}.

$R^2 \times 100$	Height	Advancement	Rounding
HZ	82	88	61
GERSTMAN	88	90	70
LOBANOV	89	91	72
CLIH_{i4}	88	91	65

$$\begin{array}{lcrrr} \text{Height} & = & -0.82D_1^{\text{lobanov}} & +0.24D_2^{\text{lobanov}} & -0.14D_3^{\text{lobanov}} \\ \text{Advancement} & = & 0.17D_1^{\text{lobanov}} & -0.91D_2^{\text{lobanov}} & +0.12D_3^{\text{lobanov}} \\ \text{Rounding} & = & -0.45D_1^{\text{lobanov}} & -0.70D_2^{\text{lobanov}} & -0.28D_3^{\text{lobanov}} \end{array} \tag{9.7}$$

The regression models for acoustic data transformed following GERSTMAN are displayed in the equations in (9.8).

$$\begin{array}{lcrrr} \text{Height} & = & -0.84D_1^{\text{gerstman}} & +0.24D_2^{\text{gerstman}} & -0.16D_3^{\text{gerstman}} \\ \text{Advancement} & = & 0.17D_1^{\text{gerstman}} & -0.90D_2^{\text{gerstman}} & +0.11D_3^{\text{gerstman}} \\ \text{Rounding} & = & -0.44D_1^{\text{gerstman}} & -0.69D_2^{\text{gerstman}} & -0.28D_3^{\text{gerstman}} \end{array} \tag{9.8}$$

The regression models for acoustic data transformed following CLIH_{i4} are displayed in the equations in (9.9).

$$\begin{array}{lcrrr} \text{Height} & = & -0.88D_1^{\text{clihi4}} & +0.24D_2^{\text{clihi4}} & -0.09D_3^{\text{clihi4}} \\ \text{Advancement} & = & & -0.90D_2^{\text{clihi4}} & +0.28D_3^{\text{clihi4}} \\ \text{Rounding} & = & -0.36D_1^{\text{clihi4}} & -0.66D_2^{\text{clihi4}} & -0.29D_3^{\text{clihi4}} \end{array} \tag{9.9}$$

In these nine regression models, the same pattern can be observed in results for the baseline data and for the vowel-intrinsic/formant-intrinsic models: Height's primary predictor is D_1, while D_2 is the most important predictor for Advancement and for Rounding. A difference between the baseline model for Rounding and the models for the vowel-extrinsic/formant-intrinsic procedures for Rounding, is that D_3 contributes significantly for the vowel-extrinsic/formant-intrinsic models, however it was never the dominant cue. Finally, the role of D_0 appears to be nonexistent in the vowel-extrinsic/formant-intrinsic models.

Vowel-extrinsic/formant-extrinsic procedures

A total of nine LRAs was carried out to obtain the results for the three vowel-extrinsic/formant-intrinsic procedures. The results for MILLER, NORDSTRÖM & LINDBLOM, and CLIH$_{s4}$ are displayed in Table 9.9.

Table 9.9: $R^2 \times 100$ *for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the acoustic variables* D_0*,* D_1*,* D_2*, and* D_3 *as predictor variables, normalized following* MILLER, NORDSTRÖM & LINDBLOM, *and* CLIH$_{s4}$.

$R^2 \times 100$	Height	Advancement	Rounding
HZ	82	88	61
MILLER	85	91	46
NORDSTRÖM & LINDBLOM	84	89	62
CLIH$_{s4}$	87	91	60

Table 9.9 shows that the explained variance for Height and Rounding for the vowel-extrinsic/formant-extrinsic procedures is lower than for the vowel-extrinsic/ formant-intrinsic procedures, for Advancement, the performance is about equal. MILLER shows a very low value for Rounding (46%). Of the three procedures, CLIH$_{s4}$ has the highest scores for Height and Advancement, whereas for Rounding the value is slightly below the value for Rounding for NORDSTRÖM & LINDBLOM and HZ. Nevertheless, the performance for CLIH$_{s4}$ was considerably poorer than the performance of CLIH$_{i4}$.

The regression models for acoustic data transformed to MILLER are displayed in the equations in (9.10).

$$\begin{array}{lcccccc} \text{Height} & = & & -0.58D_1^{\text{miller}} & +0.41D_2^{\text{miller}} & & \\ \text{Advancement} & = & -0.35D_0^{\text{miller}} & & -0.88D_2^{\text{miller}} & +0.31D_3^{\text{miller}} & \\ \text{Rounding} & = & & -0.43D_1^{\text{miller}} & +0.56D_2^{\text{miller}} & & \end{array} \tag{9.10}$$

In equation (9.10), it can be observed that D_1^{miller} is the most relevant contributor for Height (-0.58), followed by D_2^{miller} (0.41). For Rounding, only D_1^{miller} and D_2^{miller} contributes significantly (-0.43 and -0.88, respectively). For Advancement, all three of MILLER's dimensions contribute significantly (-0.35, -0.88, and 0.31, respectively). For Rounding, again only D_1^{miller}, D_2^{miller} contributes significantly (-0.43 and 0.56 respectively).

The regression models for acoustic data transformed to NORDSTRÖM & LINDBLOM are displayed in formula (9.11).

$$\begin{array}{lcccccc} \text{Height} & = & 0.11D_0^{\text{n\&l}} & -0.80D_1^{\text{n\&l}} & +0.25D_2^{\text{n\&l}} & -0.10D_3^{\text{n\&l}} & \\ \text{Advancement} & = & & 0.18D_1^{\text{n\&l}} & -0.90D_2^{\text{n\&l}} & +0.14D_3^{\text{n\&l}} & \\ \text{Rounding} & = & & -0.43D_1^{\text{n\&l}} & -0.68D_2^{\text{n\&l}} & -0.14D_3^{\text{n\&l}} & \end{array} \tag{9.11}$$

For NORDSTRÖM & LINDBLOM's results, the pattern is as follows. For Height $D_1^{\text{n\&l}}$ is the dominant predictor (-0.80), followed by $D_2^{\text{n\&l}}$ (0.25), $D_0^{\text{n\&l}}$ (0.11), and finally $D_3^{\text{n\&l}}$ (0.10). For Advancement, three out of four predictors were significant, where $D_2^{\text{n\&l}}$ is dominant (-0.90), followed by $D_1^{\text{n\&l}}$ (0.18), and $D_3^{\text{n\&l}}$ (0.14). For Rounding, three predictors were contributed significantly, $D_2^{\text{n\&l}}$ is dominant (-0.68), followed by $D_1^{\text{n\&l}}$ (-0.43), $D_3^{\text{n\&l}}$ (0.14).

The regression models for acoustic data transformed with CLIH_{s4} are displayed in the equations in (9.12).

$$\begin{array}{lcccccc} \text{Height} & = & & -0.88D_1^{\text{clihs4}} & +0.25D_2^{\text{clihs4}} & -0.07D_3^{\text{clihs4}} & \\ \text{Advancement} & = & 0.08D_0^{\text{clihs4}} & 0.27D_1^{\text{clihs4}} & -0.90D_2^{\text{clihs4}} & & \\ \text{Rounding} & = & -0.18D_0^{\text{clihs4}} & -0.41D_1^{\text{clihs4}} & -0.64D_2^{\text{clihs4}} & -0.25D_3^{\text{clihs4}} & \end{array} \tag{9.12}$$

For CLIH_{s4}, Height's dominant predictor is D_1^{clihs4} (-0.88), followed by D_2^{clihs4} (0.25). Advancement and Rounding's dominant predictor is D_2^{clihs4} (0.27 and 0.64, respectively). Compared with the models for CLIH_{i4} (9.9), the models show the following differences. D_0^{clihi4} is not significant for Advancement or Rounding for CLIH_{i4}, and D_0^{clihs4} contributes significantly to the model for CLIH_{s4} for Advancement and Rounding. In addition, for Height, D_1^{clihi4} does not contribute significantly to the model for CLIH_{i4}, while D_1^{clihs4} contributes significantly to the model for CLIH_{s4}. Finally, for Advancement, D_3^{clihi4} contributes significantly to the model for CLIH_{i4}, while D_3^{clihs4} does not contribute significantly to the model for CLIH_{s4}.

9.3.4 Summary

Table 9.10 shows an overview of the values of $R^2 \times 100$ for all 12 procedures, for the three criterion variables Height, Advancement, and Rounding. These results were presented earlier throughout section 9.3.3. Overall, these results show that the perceived phonemic variation in the three criterion variables could be modeled satisfactorily; the percentages explained variance are between 82% and 89% for Height, between 87% and 91% for Advancement, and between 46% and 72% for Rounding. It can further be observed that applying the normalization procedures to the raw acoustic data generally improved the fit between the acoustic data and the articulatory perceptual data; nine out of 11 procedures performed better than the baseline (HZ). LOBANOV's procedure shows the highest scores. SYRDAL & GOPAL's procedure performed poorer than the baseline, although only for Rounding and Advancement. MILLER performed poorer than the baseline as well, but only for Rounding.

Table 9.10: $R^2 \times 100\%$ *for all 12 procedures for the linear regression analyses for the three perceptual criterion variables Height, Advancement, and Rounding with the transformed acoustic variables* F_0*,* F_1*,* F_2*, and* F_3 *as predictor variables.* S & G *refers to* SYRDAL & GOPAL *and* N & L *refers to* NORDSTRÖM & LINDBLOM.

$R^2 \times 100$	Height	Advancement	Rounding	Mean	Ranking
HZ	82	88	61	77	10
LOG	86	90	58	78	7.5
BARK	85	91	59	78	7.5
MEL	85	90	59	78	7.5
ERB	86	91	59	79	4.5
S & G	84	87	46	72	12
GERSTMAN	88	90	70	83	2
LOBANOV	89	91	72	84	1
CLIH$_{i4}$	88	91	65	81	3
CLIH$_{s4}$	88	91	60	79	4.5
MILLER	85	91	46	74	11
N & L	84	89	62	78	7.5

When the scores for the different classes of normalization procedures are compared, it is clear that the vowel-extrinsic/formant-intrinsic procedures performed best: LOBANOV ranks first in Table 9.10, GERSTMAN ranks second, and CLIH$_{i4}$ ranks third. Furthermore, the vowel-intrinsic/formant-intrinsic procedures performed second best: LOG, BARK, MEL, and ERB all rank 6.5$^{\text{th}}$, followed by the vowel-extrinsic/formant-extrinsic procedures: CLIH$_{s4}$ ranks 4.5$^{\text{th}}$, NORDSTRÖM & LINDBLOM ranks 7.5$^{\text{th}}$, and MILLER ranks 11$^{\text{th}}$. The vowel-intrinsic/formant-extrinsic procedure performed poorest: SYRDAL & GOPAL ranks 12$^{\text{th}}$.[67]

9.4 Modeling sociolinguistic variation

This section describes the modeling of the perceived sociolinguistic variation. The 12 normalization procedures were evaluated on how well they model the Height, Advancement, and Rounding judgments within each vowel category. The analyses were carried out on the mean

[67] All analyses described in this section were repeated for the seven individual phonetically-trained listeners. The results for these analyses showed the same pattern, i.e., the highest scores for LOBANOV's procedure, followed by GERSTMAN and CLIH$_{i4}$.

values for the seven phonetically-trained listeners, as was done in the previous section. The sociolinguistic variation was modeled to be able to describe differences between realizations of vowel tokens belonging to the same vowel category (for instance the differences in the perceived tongue height of /ɪ/ between a male and a female speaker from the N-R region in the Netherlands).

9.4.1 Models for baseline data

The first step was to calculate the models for the raw data per vowel category. A series of linear regression analyses was carried out with F_0, F_1, F_2, and F_3 in Hz as the predictor variables, using Height, Advancement, and Rounding, respectively, as the criterion variables. Table 9.11 lists the values of $R^2 \times 100$ per vowel. It can be observed that overall the values for $R^2 \times 100$ are considerably lower than the percentages that were found when the analyses were carried out across vowels (see Table 9.5). In addition, some vowel categories show no significant predictors (/ɛ/, /ɔ/, and /y/), or only significant predictors for one of the three variables. Furthermore, the significant predictors do not show the same pattern as was found for the models of phonemic variation, which showed that F_1 was generally the most relevant predictor for Height and that F_2 was the dominant predictor for both Advancement and Rounding. Here, no such pattern can be observed; for instance, F_0 is the dominant predictor for Rounding and Advancement for /i/, F_1 is the most relevant predictor for Rounding for /u/. No such patterns were observed for the comparisons across vowels (i.e., for the phonemic variation).

Table 9.11: $R^2 \times 100\%$ *for the baseline data for Height, Advancement, and Rounding with the acoustic variables* F_0, F_1, F_2, *and* F_3 *as predictor variables per vowel category. Only values significantly different from zero at the p<0.01 level are given.*

$R^2 \times 100$	Height	Advancement	Rounding
/ɑ/	-	49 (F_2, F_3)	21 (F_2)
/a/	-	60 (F_2, F_0)	-
/ɛ/	-	-	-
/ɪ/	64 (F_0)	-	-
/i̇/	-	23 (F_0)	65 (F_0, F_1)
/ɔ/	-	-	-
/u/	-	62 (F_2, F_1)	35 (F_1)
/ʏ/	60 (F_1, F_0, F_2)	-	-
/y//	-	-	-

9.4.2 Models for normalized data

This section discusses the results for the sociolinguistic models (within vowel categories) for all normalization procedures. For each procedure, linear regression analyses were carried out for each vowel separately, using Height, Advancement, and Rounding as criterion variables and the transformed acoustic variables as predictor variables.

However, not all results are presented here. When the results of the normalization procedures were investigated, a pattern similar to the baseline data (Table 9.11) was observed: relatively low scores for R^2, and for some vowels no models could be created. Furthermore, the most important predictors for each vowel category were not the same as the ones found for the phonemic comparisons (across vowels); no systematic pattern could be observed in the relevant acoustic predictors per vowel. Instead, for each vowel, a different pattern was found (e.g., /ɑ/ showed different predictors than /ʏ/) as well as across normalization procedures (e.g., Height's dominant (and only) predictor variable for /ɪ/ is F_0 for NORDSTRÖM & LINDBLOM and F_1 for CLIH$_{s4}$).

An example is provided by showing the results for two normalization procedures that show patterns that were exemplary for those found across the results for all procedures. Tables 9.12 and 9.13 show the results for LOG, one of the vowel-intrinsic/formant-intrinsic procedures, and for CLIH$_{s4}$, one of the vowel-extrinsic/formant-extrinsic procedure, respectively[68].

Table 9.12: $R^2 \times 100\%$ *for the LRAs for Height, Advancement, and Rounding with the acoustic variables transformed to* LOG *as predictor variables per vowel category.*

$R^2 \times 100$	Height	Advancement	Rounding
/ɑ/	-	50 (D_3^{L}, D_2^{L})	-
/a/	-	64 (D_2^{L}, D_0^{L})	-
/ɛ/	-	-	-
/ɪ/	60 (D_0^{L})	57 (D_0^{L})	-
/i̇/	-		59 (D_0^{L}, D_1^{L})
/ɔ/	-	43 (D_2^{L}, D_1^{L})	-
/u/	-	0.61 (D_2^{L}, D_1^{L})	36 (D_1^{L})
/ʏ/	45 (D_1^{L}, D_0^{L})	-	-
/y/	-	-	-

[68] The results for CLIH$_{s4}$ are in fact the most interpretable of all normalization procedures; for the majority of the procedures, less significant predictor variables were found than was the case for CLIH$_{i4}$.

Table 9.13: $R^2 \times 100\%$ *for the LRAs for Height, Advancement, and Rounding with the acoustic variables transformed to* CLIH_{s4} *as predictor variables per vowel category.*

$R^2 \times 100$	Height	Advancement	Rounding
/ɑ/	-	50 (D_2^{clihs4}), D_3^{clihs4})	-
/a/	-	38 (F_2)	-
/ɛ/	33 (D_1^{clihs4})	38 (D_1^{clihs4})	-
/ɪ/	40 (D_{10}^{clihs4})	54 (D_0^{clihs4})	-
/i̇/	-	-	51 (D_1^{clihs4})
/ɔ/	46 (D_2^{clihs4})	56 (D_2^{clihs4}, D_1^{clihs4})	41 (D_2^{clihs4})
/u/	-	-	-
/ʏ/	41 (F_1)	-	-
/y/	-	-	-

No systematic significant differences were found between the 12 normalization procedures, grouped into the four classes of procedures. Still, some general observations across classes could be made. First, it appeared that none of the procedures produced data that allowed more vowel categories to be modeled through linear regression analysis than the raw baseline data. Second, there appeared to be an inverse relationship between the success of the procedure in modeling phonemic variation and the number of vowels that could be modeled; if the procedure modeled phonemic variation very well (e.g., LOBANOV) then the number of vowel categories that could be modeled was low (for LOBANOV, only two vowel categories could be modeled). If the procedure performed poorly on modeling phonemic category variation, then the number of categories that could be modeled was high (for SYRDAL & GOPAL significant predictors could be found for seven out of nine vowel categories), although the number of categories that could be modeled was never higher than those for the baseline.

To establish whether the data contained variation that could be modeled systematically, the perceptual data was pooled across the (intended) vowel categories and the LRAs were run again. This time, the data for the 180 vowel tokens for Height, Advancement, and Rounding were corrected for their respective vowel. This was done as follows, first the mean Height, Advancement, and Rounding per vowel category was calculated. Second, this mean value was subtracted from every vowel token in that vowel category. This process was repeated for all nine vowel categories. This way only the within-vowel (sociolinguistic) variation remained in the data. This was done to be able to use all data and to remove as much between-vowel (phonemic) variation as possible, so that only the sociolinguistic variance would remain. If listeners use the dimensions Height, Advancement, and Rounding systematically, they should have do so in the same way within and between vowels. The results for these LRAs are shown in Table 9.14.

Table 9.14: $R^2 \times 100\%$ *for Height, Advancement, and Rounding, corrected for each vowel category's mean value, with the transformed acoustic variables* D_0, D_1, D_2, *and* D_3 *as predictor variables. Only values significantly different from zero at p<0.001 were included. For each percent, the corresponding significant predictor variable(s) is/are listed between brackets.* S & G *refers to* SYRDAL & GOPAL *and* N & L *refers to* NORDSTRÖM & LINDBLOM.

$R^2 \times 100$	Height	Advancement	Rounding
HZ	6 (F_0)	17 (F_2, F_1)	3 (F_0)
LOG	6 (D_0)	18 (D_2, D_1)	3 (D_3)
BARK	6 (D_0)	15 (D_2)	3 (D_3)
MEL	6 (D_0)	18 (D_2, D_1)	3 (D_0)
ERB	6 (D_0)	18 (D_2, D_1)	3 (D_3)
S & G	-	16 (D_2, D_1)	-
GERSTMAN	-	14 (D_2, D_1)	-
LOBANOV	-	13 (D_2, D_1)	-
CLIH$_{i4}$	7 (D_0)	15 (D_2, D_1)	-
CLIH$_{s4}$	-	13 (D_2, D_1)	-
MILLER	-	18 (D_3, D_1)	-
N & L	7 (D_0)	17 (D_2, D_1)	-

In Table 9.14, it is easier to observe a pattern in the results than for the individual vowels. The results for Advancement are almost identical across normalization procedures. Overall, between 13% and 18% in the variation in the data for Advancement could be explained using D_2 and D_1. However, for Height and Rounding the results are less univocal, for Height only 6% can be predicted using D_0. D_0 and D_3 appear to be the only variables that can account for (very little of) the variance in Rounding. Furthermore, none of the procedures performed better than the baseline. The vowel-extrinsic/formant-intrinsic and the vowel-intrinsic-formant-extrinsic procedures do not appear to model Height or Rounding at all. Although Table 9.14 shows a more systematic pattern than was found in Table 9.11, the results are still lower than, and not as systematic as, the results for the models for the phonemic variation. However, overall, F_2 or D_2 and F_1 or D_1 were the most relevant predictor variables, a pattern consistent with that found for most of the phonemic models in section 9.3. Nevertheless, in Table 9.14 for Height (and in some cases for Rounding as well), it was generally found that F_0 or D_0 was the most relevant predictor. Generally, D_0

showed the smallest coefficients of all significant predictors (or it was not significant at all, as was found for the phonemic model for CLIH$_{i4}$). In sum, given the results presented in this section, it must be concluded that the perceived sociolinguistic variation could not be modeled satisfactorily[69].

9.5 Elaborating on the results

9.5.1 Phonemic modeling

The results in section 9.3.3 show that the vowel-extrinsic/formant-intrinsic procedures were the most suitable option for modeling perceived phonemic variation, followed by the vowel-intrinsic/formant-intrinsic, the vowel-extrinsic/ formant-extrinsic, and the vowel-intrinsic/formant-extrinsic procedures, respectively. Apparently, the variation in the perceptual variables Height, Advancement, and Rounding can be accounted for best when the measured formant frequencies are transformed to z-scores per speaker (LOBANOV), or scaled relative to the minimum and maximum formant frequencies for a speaker (GERSTMAN), or when the formant frequencies are log-transformed and subsequently expressed by their distance relative to the log-mean for the formant frequencies for a single speaker (CLIH$_{i4}$). The results of the phonemic comparisons thus showed that vowel-extrinsic measures such as the mean, standard deviation, and the maximum and minimum values per speaker can be useful for modeling perceptual articulatory judgments[70].

[69] As was done for the phonemic variation (across vowels), I attempted to model the sociolinguistic variation (within vowels) for individual listeners as well, to find out if one or more of them showed a different pattern than the one found for the mean values. However, only for a few listeners a model could be constructed at all. For these listeners, the same results were found as for the mean data, as was displayed in Table 9.14. Furthermore, in addition to the analyses described in this section for the data from sub-experiment 1, the analyses were repeated for the data for sub-experiment 3. The analyses were carried out for the data of the individual listeners as well as for the mean values across these seven listeners. The results are not reported here in detail; it suffices to say that the results appeared to be remarkably similar; the results for the within-vowel modeling did not improve when the data from sub-experiment 3 were used instead of those from sub-experiment 1.

[70] Because it was concluded in section 9.3.3 that LOBANOV performed best at modeling perceived phonemic variance, the model with the (unstandardized) b-coefficients including the intercept is presented. This model can be used to transform raw acoustic measurements in such a way that a large part of the perceived perceptual variation can be accounted for. More specifically, when raw (formant) data in Hz is entered into the equation presented here, 89% of the variance found in the Height judgments can be accounted for, 91% of the Advancement judgments, and 72% of the Rounding judgments.

$$
\begin{array}{lcrrrr}
\text{Height} & = & -31.05D_1^{\text{lobanov}} & +8.72D_2^{\text{lobanov}} & -5.67D_3^{\text{lobanov}} & +60 \\
\text{Advancement} & = & 5.55D_1^{\text{lobanov}} & -29.93D_2^{\text{lobanov}} & +4.03D_3^{\text{lobanov}} & +49 \\
\text{Rounding} & = & -13.50D_1^{\text{lobanov}} & -20.82D_2^{\text{lobanov}} & -8.26D_3^{\text{lobanov}} & +49
\end{array}
$$

Combinations of successful procedures

It may be possible that the models for the perceptual representation improve when a combination of successful procedures is used. The vowel-intrinsic/formant-intrinsic procedures LOG, BARK, MEL, and ERB performed second best, after the vowel-extrinsic/formant-intrinsic procedures. Of the scale transformations, ERB performed best; 86% for Height, 91 % for Advancement, and 59% for Rounding.[71]

Some of the normalization procedures consist of a scale transformation in combination with another transformation. For instance, CLIH$_{i4}$, CLIH$_{s4}$, and MILLER use a log-transformation, whereas SYRDAL & GOPAL includes a bark-transformation. The procedures that include a log-transformation were reasonably successful, while the only procedure that incorporates a bark-transformation performed poorly.

Because the class of the vowel-extrinsic/formant-intrinsic procedures performed best and the class of the vowel-intrinsic/formant-intrinsic procedures performed second best, it seemed plausible to try if a combination of these two types of procedures would lead to higher percentages explained variance in the perceptual variables. The two obvious candidates for such a combination would be LOBANOV and ERB, as LOBANOV performed best of all 12 procedures, and because ERB was the best option among the scale transformations. However, as more combinations of z-scores and scale transformations are possible and as there are relatively small differences between the scores for the raw scale transformations, all four scale transformation were investigated. A series of linear regression analyses (LRAs) was carried out on the raw acoustic variables F_0, F_1, F_3, and F_3 that were first converted to ERB and subsequently to LOBANOV (referred to as ERB$\times$LOBANOV). A comparable procedure was carried out for LOG, BARK, and MEL.

Table 9.15 shows the results for the LRAs, as well as the original results for HZ and for LOBANOV. It can be observed, that all four combinations of scale transformations $\times$ LOBANOV show slightly higher percentages than LOBANOV, but only for Height and Advancement; the percentages for Rounding are overall lower (lowest for LOG$\times$LOBANOV). The results for ERB$\times$LOBANOV are slightly higher than the other combinations for Height and Rounding. It appears that combining the best procedure of the vowel-extrinsic/formant-intrinsic procedures and the vowel-intrinsic/formant-intrinsic procedures leads to a (small) improvement of the fit of the models for Height and Advancement.

Interestingly, the results for LOG$\times$LOBANOV show the lowest results for Rounding (67%). Earlier, in Table 9.10, it was found that CLIH$_{i4}$, CLIH, and MILLER – procedures incorporating a log-transformation – showed considerably lower scores for Rounding than other procedures that show comparable scores for Height and Advancement. Apparently, using a procedure that incorporates a log-transformation results in relatively lower scores for Rounding. It is

[71] It should be noted that the percentage for ERB for Rounding was slightly lower than the baseline percentage for Rounding (for HZ this was 61%).

Table 9.15: $R^2 \times 100\%$ *for* HZ, LOBANOV, ERB×LOBANOV, LOG×LOBANOV, BARK×LOBANOV, *and* MEL × LOBANOV *for the LRAs for the three criterion variables Height, Advancement, and Rounding, with the transformed acoustic variables* F_0, F_1, F_2, *and* F_3 *as predictor variables.*

$R^2 \times 100$	Height	Advancement	Rounding	Mean
HZ (baseline)	82	88	61	77
LOBANOV(original)	89	91	72	84
ERB×LOBANOV	91	93	69	84
LOG×LOBANOV	91	92	67	83
BARK×LOBANOV	90	93	69	84
MEL×LOBANOV	90	93	69	84

unclear why this is the case. Nevertheless, because not one of the combinations of LOBANOV and a scale transformation show an improvement over all three perceptual variables, it must be concluded that original LOBANOV remains the best option.

Evaluation of formant-extrinsic transformations and F_3

The vowel-intrinsic/formant-extrinsic procedure, SYRDAL & GOPAL, performed poorest of all. Earlier findings suggest that this poor performance cannot be contributed to the fact that SYRDAL & GOPAL incorporates a bark-transformation. In Table 9.10, it can be seen that the scores for BARK are higher than those for the baseline. Perhaps the poor results for SYRDAL & GOPAL are, instead, due to the fact that this procedure uses formant-extrinsic transformations.

A series of LRAs was carried out to investigate whether the formant-extrinsic transformation in SYRDAL & GOPAL could (partially) explain the low scores for this procedure. In these LRAs, the performance of various (combinations of) predictors was evaluated. This procedure was also carried out on data in HZ: SYRDAL & GOPAL×HZ, to verify that the bark-transformation was not the cause of the poor performance of SYRDAL & GOPAL. Furthermore, a series of LRAs with two predictors was carried out: F_1 and F_2, F_1 and $F_3 - F_2$, and $F_1 - F_0$ and F_2. The new sets of predictors were chosen, to verify whether using a formant-extrinsic transformation leads to a deterioration. The combination F_1 and F_2 was compared with combinations that involved the use of formant-extrinsic transformations such as F_0 and F_3, as is the case for $F_1 - F_0$ and F_2, and F_1 and $F_3 - F_2$, and to evaluate the role of F_0 and F_3. Through a comparison of these three sets, it could be established whether using

$F_1 - F_0$ improves the performance compared with using F_1, and whether using $F_3 - F_2$ improves the performance compared with using F_2. Table 9.16 shows the results of these LRAs. In this table, the results for HZ (i.e., F_0, F_1, F_2, and F_3), and SYRDAL & GOPAL, from Table 9.10, are presented as well.

Table 9.16: $R^2 \times 100$ *resulting from the LRAs with various combinations of predictors, with Height, Advancement, and Rounding as criterion variables.*

$R^2 \times 100$	Height	Advancement	Rounding
HZ (baseline)	82	88	61
SYRDAL & GOPAL (original)	84	87	46
SYRDAL & GOPAL×HZ	82	77	41
F_1 and F_2	79	85	57
F_1 and $F_3 - F_2$	78	77	41
$F_1 - F_0$ and F_2	81	84	60

An analysis of the results displayed in Table 9.16 shows the relative effect of sets of predictor variables. First, given the difference between results for SYRDAL & GOPAL×BARK and SYRDAL & GOPAL × HZ for all three criterion variables, it can thus be confirmed that the poor performance of SYRDAL & GOPAL was not due to the use of a bark-transformation.

Second, it can be seen in Table 9.16 that when $F_3 - F_2$ was entered as a predictor instead of F_2, a considerable deterioration in the fit of the model for Advancement was found (77% for F_1 and $F_3 - F_2$ and 85% for F_1 and F_2). The same pattern was found in the results for Rounding (41% for F_1 and $F_3 - F_2$ and 57% for F_1 and F_2), and for Height, although this difference was smaller (78% for F_1 and $F_3 - F_2$ and 79% for F_1 and F_2). However, a different pattern in the results was found when $F_1 - F_0$ is used as a predictor instead of F_1. For Height, a small improvement is found when $F_1 - F_0$ was used (84% for $F_1 - F_0$ and F_2 and 85% for F_1 and F_2). For Rounding, the same pattern is found (57% for $F_1 - F_0$ and F_2 and 60% for F_1 and F_2). For Advancement, no improvement is found when $F_1 - F_0$ was used instead of F_1 (81% for $F_1 - F_0$ and F_2 and 79% for F_1 and F_2).

The results of Table 9.16 can be interpreted such that, the bark-transformation in SYRDAL & GOPAL, led to an improvement in the fit of the models for Height, Advancement, and Rounding. Second, the poor results for Syrdal & Gopal compared with the baseline are caused by the use of $F_3 - F_2$ as their second dimension. Apparently, using F_3 in a normalization procedure could lead to a considerable deterioration in the fit of the models for Advancement and Rounding. However, because the results for the combination of $F_1 - F_0$ and F_2 showed an improvement for Height compared with the combination F_1 and F_2, it cannot be concluded

that the poor results for SYRDAL & GOPAL are due to the fact that this procedure incorporates any formant-extrinsic transformation.

Nevertheless, the finding that using F_3 in a formant-extrinsic transformation lead to a deterioration of the fit of the perceptual models is useful for explaining the results for MILLER and NORDSTRÖM & LINDBLOM. The two formant-extrinsic procedures performed slightly poorer than the formant-intrinsic procedures, but they performed better than the baseline. As discussed before, MILLER's low ranking could partially be explained by the low scores for Rounding. Given the results Table 9.16, it can be assumed that these low scores are possibly caused by the use of F_3 in MILLER's third dimension.

Role of F_0

In section 9.3.2, it was found that F_0 (or D_0) did not contribute significantly to the models for Height for all three vowel-extrinsic/formant-intrinsic procedures, the procedures that were most successful at modeling phonemic variation in perceived Height. This seems in contradiction with the results found by Traunmüller (1981). He found that the value of F_0 could be used to predict perceived variation in the tongue height of vowels. On the other hand, F_0 did contribute significantly to the results for Height for HZ, BARK, MEL, ERB, and NORDSTRÖM & LINDBLOM, which corroborate the findings of Traunmüller. However, the fit of the models for Height for these procedures was considerably lower than for the vowel-extrinsic/formant-intrinsic procedures LOBANOV, GERSTMAN, and CLIH$_i$4.

It is possible, that the effects found for F_0 for HZ, BARK, MEL, ERB, and NORDSTRÖM & LINDBLOM were caused by the fact that F_0 contains considerable anatomical/physiological variation related to the speaker's sex, as was the case in Chapter 7. To investigate this, several combinations of predictors for LRAs were evaluated for the data transformed following LOBANOV, the procedure performed best at modeling perceived phonemic variation (this chapter) and best at reducing variation in the acoustic signal related to the speaker's sex (Chapter 7).

Table 9.17 shows the results for the LRAs for LOBANOV. The following combinations of predictors were used. D_0 was entered as the sole predictor and D_1 was entered as the sole predictor. The results for these two analyses are compared with the results for a model with two predictors: D_0 and D_1. Furthermore, it was evaluated whether using a formant-extrinsic transformation for LOBANOV led to a deterioration in the results for $D_1 - D_0$.
In Table 9.17, it can be seen that using a formant-extrinsic transformation again leads to a deterioration in the fit of the model. When D_1^{lobanov} is entered as the sole predictor, a value of 82% is found, whereas for $D_1^{\text{lobanov}} - D_0^{\text{lobanov}}$, a value of 67% is found[72]. Apparently the effect

[72] I verified whether deterioration would be found when $D_3^{\text{lobanov}} - D_2^{\text{lobanov}}$ was entered as the sole predictor for Advancement. I found that this was indeed the case, R^2 was 0.56. When D_2^{lobanov} was entered as the sole predictor for Advancement, R^2 was 0.86.

Table 9.17: $R^2 \times 100\%$ *for combinations of* F_0 *and* F_1 *transformed following* LOBANOV *for the linear regression analyses for Height.*

$R^2 \times 100$	Height
LOBANOV	89
D_0^{lobanov}	24
D_1^{lobanov}	82
D_0^{lobanov} and D_1^{lobanov}	83
$D_1^{\text{lobanov}} - D_0^{\text{lobanov}}$	67

found for $D_1 - D_0$ found in Table 9.16 disappears through the LOBANOV transformation, which was also found to eliminate the sex-related anatomical/physiological variation.

Furthermore, the results in Table 9.17 show that D_0^{lobanov} accounts for 24% of the variation found in Height when D_0^{lobanov} is the single predictor. The dominant predictor for Height, D_1^{lobanov}, explains 82%. However, when the two predictors D_1^{lobanov} and D_0^{lobanov} are combined, a percentage of 83% is found, a difference of only 1%. This is surprisingly low, given the relatively high percentage for D_0^{lobanov} alone. The small difference indicates that D_0^{lobanov} and D_1^{lobanov} show a great deal of correlation.

However, the correlation coefficients for the raw data displayed in Table 9.1 show that no (significant) correlation between F_0 and F_1 exists. This result raises the question whether the transformation to z-scores caused F_0 and F_1 to correlate. To investigate this question, Pearson's r was calculated for the four acoustic variables transformed to z-scores. Table 9.18 shows the results.

Table 9.18: *Pearson's* r*, calculated between the four acoustic dimensions normalized using* LOBANOV. ** p<0.05, **p<0.01.*

r	D_0^{lobanov}	D_1^{lobanov}	D_2^{lobanov}	D_3^{lobanov}
D_0^{lobanov}	-	-0.46**	0.06	-0.25**
D_1^{lobanov}		-	-0.16*	0.15*
D_2^{lobanov}			-	0.04
D_3^{lobanov}				-

Table 9.18 shows that transforming the raw data using LOBANOV causes D_0^{lobanov} to correlate with D_1^{lobanov} (-0.46). This result can be explained by the fact that F_0 is known to vary considerably across speakers (especially between female and male speakers), which intro-

duces noise in the raw data, hence the low correlation for HZ. It is not unreasonable to think that, for F_0, this variation completely overshadows the relationship that exists between the fundamental frequency and the first formant.

The relationship between the first formant and the fundamental frequency became evident only after the data is normalized. When all sex-related anatomical/physiological variation is eliminated from the normalized vowel data, it seems plausible that the remaining variation in F_0 is more phonemic in nature; the remaining F_0-variation is probably vowel-dependent (i.e., vowel intrinsic pitch).

To summarize, the results for phonemic modeling display a pattern in the performance of the 12 procedure similar to the one reported in Chapter 7. The vowel-extrinsic/formant-intrinsic procedures performed best, followed by the vowel-intrinsic/formant-intrinsic and the vowel-extrinsic/formant-extrinsic procedures, the performance of the vowel-intrinsic/formant-extrinsic procedures was very poor. In addition, the results found in this chapter and in Chapter 7 mirror results reported in the studies on vowel normalization procedures by Syrdal (1984) and Deterding (1990); again Lobanov's (1971) and Nearey's (1978)'s CLIH$_{i4}$ procedures performed best, and there was not much difference in the performance of the scale transformations (LOG, BARK, ERB and MEL). It was further argued that the poor performance of SYRDAL & GOPAL and MILLER was partially caused by the fact that these procedures incorporate F_3 in their second dimension. It was furthermore found in the present chapter that using a formant-extrinsic transformation leads to a deterioration of the fit of the models for Height and Advancement. Finally, in contrast to results reported by Traunmüller (1981), I found that F_0 did not contribute to the model for phonemic variation in Height when the anatomical/physiological variation related to the speaker's sex was eliminated.

9.5.2 Sociolinguistic modeling

In section 9.4.2, I concluded that it was not possible to satisfactorily model perceived sociolinguistic variation. Only a low percentage of the variation found for Advancement could be accounted for systematically. A plausible explanation for this finding is that there was not enough systematic sociolinguistic variation present[73].

It seems justified to assume that the acoustic data did not contain enough systematic sociolinguistic variation, despite the finding reported in section 8.5 that the sociolinguistic variation was judged reliably for a subset of three vowels (although this effect was relatively small; the mean value for Cochran's α across all three perceptual variables was 0.46 for /ɛ/, 0.477 for /ɔ/, 0.423 for /ʏ/, 0.423 for /a/, 0.299 for /ɑ/, and 0.199 /u/ see Table 8.16). Given the results for these three vowels, it was concluded in Chapter 8 that the perceptual data of these three vowels contains systematic sociolinguistic variation and that the listeners were able to systematically judge sociolinguistic variation. It could be expected that it would be easier

[73] An alternative explanation for the results of sociolinguistic modeling could be that the behavior of the listeners could not be modeled using a simple linear function as estimated using linear discriminant analysis.

to fit the models for Height, Advancement, and Rounding using the acoustic representations for these three vowels than for the other six vowels. Nevertheless, in section 9.4.2, it was found that no complete model (i.e., for Height, Advancement, and Rounding) could be fit for any of the vowels. It must therefore be concluded that these three vowels did not contain enough variation that could be reliably judged and subsequently modeled using the acoustic representations; the values of Cochran's α for /ɛ/, /ɔ/, /ʏ/, /a/, /ɑ/, and /u/ were too low.

Three additional factors can be put forward that may have attributed to the low reliability scores for the individual vowels, in addition to the aforementioned low sociolinguistic variability displayed in the vowel tokens. First, I concluded in Chapter 8 that the listeners appeared to be heavily influenced by the response vowel category. It seems plausible that the listeners used a vowel-specific reference point. It can be hypothesized that they judged a vowel token's Height and Advancement by determining the vowel category for that vowel token, and subsequently by placing that vowel token in the (designated) area in the quadrilateral for that vowel category. They placed the judgment closer or less close to their reference point for the specific response vowel category. In addition to being influenced primarily by the response vowel category, the listeners apparently chose not to use the entire area possible for their Height and Advancement judgments. A possible explanation for this behavior is that they judged the differences in the vowel tokens relative to the entire range of pronunciation variation possible. This means that they judged the vowel tokens from speakers from the N-R region, while somehow keeping in mind the entire variation spectrum that is possible for Dutch. In other words, the listeners judged the relatively subtle differences as if they were going to be presented with more variation during the experiment. However, it is not possible to test this hypothesis using the results of the experiment presented in the present research, it would take another experiment in which the N-R data was also presented as vowel tokens that display more sociolinguistic variation. Such a judgment strategy may have caused lower variability in the judgments.

A second factor that may have contributed to the low reliability scores is that the listeners did not receive information about their previous judgments in the experiment. Perhaps the listeners were unsure about the exact location of their reference point (for each response vowel) across experimental trials. This insecurity possibly introduced noise in the judgments and caused lower scores for Cochran's α[74].

A third factor that may have affected the reliability scores is that the judgment task was

[74]It should be noted that the explanation for the low reliability scores refers exclusively to the judgments of Height and Advancement. For Rounding, the judgment behavior can be argued to be different. First, when judging Rounding, the listeners seemed to be influenced less by the response vowel category than when judging Height or Advancement. This was found in Chapter 8, when possible differences between the standard deviations across the three experimental conditions were studied. It seemed that, when listeners were presented with a judgment area that did not allow them to reserve a specific area for a specific vowel (as was the case for Rounding), then their judgments display less phonemic variation. This hypothesis can explain why it was more difficult to fit models for Rounding; all models for Rounding showed a lower percentage of explained variance than those for Height and Advancement.

intensive and perceived as fatiguing by the majority of the listeners. This may have led to noise in the judgments which in turn may have decreased the reliability scores.

In sum, the results for the Height and Advancement judgments can be interpreted as that the listeners preferred to maximize phonemic variation, at the cost of sociolinguistic variation in their judgments. In addition, while it was concluded in Chapter 8 that the listeners did reliably judge the sociolinguistic variation, the reliability scores were too low for the perceived sociolinguistic variation to be modeled through linear regression analysis. The present results do not allow me to conclude which one of the normalization procedures is most suitable for representing sociolinguistic differences in agreement with phonetically-trained listeners.

9.6 Conclusions

In this chapter, a comparison was carried out between the acoustic data described in Chapter 7 and the articulatory perceptual data described in Chapter 8. The 12 normalization procedures that are evaluated in the present research were compared on how well they produced data that could be used to model phonemic and sociolinguistic variation perceived by phonetically-trained listeners. The comparison was carried out for phonemic variation and for sociolinguistic variation. First, it was found that phonemic perceived articulatory differences could be modeled very effectively using (transformed) acoustic data. The vowel-extrinsic/formant-intrinsic procedures performed best, followed by the vowel-intrinsic/formant-intrinsic, vowel-extrinsic/formant-extrinsic, and the vowel-intrinsic/formant-extrinsic procedures, respectively. The success of the vowel-extrinsic/formant-intrinsic procedures was attributed to the fact that all three of these procedures make use of speaker-specific information, such as the mean across a speaker's vowels, to re-scale the acoustic data. The poor performance of the vowel-intrinsic/formant-extrinsic and (two of the) vowel-extrinsic/formant- extrinsic procedures was attributed to the fact that these procedures use F_3 to model their second dimensions. For the sociolinguistic variation, it was concluded that it was not possible to model perceived sociolinguistic variation in the perceptual data satisfactorily, because the stimulus vowel tokens did not reflect enough variation to be judged reliably enough by the listeners. Because no models could be fit, it was not possible to establish which one of the normalization procedures was most suitable for representing sociolinguistic differences in agreement with phonetically-trained listeners.

Chapter 10

General discussion and conclusions

10.1 Introduction

The research described in this thesis aimed to establish which procedure for acoustic vowel normalization meets the following criterion best. The procedure must preserve the phonemic variation and the sociolinguistic speaker-related variation, while minimizing the anatomical/physiological speaker-related variation in the transformed acoustic representation of vowel tokens. As was argued in Chapter 1, this criterion must be met so that the normalization procedure is considered suitable for use in sociolinguistics.

The present chapter discusses the performance of the normalization procedures, given the results of three sets of comparisons. The first are the acoustic comparisons (Chapter 7), the second are the perceptual comparisons, which involved a comparison of the performance of phonetically-trained experts (Chapter 8), and the third are the perceptual-acoustic comparisons of the acoustic normalization procedures to the experts' judgments (Chapter 9). This chapter aims further to provide a discussion of the results and to present the conclusions of the present research.

This chapter is set up as follows. Section 10.2 evaluates how well each of the 12 procedures performed at the three types of comparisons and discusses the core research question. Section 10.3 sums up the implications of the findings of the present research for sociolinguistics, while section 10.4 discusses the implications of the present research for phonetics. Finally, in section 10.5, a short discussion is provided of the limitations of the present research, suggestion for further research are given, and some concluding remarks are made.

10.2 Results

10.2.1 Results for individual procedures

This section evaluates how well each normalization procedure performed in the acoustic comparisons (Chapter 7) and the perceptual-acoustic comparisons (Chapter 9). The acoustic comparisons were carried out using data from all 160 speakers in the sociolinguistically balanced data set (Chapter 5). The perceptual-acoustic comparisons were carried out using data from 20 speakers from the N-R region, a subset of the sociolinguistically balanced data set. To summarize, a normalization procedure is considered to be the most successful when it performs best at the three tasks of the acoustic comparisons (preserve phonemic variation, preserve sociolinguistic variation, minimize anatomical/physiological variation) as well as at representing perceived phonemic variation, tested in the perceptual-acoustic comparisons.

Table 10.1 summarizes per procedure the results presented previously in Tables 7.5 and 9.10 for each individual procedures, as grouped into the four classes of procedures: the vowel-intrinsic/formant-intrinsic, vowel-intrinsic/formant-extrinsic, vowel-extrinsic/formant-intrinsic, and finally vowel-extrinsic/formant-extrinsic procedures.

HZ was the baseline procedure. In the acoustic comparisons, five procedures performed better than HZ at preserving phonemic variation in the acoustic data (F_0, F_1, F_2, and F_3), two performed poorer, and four procedures performed equally well. In addition, a considerable amount of anatomical/physiological variation was present in the data in HZ; nearly all other procedures (except for the four scale transformations, which performed as poorly as HZ) reduced more anatomical/physiological variation. Sociolinguistic variation was preserved in the HZ data, as well as in the data transformed following LOG, BARK, MEL, ERB, CLIH$_{s4}$, NORDSTRÖM & LINDBLOM, and MILLER. The three vowel-extrinsic/formant-intrinsic procedures, LOBANOV, GERSTMAN, and CLIH$_{i4}$ reduced some of the sociolinguistic variation. For the perceptual-acoustic comparisons, nine of the 11 procedures modeled the phonemic variation in the perceptual articulatory judgments better than HZ, while two procedures performed poorer.

LOG, BARK, MEL, and ERB performed very similarly. For the acoustic comparisons, it was concluded that applying these four procedures to the raw acoustic variables improved the preservation of phonemic variation compared with HZ. Second, data that was transformed following LOG, BARK, MEL, and ERB contained an equal amount of anatomical/physiological variation as in the raw data. The same pattern was found for the preservation of sociolinguistic variation; only a small deterioration in performance was found between the four procedures and HZ. For the perceptual-acoustic comparisons, all four scale transformations modeled the phonemic variation in the articulatory variables (Height, Advancement, and Rounding) slightly better than HZ. Of these four transformations, ERB performed best.

SYRDAL & GOPAL was the only vowel-intrinsic/formant-intrinsic procedure evaluated. Regarding the acoustic comparisons, SYRDAL & GOPAL performed poorest of all procedures

Table 10.1: *Relative performance of each normalization procedure. '0': the procedure did not perform better or poorer than* HZ *(the baseline). '+': the procedure performed better than* HZ*. '++': considerably better than* HZ*. '−': poorer than* HZ*. '−−': considerably poorer than* HZ*. The acoustic comparisons were carried out on the acoustic data of all 160 speakers, and the perceptual-acoustic comparisons used data from 20 speakers. 'Vow': vowel, 'form': 'formant', 'in': intrinsic, and 'ex': extrinsic*

Comparisons		Acoustic			Perceptual-acoustic
Class	Procedure	Preserve phonemic	Reduce anatomical/physiological	Preserve sociolinguistic	Meet perceptual benchmark
Vow-in/ form-in	LOG	0	0	0	+
	BARK	0	0	0	+
	MEL	0	0	0	+
	ERB	0	0	0	+
Vow-in/ form-ex	SYRDAL & GOPAL	−−	−−	−−	−−
Vow-ex/ form-in	LOBANOV	++	++	−	++
	GERSTMAN	++	++	−	++
	CLIH_{i4}	++	++	−	++
Vow-ex/ form-in	CLIH_{s4}	+	+	0	+
	NORDSTRÖM & LINDBLOM	+	+	0	+
	MILLER	−−	−	0	−

at preserving phonemic variation. SYRDAL & GOPAL reduced least of the anatomical/physiological variation. It thus reduced the anatomical/physiological variation without improving the preservation of phonemic variation in the transformed acoustic signal. This is remarkable, as for the majority of the procedures a reduction in the anatomical/physiological variation was accompanied by an improved representation of phonemic variation. Furthermore, SYRDAL & GOPAL performed poorest of all procedures at the task of preserving sociolinguistic variation. Regarding the perceptual-acoustic comparisons, SYRDAL & GOPAL performed poorest of all 12 procedures. After closer inspection of the properties of SYRDAL & GOPAL, it was concluded that the poor performance was probably due to the fact that SYRDAL & GOPAL uses F_3 in its second dimension.

LOBANOV performed best of all 12 procedures at representing phonemic variation in the transformed acoustic vowel data. This procedure effectively reduced the anatomical/physi-

ological variation in the vowel data. However, LOBANOV reduced some (very little) of the sociolinguistic variation, although this reduction was smaller than was found for SYRDAL & GOPAL and for GERSTMAN. For the perceptual-acoustic comparisons, LOBANOV produced acoustic data that could be used to model the variation in the three perceptual variables more effectively than all other procedures.

For GERSTMAN, the results showed that only two procedures, LOBANOV and CLIH$_{i4}$, represented phonemic variation in the transformed acoustic vowel data better and that nine performed worse. Furthermore, the results show that GERSTMAN effectively reduced the anatomical/physiological variation in the acoustic vowel data, only two procedures, LOBANOV and CLIH$_{i4}$, performed better, all other procedures performed worse than GERSTMAN. GERSTMAN further reduced some of the sociolinguistic variation, although this reduction is smaller than the reduction found for SYRDAL & GOPAL. For the perceptual-acoustic comparisons, GERSTMAN performed only slightly poorer than LOBANOV at modeling the perceived phonemic variation in the three perceptual variables.

CLIH$_{i4}$ performed second best of all 12 procedures at preserving phonemic variation in the transformed acoustic vowel data, only LOBANOV performed better. CLIH$_{i4}$ effectively minimized the anatomical/physiological variation. CLIH$_{i4}$ also reduced some of the sociolinguistic variation, although SYRDAL & GOPAL, LOBANOV, and GERSTMAN showed larger reductions. The perceptual-acoustic comparisons show that CLIH$_{i4}$ modeled the perceived phonemic variation in the three perceptual variables better than nine other procedures; only LOBANOV and GERSTMAN performed better.

The results for the acoustic comparisons show that CLIH$_{s4}$ performed considerably better than the baseline at preserving phonemic variation in the transformed acoustic vowel data; only the three vowel-extrinsic/formant-intrinsic procedures, LOBANOV, CLIH$_{i4}$, and GERSTMAN, performed better. However, CLIH$_{s4}$ hardly reduced the anatomical/physiological variation in the vowel data; it performed this task only slightly better than the baseline. This result deviates from the general finding that a reduction of anatomical/physiological variation leads to an improved phonemic representation. Moreover, CLIH$_{s4}$ preserved the sociolinguistic variation in the transformed vowel data relatively well, only in the data in HZ more variation was preserved. In addition, the perceptual-acoustic comparisons reveal that CLIH$_{s4}$ was relatively successful at modeling the perceived phonemic variation in the three perceptual variables; only the three vowel-extrinsic/formant-intrinsic procedures performed better. As argued in Chapter 9, CLIH$_{s4}$'s relatively poor performance in this last task could also be explained by the fact that CLIH$_{s4}$ uses formant-extrinsic transformations. The fact that CLIH$_{s4}$ considerably poorer than CLIH$_{i4}$ can probably be explained by the fact that CLIH$_{s4}$ uses a formant-extrinsic transformation, whereas CLIH$_{i4}$ uses a formant-intrinsic transformation.

Table 10.1 shows that MILLER performed poorly at preserving phonemic variation in the acoustic vowel data. Only SYRDAL & GOPAL performed poorer. For the second task

in the acoustic domain, reducing the anatomical/physiological variation, MILLER performed slightly better than the baseline. MILLER preserved the sociolinguistic variation in the transformed data. For the perceptual-acoustic comparisons, it was concluded that MILLER performed poorly, again only one procedure (SYRDAL & GOPAL) performed poorer at modeling the phonemic variation in the perceptual variables. In Chapter 9, it was concluded that MILLER's poor performance was due to the fact that MILLER's third dimension incorporates F_3.

Finally, NORDSTRÖM & LINDBLOM preserved the phonemic variation in the acoustic vowel data better than HZ. Only the three vowel-extrinsic/formant-intrinsic procedures performed better. Second, NORDSTRÖM & LINDBLOM did not perform very well at reducing anatomical/physiological variation in the acoustic measurements; its performance is only slightly above the performance of the baseline. However, closer inspection in Chapter 7 of the performance of this procedure shows that this relatively poor performance can almost entirely be attributed to the variation in F_0 related to the speaker's sex. When F_0 was excluded from the analysis, NORDSTRÖM & LINDBLOM minimized all variation related to the speaker's sex in the acoustic data. NORDSTRÖM & LINDBLOM was not appropriate for use on F_0-measurements, presumably because it was designed to correct for vocal-tract differences between (male and female) speakers. NORDSTRÖM & LINDBLOM preserved the sociolinguistic variation in the transformed vowel data. For the perceptual-acoustic comparisons, NORDSTRÖM & LINDBLOM model the perceived phonemic variation in the three perceptual variables slightly better than the baseline, although six procedures performed better than NORDSTRÖM & LINDBLOM.

10.2.2 Balancing the three variation sources

The results described in Chapter 7 of the present research show that, after normalization following the vowel-extrinsic/formant-intrinsic procedures, the phonemic variation is most prominent in the acoustic measurements, followed by the sociolinguistic variation, whereas the anatomical/physiological variation appeared to be minimized. In the raw data, the anatomical/physiological variation is the most prominent, followed by the phonemic variation, and finally the sociolinguistic variation.

The core research question was formulated in Chapter 1: which existing procedure for acoustic vowel normalization succeeds best at separating the three types of variation conveyed in the acoustic signal: phonemic variation, sociolinguistic speaker-related variation and anatomical/physiological speaker-related variation?

The results of the present research, as summarized in section 10.2, indicate that the vowel-extrinsic/formant-intrinsic procedures performed best at separating the three variation sources. These procedures achieved this by effectively reducing the acoustic consequences of anatomical/physiological sources of variation, while preserving the most of the acoustic

consequences of sociolinguistic sources of variation. The vowel-extrinsic/formant-extrinsic procedures performed second best, followed by the vowel-intrinsic/formant-intrinsic procedures[75], whereas the vowel-intrinsic/formant-extrinsic procedure did not perform very well at all.

Of the three vowel-extrinsic/formant-intrinsic procedures, LOBANOV performed best, followed by GERSTMAN and CLIH$_{i4}$. Therefore, it must be concluded that transforming a set of vowel data using LOBANOV resulted in a separation of acoustic consequences of variation sources in the acoustic signal that could best be used in sociolinguistic research.

Close examination of the performance of the four classes of procedures reveals a pattern in the results as can be observed in Table 10.2. Procedures that use a vowel-extrinsic transformation performed better than procedures that use a vowel-intrinsic transformation; while procedures that use a formant-intrinsic transformation performed better than procedures that use a formant-extrinsic transformation. This pattern can be illustrated using the difference in performance between CLIH$_{i4}$ and CLIH$_{s4}$. Both are vowel-extrinsic procedures, but CLIH$_{s4}$ incorporates a formant-extrinsic transformation. Furthermore, MILLER as well as SYRDAL & GOPAL use a formant-extrinsic transformation (as well as a dimension incorporating F_3), but MILLER performed better, because it uses a vowel-extrinsic information instead of merely vowel-intrinsic information.

Table 10.2: *Results for each of the four types of information that were combined to form four classes of normalization procedures. '0': not better or poorer than the baseline (*HZ*). '+': better than the baseline. '++': best of all four classes. '−': poorer than the baseline. '−−': worst of all four classes.*

Information	Vowel-intrinsic	Vowel-extrinsic
Formant-intrinsic	0	++
Formant-extrinsic	−−	+

10.3 Sociolinguistic considerations

The results of the comparison of the normalization procedures, summarized in section 10.2.1, show that the three vowel-extrinsic/formant-intrinsic procedures were most suitable for use in sociolinguistics. My results suggest that sociolinguistic differences in the pronunciation of vowels between groups of speakers can be investigated acoustically as follows.

First, the fundamental frequency and the frequencies of the first three formants should

[75] Although the vowel-intrinsic/formant-intrinsic procedures performed slightly better in the perceptual-acoustic comparisons.

be measured. Although it was found in Chapter 9 that the models for the perceived articulatory characteristics could be fit best using only F_1 and F_2, and that F_0 and F_3 did not improve the fit of the models, I do not recommend to exclude F_0 and/or F_3 a priori[76]. In tone languages or dialects such as, Mandarin Chinese or for Dutch dialects including tonal differences (e.g. some dialects spoken in the S-N region in the present research), F_0 can have a contrastive function for tones. F_3 helps listeners to distinguish between certain classes of front unrounded vowels in languages such as Swedish (cf. Fujimura, 1967) and for American English, F_3 is necessary to classify rhotacized vowels (Ladefoged, 2001).

Second, the measurements in HZ should be transformed to z-scores using LOBANOV. Nevertheless, it seems advisable to compare the mean values per vowel in LOBANOV with the mean values in Hz, because the possibility that LOBANOV introduces results that are inherent consequences of the transformation itself cannot be excluded given the results found in Chapter 7. Any differences in the mean values found in the data transformed following LOBANOV and the mean values for HZ should be investigated further, because they may be indications of language variation.

I would like to emphasize that Disner's (1980) recommendations are still valid; it is not advisable to carry out normalization procedures on data sets that are not fully phonologically comparable. Disner found that each procedure makes 'implicit assumptions' (i.e., the mean values and the standard deviations of both speaker groups have to be comparable) about the underlying vowel system when languages or dialects are compared on the basis of the normalized vowel frequencies. When the phonological vowel systems differ, the scale factors used in the vowel-extrinsic/formant-intrinsic procedure become biased. This may result in artificial differences between the data sets to be compared, or in the elimination of relevant differences.

One of the disadvantages of the three vowel-extrinsic/formant-intrinsic procedures is that they require information across all (monophthongal) vowels of a single speaker to estimate the scale factors (e.g., the mean formant frequency across all vowels for LOBANOV) that are used to normalize the raw measurements. Some sociolinguistic studies do not include all vowels from each informant. Nevertheless, the results in Chapter 7 indicate that it may suffice to obtain recordings and measurements of the three point vowels per speaker for the vowel system in question. Using only these three vowels per speaker, the scale factors can be estimated satisfactorily. I found that measurements obtained from scale factors estimated with the three point vowels per speaker correlate strongly with measurements obtained with scale factors estimated using all nine vowels per speaker. It should be noted that, because the correlations were not 100%, and because the results of the linear discriminant analysis showed a deterioration for the three-vowel case, measurements transformed using three vowels per speaker are of lower quality than the measurements that were transformed using all vowels for a speaker[77].

[76] It is advisable to also measure duration, for instance for distinguishing contrastive length differences.

[77] Although it should be noted that the measurements normalized using scale factors estimated using only three

In Chapter 9, it was concluded that it was not possible to model the sociolinguistic variation in the perceptual data. This means that it is not clear to what extent (transformed) formant frequencies must differ to be perceived by phonetically-trained listeners as a sociolinguistic difference. In addition, in Chapter 7 it was established that applying vowel-extrinsic/formant-intrinsic procedures to data in HZ may lead to differences between mean values that were not present in the data in HZ. However, some of these (additional) differences between the mean formant frequencies of the normalized data and the raw data can be excluded beforehand. This can be accomplished using the just noticeable difference for these formant frequencies. Kewley-Port & Watson (1994) measured (language-independent) discrimination difference limens for pairs of formant frequencies for isolated synthetic vowels (simulating a female voice). They stated that for the F_1 region a constant difference of 14.9 Hz between two formant frequencies is just noticeable. For the F_2 region, a linear difference of 1.5% is necessary to be noticeable. Kewley-Port & Watson suggested a 'piecewise-linear' function in which the F_1 region is defined as $<$ 800 Hz and the F_2 region is defined as $>$ 800 Hz. However, Kewley-Port & Watson's formula is discontinuous in the region between 800 and 1000 Hz. I therefore suggest that the F_1 region is set to $<$ 1000 Hz and the F_2 region is set to $>$ 1000 Hz, to avoid formant means with values between 800 and 1000 Hz showing difference limens that are too low.

By applying the modified version of the formula proposed by Kewley-Port & Watson, the differences between mean formant frequencies smaller than the difference limen can a priori be excluded as possible sociolinguistic differences, for the raw mean values in HZ. Before subjecting the differences between mean values to further analysis, it must be verified that these differences are larger than the difference limen. Differences that are smaller than the difference limen cannot reliably be perceived and are therefore not likely to represent a sociolinguistic difference. Note that this procedure is only to be applied to mean formant frequencies, and not to formant frequencies from two single vowel tokens. This issue deserves to be investigated further[78].

10.4 Phonetic considerations

The results described in Chapter 9 show that perceived phonemic variation in articulatory characteristics of vowels were modeled satisfactorily using the acoustic characteristics of those vowels. For perceived tongue height, a considerable proportion of the perceived phonemic variation was modeled primarily using F_1. For perceived tongue advancement, this was modeled best using F_2. Perceived lip rounding and spreading was modeled best using F_2, and to a lesser extent, F_1, although it should be noted that the proportion of explained variance

vowels per speaker are still of a better quality that raw measurements.

[78] For information on the perception of formant transitions see Van Wieringen & Pols (1993; 1995).

was lower for rounding than for tongue height and tongue advancement. These results are generally in line with those reported by Assmann (1979).

Moreover, the results described in Chapter 9 indicate that if more information about the speaker is incorporated in the scale factors that are used for transforming the acoustic measurements, then more of the variation in the perceived articulatory characteristics of vowels can be modeled. This result was found for acoustic measurements normalized using LOBANOV, CLIH$_{i4}$ and GERSTMAN.

The results in Chapter 9 draw a clearer picture of the role of F_0 in the perception of tongue height. Chapter 9 shows that including F_0 did not improve the model for perceived height when the anatomical/physiological variation was eliminated from the data. This result does not support earlier findings by Traunmüller (1981), who found that vowel-intrinsic F_0 did affect perceived vowel height. Nevertheless, it should be noted that the difference between my results and Traunmüller's could possibly be accounted for by the fact that I used read vowels from standard Dutch, while Traunmüller used synthetic vowels from a Bavarian dialect.

F_3 showed a result that was similar to the one found for F_0. F_3 was also not useful for modeling the perceived articulation of vowels of standard Dutch.

The research in Chapter 8 and 9 provided the following insight into the performance and behavior of phonetically-trained listeners when categorizing vowel tokens and when judging the articulatory characteristics of those vowel tokens. First, it was concluded that the phonetically-trained listeners had judged the phonemic variation reliably and consistently. This was found for the category judgments and for the articulatory judgments. Second, the listeners were influenced primarily by the response vowel category when making their judgments of the vowel token's tongue height, tongue advancement, and – to a lesser extent – of the vowel token's lip rounding. Third, it seemed plausible that listeners did not need to be exposed to more than one vowel produced by a single speaker; instead, listeners based their articulatory judgments on information contained within a single vowel token.

Chapter 8 shows that the availability of additional speaker-related information hardly affected listeners' judgments. It was found that the listeners did benefit somewhat from this information when categorizing vowel tokens but not when making articulatory judgments. I hypothesized that listeners make an estimation of the speaker's vowel system on the first presentation of a stimulus token. This estimation becomes more accurate when more information about that speaker's vowel system becomes available.

However, the results in Chapter 8 did not provide cues about which information in the stimuli, besides F_0, F_1, F_2, and F_3, could be used by the listeners to arrive at an estimation of the acoustic aspects of the speaker's vowel system. Johnson (1990a; 1990b; 1997) and Nusbaum & Magnuson (1997) suggested the following candidates.
Johnson (1990a; 1997) suggested that vowel-intrinsic F_0 can be used as an "indirect"[79] factor in vowel categorization. When vowel-intrinsic F_0 is used indirectly, it serves as a cue to the

[79] Cf. my description of Johnson's 1990a study in Chapter 3.

type of speaker (i.e., male, female, or a child). Johnson claims that Nearey's (1978) procedure used F_0 indirectly. Nusbaum & Magnuson share Johnson's view. They envision a process in which listeners, when perceiving ambiguous values of F_1 and F_2 due to speaker variability, direct their attention to F_0 and F_3 to obtain information about the larynx or vocal tract of the speaker.

Johnson (1990a) and Nusbaum & Magnuson (1997) thus suggested that listeners use F_0 and F_3 to estimate the speaker's vowel system. Although these authors referred to perceptual processes involved in vowel categorization, my results from the comparison with the acoustic data in Chapter 9 suggest that similar processes could operate in articulatory judgments of vowel tokens. Consequently, listeners may have used vowel-intrinsic F_0 and F_3 to make their articulatory judgments as well as their category judgments.

The results in Chapter 9 show that neither F_0 nor F_3 improved the regression models of the perception of phonemic variation. These results can be accounted for when it is assumed that the listeners did not directly use the vowel-intrinsic values of F_0 and F_3 for their judgments, but instead they used them indirectly, by making an estimation of the anatomical/physiological characteristics of the speaker, as suggested by Nusbaum & Magnuson. Such an estimation could allow the listeners to evaluate the stimulus vowel tokens relative to the expected characteristics of the vowel system of the speaker in question. When more information about other vowels produced by the same speaker became available (as was the case in my speaker-blocked presentation of the stimuli in the experiment in Chapter 8), this new information was incorporated in the initial estimation of the listeners. The reliability with which the stimuli were judged, improved as a result of this more accurate estimation of the speaker's vowel system. It can thus be hypothesized that, when hearing a vowel token produced by an unfamiliar speaker, listeners use acoustic information such as vowel-intrinsic F_0 and/or F_3 in the "indirect" fashion suggested by Johnson (1990a; 1997). Using this acoustic information, an estimation may be made of the type of speaker (e.g. male, female, small, large, young, or old); such a speaker type could have the form of a template that is based on speech from familiar speakers. Finally, the vowel token to be recognized is compared with the estimation, as to allow classification. However, this matter needs further investigation, based on my results no conclusions can be formulated on the behavior of phonetically-trained listeners.

Finally, in Chapter 1, I referred to Thomas (2002), who argued that Nearey's (1978) procedure did not reflect human speech perception, because Nearey's procedure requires more than one vowel per speaker to calculate the scale factors, while listeners can normalize a single vowel from a speaker without hearing another vowel from that speaker. However, the present research suggests that it is not necessary for a normalization procedure to reflect or resemble processes involved in human speech perception. For a normalization procedure to be successful, it is more important whether the procedure's output (i.e., the transformed formant frequencies) resembles the output of the listeners (i.e., the articulatory judgments).

Given the results presented in Chapter 9, it can be concluded that normalization procedures that are thought to resemble processes in human speech perception more, as is the case for SYRDAL & GOPAL and MILLER, produce output that did not resemble the judgments of listeners any better than the raw acoustic data. On the other hand, the processing mechanisms of listeners and the vowel-extrinsic/formant-intrinsic normalization procedures, such as Nearey's CLIH$_{i4}$, are very different, but their output was very similar. From this point of view, Thomas' argument does not seem very convincing.

On the other hand, my findings seem generally in line with Thomas' remark that listeners can normalize a single vowel from a speaker without hearing another vowel from that speaker; listeners appeared to be able to categorize a vowel token using only the information contained in that vowel token itself. Nevertheless, the fact should not be overlooked that listeners have had years of exposure to different speakers' voices before being able to categorize vowel tokens as effectively as found in studies on vowel perception discussed in Chapter 3. It seems possible that Nearey's CLIH$_{i4}$ procedure (as well as LOBANOV and GERSTMAN) accounts for the listeners' experience by using a scaling factor that is calculated using information distributed across other vowels produced by the same speaker.

10.5 Prospects

The comparison of the acoustic representations with the perceptual representation described in Chapter 9 revealed that it was not possible to satisfactorily model the perceived sociolinguistic variation using the acoustic representations. I argued that this was probably due to the small amount of sociolinguistic variation present in the speech material. For the present research, I decided to use speech material that displayed a moderate amount of variation to be able to establish whether the normalization procedures preserve subtle perceived sociolinguistic differences between vowel tokens. As argued in Chapter 1, such subtle sociolinguistic differences could be indications of possible language changes. Furthermore, presenting phonetically-trained listeners with vowel stimuli that reflect fine-grained to moderate sociolinguistic differences was thought to induce them to use their perceptual scale optimally. If listeners had been presented with stimuli in which large sociolinguistic differences between vowel tokens were present, it would not have been possible to establish whether phonetically-trained listeners can perceive (and reliably record) subtle sociolinguistic differences.

The present research shows that the sociolinguistic difference between two vowel tokens should be larger than the differences that were presented to the listeners in the listening experiment, for these differences to be modeled successfully. To evaluate how much two vowels should differ minimally in their sociolinguistic characteristics, a listening experiment should be carried out with stimuli that display considerable sociolinguistic differences. This can be accomplished by carrying out an experiment that is, by and large, identical to the experiment described in Chapter 8. However, this experiment must be carried out using

stimulus vowel tokens that were sampled across all eight regions of the sociolinguistic data set (described in Chapter 5). This way, the sociolinguistic variation in the vowel tokens related to the speaker's regional background is maximized and Disner's (1980) criterion (that the speakers share comparable phonological vowel systems) is still met. The sociolinguistic variation in the vowel data can be increased even further when speech material from speakers who differ more in their sociological characteristics (as was done in Adank, Van Heuven & Van Hout, 1999).

One of the limitations of the present research is that it did not test whether the category judgments and the articulatory judgment were language-independent. It could not be concluded that the results are language-independent, because the listening experiment was not carried out with phonetically-trained listeners who were not speakers of Dutch. However, it was reported that vowel categorization is affected by the language background of phonetically-trained listeners (e.g., Van Heuven & Van Houten, 1989). Findings by Dioubina & Pfitzinger (2002) indicate that articulatory judgments also vary with the language background of the phonetically-trained listener. It seems plausible that the normalization skills necessary for the task of making articulatory judgments are learned during the acquisition of the phonetically-trained expert's native language. It would be interesting to investigate further how the language background influences the judgment behavior of experts, and to establish if the results found for Dutch can be replicated for other languages.

It can be concluded that normalization procedures can be useful for research investigating language variation and change, provided that some restrictions are taken into account. When normalization procedures are applied to the frequencies of the fundamental frequency and the frequencies of the first three formants produced by a small database of speakers, the anatomical/physiological variation is discarded of, and it can be assumed that the variation that remains in the data is either of a phonemic or a sociolinguistic nature. Normalization is especially useful when data from male and female speakers is compared, as the normalization procedures that were most successful eliminated all variation related to speaker-sex.

Finally, the present research shows that perceived articulatory variation between vowels can be modeled satisfactorily using normalized acoustic representations of those vowel tokens. It seems worthwhile for sociolinguistic research to pursue further investigation of the relationship between perceptual and acoustic representations of vowels. This may lead to more insight into the development of language changes between and within vowels. I feel that future research should not be limited to studying F_1 and F_2; variation in F_0 or in the duration should also be taken into account (e.g., to allow investigation of the dynamic specification of diphthongs). Finally, it may be fruitful to extend the sociolinguistic research in the acoustic domain to spontaneous speech, to allow variation within speakers to be recorded as well as variation between speakers.

References

Aaltonen, O. (1985). The effects of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. *Journal of Phonetics 13*, 1–9.

Abercrombie, D. (1985). Daniel Jones' teaching. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pp. 15–24. Academic Press, Orlando.

Adank, P. (1999). Acoustic vowel normalisation: dealing with different sources of variation. In W. Bergsma, M. Palmen, and M. Wester (Eds.), *Proceedings of the CLS opening of the Academic Year 1999-2000*, pp. 55–77.

Adank, P. (2000). Acoustic and perceptual speaker normalisation. Newport Beach, pp. 2479. Proc. of 140th meeting of the Acoustical Society of America, 3-8 December.

Adank, P., V. J. van Heuven, and R. van Hout (1999). Uitspraak van de Nederlandse klinkers in Noordelijk Standaardnederlands en in Zuid-Limburg; een akoestische en perceptieve studie. In E. Huls and B. Weltens (Eds.), *Artikelen van de derde sociolinguïstische conferentie*, pp. 15–26. Eburon, Delft.

Adank, P., R. van Hout, and R. Smits (2001). A comparison between human vowel normalization strategies and acoustic vowel transformation techniques. In *Proceedings of EUROSPEECH '01*, Aalborg, pp. 481–484.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant and M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech*, pp. 103–113. Academic Press, London.

Assmann, P. (1979). *The Role of Context in Vowel Perception*. Ph. D. thesis, University of Alberta.

Assmann, P., J. T. Hogan, and T. M. Nearey (1982). Vowel identification: Orthographic, perceptual and acoustic factors. *Journal of the Acoustical Society of America 71*, 975–989.

Bergem, D. R. V., L. C. W. Pols, and F. J. K. van Beinum (1988). Perceptual normalization of the vowels of a man and a child. *Speech Communication 7*, 1–20.

Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson and B. Granström (Eds.), *The representation of speech in the peripheral auditory system*, pp. 95–102. Elsevier Biomedical Press, Amsterdam.

Bladon, A. (1985). *Auditory Phonetics*. Ph. D. thesis, Oxford University.

Bladon, A. and B. Lindblom (1981). Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America 69*, 1414–1422.

Bloomfield, L. (1933). *Language*. Henry Holt, New York.

Boersma, P. (1993). Accurate short-term analysis of fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *17*, 97–110.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International 5(9/10)*, 341–345.

Boersma, P. and D. Weenink (1996). Praat, a system for doing phonetics by computer, version 3.4. Report 132.

Booij, G. (1995). *The Phonology of Dutch*. Clarendon Press, Oxford.

Broadbent, D. E. and P. Ladefoged (1960). Vowel judgements and adaptation level. *Proceedings of the Royal Society B151*, 384–399.

Cameron, D. and J. Coates (1988). Some problems in the sociolinguistic explanation of sex differences. In D. Cameron and J. Coates (Eds.), *Women in their Speech Communities New Perspectives on Language and Sex*, pp. 13–26. Longman: London and New York.

Carlson, R., G. Fant, and B. Grandström (1975). Two-formant models, pitch and vowel perception. In G. Fant and M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech*, pp. 55–82. Academic Press, London.

Chambers, J. K. (2003). *Sociolinguistic theory*. Blackwell, Oxford.

Chiba, T. and M. Kajiyama (1941). *The Vowel: Its Nature and Structure*. Tokyo-Kaiseikan, Tokyo.

Chistovich, L. A., R. L. Sheikin, and V. V. Lublinskaya (1979). ”Centres of gravity” and spectral peaks as the determinants of vowel quality. In B. Lindblom and S. Öhman (Eds.), *Frontiers of Speech Communication Research*, pp. 143–157. Academic Press, New York.

Cohen, A., I. H. Slis, and J. ’t Hart (1963). Perceptual tolerance of isolated Dutch vowels. *Phonetica 9*, 65–78.

Cohen, A., I. H. Slis, and J. ’t Hart (1967). On tolerance and intolerance in vowel perception. *Phonetica 16*, 65–70.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 37–46.

Cucchiarini, C. (1993). *Phonetic Transcription: a Methodological and Empirical Study*. Ph. D. thesis, University of Nijmegen.

Daan, K. and D. P. Blok (1967). *Van Randstad tot Landrand. Toelichting bij de kaart: dialecten en naamkunde*. B.V. Noordhollandse Uitgevers Maatschappij, Amsterdam.

Davis, S. B. (1976). *Computer evaluation of laryngeal pathology based on inverse filtering of speech*. SCRL Monograph No. 13 Santa Barbara, CA.

Delattre, P. (1952). Some factors of vowel duration and their cross-linguistic validity. *Journal of the Acoustical Society of America 34*, 1141–1143.

Deterding, D. (1990). *Speaker Normalisation for Automatic Speech Recognition*. Ph. D. thesis, University of Cambridge.

Dioubina, O. I. and H. R. Pfitzinger (2002). An IPA vowel diagram approach to analysing L1 effects on vowel production and perception. In *Proc. ICSLP 2002*, Denver, pp. 2265–2268.

Disner, S. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America 67*, 253–261.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. Speech Transmission Quarterly Progress and Status Report 4/1966.

Fant, G. (1973). *Speech Sounds and Features*. MIT Press, Cambridge.

Fant, G. (1975). Non-uniform vowel normalization. In *Speech Transmission Laboratory Quarterly Progress and Status Report*, Volume 2-3, pp. 1–19. Royal Institute of Technology Stockholm.

Fant, G. (1982). Feature analysis of Swedish vowels -a revisit. In *Speech Transmission Laboratory Quarterly Progress and Status Report*, Volume 2-3, pp. 1–19. Royal Institute of Technology Stockholm.

Fant, G., R. Carlson, and B. Granstrøm (1974). The [e]-[ø] ambiguity. *Speech Communication Seminar*, 117–121.

Fant, G., G. Hennigson, and U. Stålhammer (1969). Formant frequencies of Swedish vowels. In *Speech Transmission Laboratory Quarterly Progress and Status Report*, Volume 4, pp. 26–31. Royal Institute of Technology Stockholm.

Fischer-Jørgensen, E. (1967). Perceptual dimensions of vowels. In *To honor Roman Jakobson*, pp. 667–671. Mouton, The Hague.

Fischer-Jørgensen, E. (1972). Formant frequencies of long and short Danish vowels. In E. Firchow, K. Grimstad, N. Hasselmo, and W. O'Neil (Eds.), *Studies for Einar Haugen*, pp. 189–213. Mouton, The Hague.

Fletcher, H. (1940). Auditory patterns. *Review of Modern Physics 12*, 47–65.

Fujimura, O. (1967). On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants. *Language and Speech 10*, 181–193.

Fujisaki, H. and T. Kawashima (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics AU-16*, 73–77.

Gamnes, H. (1965). *En akustisk undersökelse og maling av formantfrekvense i östnorske vokaler*. Ph. D. thesis, University of Oslo.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustic. AU-16*, 78–80.

Glasberg, B. R. and B. C. J. Moore (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research 47*, 103–138.

Harshman, R. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. In *Working Papers in Phonetics*, Volume 16. Phonetics Lab, UCLA.

Hermansky, H., B. A. Hanson, and H. Wakita (1985a). Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain. *Speech Communication 4*, 181–187.

Hermansky, H., B. A. Hanson, and H. Wakita (1985b). Perceptually based processing in automatic speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 509–512.

Hieronymus, J. L. (1991). Formant normalization for speech recognition and vowel studies. *Speech Communication 10*, 471–478.

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (Ed.), *Linguistic Variation, Models and Methods*, pp. 161–171. Academic Press, New York.

Hirahara, T. and H. Kato (1992). The effect of F0 on vowel identification. In Y. Tokhura, E. Vatikiotis-Bateson, and Y. Sagisaka (Eds.), *Speech Perception, Production, and Linguistic Structure*, pp. 89–112. Ohmusha Ltd, Tokyo.

Holmes, J. (1997). *An Introduction to Sociolinguistics*. Longman: London and New York.

International Phonetic Association (1999). *Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Jaberg, K. and J. Judd (1927). Transkriptionsverfahren, Aussprache und Gehöhrsschwankungen (Progelomena zum "Sprach- und Sachatlas Italiens und der Südschweiz"). *Zeitschrift für romanische Filologie 47*, 170–215.

Johnson, K. (1990a). Contrast and normalization in vowel perception. *Journal of Phonetics 18*, 229–254.

Johnson, K. (1990b). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America 88*, 642–654.

Johnson, K. (1997). Speech Perception without Speaker Normalization. In K. Johnson and J. Mullennix (Eds.), *Talker Variability in Speech Processing*, pp. 145–165. Academic Press, San Diego.

Jones, D. (1917). *An English Pronouncing Dictionary*. Dent, London.

Joos, M. (1948). Acoustic Phonetics. Language Monograph No. 23.

Jørgensen, E. (1969). Die gespannten und ungespannten Vokale in der Norddeutschen Hochsprache mit einer spzifischen Untersuchung der Struktur inher Formantfrequenzen. *Phonetica 19*, 217–245.

Kewley-Port, D. and C. S. Watson (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America 95*, 485–496.

Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical-band spectra: a first step. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1278–1281. Institute of Electrical and Electronics Engineers, New York.

Klein, W., R. Plomp, and L. W. C. Pols (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America 48*, 999–1009.

Koenig, W. (1949). A new frequency scale for acoustic measurements. *Bell Laboratories Record 27*, 299–301.

Koolhoven, H. (1968). *Dutch*. The English Universities Press, London.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change 2/2*, 205–254.

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Language in Society, volume 20. Oxford: Blackwell.

Labov, W., M. Yaeger, and R. Steiner (1972). A quantitative study of sound change in progress. U.S Regional Survey, Philadelphia.

Ladefoged, P. (1960). Vowel judgments and adaptation level. *Proceedings Royal Society B151*.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. Oxford U.P., London.

Ladefoged, P. (1975). *A course in Phonetics*. Harcourt Brace Jovanovich, New York.

Ladefoged, P. (2001). *Vowels and consonants*. Oxford: Blackwell.

Ladefoged, P. and D. E. Broadbent (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America 29*, 98–104.

Ladefoged, P., R. Harshman, L. Goldstein, and L. Rice (1978). Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America 64*, 1027–1035.

Laver, J. D. M. (1965). Variability in vowel perception. *Language and Speech 8*, 95–121.

Lawrence, W. (1953). The synthesis of signals which have a low information rate. In W. Jackson (Ed.), *Communication Theory*, pp. 460. New York and London.

Lehiste, I. and G. E. Peterson (1959). The identifiability of filtered vowels. *Phonetica 4*, 161–177.

Lieberman, P., E. S. Crelin, and D. H. Klatt (1972). Phonetic ability and related anatomy of the newborn and adult man, Neanderthal man, and the chimpanzee. *American Anthropologist 74*, 287–307.

Lindblom, B. E. F. and J. E. F. Sundberg (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *Journal of the Acoustical Society of America 4*, 1166–1179.

Lisker, L. and M. Rossi (1992). Auditory and visual cueing of the [+/-rounded] feature of vowels. *Language and Speech 35(4)*, 391–417.

Lloyd, R. J. (1890a). *Some researches into the Nature of Vowel-Sound*. Turner and Dunnett, Liverpool.

Lloyd, R. J. (1890b). Speech sounds: Their nature and causation (I). *Phonetische Studien 3*, 251–278.

Lloyd, R. J. (1891). Speech sounds: Their nature and causation (II-IV). *Phonetische Studien 4*, 37–67, 183–214, 275–306.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America 49*, 606–608.

Macchi, M. J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America 68*, 1636–1642.

Markel, J. D. and A. H. Gray (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin.

Miller, J. D. (1989). Auditory perceptual interpretation of the vowel. *Journal of the Acoustical Society of America 85*, 2114–2134.

Miller, J. D., A. M. Engebretson, and N. M. Vemula (1980). Vowel normalization: Differences between vowels spoken by children, women, and men. *Journal of the Acoustical Society of America Supplement 68*, s33(A).

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America 25*, 114–121.

Moore, B. C. J. (1977). *The Psychology of Hearing*. Macmillan.

Moore, B. C. J. and B. R. Glasberg (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America 74*, 750–753.

Mullennix, J. W., D. B. Pisoni, and C. S. Martin (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America 85*, 365–378.

Nearey, T. M. (1978). Phonetic Feature Systems for Vowels. Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in speech perception. *Journal of the Acoustical Society of America 85*, 2088–2113.

Nearey, T. M. (1992). Applications of Generalized Linear Modeling to vowel data. In *Proceedings of 1992 International Conference on Spoken Language Processing*, pp. 583–587. ICSLP.

Nearey, T. M., P. F. Assmann, and J. M. Hillenbrand (2002). Evaluation of a strategy for automatic formant tracking. *Journal of the Acoustical Society of America 112(5)*, 2323.

Nordström, P. E. (1977). Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics 5*, 81–92.

Nordström, P. E. and B. Lindblom (1975). A normalization procedure for vowel formant data. In *Proceedings of the 8^{th} International Congress of Phonetic Sciences*, Leeds, Paper 212.

Nusbaum, H. and J. Magnuson (1997). Talker Normalization: Phonetic Constancy as a Cognitive Process. In K. Johnson and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*, pp. 109–132. Academic Press, San Diego.

Oller, D. K. and R. E. Eilers (1975). Phonetic expectation and transcription validity. *Phonetica 31*, 288–304.

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In W. Labov (Ed.), *Locating Language in Time and Space*, pp. 143–178. Academic Press, New York.

Peterson, G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research 4*, 10–29.

Peterson, G. E. and H. L. Barney (1952). Control methods used in the study of the vowels. *Journal of the Acoustical Society of America 24*, 175–184.

Pickering, J. B. (1986). *Auditory Vowel Formant Variability*. Ph. D. thesis, Oxford University.

Pols, L. C. W. (1977). *Spectral analysis and identification of Dutch vowels in monosyllabic words*. Ph. D. thesis, Institute for Perception TNO, Soesterberg, the Netherlands.

Pols, L. C. W., H. R. C. Tromp, and R. Plomp (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America 53*, 1093–1101.

Potter, R. K. and G. E. Peterson (1948). The representation of vowels and their movements. *Journal of the Acoustical Society of America 20*, 528–535.

Potter, R. K. and J. C. Steinberg (1950). Towards the specification of speech. *Journal of the Acoustical Society of America 22*, 807–820.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: the Art of Scientific Computing*. Cambridge University Press, Second Edition, Cambridge.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin 92*, 81–110.

Rietveld, A. C. M. and V. J. van Heuven (2001). *Algemene fonetiek*. Coutinho, Bussum.

Rietveld, T. and R. van Hout (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin, New York.

Rosner, B. S. and J. B. Pickering (1994). *Vowel perception and production*. Oxford University Press, Oxford.

Ryals, J. H. and P. Lieberman (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America 72*, 1631–1634.

Sankoff, D., R. W. Shorrock, and W. McKay (1974). Normalization of formant space through the least squares affine transformation. Unpublished program and documentation.

Scharf, B. (1970). Critical bands. In J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory*, Volume 1, pp. 157–200. Academic, New York.

Schwartz, J. L. and P. Escudier (1987). Does the human auditory system include large scale perceptual integration? In M. E. H. Schouten (Ed.), *The psychophysics of speech perception*, pp. 284–292. Martinus Nijhoff, Dordrecht.

Scovel, T. (1988). *A Time to Speak: A Psycholinguistic Inquiry into the Critical Period for Human Speech*. Newbury House, Rowley.

Stevens, K. N. (1972). The quantal nature of speech; evidence from articulatory-acoustic data. In P. Denes and E. David (Eds.), *Human Communication: A Unified View*, pp. 51–66. McGraw-Hill, New York.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, Cambridgre.

Stevens, S. S. and J. Volkmann (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology 53*, 329–353.

Strange, W., R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman (1976). Consonant environment identifies vowel identity. *Journal of the Acoustical Society of America 60*, 213–214.

Syrdal, A. K. (1984). Aspects of a model of the auditory representation of American English vowels. *Speech Communication 4*, 121–135.

Syrdal, A. K. and H. S. Gopal (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America 79*, 1086–1100.

Thomas, E. R. (2002). Instrumental Phonetics. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change*, pp. 168–200. Blackwell Publishers, Oxford.

Traunmüller, H. (1981). Perceptual dimensions of openness in vowels. *Journal of the Acoustical Society of America 69*, 1465–1475.

Traunmüller, H. (1990). Analytic expressions for the tonotopic frequency scale. *Journal of the Acoustical Society of America 88*, 97–100.

Trudgill, P. (1983). *On Dialect: Social and Geographic Factors*. Blackwell: Oxford.

Van de Velde, H. (1996). *Variatie en Verandering in het gesproken Standaard-Nederlands (1935-1993)*. Ph. D. thesis, University of Nijmegen.

Van de Velde, H. and M. Houtermans (1999). Vlamingen en Nederlanders over de uitspraak van nieuwslezers. In E. Huls and B.Weltens (Eds.), *Artikelen van de Derde Sociolinguistische Conferentie*, pp. 451–462. Eburon, Delft.

Van de Velde, H. and R. van Hout (2000). N-deletion in reading style. In H. de Hoop and T. van der Wouden (Eds.), *Linguistics in the Netherlands 2000*, pp. 209–219. John Benjamins Publishing Company, Amsterdam.

Van Heuven, V. J. and J. E. van Houten (1989). Vowel labelling consistency as a measure of familiarity with the phonetic code of a language or dialect. In M. Schouten and P. van Reenen (Eds.), *New methods in Dialectology: Proceedings of a Workshop Held at the Free University*, pp. 79–86. Foris: Dordrecht.

Van Hout, R., G. de Schutter, E. de Crom, W. Huick, H. Kloots, H. Van de Velde, and M. Houtermans (1999). De uitspraak van het Standaard-Nederlands: Variatie en varianten in Vlaanderen en Nederland. In E. Huls and B.Weltens (Eds.), *Artikelen van de Derde Sociolinguistische Conferentie*, pp. 183–196. Eburon, Delft.

Van Nierop, D. J. P. J., L. W. C. Pols, and R. Plomp (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica 29*, 110–118.

Van Rie, J., R. van Bezooijen, and W. H. Vieregge (1995). Het Standaard Nederlands: een verkennend empirisch onderzoek. In E. Huls and J. Klatter Former (Eds.), *Artikelen van de Tweede Sociolinguistische Conferentie*, pp. 491–505. Eburon, Delft.

Van Wieringen, A. and L. C. W. Pols (1993). Frequency and duration discrimination of short first-formant speechlike transitions. *Journal of the Acoustical Society of America 95*, 502–511.

Van Wieringen, A. and L. C. W. Pols (1995). Discrimination of single and complex CV- and VC-like formant transitions. *Journal of the Acoustical Society of America 98*, 1304–1312.

Verbrugge, R. R., W. Strange, D. P. Shankweiler, and T. R. Edman (1976). What information enables a listener to map a speaker's vowel space? *Journal of the Acoustical Society of America 60*, 198–212.

Wakita, H. (1977). Normalization of vowels by vocal tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-25*, 183–192.

Watson, C. I., M. Maclagan, and J. Harrington (2000). Acoustic evidence for vowel change in New Zealand English. *Language Variation and Change 12*, 51–68.

Weenink, D. J. M. (1986). The identification of vowel stimuli from men, women, and children. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 10*, 41–54.

Weenink, D. J. M. (1993). Modelling speaker normalization by adapting the bias in a neural net. *Proceedings Eurospeech93, Berlin*, 2259–2262.

Weenink, D. J. M. (1997). Category ART: A variation on adaptive resonance theory neural networks. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 21*, 117–129.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands as a function of frequency. *Journal of the Acoustical Society of America 33*, 248.

Zwicker, E. and R. Feldtkeller (1967). *Das Ohr als Nachrichtenempfänger*. Hirzel, Stuttgart.

Zwicker, E. and E. Terhardt (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America 68*, 1623–1625.

Appendix A

Mean acoustic values

Ellipse plots for N-R and N-S

Figures A.1 and A.2 show the data for the N-R and N-S region, in HZ and normalized following LOBANOV, GERSTMAN, and CLIH$_{i4}$, in so-called 'ellipse plots'. In each of these ellipses, the mean value per vowel category is plotted as the centroid value. Each centroid represents the measurements of 40 tokens. The values used to draw the ellipses per vowel were obtained through a discriminant analysis on the measurements of F_1/D_1 and F_2/D_2, carried out using the built-in tools of the program Praat (Boersma, 2001). The radius of the ellipses was set for the entire ellipse to cover 68% of the data in that vowel category.

Plots of mean values

In Figures A.3, the mean values per vowel category for the eight regions N-R (Netherlands-Randstad), N-M (Netherlands-Middle), N-S (Netherlands-South), N-N (Netherlands-North), F-B (Flanders-Brabant), F-E (Flanders-East), F-L (Flanders-Limburg), and F-W (Flanders-West), as described in Chapter 5. The means are given for F_1 and F_2 for the 2880 vowel tokens in Hz, the data transformed (D_1 and D_2) following LOBANOV, GERSTMAN, and CLIH$_4$, respectively. Each mean value represents the measurements of 40 vowel tokens.

Tables of mean values

Tables A.1, A.2, A.3, and A.4 list the mean values of F_0/D_0, F_1/D_1, F_2/D_2, and F_3/D_3 for each monophthongal vowel, for each of the eight regions. All mean values are transformed following HZ, LOBANOV, CLIH$_{i4}$, and GERSTMAN, respectively. The values for F_1/D_1 and F_2/D_2 are displayed as well in Figure A.3.

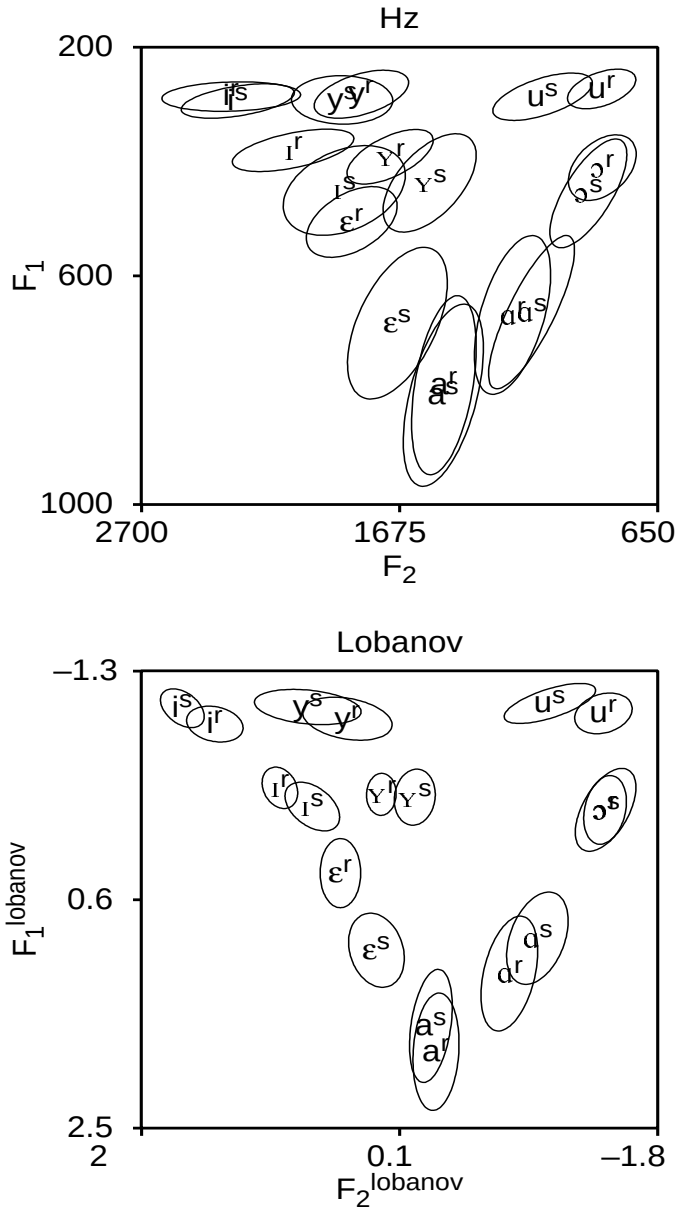

Figure A.1: *Ellipse plot for the 20 speakers in the N-R region,* ^{r}*, and the N-S region,* ^{s}*, for* HZ *and* LOBANOV. *Each centroid represents the measurements of 40 tokens.*

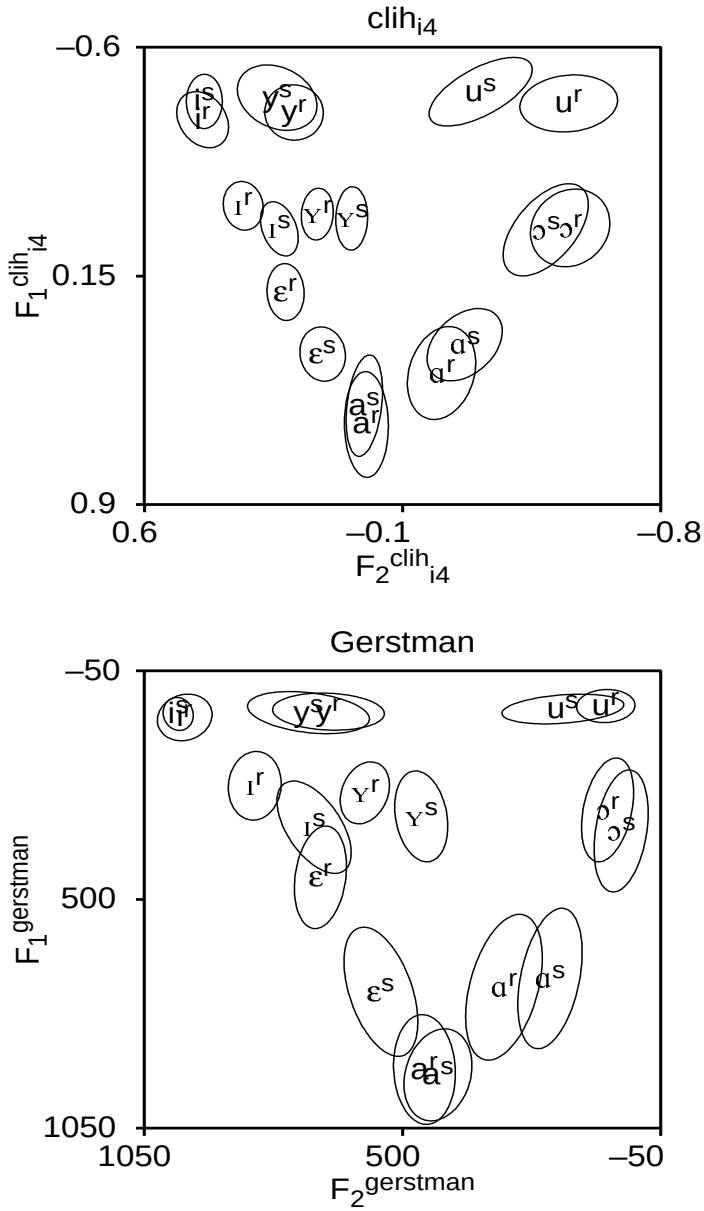

Figure A.2: *Ellipse plot for the 20 speakers in the N-R region,* ^{r}*, and the N-S region,* ^{s}*, for* CLIH$_{i4}$*, and* GERSTMAN*. Each centroid represents the measurements of 40 tokens.*

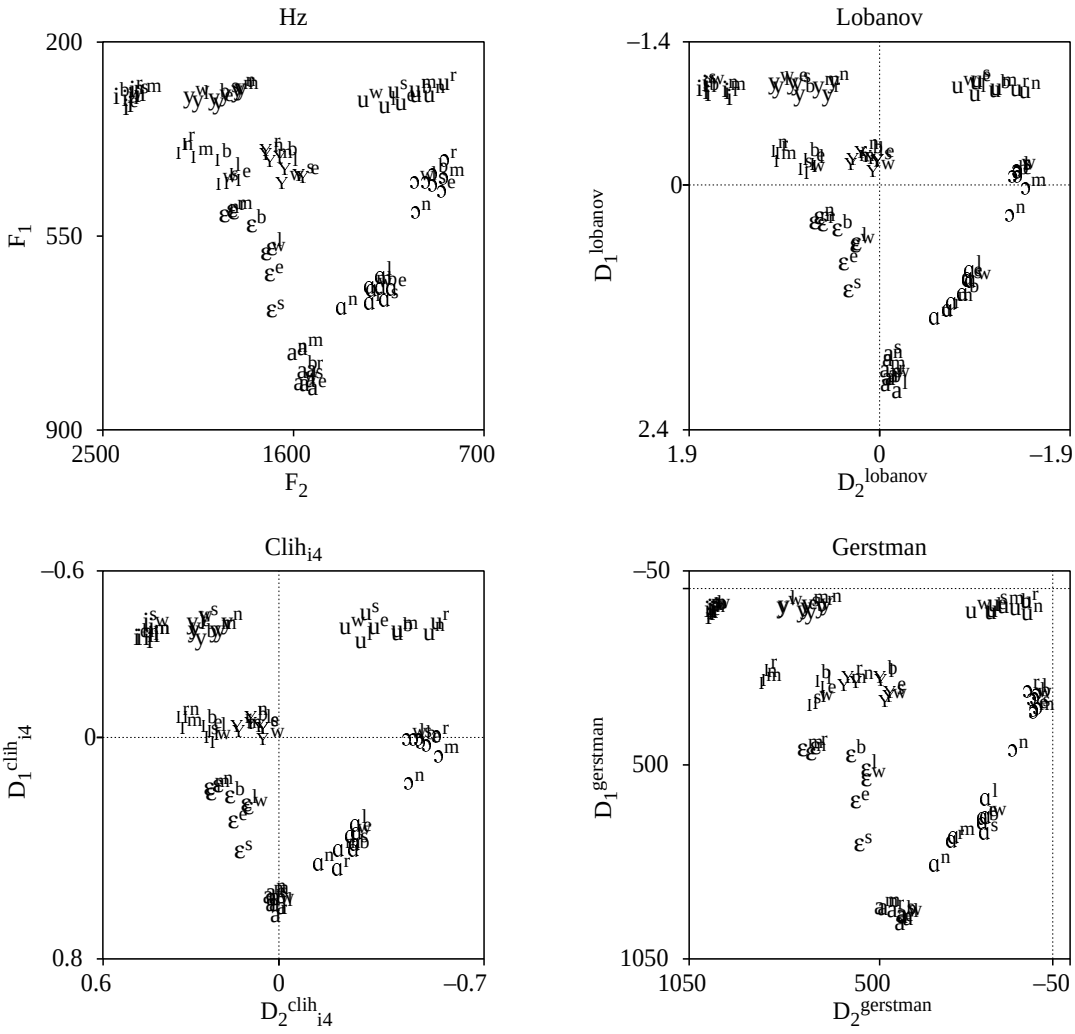

Figure A.3: *Un-normalized mean values of F_1 and F_2 per vowel category per region in Hz,* LOBANOV, CLIH$_{i4}$ *and* GERSTMAN. ^{r}*= N-R,* ^{m}*= N-M,* ^{s}*= N-S,* ^{n}*= N-N,* ^{b}*= F-B,* ^{e}*= F-E,* ^{l}*= F-L, and* ^{w}*= F-W.*

Table A.1: *Mean values in hertz for the four acoustic variables F_0, F_1, F_2, and F_3 for the nine monophthongal vowels of Dutch, for the four regions in the Netherlands (N-R, N-M, N-S, N-N) and Belgium (F-B, F-E, F-L, F-W). The number of vowel tokens per mean value is 40.*

HZ		N-R	N-M	N-S	N-N	F-B	F-E	F-L	F-W
/ɑ/	F_0	187	180	187	197	175	179	179	178
	F_1	668	639	663	677	640	642	621	644
	F_2	1226	1202	1150	1351	1164	1115	1177	1197
	F_3	2665	2655	2670	2619	2848	2790	2836	2732
/a/	F_0	164	163	170	181	159	152	158	164
	F_1	791	751	808	760	792	823	816	813
	F_2	1499	1520	1501	1581	1534	1488	1534	1541
	F_3	2669	2598	2699	2747	2841	2805	2804	2729
/ɛ/	F_0	187	181	185	204	176	179	178	179
	F_1	505	503	682	511	528	617	570	579
	F_2	1865	1853	1683	1903	1774	1691	1689	1695
	F_3	2681	2680	2693	2691	2775	2799	2803	2729
/ɪ/	F_0	187	193	197	212	195	198	194	188
	F_1	380	405	450	396	409	445	435	454
	F_2	2098	2031	1894	2114	1930	1838	1879	1917
	F_3	2709	2696	2734	2722	2757	2801	2815	2783
/i/	F_0	202	190	210	213	191	202	206	196
	F_1	286	292	294	295	298	318	314	302
	F_2	2343	2275	2317	2345	2413	2369	2356	2361
	F_3	2788	2756	2840	2773	3050	2975	3058	2987
/ɔ/	F_0	185	184	195	201	185	189	195	185
	F_1	410	443	455	505	436	465	450	450
	F_2	869	854	925	1000	919	881	961	995
	F_3	2932	2662	2697	2633	2899	2891	2803	2769
/u/	F_0	206	190	210	216	193	210	208	200
	F_1	273	285	286	294	294	309	315	303
	F_2	872	987	1107	935	998	1072	1156	1237
	F_3	2495	2499	2562	2570	2647	2589	2544	2490
/ʏ/	F_0	200	190	200	209	194	201	185	190
	F_1	391	411	437	399	405	440	425	451
	F_2	1713	1676	1555	1705	1639	1532	1626	1620
	F_3	2528	2536	2551	2480	2699	2703	2712	2691
/y/	F_0	204	189	204	213	190	197	199	188
	F_1	282	286	292	281	301	307	304	297
	F_2	1826	1827	1903	1830	1951	1933	2038	2058
	F_3	2420	2382	2513	2407	2491	2451	2502	2496

Table A.2: *Mean values in hertz for the four acoustic variables D_0, D_1, D_2, and D_3, transformed following* LOBANOV. *for the nine monophthongal vowels of Dutch. for the four regions in the Netherlands (N-R. N-M. N-S. N-N) and Belgium (F-B. F-E. F-L. F-W). The number of vowel tokens per mean value is 40.*

LOBANOV		N-R	N-M	N-S	N-N	F-B	F-E	F-L	F-W
/ɑ/	D_0	-0.22	-0.24	-0.41	-0.35	-0.30	-0.43	-0.39	-0.39
	D_1	1.21	1.14	0.92	1.29	1.06	0.90	0.83	0.93
	D_2	-0.71	-0.79	-0.92	-0.60	-0.88	-0.92	-0.93	-0.98
	D_3	0.09	0.22	-0.01	-0.02	0.32	0.19	0.33	0.13
/a/	D_0	-1.18	-1.27	-1.31	-1.22	-1.26	-1.43	-1.34	-1.22
	D_1	1.86	1.81	1.65	1.70	1.94	1.89	200.01	1.89
	D_2	-0.17	-0.13	-0.13	-0.13	-0.11	-0.12	-0.20	-0.18
	D_3	0.02	0.00	0.18	0.66	0.30	0.18	0.11	0.05
/ɛ/	D_0	-0.19	-0.12	-0.46	-0.07	-0.29	-0.39	-0.43	-0.33
	D_1	0.38	0.36	1.02	0.31	0.43	0.76	0.56	0.58
	D_2	0.54	0.57	0.27	0.55	0.37	0.31	0.20	0.17
	D_3	0.16	0.43	0.15	0.34	0.03	0.21	0.18	0.12
/ɪ/	D_0	-0.10	0.37	0.17	0.32	0.48	0.34	0.22	0.18
	D_1	-0.33	-0.25	-0.17	-0.35	-0.27	-0.23	-0.21	-0.13
	D_2	0.98	0.94	0.74	1.00	0.68	0.62	0.61	0.65
	D_3	0.26	0.51	0.45	0.50	-0.05	0.24	0.22	0.41
/i/	D_0	0.42	0.28	0.63	0.34	0.33	0.38	0.64	0.67
	D_1	-0.86	-0.92	-0.99	-0.94	-0.92	-0.95	-0.90	-0.98
	D_2	1.46	1.44	1.70	1.49	1.68	1.75	1.68	1.65
	D_3	0.57	0.83	1.01	0.70	1.20	1.00	1.31	1.38
/ɔ/	D_0	-0.14	0.05	-0.04	-0.07	0.04	0.02	0.14	0.06
	D_1	-0.15	0.02	-0.15	0.28	-0.10	-0.10	-0.14	-0.16
	D_2	-1.41	-1.53	-1.42	-1.35	-1.38	-1.42	-1.41	-1.43
	D_3	1.14	0.25	0.29	0.02	0.52	0.69	0.20	0.28
/u/	D_0	0.63	0.41	0.69	0.47	0.35	0.79	0.72	0.83
	D_1	-0.95	-0.96	-1.03	-0.93	-0.94	-1.01	-0.90	-0.97
	D_2	-1.40	-1.24	-1.01	-1.50	-1.20	-1.01	-0.98	-0.85
	D_3	-0.66	-0.55	-0.52	-0.22	-0.61	-0.76	-0.92	-1.14
/ʏ/	D_0	0.21	0.36	0.26	0.22	0.44	0.40	0.09	0.16
	D_1	-0.27	-0.23	-0.25	-0.34	-0.30	-0.26	-0.28	-0.15
	D_2	0.23	0.20	-0.01	0.13	0.09	-0.03	0.06	-0.01
	D_3	-0.51	-0.36	-0.61	-0.80	-0.39	-0.25	-0.23	-0.08
/y/	D_0	0.58	0.15	0.46	0.37	0.22	0.32	0.33	0.04
	D_1	-0.90	-0.97	-1.00	-1.02	-0.90	-1.01	-0.97	-1.01
	D_2	0.48	0.53	0.78	0.41	0.75	0.82	0.98	0.98
	D_3	-1.07	-1.34	-0.93	-1.20	-1.32	-1.49	-1.20	-1.14

Table A.3: *Mean values in hertz for the four acoustic variables D_0, D_1, D_2, and D_3, transformed following* CLIH$_{i4}$. *for the nine monophthongal vowels of Dutch. for the four regions in the Netherlands (N-R. N-M. N-S. N-N) and Belgium (F-B. F-E. F-L. F-W). The number of vowel tokens per mean value is 40.*

CLIH$_{i4}$		N-R	N-M	N-S	N-N	F-B	F-E	F-L	F-W
/ɑ/	D_0	-0.02	-0.02	-0.04	-0.04	-0.05	-0.07	-0.05	-0.05
	D_1	0.47	0.40	0.38	0.45	0.40	0.34	0.31	0.35
	D_2	-0.21	-0.23	-0.27	-0.15	-0.27	-0.28	-0.27	-0.27
	D_3	0.00	0.02	0.00	0.00	0.03	0.01	0.03	0.01
/a/	D_0	-0.15	-0.12	-0.14	-0.13	-0.15	-0.21	-0.17	-0.12
	D_1	0.64	0.57	0.58	0.56	0.60	0.58	0.61	0.60
	D_2	0.00	0.01	0.00	0.01	0.01	0.01	-0.02	-0.02
	D_3	0.01	0.00	0.01	0.05	0.02	0.02	0.00	0.01
/ɛ/	D_0	-0.02	-0.01	-0.05	0.00	-0.05	-0.05	-0.05	-0.03
	D_1	0.20	0.18	0.41	0.17	0.21	0.30	0.24	0.25
	D_2	0.22	0.21	0.12	0.19	0.15	0.14	0.10	0.08
	D_3	0.01	0.03	0.01	0.03	0.00	0.02	0.02	0.01
/ɪ/	D_0	-0.02	0.05	0.02	0.04	0.05	0.06	0.03	0.02
	D_1	-0.08	-0.04	-0.01	-0.08	-0.05	-0.03	-0.02	0.01
	D_2	0.33	0.30	0.23	0.30	0.24	0.22	0.20	0.20
	D_3	0.02	0.04	0.03	0.04	-0.01	0.02	0.02	0.03
/i/	D_0	0.06	0.03	0.07	0.04	0.05	0.06	0.09	0.06
	D_1	-0.36	-0.37	-0.42	-0.37	-0.36	-0.36	-0.35	-0.40
	D_2	0.44	0.41	0.44	0.40	0.46	0.47	0.43	0.41
	D_3	0.05	0.06	0.07	0.06	0.09	0.08	0.11	0.10
/ɔ/	D_0	-0.03	0.00	-0.01	-0.02	0.01	0.00	0.03	-0.01
	D_1	-0.01	0.06	0.00	0.16	0.01	0.02	0.00	0.00
	D_2	-0.55	-0.57	-0.49	-0.46	-0.50	-0.52	-0.47	-0.46
	D_3	0.10	0.02	0.01	0.00	0.04	0.05	0.01	0.02
/u/	D_0	0.08	0.03	0.08	0.05	0.06	0.11	0.09	0.07
	D_1	-0.42	-0.39	-0.45	-0.38	-0.38	-0.40	-0.35	-0.40
	D_2	-0.55	-0.43	-0.31	-0.53	-0.42	-0.34	-0.29	-0.25
	D_3	-0.06	-0.04	-0.04	-0.03	-0.05	-0.06	-0.08	-0.08
/ʏ/	D_0	0.04	0.03	0.03	0.02	0.05	0.07	-0.03	0.04
	D_1	-0.05	-0.03	-0.04	-0.08	-0.06	-0.04	-0.05	0.00
	D_2	0.13	0.11	0.04	0.08	0.08	0.04	0.06	0.03
	D_3	-0.05	-0.03	-0.04	-0.06	-0.03	-0.02	-0.01	-0.01
/y/	D_0	0.07	0.01	0.04	0.04	0.04	0.04	0.06	0.02
	D_1	-0.39	-0.39	-0.44	-0.42	-0.36	-0.40	-0.39	-0.42
	D_2	0.19	0.19	0.24	0.16	0.25	0.27	0.28	0.27
	D_3	-0.09	-0.09	-0.06	-0.09	-0.11	-0.12	-0.10	-0.08

Table A.4: *Mean values in hertz for the four acoustic variables D_0, D_1, D_2, and D_3, transformed following* GERSTMAN. *for the nine monophthongal vowels of Dutch, for the four regions in the Netherlands (N-R. N-M. N-S. N-N) and Belgium (F-B. F-E. F-L. F-W). The number of vowel tokens per mean value is 40.*

GERSTMAN		N-R	N-M	N-S	N-N	F-B	F-E	F-L	F-W
/ɑ/	D_0	469	433	377	370	412	416	414	378
	D_1	711	700	690	779	659	647	592	643
	D_2	284	265	186	328	188	191	185	173
	D_3	467	595	529	502	564	581	580	516
/a/	D_0	199	183	135	141	168	153	161	182
	D_1	909	902	922	905	923	945	933	926
	D_2	453	478	424	476	422	429	410	412
	D_3	458	528	577	688	549	578	520	490
/ɛ/	D_0	466	468	362	446	418	430	401	416
	D_1	446	453	722	463	469	601	511	537
	D_2	675	698	546	683	568	556	530	517
	D_3	493	657	566	612	483	587	536	510
/ɪ/	D_0	489	596	534	556	623	640	578	550
	D_1	226	262	326	253	256	298	272	317
	D_2	815	815	688	820	664	647	656	663
	D_3	519	682	656	646	460	590	546	592
/i/	D_0	634	579	666	558	593	652	707	686
	D_1	62	58	53	68	59	78	66	57
	D_2	964	973	978	969	963	981	979	961
	D_3	605	763	814	703	804	794	845	865
/ɔ/	D_0	471	514	482	447	514	541	558	505
	D_1	284	346	335	455	306	340	296	309
	D_2	63	33	34	101	37	43	38	40
	D_3	760	605	612	520	613	710	541	562
/u/	D_0	689	612	678	593	611	767	725	727
	D_1	34	47	42	67	53	59	66	61
	D_2	67	120	158	59	94	163	171	215
	D_3	256	352	374	448	307	325	227	147
/ʏ/	D_0	577	603	558	525	617	650	548	552
	D_1	243	269	300	258	245	289	253	312
	D_2	580	581	460	553	483	457	487	463
	D_3	300	410	345	287	364	464	419	450
/y/	D_0	671	539	616	567	565	633	619	515
	D_1	48	42	50	40	65	59	45	48
	D_2	658	685	700	640	683	709	768	760
	D_3	144	137	259	175	106	135	153	147

Appendix B

Questionnaire for phonetically-trained listeners

This questionnaire was translated from Dutch.

Dear listener,

Thanks again for participating in my experiment. The goals of this experiment are twofold. I, first, aim to gain more insight in the relation between the acoustic and perceived dimensions of vowel sounds and, second, I aim to learn more about the strategies used by expert listeners in judging vowels. Some additional information about the participants is required, to allow me to optimally interpret the results from this experiment. For this purpose, I have compiled a list of questions to gather more information about your expertise and personal background. Would you be so kind as to answer these questions and send the questionnaire back to me through e-mail? All information will be treated confidentially and will only appear in a coded form in publications (e.g., thesis). Thanks for your co-operation.

With kindest regards,

Patti Adank

- Name:
- Age:
- Place of birth:
- Additional cities/countries you have lived in (especially during your childhood years and longer residencies in foreign countries (please cite the years)):
- Native tongue, languages/dialects parents:
- Additional languages (please also cite your level of proficiency for each language and how often you use each language):
- City and institute where you received your phonetic (transcription) training:
- Name of your transcription teacher:
- How many years of professional listening experience would you estimate you have?
- Please cite some names of other expert listeners who you have worked with:
- Do you transcribe a lot of speech material from other languages/dialects? If the answer is yes: which ones?
- Are you more involved with the transcription of vowels or of consonants?
- Are you more involved with the transcription of segmental or supra-segmental phenomena?
- When you are listening to a vowel sound, to judge its linguistic quality (a very open articulated /E/ or a very fronted /y/ for instance), are you aware of the strategy you use? Do you try to imitate the vowel sound or do you listen to the sound very often?
- Vowel height, place of constriction, and rounding are used often as dimension to describe vowel quality. Do you primarily use these specific dimensions too, or do you use additional ones as well?
- Which vowel scheme of vowel quadrilateral do you usually use when judging and identifying vowel tokens, the traditional IPA-quadrilateral or another scheme?

Appendix C

Instructions for the experiment

These instructions were translated from Dutch.

General Instructions

Dear listener,

The experiment you are about to participate in, aims at providing more information on the strategies used by phonetically-trained listeners when judging the linguistic quality of vowel tokens. This experiment consists of three parts. The design of these three sub-experiments is for the largest part identical. In every sub-experiment, nine white buttons on the left side of the screen are presented. These are the ‘vowel buttons‘. These vowel buttons depict the phonetic symbols for the nine monophthongal Dutch vowels: /ɑ/ (as in kat), /a/ (kaas), /ɛ/ (hek), /ɪ/ (kip), /i/ (lied), /ɔ/ (bok), /u/ (hoed), /ʏ/ (put), and /y/ (fuut). The long mid vowels (/e/, /o/, /ø/, and the diphthongs (/ɔu/, /ɛi/, and /œy/) are not included.

On the right hand side of the screen, a white square shaped like the vowel quadrilateral is depicted. This quadrilateral is a scaled copy of the quadrilateral of the International Phonetic Alphabet (the 1996 corrected version). Almost all of the phonetic symbols have been erased, except for the phonetic symbols for the point vowels in the four corners of the quadrilateral. Note that the point vowels at the corners are to be regarded as theoretical end points of the quadrilateral. The horizontal axis depicts the articulatory dimension ‘place of constriction’. Along the vertical axis, the articulatory dimension ‘vowel height’ is displayed. The quadrilateral is regarded as a two-dimensional space in this experiment. Lip rounding and spreading is judged outside the quadrilateral. The white rectangle directly above the vowel quadrilateral is used for the depicting the perceived amount of lip rounding and spreading.

This rectangle contains a scale, with a minus sign on the left (spreading) and a plus sign on the right (rounding). This scale is used in the experiment to judge rounding and spreading of the vowel token. The experimental screen looks like this:

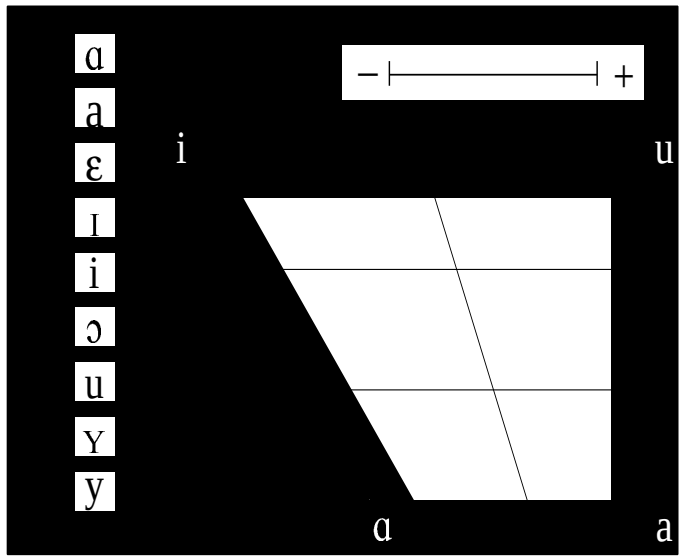

Figure C.1: The experimental response screen used

The speech material you are going to listen to consists of vowel tokens that were pronounced in a neutral context: /sVs/. The vowel tokens that are played were produced in the carrier sentence “In sVs and in sVsse zit de V‘’ [In sVs and in sVsse is the V], (e.g., “In sas en in sasse zit de a”). The speakers are all speakers of Standard Dutch. In each sub-experiment you are requested to perform two tasks:

1. The vowel tokens have to be identified. You can do this by clicking the mouse on the vowel button that displays the phonetic character of your choice. You are asked to make this choice every time you judge a vowel token.

2. The vowel tokens have to be judged on their linguistic quality relative to the linguistic quality of the extremes or end points of the quadrilateral. This judging is a two-step process: first, you are asked to judge the rounding/spreading of the vowel token. You can do this by clicking with the mouse on the scale in the rectangle above the quadrilateral, on the point of your choice relative to the extreme points; the left side of the scale represents a maximally spread vowel token and the right side represents a maximally rounded vowel token. Second, you are asked to point out the place of constriction and the vowel height in the vowel quadrilateral. You can do this by positioning the mouse in the vowel quadrilateral, on a place that resembles the place of constriction and vowel height you perceived, relative to the extremes (the point vowels) of the quadrilateral. Clicking the mouse on the place you selected causes the program to write the corresponding co-ordinates to a file.

All three sub-experiments show (roughly) similar procedures. The exact procedure of each sub-experiment is explained in separate instructions.

Thanks for your participation in this experiment,

Patti Adank

Instructions for sub-experiment 1

In this sub-experiment, 207 vowel tokens in a /sVs/ context are played. In this specific sub-experiment, different vowel categories from different speakers are played in a randomized order.

You are expected to both identify (choose a phonemic symbol) and judge (indicate the amount of rounding/spreading and indicate the place of constriction and the vowel height) each vowel token. You can indicate which Dutch vowel category was pronounced by clicking with the mouse on one of the vowel category buttons. After you have identified the vowel token, you are to judge the linguistic quality of the vowel token by first clicking the mouse in the rounding/spreading area on the position corresponding to the amount of rounding or spreading of that particular vowel and second by clicking the mouse on the position of your choice in the vowel quadrilateral. Once you have done these two things, the next vowel token to be judged is played.

The process of one experimental cycle (in which only one vowel token is to be identified and judged) is as follows. First, a short tone signal is played. After the tone signal, the vowel token in the nonsense syllable is played. The nonsense syllable is played 10 times maximally, with one to one-and-a-half second intervals. After the vowel token was played 10 times, the program waits until you have identified and judged the vowel (by clicking on one of the vowel buttons, clicking in the rounding/spreading rectangle, and clicking in the

vowel quadrilateral), before the next experimental cycle starts. This is a fixed order: first the vowel token has to be identified and then judged. In this sub-experiment, the next cycle starts immediately after you have clicked in the vowel quadrilateral.

You do not have to wait until the vowel token was played the full 10 times, you can start making your choice after the first time the vowel token was played.

The sub-experiment consists of nine blocks of 20 vowel tokens and a block of 27 vowel tokens (the first block). Between these blocks, a screen appears with the word 'pauze' [pause], plus a number. This number corresponds to the number of the pause screen (there are nine pause screens). Whenever this screen appears you have an option of taking a short break. If you want to commence, all you have to do is click with the mouse, and the next block starts.

The first five stimuli are practice trials. After these five trials the real experiment starts. You can take a short break after the practice trails, to ask any questions, if you wish. Thanks.

Instructions for sub-experiment 2

In this sub-experiment, 207 vowel tokens are played in a sVs-context. The vowel tokens are blocked by vowel type. Each block consists therefore of vowel tokens of the same vowel type, uttered by different speakers.

In this sub-experiment you are asked to judge the vowel tokens by clicking in the vowel quadrilateral (and in the rounding/spreading rectangle). This means that you do not have to identify each vowel token by clicking one of the vowel buttons, this was done for you. You do, however, have the possibility of indicating that you disagree with the 'labeling' of a certain vowel token. If you perceive a vowel token as an /ɛ/ (as in 'pet' (hat), while the vowel token was supposed to be pronounced as /ɪ/ (as in 'pit' (pit)), you can indicate this by clicking the vowel button "ɛ". If this is the case, you are expected to click on the vowel button before you click in the rounding/spreading rectangle and the vowel quadrilateral. In this sub-experiment, the next cycle starts immediately after you have clicked in the vowel quadrilateral.

The nonsense syllable that the speakers were instructed to pronounce, is displayed between the nine vowel buttons and the vowel quadrilateral in each trial. In the block in which the vowels in the syllables " sis " (sɪs) are to be judged, the syllable " sis " is displayed in the middle of the screen.

The sub-experiment consists of nine blocks of 23 vowel tokens. Between these blocks, a screen appears with the word 'pauze' [pause], plus a number. This number corresponds to the number of the pause screen (there are eight pause screens). Whenever this screen appears you have an option of taking a short break. If you want to commence, all you have to do is click with the mouse, and the next block starts.

The first five stimuli are practice trials. After these five trials the real experiment starts.

You can take a short break after the practice trails, to ask any questions, if you wish. Thanks.

Instructions for sub-experiment 3

In this sub-experiment, 207 vowel tokens in a sVs-context are played. These vowel tokens are blocked by speaker. Each block consists therefore of all the vowel tokens of only one speaker.

As was the case in experiment 1, you are required to both identify (choose a phonemic symbol) and judge (indicate the amount of rounding/spreading and indicate the place of constriction and the vowel height) each vowel token. You are expected to indicate which Dutch vowel was pronounced by clicking with the mouse on one of the vowel buttons. After you have identified the vowel token, you are to judge the linguistic quality of the vowel token by first clicking the mouse in the rounding/spreading area on the position corresponding to the amount of rounding or spreading of that particular vowel token and second by clicking the mouse on the position of your choice in the vowel quadrilateral. Once you have done these two things, the next vowel token to be judged is played.

The process of one experimental cycle (in which only one vowel token is to be identified and judged) is as follows. First, a short tone signal is played. After the tone signal, the vowel in the nonsense syllable is played. The nonsense syllable is played 10 times maximally, with one to one-and-a-half second intervals. After the vowel is played 10 times, the program waits until you have identified and judged the vowel token (by clicking on one of the vowel buttons, clicking in the rounding/spreading rectangle, and clicking in the vowel quadrilateral), before the next experimental cycle starts. This is a fixed order: first the vowel token has to be identified and then judged. In this sub-experiment, the next cycle starts immediately after you have clicked in the vowel quadrilateral.

You do not have to wait until the vowel token was played the full 10 times, you can start making your choice after the first time the vowel was played.

The sub-experiment consists of 23 blocks of nine vowel tokens. Between these blocks, a screen appears with the word ‘pauze’ [pause], plus a number. This number corresponds to the number of the pause screen (there are eight pause screens). Whenever this screen appears you have an option of taking a short break. If you want to commence, all you have to do is click with the mouse, and the next block commences.

The first five stimuli are practice trials. After these trials the real experiment starts. You can take a short break after the practice trails, to ask any questions, if you wish. Thanks.

Appendix D

Reliability per listener

Table D.1: *Intra-listener reliability for listeners 1-92 per vowel category across the three sub-experiments; Cochran's α calculated per listener per vowel category for Height, Advancement, and Rounding.*

Listener 1				
Cochran's α	Height	Advancement	Rounding	Mean
/ɑ/	0.141	0.827	-0.581	0.129
/a/	0.672	0.598	0.219	0.496
/ɛ/	0.415	0.164	-0.181	0.133
/ɪ/	0.725	0.597	0.416	0.579
/i̇/	-0.144	0.458	-0.332	-0.006
/ɔ/	0.456	0.476	0.494	0.475
/ʊ/	0.153	0.560	0.207	0.307
/ʏ/	0.554	0.399	0.132	0.362
/y/	0.211	0.325	0.154	0.230
Mean	0.354	0.489	0.0587	0.301
Listener 2				
/ɑ/	0.472	0.760	0.364	0.532
/a/	0.693	0.860	-0.152	0.467
/ɛ/	0.869	0.477	0.007	0.451

Continued on next page

/ɪ/	0.295	-0.054	-0.119	0.041
/i̇/	-0.210	0.670	0.450	0.303
/ɔ/	0.889	0.361	0.655	0.635
/ʉ/	0.699	0.392	0.417	0.503
/ʏ/	0.474	0.682	0.629	0.595
/y/	0.485	0.662	0.254	0.467
Mean	0.518	0.534	0.2782	0.443
Listener 3				
/ɑ/	0.561	0.730	0.605	0.632
/a/	0.266	0.725	0.648	0.546
/ɛ/	0.718	0.672	0.533	0.641
/ɪ/	0.465	0.473	0.283	0.407
/i̇/	0.071	0.690	0.491	0.417
/ɔ/	0.137	0.890	0.823	0.617
/ʉ/	0.459	0.112	0.450	0.340
/ʏ/	0.549	0.099	0.318	0.322
/y/	0.599	0.693	0.613	0.635
Mean	0.425	0.565	0.529	0.506
Listener 4				
/ɑ/	-1.016	-0.969	0.289	-0.565
/a/	0.291	0.626	0.125	0.347
/ɛ/	-1.035	0.037	-0.242	-0.413
/ɪ/	-0.008	0.535	0.189	0.239
/i̇/	0.397	0.156	0.082	0.212
/ɔ/	-0.152	0.158	-0.062	-0.019
/ʉ/	-0.792	0.221	0.211	-0.120
/ʏ/	-0.458	-0.040	-0.031	-0.176
/y/	-0.772	0.480	-0.087	-0.126
Mean	-0.394	0.134	0.0528	-0.069
Listener 5				
/ɑ/	0.632	0.785	0.782	0.733
/a/	0.615	0.759	0.264	0.546

Continued on next page

/ɛ/	0.890	0.672	0.266	0.609
/ɪ/	0.769	-0.111	0.089	0.249
/i̇/	0.480	0.568	0.279	0.442
/ɔ/	0.845	0.834	-0.134	0.515
/ʉ/	0.463	0.695	0.030	0.396
/ʏ/	0.659	0.322	0.366	0.449
/y/	0.341	0.239	-0.055	0.175
Listener 6				
/ɑ/	0.499	0.919	0.650	0.689
/a/	0.615	0.255	0.279	0.383
/ɛ/	0.563	0.445	0.737	0.582
/ɪ/	-0.341	0.013	0.170	-0.053
/i̇/	0.599	0.046	-0.049	0.199
/ɔ/	0.650	0.645	-0.332	0.321
/ʉ/	0.574	0.666	0.385	0.542
/ʏ/	0.724	0.262	0.121	0.369
/y/	-0.120	0.057	0.557	0.165
Mean	0.418	0.367	0.280	0.355
Listener 7				
/ɑ/	0.434	0.815	-0.202	0.349
/a/	0.379	0.686	0.628	0.564
/ɛ/	0.688	0.559	0.095	0.447
/ɪ/	0.540	0.627	-0.624	0.181
/i̇/	-0.043	-0.627	-0.047	-0.239
/ɔ/	0.053	0.444	0.460	0.319
/ʉ/	0.445	0.332	-0.321	0.152
/ʏ/	0.623	0.453	0.010	0.362
/y/	0.355	0.823	0.013	0.397
Mean	0.386	0.457	0.001	0.281
Listener 8				
/ɑ/	0.041	0.308	0.085	0.145
/a/	-0.287	0.245	-0.331	-0.124

Continued on next page

/ɛ/	0.797	0.619	-0.553	0.288
/ɪ/	0.340	0.380	0.142	0.287
/i̇/	0.432	0.076	-0.514	-0.002
/ɔ/	-0.051	0.486	-0.054	0.127
/ʉ/	0.211	-0.740	0.506	-0.008
/ʏ/	0.519	0.246	-0.197	0.189
/y/	0.279	-0.027	-1.216	-0.321
Mean	0.254	0.177	-0.237	0.065
Listener 9				
/ɑ/	0.356	-0.150	0.074	0.093
/a/	0.157	0.631	0.451	0.413
/ɛ/	0.576	0.195	0.310	0.360
/ɪ/	0.614	0.208	0.803	0.542
/i̇/	0.490	0.525	0.108	0.374
/ɔ/	0.476	0.901	0.002	0.460
/ʉ/	0.478	0.533	-0.302	0.236
/ʏ/	0.424	0.542	0.549	0.505
/y/	-0.393	-0.008	0.722	0.107
Mean	0.353	0.375	0.302	0.343

Samenvatting (Summary in Dutch)

Het in dit proefschrift beschreven onderzoek werd uitgevoerd in het kader van een overkoepelend sociolinguïstisch onderzoeksproject, het zogenaamde VNC-project (Vlaams-Nederlands Comité). Eén van de doelen van het VNC-project was het in kaart brengen van de akoestische variatie in de uitspraak van Standaardnederlandse klinkers in Nederland en Vlaanderen.

De akoestische representatie van gesproken klinkers variëert doorgaans afhankelijk van de bedoelde klinker en de spreker. Deze representatie bevat onder andere de volgende drie akoestische variatiebronnen:

• *Fonemische* variatie (variatie tussen verschillende klinkers, gerelateerd aan de door de spreker bedoelde klinker),
• *Sociolinguïstische* sprekerspecifieke variatie (gerelateerd aan sociologische eigenschappen van de spreker zoals de regionale herkomst),
• *Anatomisch/fysiologische* sprekerspecifieke variatie (gerelateerd aan bijvoorbeeld de sekse, lichaamsgrootte of leeftijd van de spreker).

Fonemische variatie kan ook wel omschreven worden als tussenklinkervariatie (bijvoorbeeld het verschil tussen de klinkers “i” als in ‘pit’ en “e” als in ‘pet’) en de sociolinguïstische variatie als binnenklinkervariatie (bijvoorbeeld het variatie in uitspraak van de klinker “i” in ‘pit’)[80]. In het VNC-project was men vooral geïnteresseerd in de eerste twee variatiebronnen in het akoestisch signaal, dwz. de fonemische en de sociolinguïstische variatie, en minder in de anatomisch/fysiologische variatie.

In sociolinguïstisch onderzoek is het gebruikelijk al luisterend fonetische transcripties te maken om linguïstisch relevante verschillen in uitspraak binnen en tussen klinkers vast te stellen. Deze transcripties worden gemaakt door fonetisch getrainde expertluisteraars. Deze

[80] Als dit bijvoorbeeld wordt uitgesproken door iemand uit (Nederlands) Zuid-Limburg, klinkt de “i” meer als de “e” in ‘pet’).

expertluisteraars kunnen in spraak zowel variatie tussen als variatie binnen klinkers duiden.

Fonetische transcriptie heeft echter als nadelen dat het handwerk is en daardoor zeer tijdrovend en dat weinig bekend is over de repliceerbaarheid van de resultaten. Een alternatief voor de handmatige transcriptie is het semi-automatisch meten van relevante frequentiewaarden in de akoestisch representatie van de klinker. Deze semi-automatische methoden worden in de fonetiek gebruikt om uitspraakvariatie binnen en tussen klinkers te meten. Deze methoden zijn minder tijdrovend en optimaal repliceerbaar. Zo kan een klinker akoestisch worden gerepresenteerd als een verzameling relevante frequentiewaarden, ook wel *formantfrequenties* genoemd, gemeten aan het akoestisch signaal. Bijvoorbeeld, voor een "aa" uitgesproken door een mannelijke spreker zijn de formantfrequenties ruwweg 700 hertz, voor F_1, de eerste formant, 1200 hertz, voor F_2, en 2200 hertz, voor F_3. De toonhoogte, of F_0 is meer afhankelijk van de spreker dan van de klinker, en varieert voor een man ruwweg tussen de 70 en 150 hertz.

Eén van de nadelen van formantmeetmethoden is echter dat de gemeten formantfrequenties niet altijd direct gebruikt kunnen worden, omdat onder andere de anatomisch/fysiologische variatie het zicht op de andere twee typen variatie verstoort. In de fonetiek zijn daarom zogenaamde formantnormalisatieprocedures ontwikkeld. Het doel van deze procedures is de formantwaarden zo om te rekenen, dat de sprekergerelateerde variatie wordt geëlimineerd en de fonemische variatie bewaard blijft. Echter, het is niet duidelijk hoe deze procedures over het algemeen omgaan met de drie variatiebronnen in het akoestisch signaal.

Dit proefschrift beschrijft hoe ik heb onderzocht welke in de fonetische literatuur reeds voorgestelde normalisatieprocedures het meest geschikt zijn om toe te passen in sociolinguïstisch onderzoek naar taalvariatie. Dit heb ik gedaan door uit te zoeken hoe deze procedures omgaan met de drie variatiebronnen. De centrale vraag is: welke normalisatieprocedure transformeert formantfrequenties dusdanig dat de fonemische variatie en de sociolinguïstische sprekerspecifieke variatie het best bewaard blijven en de anatomisch/fysiologische sprekerspecifieke variatie geminimaliseerd wordt?

De normalisatieprocedures zijn geëvalueerd middels een akoestische vergelijking en een perceptueel-akoestische vergelijking. De akoestische vergelijking werd uitgevoerd door de procedures toe te passen op de waarden van de eerste drie formanten (F_1, F_2 en F_3) en van de fundamentele frequentie (F_0) van een grote hoeveelheid klinkerdata (2880 klinkers in totaal), afkomstig van de in totaal 160 sprekers uit het VNC-project. Vervolgens is er bepaald welke procedure het best scoorde in het bewaren van fonemische en sociolinguïstische variatie en het elimineren van anatomisch/fysiologische variatie. De perceptueel-akoestische vergelijking hield in dat resultaten van de twaalf normalisatieprocedures (dwz., genormaliseerde formantfrequenties) van een set klinkerdata werden vergeleken met luisteroordelen van expertluisteraars op basis van diezelfde set klinkerdata. De procedure die resultaten produceerde die het meest leek op de luisteroordelen was de beste optie voor gebruik in sociolinguïstisch onderzoek, aangezien expertluisteraars heel goed in staat zijn

de anatomisch/fysiologische variatie te negeren en oordelen te geven over de fonemische en de sociolinguïstische variatie. De procedure die het beste presteerde in zowel de akoestische als de perceptief-akoestische vergelijking was het meest geschikt voor sociolinguïstisch onderzoek. In de rest van deze samenvatting beschrijf ik hoe het onderzoek is uitgevoerd.

Selectie normalisatieprocedures

Hoofdstuk 2 beschrijft de selectie van de normalisatieprocedures. Bij deze selectie werden twee criteria gebruikt. Ten eerste moeten de procedures klinkers normaliseren door de formantfrequenties te transformeren (en niet door, bijvoorbeeld, een transformatie van het gehele amplitudespectrum). Ten tweede moeten de procedures al eens eerder vergeleken zijn met andere procedures. In totaal werden elf procedures geselecteerd, plus de zogenaamde baseline-procedure, de ruwe ongenormaliseerde data in hertz. Deze elf procedures plus de baseline werden vervolgens ingedeeld volgens het type informatie dat wordt gebruikt voor de normalisatie:

- *klinkerintrinsiek*: informatie van één klinker,
- *klinkerextrinsiek*: informatie van meerdere klinkers van een spreker,
- *formantintrinsiek*: informatie van één formant,
- *formantextrinsiek*: informatie van meerdere formanten binnen één klinker.

De eerste van de vier groepen bestaat uit de klinkerintrinsieke/formantintrinsieke procedures. De procedures in deze groep maken dus gebruik van informatie van één klinker en van één formant. In totaal heb ik vier van deze procedures getest. Deze procedures, LOG, BARK, MEL en ERB, zijn zogenaamde schaaltransformaties.

De tweede groep zijn de klinkerintrinsieke/formantextrinsieke procedures. Deze procedures maken gebruik van informatie van één klinker en van meerdere formanten van die klinker (bijvoorbeeld het verschil in frequentie tussen twee opeenvolgende formantfrequenties: F_2 minus F_1). Ik heb één klinkerintrinsieke/formantextrinsieke procedures geëvalueerd. In deze procedure, SYRDAL & GOPAL, wordt eerst een barktransformatie (met de eerder genoemde procedure BARK) toegepast. Het kenmerkende is dat de klinkers worden gerepresenteerd als afstanden tussen formanten: $F_1 - F_0$ in barks representeert eerste dimensie en $F_3 - F_2$ representeert de tweede dimensie van de klinker.

De derde groep zijn de klinkerextrinsieke/formantintrinsieke procedures. Deze procedures maken gebruik van informatie op basis van meerdere klinkers van een spreker en informatie van één formant. Deze groep bevat drie procedures: LOBANOV, GERSTMAN en CLIH_{i4}. Zo wordt een klinker in LOBANOV's procedure als volgt gerepresenteerd. Eerst worden per formant voor alle (beschikbare) klinkers van één spreker het gemiddelde en de standaarddeviatie uitgerekend. Vervolgens worden de individuele formantfrequenties uitge-

drukt in gestandaardiseerde z-scores ten opzichte van dat gemiddelde.

De vierde groep zijn de klinkerextrinsieke/formantextrinsieke procedures. Deze procedures maken gebruik van informatie van meerdere klinkers van een spreker en van informatie van meerdere formanten binnen een klinker. Deze groep bestaat uit drie procedures: NORDSTRÖM & LINDBLOM, MILLER en CLIH$_{s4}$. Zo normaliseert NORDSTRÖM & LINDBLOM een klinker door elke formant apart te transformeren met een correctiefactor. Deze correctiefactor corrigeert voor de verschillen in de lengte van het spraakkanaal (dwz. de keel- neus- en mondholte). Deze correctiefactor is zelfs gebaseerd op de gemiddelde waarden van de F_3 van alle sprekers in de groep.

Literatuurstudies

In hoofdstuk 2 wordt naast de beschrijving van de normalisatieprocedures ook een aantal artikelen uit de fonetische literatuur besproken waarin eerdere vergelijkingen van normalisatieprocedures beschreven staan. Deze literatuurstudie werd uitgevoerd om een eerste indruk te verkrijgen over hoe goed de procedures presteerden in eerdere akoestische vergelijkingen. De conclusie van deze literatuurstudie was dat er een uitputtende akoestische vergelijking van alle twaalf de normalisatieprocedures nodig was. Deze vergelijking diende uitgevoerd te worden op een database van spraakmateriaal waar naast de fonemische en de anatomisch/fysiologische variatie ook sociolinguïstische variatie aanwezig is.

Hoofdstuk 3 bespreekt experimenteel-fonetische en -psychologische literatuur over perceptuele klinkernormalisatie. Ten eerste worden artikelen besproken die de rol beschrijven van F_0, F_1, F_2 en F_3 voor luisteraars in luisterexperimenten met als taak klinkercategorisatie. Al deze experimenten maakten gebruik van gesynthetiseerde stimuli. Ten tweede wordt in dit hoofdstuk een overzicht gegeven van een aantal artikelen waarin experimenten met natuurlijke (voorgelezen) stimuli worden beschreven. In deze experimenten moesten de proefpersonen klinkers categoriseren in experimentele sprekergeblokte en sprekergemengde condities waarin ze meer (gegroepeerd per spreker) of minder (verschillende sprekers door elkaar) informatie over de spreker van de klinkers tot hun beschikking hadden. Ten derde werd een aantal artikelen besproken waarin het beoordelingsgedrag van expertluisteraars centraal stond. Op basis van dit literatuuronderzoek werd geconcludeerd dat het nodig was vast te stellen hoe betrouwbaar expertluisteraars de (sociolinguïstische) variatie binnen klinkers kunnen beoordelen en hoe de beschikbaarheid van meer of minder informatie over de spreker de luisteroordelen beïnvloedt.

Hoofdstuk 4 beschrijft het onderzoeksdesign van dit proefschrift. Er wordt beschreven hoe de conclusies uit de twee literatuurstudies uit de hoofdstukken 2 en 3 zijn meegenomen in dit design.

Spraakmateriaal

Het spraakmateriaal werd verzameld in het kader van het VNC-project onderzoek en staat in detail beschreven in Hoofdstuk 5. Het spraakmateriaal bestaat uit de negen monoftonge Nederlandse klinkers, /ɑ, a, ɛ, ɪ, i, ɔ, u, ʏ, y/ in een "neutrale" doelsyllabe: /sVs/ ('V' representeert één van de negen klinkers). Deze klinkers waren voorgelezen door 160 sprekers, jonger en ouder, mannen en vrouwen, allen leraren Nederlands. Van de 160 sprekers waren 80 afkomstig uit vier regio's in Vlaanderen en 80 uit vier regio's uit Nederland. Per regio waren er 20 sprekers, vijf jongere vrouwen, vijf jongere mannen, vijf oudere vrouwen en vijf oudere mannen. Elk van deze sprekers sprak de doelsyllabe twee keer uit (9 klinkers $\times$ 160 sprekers $\times$ 2 keer uitspreken = 2880 klinkers in totaal).

Akoestische meetwaarden

Hoofdstuk 6 bespreekt hoe de waarden van de formanten (F_1, F_2 en F_3) en de toonhoogte (F_0) werden gemeten voor elke klinker. De toonhoogtemetingen werden verricht met het programma Praat (Boersma & Weenink, 1996). De formantmetingen werden verkregen met een ander programma, geschreven door Nearey (2002). Voor elke klinker werden de formantwaarden gemeten op het temporele midden.

Akoestische vergelijking

Hoofdstuk 7 beschrijft de akoestische vergelijking van de twaalf normalisatieprocedures toegepast op de meetwaarden van de klinkers van alle 160 sprekers. Drie series vergelijkingen werden uitgevoerd.

De eerste serie vergelijkingen werd uitgevoerd om te vast te stellen hoe goed de fonemische variatie bewaard bleef in de getransformeerde formant- en toonhoogtewaarden. Dit werd getest middels een serie lineaire discriminantanalyses. In deze discriminantanalyses werd vastgesteld hoe goed de 2880 genormaliseerde klinkers in te delen waren in hun bijbehorende klinkercategorie. Als het percentage correct ingedeelde klinkers hoog is, dan betekent dit dat de procedure er goed in geslaagd is, de fonemische variatie in de akoestische representatie van klinkers te bewaren.

In de tweede serie vergelijkingen werd vastgesteld hoe goed de procedures de anatomisch/fysiologische variatie minimaliseerden. In deze serie werd wederom een aantal discriminantanalyses uitgevoerd op de data, genormaliseerd volgens elke procedure. In deze discriminantanalyses werd gekeken of er na normalisatie nog anatomisch/fysiologische variatie in de data aanwezig was. Hiertoe werd gekeken hoe goed de genormaliseerde data te groeperen was op basis van de sekse en leeftijd van de spreker (beide bronnen van anatomisch/fysiologische variatie). Als de genormaliseerde data op kansniveau scoorde, betekende dit dat alle anatomisch/fysiologische variatie in het signaal geminimaliseerd was.

De derde serie vergelijkingen had als doel vast te stellen hoe goed de sociolinguïstische variatie bewaard bleef. Dit werd vastgesteld met discriminantanalyses en met multivariate variantieanalyses. In de discriminantanalyses werd geëvalueerd in hoeverre de procedure regionale variation (een bron van sociolinguïstische variatie) in het signaal bewaarde. Dit werd getest door te bepalen hoe goed de data zoals genormaliseerd volgens elk van de elf procedures in te delen was in de bijbehorende regionala achtergron van de spreker. Hoe hoger het percentage correct ingedeelde klinkers was, des te beter bewaarde de procedure de sociolinguïstische variatie. In de multivariate variantieanalyses werd ongeveer hetzelfde getest; hier werd vastgesteld welke proportie van de variatie in de akoestische meetwaarden te verklaren viel op basis van de genormaliseerde akoestische data. Als de procedure bijvoorbeeld alle sociolinguïstische variatie bewaarde, dan was de proportie regionale variatie relatief hoog.

Uit de resultaten bleek dat de klinkerextrinsieke/formantintrinsieke procedures over het algemeen het beste presteerden, gevolgd door de klinkerextrinsieke/formantextrinsieke, klinkerintrinsieke/formantintrinsieke en klinkerintrinsieke/formantextrinsieke procedures.

Luisterexperiment

Het doel van het in Hoofdstuk 8 besproken luisterexperiment was tweeledig: ten eerste het genereren van luisteroordelen voor de perceptueel-akoestische vergelijking in Hoofdstuk 9 en ten tweede om inzicht te verkrijgen in de betrouwbaarheid van expertluisteraars.

In het luisterexperiment hadden in totaal 10 luisteraars de volgende taak. Ze kregen de stimulusklinker in de /sVs/-context te horen. Ze moesten eerst de klinker categoriseren als een van de negen Nederlandse monoftonge klinkers. Daarna moesten ze de waargenomen klinkerhoogte (graad van mondopening), tongpositie en lipronding beoordelen door achtereenvolgens in twee responsievelden met de muis te klikken op de volgens hen linguïstisch meest toepasselijke plaats. Eén van deze responsievelden was de IPA-vierhoek (International Phonetic Association), waarin de klinkerhoogte en de tongpositie beoordeeld werd. Klinkerronding werd beoordeeld in een tweede responsieveld. De gekozen klinkercategorie en de coördinaten van de twee responsievelden waren de luisteroordelen, ofwel de perceptuele representaties van de klinkers. De 180 stimuli waren de negen monoftonge klinkers uitgesproken door 20 sprekers uit de database van 160 sprekers. Alle 20 sprekers waren afkomstig uit dezelfde regio in Nederland (de Randstad).

De luisteraars moesten de 180 klinkers beoordelen in drie experimentele condities. In conditie 1 werden de klinkers en sprekers random aangeboden, de stimuli werden dus niet gegroepeerd per klinkercategorie of per spreker. In conditie 2 werden de stimuli per klinker geblokt aangeboden en de sprekers werden door elkaar gemengd aangeboden. In conditie 3 werden de stimuli gegroepeerd per spreker en gemixt per klinker gepresenteerd. Het idee was dat de luisteroordelen zouden worden beïnvloed door de hoeveelheid informatie over

de stem-en-spraakkarakteristieken van de spreker en over de identiteit van de klinker. Als de luisteraar bijvoorbeeld minder informatie over de spreker zou krijgen, zou de betrouwbaarheid van de oordelen lager kunnen worden en andersom. Daarnaast zouden de oordelen ook kunnen variëren onder invloed van de informatie over de klinker. Als de luisteraar meer informatie over de identiteit van de klinker heeft, dan zou de luisteraar meer ruimte kunnen gaan gebruiken in de responsievelden, en andersom. De expertluisteraars moesten dus het experiment in totaal drie keer uitvoeren.

De oordelen van de zeven betrouwbaarste expertluisteraars werden geselecteerd als perceptuele representatie. De gemiddelde waarden van de luisteroordelen (waargenomen klinkerhoogte, tongpositie en lipronding) voor de 180 beoordeelde klinkers, werden gebruikt voor de in Hoofdstuk 9 beschreven perceptueel-akoestische vergelijking.

De resultaten van het experiment lieten zien dat de luisteraars de tussenklinkervariatie zeer betrouwbaar beoordeelden, terwijl de betrouwbaarheid van de oordelen van de binnenklinkervariatie lager was. Dit laatste kwam waarschijnlijk doordat de aangeboden stimuli te weinig sociolinguïstische variatie bevatten, aangezien alle sprekers uit dezelfde regio kwamen. Daarnaast was het experiment redelijk vermoeiend voor de expertluisteraars, en hadden de luisteraars geen mogelijkheid hun oordelen met eerder in het experiment gegeven oordelen te vergelijken. Deze laatste twee factoren kunnen geleid hebben tot meer ruis in de data.

Met betrekking tot de drie experimentele condities waren de resultaten als volgt. De statistische analyse liet zien dat de luisteraars zich niet substantieel hadden laten beïnvloeden door het beschikbaar maken van meer of minder informatie over de stem-en-spraakkarakteristieken van de spreker en over de identiteit van de klinker. Deze resultaten wijzen erop dat luisteraars in staat zijn betrouwbare oordelen over waargenomen klinkerhoogte, tongpositie en lipronding te geven op basis van de informatie die besloten ligt in één enkele klinker.

Perceptueel-akoestische vergelijking

Hoofdstuk 9 beschrijft de perceptueel-akoestische vergelijking van de normalisatieprocedures met de perceptuele oordelen. De perceptuele representatie van de 180 klinkers is dezelfde als die uit Hoofdstuk 8. De akoestische representatie werd verkregen door elk van de twaalf procedures toe te passen op de 180 klinkers die beoordeeld waren door de expertluisteraars. Op deze wijze werden twaalf sets genormaliseerde waarden van F_0, F_1, F_2 en F_3 (gemeten in Hoofdstuk 6) verkregen. Elk van deze sets werd vergeleken met de perceptuele oordelen middels lineaire regressieanalyse.

De resultaten voor de fonemische variatie, of tussenklinkervariatie laten zien dat formantfrequenties, genormaliseerd volgens de klinkerextrinsieke/formantintrinsieke procedures het meest leken op de perceptuele oordelen, gevolgd door de klinkerextrinsieke/formantextrinsieke, klinkerintrinsieke/formantintrinsieke en tenslotte de klinkerintrinsieke/formantextrinsieke procedures. De resultaten voor de sociolinguïstische variatie laten daarentegen zien dat

de door de expertluisteraars beoordeelde variatie binnen klinkers veel minder goed viel te voorspellen op basis van de twaalf akoestische akoestische representaties dan de fonemische variatie.

Algemene conclusies

De algemene conclusie luidt dat normalisatieprocedures zeer goede resultaten opleveren voor sociolinguïstisch onderzoek naar taalvariatie in klinkers. De beste normalisatieprocedures uit de hoofdstukken 7 en 9, de drie klinkerextrinsieke/formantintrinsieke procedures LOBANOV, GERSTMAN en CLIH$_{i4}$, bewaarden de fonemische en de sociolinguïstische variatie en minimaliseerden de anatomisch/fysiologische variatie.

De resultaten van de perceptueel-akoestische vergelijking laten zien dat waargenomen fonemische variatie goed voorspeld kan worden op basis van akoestische meetwaarden die genormaliseerd zijn met procedures die gebruik maken van informatie van meerdere klinkers en binnen één formant.

Tenslotte geven de resultaten van het onderzoek inzicht over welke informatie nuttig is voor het normaliseren van ruwe akoestische meetwaarden voor gebruik in sociolinguïstisch onderzoek. Het gebruik van klinkerextrinsieke informatie werkt beter dan klinkerintrinsieke informatie en formantintrinsieke informatie werkt aanzienlijk beter dan formantextrinsieke informatie.

Het onderzoek gaf ook meer inzicht in het gebruik van klinkerintrinsieke (binnen één klinker) en klinkerextrinsieke (meerdere klinkers) informatie door expertluisteraars. Luisteraars lijken alleen gebruik te maken van klinkerintrinsieke informatie voor het geven van luisteroordelen. De klinkerextrinsieke informatie leek hun oordelen niet direct te beïnvloeden.

Curriculum Vitae

Patti Adank was born December 8th, 1972 in Ommen, the Netherlands. She spent her childhood in Ommen, Almelo, Heemstede, and Haarlem, all in the Netherlands. In 1992 she graduated from the Atheneum College Hageveld in Heemstede. She studied Phonetics at the University of Utrecht, the Netherlands, and got her "Doctoral Diploma" (equivalent to an MA degree) in Phonetics in October 1997. From February 1998 to September 2002, she was employed as a PhD student (OiO) at the Department of General Linguistics and Dialectology at the University of Nijmegen, the Netherlands. She worked on the Flemish-Dutch cooperation project (VNC) "The pronunciation of Standard Dutch: variation and varieties in Flanders and The Netherlands". Parts of the research described in this thesis were carried out at the Department of Linguistics of the University of Alberta, Edmonton, Canada. She stayed there from July 2000 to January 2001 as a visiting researcher and worked with Prof. T. Nearey. From September 2002, she has been employed as a researcher at the Speech Processing Expertise Center (SPEX) at the Department of Language and Speech, at the University of Nijmegen.