

# Inside the black box of human vocal learning: A simulation approach

Anqi Xu

A dissertation submitted in  
fulfilment of the requirements for the degree of  
Doctor of Philosophy

to

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

University College London (UCL)

2022

## **Declaration**

I, Anqi Xu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Anqi Xu



## **Abstract**

Children learn to speak despite age-related anatomical differences that give rise to significant discrepancies between their vocalisations and those of adults. How children overcome these obstacles without explicit instructions remains unclear. Influential accounts suggest that vocal learning is achieved by producing sounds to match auditory memory in both songbirds and humans. However, observational studies alone cannot determine whether auditory-guided learning is the key mechanism. Here, I use computational modelling to test the feasibility of the hypothesis, by training an articulatory synthesiser with three-dimensional vocal tract models of an adult and children at different ages to simulate the learning of English words. The model involves two kinds of auditory guidance: 1) acoustic features to simulate universal perception of phonetic differences in all languages, and 2) a deep-learning-based automatic phoneme recogniser to simulate language-specific perception of sound contrasts in native languages. The results show that words trained by the automatic phoneme recogniser were more intelligible than those trained by acoustic features in the listening experiments, showing that language-specific perception can resolve the long-standing problem of anatomical differences between speakers. It demonstrates that auditory-guided learning is indeed feasible. In contrast with previous simulation attempts that were limited to vowel acquisition, the current model learned words containing consonant-vowel sequences that approach the intelligibility of natural speech. It has also found that the embodied articulatory dynamics limited the scope of vocal practice and somatosensory feedback provided additional benefits. Yet, learning was better and easier by the adult than by the child articulatory systems. The model experienced challenges in learning certain speech sounds, resembling the patterns of child speech development. The study further suggests that it is the vocal learning process that helps forge the link between speech perception and production. The computational approach opens a new path towards examining the cognitive mechanisms behind vocal learning.

## **Impact statement**

The study presents a highly effective vocal learning model that shows how an infant may learn to speak without explicit instructions. I constructed a biologically-plausible model that realistically simulates vocal learning at different ages. The methodology contains two key innovations: 1) sensory feedback and articulatory dynamics were explicitly modelled, which has been largely overlooked in previous computation models and, 2) listening experiments were used to evaluate the simulation performance, setting a new standard for further simulation studies, and showing a potential for quantitative hypothesis testing. By this framework, it is demonstrated that learners can use speech perception that encodes native-language sound categories to self-guide vocal learning, showing a striking parallel to songbirds and some mammals. More broadly, the findings have opened a window into the cognitive mechanism underpinning language acquisition, one of the most mysterious aspects of being human. The modelling approach has demonstrated the feasibility of a non-invasive way of investigating vocal development in all animals that show a vocal learning behaviour by implementing vocal systems with different anatomy. The mechanism-driven simulation can make use of recent advances in artificial intelligence to reveal mysteries in human intelligence.

In addition, the study has important implications for non-academic fields as well. First of all, these findings have revealed the indispensable role of auditory experience in speech acquisition, that is, the capability to perceive phonetic contrast is the key to production learning. This may carry implications for computational tools that predict speech development, advancing early diagnosis and intervention for speech pathology, such as autism spectrum disorders and special language impairment. Secondly, the study offers new insight into the development of assistive communication devices for people with motor speech disorder such as anarthria, dysarthria, and apraxia. The model transforms phonetic goals to speech sounds by articulatory synthesis, which can be combined with neural decoding techniques to restore speech. In contrast with the state-of-art speech synthesis, the present model is less reliant on huge

amounts of speech data and extensive computation time. The success in generating intelligible words suggests its potential application in speech synthesis of low-resource languages. Moreover, the study directly contributes to the development of articulatory synthesis, showing a possibility to be incorporated into commercial speech synthesis systems. Finally, this work is one of many contributions to open science. An online repository has been created to host a demonstration video, experiment stimuli and codes used to reproduce the results: [https://gitlab.com/Anqi\\_Xu/evoc\\_learn](https://gitlab.com/Anqi_Xu/evoc_learn) (Appendix Figure A).

## Acknowledgements

I would like to express my gratitude to my principal supervisor Prof. Yi Xu for his patience and continuous guidance throughout my PhD. Your enthusiasm towards unknown scientific fields and your determination to challenge conventional research have inspired me a lot. The journey would have been overwhelming without your tremendous support. Words cannot express my appreciation. Thank you, Prof. Xu.

I also wish to thank the entire Evoc-learn (electronic/early vocal learning) team, including Daniel, Branko, Paul, Peter, San and Lorna, who have offered extensive help with constructing and optimising the vocal learning model. I would like to say a special thank you to Peter for developing VocalTractLab. It is an amazing tool for Speech Sciences, which has made both my Master and PhD project possible.

My sincere thanks go to Mark, my secondary supervisor for the valuable suggestions. I enjoyed being a teaching assistant at your course and I have learned so much from you. Thanks to all the faculty members at SHaPS for all the scientific advice and help.

Many thanks to Albert and Yuyin for guiding me to do research step by step and for showing me how to be a proper researcher. I may not have the opportunity to pursue my PhD without your help.

Thanks to my friends and colleagues at Chandler House, including Shego, Clara, Gwijde, Julie, Tony, Xiao, Han, Chengxia, Yue, Rachel, Ana, Anna, Shiran, Katharina, Max, Elise, Bryony, Jonas, Begona, Hannah, Magda, Faith, Dan, Gwen, Giulia, and many more.

I would like to thank James White and Roger Moore for accepting the viva invitation and spending your precious time reading the thesis.

Thanks to my dearest friends, Clara, Shego, Wentian, Lake, Han, Chengsheng, Chengde, Tingting, Xiaoqing, Wanling and Shanshan. Without your emotional

support, I could not have survived the lockdown during the COVID-19 pandemic. Meeting all of you is one of the best things that ever happened to me.

I would like to thank UCL Overseas Research Scholarship and China Scholarship Council for the financial awards.

This thesis is dedicated to my family, Jianyu and Aiqing for always being there to support me. I love you.

# TABLE OF CONTENTS

<b>LIST OF FIGURES .....</b>	<b>12</b>
<b>LIST OF TABLES .....</b>	<b>19</b>
<b>CHAPTER 1 INTRODUCTION .....</b>	<b>20</b>
<b>1.1 Background.....</b>	<b>20</b>
<b>1.2 Vocal learning in animals .....</b>	<b>22</b>
1.2.1 Songbirds .....	22
1.2.2 Mammals .....	24
1.2.3 Summary.....	27
<b>1.3 Child speech development .....</b>	<b>28</b>
1.3.1 Background .....	28
1.3.2 Perception development .....	29
1.3.3 Production development.....	32
1.3.3.1 Background.....	32
1.3.3.2 Development stages .....	34
1.3.3.3 Vowel development.....	36
1.3.3.4 Consonant development .....	38
1.3.4 Summary.....	39
<b>1.4 Sensorimotor interaction .....</b>	<b>40</b>
1.4.1 Sensorimotor learning .....	40
1.4.2 Imitative learning.....	41
1.4.3 Summary.....	43
<b>1.5 Research questions and thesis outline.....</b>	<b>44</b>

<b>CHAPTER 2</b>	<b>REVIEW OF VOCAL LEARNING MODELS .....</b>	<b>47</b>
<b>2.1</b>	<b>Learning strategy.....</b>	<b>47</b>
2.1.1	Neurobiological models .....	48
2.1.2	Acoustic imitation .....	50
2.1.3	Infant–caregiver interaction .....	53
2.1.4	Reinforcement learning .....	56
2.1.5	Goal babbling .....	57
2.1.6	Self-organisation .....	58
2.1.7	Summary .....	60
<b>2.2</b>	<b>Speech sensory system.....</b>	<b>61</b>
2.2.1	Background .....	61
2.2.2	Mechanisms behind speech perception .....	62
2.2.3	Simulation of speech sensory system .....	65
<b>2.3</b>	<b>Speech motor control .....</b>	<b>68</b>
2.3.1	Background .....	69
2.3.2	Coarticulation.....	70
2.3.3	The synchronised dimension-specific sequential target approximation model	72
<b>CHAPTER 3</b>	<b>SIMULATION OF VOCAL LEARNING.....</b>	<b>73</b>
<b>3.1</b>	<b>Model overview .....</b>	<b>74</b>
<b>3.2</b>	<b>Motor control.....</b>	<b>76</b>
3.2.1	Articulatory synthesis .....	76
3.2.2	Articulator dynamics .....	78
<b>3.3</b>	<b>Sensory system .....</b>	<b>81</b>
3.3.1	Acoustic features .....	82
3.3.2	Automatic word recognition .....	83

3.3.3	Automatic phoneme recognition.....	84
3.3.4	Somatosensory feedback.....	87
<b>3.4</b>	<b>Optimisation algorithm .....</b>	<b>88</b>
<b>3.5</b>	<b>Model evaluation.....</b>	<b>90</b>
3.5.1	Listening experiments.....	90
3.5.2	Participants.....	91
3.5.3	Analysis .....	92
<b>CHAPTER 4</b>	<b>SENSORY AND MOTOR SYSTEMS IN VOCAL LEARNING.....</b>	<b>93</b>
<b>4.1</b>	<b>Sensory feedback.....</b>	<b>93</b>
4.1.1	Acoustic features: Log Mel spectrograms vs. MFCCs .....	94
4.1.2	MFCCs vs. Automatic phoneme recogniser .....	97
4.1.3	Somatosensory feedback.....	104
4.1.4	Discussion .....	105
<b>4.2</b>	<b>Coarticulatory control .....</b>	<b>108</b>
4.2.1	Learned articulatory kinematics .....	109
4.2.2	Discussion .....	114
<b>CHAPTER 5</b>	<b>HUMAN VOCAL LEARNING.....</b>	<b>116</b>
<b>5.1</b>	<b>Adult vocal learning .....</b>	<b>116</b>
5.1.1	CVC words.....	117
5.1.2	CVCV words .....	124
<b>5.2</b>	<b>Child vocal learning.....</b>	<b>128</b>
5.2.1	CVC words.....	128
5.2.2	CVCV words .....	135
<b>5.3</b>	<b>Age-related vocal tract differences.....</b>	<b>139</b>



5.4	Discussion .....	145
CHAPTER 6	GENERAL DISCUSSION .....	152
APPENDIX.....		156
REFERENCES .....		173

## LIST OF FIGURES

Figure 1 Timeline of vocal learning in infants and songbirds (Kuhl, 2003). Image reproduced with permission of the rights holder, Annual review of neuroscience. ....	23
Figure 2 Child speech perception development. ....	31
Figure 3 Average F1–F2 acoustic space for American English males aged 4 years through adulthood (Vorperian & Kent, 2007). Image reproduced with permission of the rights holder, Journal of speech, language, and hearing research : JSLHR.....	34
Figure 4 Child speech production development. ....	36
Figure 5 English vowel development in 204 children aged 2–6 years using correct production by 75% of children in a particular age group as the criterion (Wellman et al., 1931).....	38
Figure 6 Average age of acquisition of American English consonants, adapted from Table 2 in Crowe & McLeod (2020). The arrow starts from 50% and ends with 90% criterion.....	39
Figure 7 Illustration of the DIVA model (Tourville & Guenther, 2011) © copyright 2022, reprinted by permission of Informa UK Limited, trading as Taylor & Taylor & Francis Group, <a href="http://www.tandfonline.com">http://www.tandfonline.com</a> . ....	49
Figure 8 Architecture of an acoustic imitation model (Prom-On et al., 2014a, Figure 1).....	53
Figure 9 Illustration of imitative learning paradigm for Elija (Howard & Messum, 2014).....	55
Figure 10 Illustration of goal babbling for speech acquisition (Philippsen, 2021, Figure 1).....	58
Figure 11 Simplified spectrograms of consonant /d/ followed by vowel /i/ and /u/, adapted from Liberman et al. (1954) .....	62
Figure 12 Narrow-band spectrograms of /bad/ produced by a female and a male native British speaker respectively. ....	62

Figure 13 Bark and Mel scale as a function of frequency from 0 to 8000 Hz. X axis shows the physical frequency in hertz and Y axis shows the normalised scale.....	66
Figure 14 Overview of the vocal learning model.....	75
Figure 15 Workflow of the motor control system. The system takes articulatory targets such as consonant or vowel targets as input and returns 17-dimensional vocal tract parameter trajectories (Table 2) to be passed to the articulatory synthesiser.....	79
Figure 16 Illustration of the synchronised dimension-specific sequential target approximation model in the case of bilabial stop-vowel sequences. Dashed lines represent the articulatory trajectories of the consonant target and solid lines represent the articulatory trajectories of the vowel target. ....	80
Figure 17 Confusion matrices of CVC words produced by a female native speaker, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. ....	84
Figure 18 Schematic diagram of an automatic phoneme recogniser (Niekerk et al., 2022; Xu et al., 2022).....	86
Figure 19 Illustration of the two-step optimisation process. Step 1 Exploration: Uniformed random parameter search; Step 2 Refinement: Random parameters search around good solutions.....	90
Figure 20 Confusion matrices of CVC words learned by adult vocal tract model when guided by Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.....	94
Figure 21 Confusion matrices of words learned by adult vocal tract model when guided by MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.....	95
Figure 22 Confusion matrices of consonants trained by Log Mel spectrograms and MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. ....	96
Figure 23 Confusion matrices of vowels trained by Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.....	97

Figure 24 Confusion matrices of vowels trained by MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.....	97
Figure 25 Confusion matrices of words learned by adult vocal tract model when guided by an automatic phoneme recogniser, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. ....	98
Figure 26 By-listener mean phoneme accuracy rates of CVC words learned by different vocal tract models in an open-vocabulary transcription experiment and a close-set transcription experiment. **** $P \leq 10^{-4}$ .....	99
Figure 27 Distribution of by-listener phoneme accuracy rates for synthetic CVC words learned by vocal tract models of different ages in three syllable positions, evaluated by an open-vocabulary transcription experiment. **** $P \leq 10^{-4}$ . ...	100
Figure 28 Distribution of by-listener phoneme accuracy rates for synthetic CVCV words learned by vocal tract models of different ages in four syllable positions, evaluated by an open-vocabulary transcription experiment. **** $P \leq 10^{-4}$ .....	101
Figure 29 Reaction time of American English listeners in an open-vocabulary transcription (A) and in a close-set transcription experiment (B). The vertical lines represent the median reaction time for two types of auditory feedback.	102
Figure 30 Relationship between phoneme identification rates and auditory feedback .....	103
Figure 31 Comparison of human identification with an automatic phoneme recogniser and MFCCs by target words. Perceptual scores are normalised based on the phoneme accuracies judged by human participants, the recogniser and MFCCs respectively. ....	104
Figure 32 Effect of somatosensory feedback on the phoneme errors of synthetic CVC words learned by an adult vocal tract model, evaluated by an automatic phoneme recogniser. ** $P \leq 10^{-2}$ .....	105
Figure 33 Midsagittal sections of the vocal tract shapes of bilabial stop-vowel sequences learned by an adult vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets. ....	110

Figure 34 Boxplots of the learned lip distance parameters of all bilabial stop-vowel sequences learned by an adult vocal tract model. ....	111
Figure 35 Midsagittal sections of the vocal tract shapes of alveolar stop-vowel sequences learned by an adult vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets. ....	112
Figure 36 Correlation between the learned tongue tip parameters in the horizontal and vertical positions of an adult vocal tract model.....	112
Figure 37 Midsagittal sections of the vocal tract shapes of velar stop-vowel sequences learned by an adult male vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets. ....	113
Figure 38 Correlation between the learned tongue body parameters in the horizontal and vertical positions of an adult vocal tract model.....	114
Figure 39 Waveforms and wide-band Mel-spectrograms of ‘bad’ produced by a native speaker and learned by an adult vocal tract model.....	117
Figure 40 Histograms and Kernel density plots of by-listener mean phoneme identification accuracy rates of CV syllables in target CVC words produced by a female native speaker and by an adult male vocal tract model in listening experiments. ....	118
Figure 41 By-listener phoneme accuracy rates of natural female speech and synthetic male speech in different syllable positions, evaluated by an open open-vocabulary transcription experiment and a close-set transcription experiment. **** $P \leq 10^{-4}$ .....	119
Figure 42 Mean identification rates of CVC words produced by a female native speaker and learned by an adult male vocal tract model in the onset position, vowel position and coda position in an open-vocabulary transcription experiment. ....	120
Figure 43 Mean identification rates of CVC words produced by a female native speaker and learned by an adult male vocal tract model in the onset position and vowel position in a close-set transcription experiment.....	121

Figure 44 Comparison between natural female speech and synthetic male speech. Confusion matrix (%) of CVC words produced by a female native speaker (a) and learned by an adult male vocal tract model (b), measured by a close-set transcription experiment. ....	123
Figure 45 Distribution of by-listener mean phoneme identification accuracy rates of CVCV words produced by a female native speaker and learned by an adult male vocal tract model in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners.....	125
Figure 46 By-listener phoneme accuracy rates of CVCV words produced by a native female speaker and learned by an adult male vocal tract model in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns $P > 0.05$ , * $P < 0.05$ , *** $P \leq 10^{-3}$ . ....	126
Figure 47 Mean identification rates of CVCV words produced by a female native speaker by an adult male vocal tract model in all the phoneme positions in an open-vocabulary transcription experiment. ....	127
Figure 48 Comparison between natural female speech and synthetic male speech. Confusion matrix (%) of CVCV words produced by a female American English native speaker (a) and learned by an adult male vocal tract model (b), measured in a close-set transcription experiment. ....	128
Figure 49 Distribution of by-listener mean phoneme identification accuracy rates of CVC words learned by a 1-year-old and a 3-year-old vocal tract model, tested in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners. ....	129
Figure 50 Boxplots of by-listener phoneme accuracy rates in different syllable positions of CVC words learned by a 1-year-old and a 3-year-old vocal tract model, measured in an open-vocabulary transcription experiment (a) and a close-set transcription experiment (b). ns $P > 0.05$ , * $P < 0.05$ , **** $P \leq 10^{-4}$ . ....	131
Figure 51 By-listener mean phoneme accuracy rates of utterances learned by a 1-year-old and a 3-year-old vocal tract models in the onset position, the vowel position, the coda position of the CVC words, measured by an open-vocabulary transcription task.....	132

Figure 52 Confusion matrices (%) of CVC words learned by a 1-year-old (a) and a 3-year-old vocal tract model (b), measured by a close-set transcription experiment. ....	134
Figure 53 Distribution of by-listener mean phoneme identification accuracy rates of CVCV words learned by a 1-year-old and 3-year-old vocal tract model, tested in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners. ....	136
Figure 54 By-listener phoneme accuracy rates of learned CVCV words learned by two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns $P > 0.05$ , *** $P \leq 10^{-3}$ . ....	137
Figure 55 Mean identification rates of CVCV words learned by two child vocal tract models in all the phoneme positions in an open-vocabulary transcription experiment. ....	138
Figure 56 Comparison between two vocal tract models. Confusion matrices of CVCV words regenerated by learned vocal tract parameters of a 1-year-old (a) and a 3-year-old vocal tract model (b), measured by a close-set transcription experiment. ....	139
Figure 57 By-listener mean phoneme accuracy rates of CV syllables learned by a 1-year-old, a 3-year-old and an adult male vocal tract model, evaluated by an open-vocabulary transcription experiment and a close-set transcription experiment. Error bars show standard errors. ns $P > 0.05$ , **** $P \leq 10^{-4}$ . ....	140
Figure 58 Phoneme accuracy rates of CVC words learned by an adult and two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. **** $P \leq 10^{-4}$ . ....	141
Figure 59 By-listener phoneme accuracy rates of CVCV words learned by an adult and two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns $P > 0.05$ , **** $P \leq 10^{-4}$ . ....	142
Figure 60 Recognition error distribution of 10-best CVC words evaluated by an automatic phoneme recogniser. ns $P > 0.05$ , **** $P \leq 10^{-4}$ . ....	143
Figure 61 By-listener phoneme accuracy rates of the CV syllables in CVC words and CVCV words learned by adult and child vocal tract models in an open-vocabulary transcription experiment. **** $P \leq 10^{-4}$ . ....	144

Figure 62 Comparison of vocal tract models of different ages. First formant (F1) and Second formant (F2) of vowels in CVC words with bilabial stops learned by a 1-year-old model (a), a 3-year-old model (b) and an adult male model (c). The squared IPA labels represent the median. The F1 and F2 were based on the vowel spectra calculated by VocalTractLab 2.3 (Birkholz, 2013). ..... 145



## LIST OF TABLES

Table 1 Target words for the vocal learning model.....	76
Table 2 Vocal tract parameters in the model. ....	78
Table 3 Vocal tract parameters controlled by the consonant.....	80
Table 4 Speech data extracted from LibriSpeech corpus (Panayotov et al., 2015) for training an automatic phoneme recogniser .....	87
Table 5 Transcript examples of phoneme insertion and deletion. Phonemes are labelled using CMU pronunciation dictionary (Carnegie Mellon University, 2022).....	93
Table 6 Mean and standard deviation (in parentheses) of natural and synthetic CVCV words in each phoneme position, evaluated by an open-vocabulary transcription experiment.....	126
Table 7 Mean (%) and standard deviation (%) of CVCV words learned by the two child models in each phoneme position, evaluated by the open-vocabulary transcription experiment.....	137

# Chapter 1 INTRODUCTION

## 1.1 BACKGROUND

Speech is a highly complex cognitive activity and often considered unique to humans that requires sophisticated control over multiple articulators including the tongue, the lips, the jaw and the larynx. It seems mysterious how babies learn to speak without explicit training. The vocal apparatus of an infant is distinct from that of the adult, more closely resembling that of a non-human primate (Lieberman et al., 1972). It is much shorter in length and smaller in size, rendering consistently higher resonance frequencies (formants) than adult speech. A challenge that an infant has to face is to produce vocalisations equivalent to adult speech, just like using an entirely different musical instrument to play the same note. This is known as the speaker normalisation problem (K. Johnson, 2005; K. Johnson & Sjerps, 2021) in the field of speech perception or the correspondence problem in sensorimotor learning (Brass & Heyes, 2005; Nehaniv & Dautenhahn, 2002). In particular, unlike other actions such as hand movements, there is very limited visual information to help to tackle the problem, as most of the articulators are hidden. So, how can an infant manage to link their own vocalisations to adult speech?

Also as skilled vocal learners, songbirds share the same attribute in vocal development with humans (Brainard & Doupe, 2002; Doupe & Kuhl, 1999b; Kuhl, 2003). Considerable evidence has shown that songbirds retain tutor songs in long-term memory (Funabiki & Konishi, 2003; Phan et al., 2006), and the memory may serve as an 'auditory template' for song evaluation (Keller & Hahnloser, 2009). It has been suggested that likewise in humans, speech production can be driven by speech perception through vocal mimicry (Kuhl, 2000). According to this view, an infant endeavours to match his own vocalisations with the auditory memory of previously heard speech sounds. More importantly, for both song birds (Konishi, 1965) and humans (Oller & Eilers, 1988), the lack of auditory input can

lead to severe impairment in their vocal development. What is yet unclear is the nature of the perceptual representation that guides vocal learning in humans.

Speech acquisition requires the learning of sensorimotor association that maps movements of various articulators with sensory goals. Although a large number of studies has shown the existence of the sensorimotor coupling, the ontogeny of the coupling remains dimly understood. Several learning mechanisms such as error-based learning and reinforcement learning have been proposed (Wolpert et al., 2011). It is also suggested that associative learning of correlated sensorimotor experience forges the linkage between motor and sensory systems (Cook et al., 2014; Heyes, 2001; Keysers & Perrett, 2004). The sensorimotor experience can be gained from self-observation of actions, synchronous actions and being imitated by social partners. Studies on infant sensorimotor learning such as crawling (van Elk et al., 2008) and stepping (de Klerk et al., 2015) have shown support for this account, but much less attention has been paid to speech acquisition. Moreover, we know remarkably little about the contribution of different kinds of sensorimotor experience to vocal learning.

Although there exist extensive amounts of observations and theoretical perspectives on vocal learning and sensorimotor learning, the emerging picture is still blurry as questions remain regarding the learning mechanisms. Computational approach is constructive in delineating the underlying cognitive mechanism because it provides a platform for the verification of different assumptions. If we can recreate the learning process by simulation, then it is possible to probe any component of particular relevance, which is sometimes neither practical nor ethical in behavioural experiments. Previous research has explored various possible approaches, based on neurobiological modelling (Kröger et al., 2014; Tourville & Guenther, 2011), acoustic imitation (Howard & Huckvale, 2005; Philippsen et al., 2014; Prom-On et al., 2014a, 2014b), caregivers' feedback (Acevedo-Valle et al., 2020; Messum & Howard, 2015; Miura et al., 2012), reinforcement learning (Warlaumont & Finnegan, 2016), self-motivation (Moulin-Frier et al., 2014) and goal babbling (Philippsen, 2021a;

Philippsen et al., 2016). However, so far there has been no clear demonstration of successful learning of intelligible words (see Appendix Table A Performance). In consequence, we are unable to identify which mechanisms are at play, nor can we examine the key aspects of learning quantitatively. In this study, I aim to construct a computational model that emulates the learning of intelligible English words, allowing an in-deep investigation into both speech sensory and motor control during speech acquisition. In this chapter, I will first introduce the research on vocal learning in animals, which shows striking parallels to humans. I will then discuss speech acquisition in the context of sensorimotor learning and present remaining research questions in the field of study.

## 1.2 VOCAL LEARNING IN ANIMALS

### 1.2.1 SONGBIRDS

Though speech is unique to humans, vocal learning has been found in other animals (Catchpole & Slater, 2008). Since the first spectrogram of songbirds made by Thorpe in 1954, their vocal learning behaviour has long been of great research interest. Songbirds are often regarded as an ideal model for studying human vocal learning (Brainard & Doupe, 2002; Doupe & Kuhl, 1999; Kuhl, 2003; Marler, 1970). Although songbirds learn species-specific notes, syllables, and prosodic features different from phonetic units in human speech, birdsong learning shows striking parallels to human speech learning in terms of the developmental stages (Marler, 1970; Doupe & Kuhl, 1999). As shown in Figure 1, both infants (J. S. Johnson & Newport, 1989; Lenneberg, 1967; Scovel, 2000) and songbirds (Marler & Tamura, 1964) have a critical period for learning, after which the acquisition of new sound sequences becomes difficult because the sensitivity to sensory experience is lowered (Doupe & Kuhl, 1999)

Song and speech acquisition both involve two main phases: a period of auditory extraction and memorisation of the auditory input, followed by a period of vocal mimicry of the auditory representations (Kuhl, 2003).

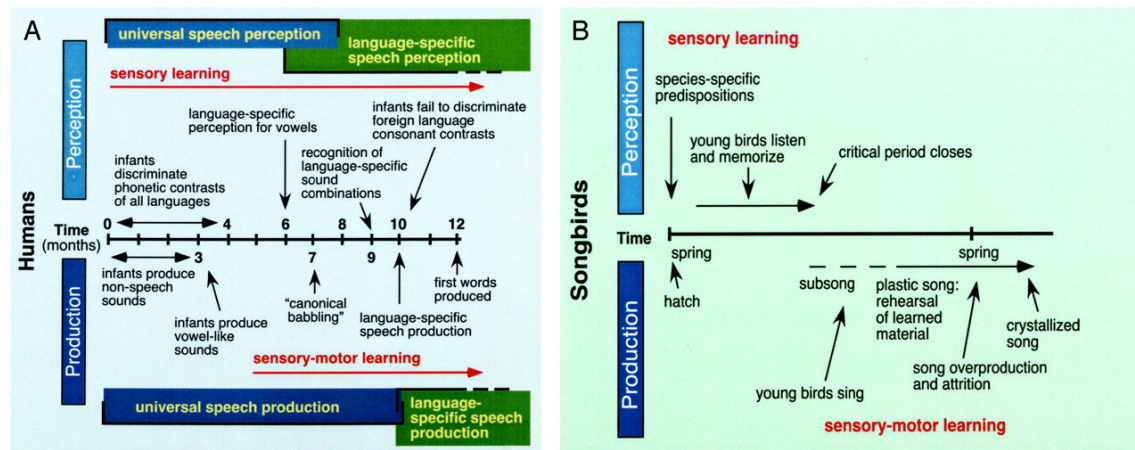


Figure 1 Timeline of vocal learning in infants and songbirds (Kuhl, 2003). Image reproduced with permission of the rights holder, Annual review of neuroscience.

During the sensory learning phase, songbirds listen to tutor songs and progressively develop an 'auditory template' (Marler, 1970). The 'auditory template' of the tutor song is retained in long-term memory, which can even survive after the perturbation of auditory feedback (Funabiki & Konishi, 2003). The memorisation of an 'auditory template' enables song evaluation to guide vocal practice, referred to as template learning (Konishi, 1965). This view has been supported by evidence from neurophysiological experiments, which demonstrates that the auditory brain area selectively reacts to the tutor song (Phan et al., 2006). More importantly, recent optogenetic techniques have shown a causal relationship between the stored auditory memory and the acoustic properties of their song mimicking (Zhao et al., 2019). The study found that the activation of synapses in the auditory brain pathway with light pulses of different durations significantly modulated the temporal elements in the songs of zebra finches.

The auditory information is indispensable not only for the sake of forming an auditory template, but also for guiding the sensorimotor learning phase. The investigation into the neural responses of auditory forebrain suggests its functionality in detecting singing errors (Keller & Hahnloser, 2009). Songbirds deafened after the early sensory learning phase, who lost the ability to hear their own songs, learned abnormal songs during the vocal practice phase, even with an intact auditory template stored in their brain (Konishi, 1965). Furthermore, it is suggested that the auditory feedback may reflect categorical perception<sup>1</sup> of species-specific songs. In swamp sparrow, for instance, auditory responsive neurons showed categorical response to varying note durations (Prather et al., 2009). The neural response boundary happened to precisely correspond with the dialectical boundary in birdsongs.

### 1.2.2 MAMMALS

Due to the limitation in audio recording and acoustic analysis techniques in the early days, less attention has been given to vocal learning in mammals. In fact, the faculty of vocal imitation is more widespread than previously thought. In addition to avian species, mammals including cetaceans, pinnipeds, elephants and bats likewise demonstrate advanced ability of vocal learning (Janik & Knörnschild, 2021; Janik & Slater, 1997). The bat was the first nonprimate mammals to be reported to share a similar vocal development trajectory with songbirds and human infants (Boughman, 1998; Knörnschild et al., 2006). Bats produce renditions of calls independent of social context in the babbling phase, and then gradually become attuned to their territorial songs (Knörnschild et al., 2010). It is worth noting that even though bats and songbirds are alike in their

---

<sup>1</sup> Categorical perception refers to the phenomenon that gradual differences along a stimulus continuum are perceived as having sharp discontinuities around categorical boundaries (Harnad, 1987). It has been observed in the perception of color (Harnad, 1987), facial discrimination (Webster et al., 2004), as well as speech (Liberman et al., 1957).

babbling behaviour, the babbles of bat pups are not sex-biased and cover the whole adult vocal repertoire (Knörnschild et al., 2006). Before the maturation of their vocalisations, bats develop an auditory template based on their respective tutors. The acoustic feature of the tutor song and pup song correlates remarkably regardless of the sex of the pup (Knörnschild et al., 2010).

Pinnipeds are likewise found to be skilled vocal learners in multiple training experiments. Since the anecdotal report of a famous harbour seal named Hoover that produced vowel sounds (Ralls et al., 1985), grey seals have intrigued researchers' interest to study their vocal behaviours. Recently, experimental studies have verified their capability of modifying fundamental frequencies and formant frequencies analogous to humans (Stansbury & Janik, 2019). The seals are able to accurately imitate artificially manipulated moan calls with shifted peak frequencies and harmonics (Stansbury & Janik, 2019). What is more striking is that seals can even copy human simple vowels such as cardinal vowels (/a/, /e/, /i/, /ɔ/, and /u/). Interestingly, auditory exposure to recorded vocalisations has been found to boost the probability of baby seals producing a matching call (Stansbury & Janik, 2021).

Cetaceans such as beluga whales (Murayama et al., 2014), humpback whales and bottlenose dolphins (Janik & Sayigh, 2013) exhibit extraordinary capability of imitating species-specific calls, artificial sounds, and sometimes even vocalisations from other species. Bottlenose dolphins produce a highly distinct vocal repertoire, known as signature whistles, to broadcast the identity of the vocalist (Janik & Sayigh, 2013). The emergence of signature whistles happens early in life and it is progressively crystallised within the first 3 months (Janik & Sayigh, 2013). Not only are they able to learn signature whistles from their biological mother (Tyack, 1997), foster mother (Tyack & Sayigh, 1997) and other members of the community (Fripp et al., 2005), but also human trainers' whistles (Miksis et al., 2002) and artificial sounds (Richards et al., 1984). Interestingly, it is suggested that due to the unstable water pressure, bottlenose dolphins extract identity information encoded by frequency modulation regardless of voice

features (Janik et al., 2006). In addition, dolphins show remarkable ability to recognise artificial signature whistles that resemble those of familiar individuals. This implies that dolphins are capable of normalising acoustic signals to a certain extent to obtain identity information.

Whether primates other than humans are vocal learners has been controversial for a long period of time (Janik & Slater, 1997). Despite extensive attempts to teach great apes to learn human speech, researchers have not seen successful training results (Fitch, 2000). Chimpanzees, one of our nearest relatives, seem to lack developmental plasticity in vocal production learning (Menzel, 1964; Owren et al., 1992). Infant squirrel monkeys born at the first day of their life were found to produce calls very close to adult calls (Winter, 1969). A follow-up study further suggests that squirrel monkeys raised by muted caregivers without species-specific auditory input also learned identical vocal repertoire to normally raised monkeys (Winter et al., 1973).

The accumulating negative evidence seems to suggest that vocal learning in non-human primates is doubtful. However, recent studies using more advanced recording and acoustic analysis techniques have shed new light on vocal learning in primates (Egnor & Hauser, 2004). Infant common marmosets (*Callithrix jacchus*) have been found to produce call types that are not present in adult vocalisations, which compose of harmonic and temporal structures outside of normal adult call range (Pistorio et al., 2006). The immature vocalisations bear resemblance to babbling in humans and songbirds. Another line of study on pygmy marmosets provides supporting evidence of early vocal practice. More than twenty years of studies in Elowson's lab show that their babbling behaviours parallel those of human babies in many respects (Elowson et al., 1998a, 1998b). First of all, the onset of rhythmic and recurring babbling appears between six and ten months. Second, the vocalisation is frequent and not limited to call types that are present in adult vocal inventories, which is suggestive of vocal practice. Last, babbling is universal and independent of social groups. Call types of infant marmoset undergo significant changes in spectral and temporal features with



reducing variability, which gradually converge to species-specific calls (common marmoset: Pistorio et al., 2006; pygmy marmoset: Elowson et al., 1998a).

The similarities in anatomy make nonhuman primates an even more relevant model for human speech acquisition than songbirds. There is a growing body of literature that investigates what factors influence their vocal development experimentally. A seminal work by Takahashi et al (2015) on marmoset monkeys demonstrates clear evidence of vocal developmental changes that is not solely due to anatomical changes. Statistical analysis has revealed no significant correlation between their vocal tract growth and the changes in acoustic parameters. They also show that the amount of contingent parental feedback influences the rate of maturation of calls. Studies on twin infant marmoset monkeys raised with different amounts of parental feedback corroborate these findings. Infant monkeys in a low-feedback group showed delayed vocal development compared with a high-feedback group (Takahashi et al., 2017). The lack of parental auditory feedback and social feedback has led to long-lasting disruption in the acoustic structure of their vocalisations (Gultekin & Hage, 2018). In a more extreme case, common marmoset (*Callithrix jacchus*) infants were deafened immediately after birth and they showed abnormal spectral-temporal features in their calls, which endured into adulthood (Roupe et al., 2003). The series of studies have clearly demonstrated that auditory feedback is crucial for vocal learning in marmoset monkeys.

### 1.2.3 SUMMARY

Across the animal kingdom, songbirds and mammals are both promising models for investigating vocal production learning. Similar to humans, an early sensory phase of perception learning, during which the learners gain experience of species-specific signals, has been observed in songbirds (Konishi, 1965). A phase of vocal practice exists in humans (Oller, 1980), songbirds (Thorpe, 1954), marmosets (Elowson et al., 1998a, 1998b) and bats (Fernandez et al., 2021). Vocal practice can be seen as early calibration of vocal systems that converts

motor commands to sound production (Marler & Peters 1982; Kuhl & Meltzoff 1996). The vocal learners seem to have certain learning phases in common: a phase of accumulating auditory experience and a phase of vocal practice. Bringing together all the experimental evidence, it becomes clear that auditory input plays two vital roles in vocal learning: 1) store auditory experience in long-term memory; 2) to detect production errors during vocal practice. For these vocal learners, eliminating the auditory information during the critical sensory period can lead to significant impairment in their vocal behaviours (songbirds: Marler & Tamura, 1964, marmosets: Roupe et al., 2003). Even after the critical period, auditory information is still essential for monitoring production learning (songbirds: Konishi, 1965).

Auditory-guided vocal learning in animals has received considerable attention and thus has been well-attested in neurobiological and behaviour experiments. As far as humans are concerned, a similar mechanism has been suggested, i.e., infants learn the structure of phonetic categories in their native languages by listening to the ambient speech during sensory learning phase (Kuhl, 1991). The derived auditory representations that contain linguistic information would then guide vocal production in the sensorimotor learning phase (Kuhl & Meltzoff, 1996). When auditory guidance is absent, the production learning becomes difficult; for example, congenitally hearing-impaired children showed disrupted development in speech production without intervention (See Osberger, & McGarr, 1982 for a review). In the next section, I will focus on introducing previous studies regarding production and perception development in humans.

## 1.3 CHILD SPEECH DEVELOPMENT

### 1.3.1 BACKGROUND

How do infants gradually learn to crack the speech code? The basic function of the infantile auditory system is present at birth, while the peripheral and central

nervous system grow continuously (Litovsky, 2015). The development of speech perception involves a transition from universal perception to language-specific perception (Kuhl, 2000; Kuhl et al., 2008). That is, children are born with the capability to discriminate the sounds of all languages, and gradually become attuned to phonetic contrast in their native languages. Infants also undergo complex anatomical restructuring of the vocal tract in the first few years of their lives (Vorperian et al., 2005; Vorperian & Kent, 2007). The development involves a rapid growth period from birth to 18 months and a period of slow but steady growth until maturity. Compared with the adult, the infant's vocal tract is shorter, with proportionally larger anterior tongue mass and narrower pharyngeal cavity (Kent, 1992). The anatomical differences contribute to the higher resonance frequencies in their speech compared with adult speech (Fitch & Giedd, 1999). The maturation of their speech production is accompanied by dramatic anatomical changes. Moreover, there is a specific order in which the infants acquire the speech sounds in their native languages.

### 1.3.2 PERCEPTION DEVELOPMENT

Not only does the development of speech perception rely on the maturation of the peripheral system devoted to encoding temporal, spectral and intensity information, but also on the central system which links auditory signals to meaning (Litovsky, 2015). At 20 weeks of gestation, foetal cochlear hair cells and their innervation are already fairly mature (Locher et al., 2013). By 27-28 weeks' gestation age, foetal heart rates already indicate responses to sounds (Litovsky, 2015). 2-5 weeks prior to birth, fetuses demonstrate sensitivity to the changes in the order of syllables, i.e., /ba/ and /bi/ vs. /bi/ and /ba/, as measured by cardiac reactivities (Lecanuet et al., 1987). After birth, infant ear canal diameter and length continue to grow during the first two years of life (Keefe et al., 1993). The development of middle-ear cavities is more prolonged, which extends to later teenage years (Eby & Nadol, 1986).

Infants undergo several stages in the perceptual development during the first year of their lives, as illustrated in Figure 2. Soon after birth, infants display startling sophistication in discriminating phonetic categories (Kuhl, 1993; Werker & Lalonde, 1988). Evidence from high-amplitude sucking experiments shows that infants as young as 1-month old are already capable of discriminating consonant pairs such as /b-p/, /d-t/, and /g-k/ based on the cues in the voice onset time (VOT) of synthetic speech (Eimas et al., 1971). Consonantal distinctions, such as place of articulation (Eimas, 1974) and manner of articulation (Eimas & Miller, 1980), were discriminable by infants between the ages of 2 and 4 months. Notice that, despite their impressive capability of perceiving acoustic cues, contrasts between fricatives can still be challenging for infants (Eilers & Minifie, 1975). More strikingly, the perception of phonetic distinction is universal for infants brought up in different linguistic environments, that is, their sensitivity to sound contrast is not limited to native languages but also in non-native languages. The phenomenon is known as universal speech perception (Kuhl, 2004; Werker & Lalonde, 1988).

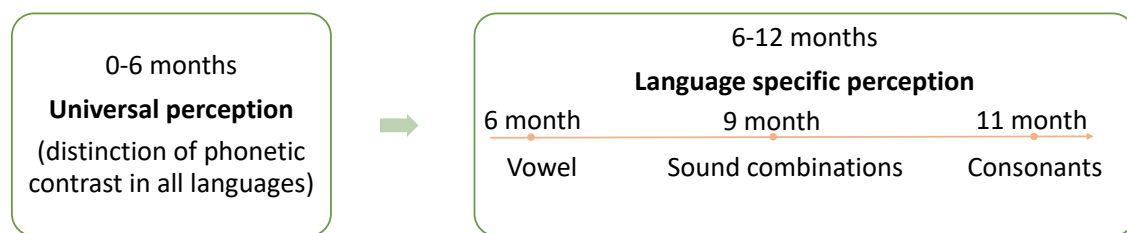


Figure 2 Child speech perception development.

As the sensory system develops, the ability of non-native language perception decays with cumulative capacity for discriminating sounds in native languages (Kuhl et al., 2006). Infants exhibit improvement in discriminating native vowels since 6-month-old in head turning experiments (Kuhl et al., 1992) and neurophysiological experiments (Cheour et al., 1998). The alternation in the perception of vowels normally precedes the change in consonant perception (Polka & Werker, 1994). As to consonants, English infants aged 6-8 months are able to discriminate between two pairs of Hindi consonants, but by 10 to 12 months of age, there is a huge decrease in their discriminability (Werker & Tees, 1984). Meanwhile, there is a significant increase in the performance for English infants between 6 and 12 months of age when discriminating /r-l/ contrast (Kuhl et al., 2006). Such developmental pattern is buttressed by the evidence that English and Mandarin infants gradually exhibit a perceptual inclination to the distinction between affricates and fricatives in their own native languages (Tsao et al., 2006). By adulthood, the ability of universal listening is obscure and non-native speech perception becomes extremely difficult (Best et al., 2001; Miyawaki et al., 1975; Strange & Jenkins, 1978). The perception of phoneme contrast in native languages is known as language-specific perception (Kuhl, 2004; Werker & Lalonde, 1988), or phonemic categorization (Hazan, & Barrett, 2000).

The accumulation of behavioural observations has led to the emergence of theories that attempt to elucidate the mechanisms underlying the developmental transitions. Early models tend to emphasise the sensitivity to phonetic distinction

is an innate ability which is maintained or lost depending on the linguistic environment. It is suggested by the phonetic feature detector account that the course of human evolution induces phonetic distinctions in speech perception (Cooper, 1974; Eimas, 1975). Liberman and Mattingly (1985)'s motor theory of speech perception argues that infants may be born to differentiate the acoustic differences of linguistic gestures and the capability is altered when the gesture is not used. Kuhl proposed the native language magnet model (NLM), in which some perceptual area can serve as a prototype that supports the formation of categorical perception<sup>2</sup> (Kuhl, 1994, 2000; Kuhl et al., 2008). The warped perceptual space later facilitates access to native categories. A similar account is the perceptual assimilation model (PAM), which argues that non-native sound contrast can be assimilated to the phonological categories of the native language (Best, 1994, 1995). According to the natural referent vowel model (NRV), however, the anchor for vowel categories is the vowels with the most extreme acoustic properties. The exposure to languages triggers the organisation of vowel categories (Polka & Bohn, 2011, 2003). Existing theoretical accounts have been controversial in explaining the behavioural observations of the developmental changes. Nevertheless, much less is known concerning the exact role of perceptual development in production development.

### 1.3.3 PRODUCTION DEVELOPMENT

#### 1.3.3.1 Background

The acoustic portrait of speech production in infants is heavily influenced by the growth of the anatomical structure of the vocal tract and the vocal folds. The infant

---

<sup>2</sup> A recent study by Kronrod, Coppess and Feldman (2016) has proposed that the categorical perception can be explained by having a Bayesian computational that quantifies meaningful to noise variance to unify categorical effects in vowel, consonant and fricative perception rather than only vowel perception in the NLM model.

vocal apparatus is not a miniature of the adult organ. The infant has a vocal tract of 6-8 cm in length, while the length of vocal tract of an adult is approximately 15 cm for female and 18 cm for male (Vorperian et al., 2005). During the first three years, the infant vocal tract grows around 3 cm (Vorperian et al., 1999). The infant's vocal tract differs from the adult's also in shape. It has 1) a proportionally larger anterior tongue mass, 2) a narrower pharyngeal cavity, 3) an adjacent velum and epiglottis, and 4) a gradual rather than right-angle bend oropharyngeal channel (Kent, 1992; Kent & Murray, 1982). The infant vocal folds are approximately 4-5 mm long, consisting of uniformly structured lamina propria (Sato et al., 2001). In contrast, the mature adult membranous vocal fold length is around 17 mm for male and 12 mm for female (Rogers et al., 2014). The maturation of vocal ligament occurs gradually along with increase of laryngeal size. Controversial results are reported with regard to the appearance of sexual dimorphism of voice production in early childhood (Crelin, 1973; Eckel et al., 1999). However, there is little dispute that during puberty the larynx undergoes significant changes, when the male vocal folds are lengthened and the larynx is descended at a much higher speed than the female's (Kahane, 1978, 1982).

The anatomical structure determines the acoustic properties of the produced speech sounds, leading to the infant's frequency ranges well above that of the adult, as illustrated in Figure 3. A classic work by Peterson and Barney in (1952) demonstrates that the fundamental frequency is much higher for children, as well as the first and second formants which determine the vowel quality. Infants are able to use vocal fry and high register resulting in a F0 ranging from 30 to 2500Hz (P. Keating & Buhr, 1978). A rapid declination of average F0 begins at age 3, and during adolescence for males when the male and female distinction of F0 emerges. For American English native speakers, the average F0 for adult male and female is 120 Hz and 220 Hz, respectively (Lee et al., 1999), while the first formant (F1) of infant vowels ranges from around 450 Hz to 1650 Hz and the second formant (F2) ranges from 1500 Hz to 4100 Hz (Kent & Murray, 1982; Kuhl & Meltzoff, 1996). In contrast, recordings of American English monolinguals show that adult male produce vowels ranging from 250 Hz to 800 Hz for F1 and from

1100 Hz to 2500 Hz for F2. F1 of adult female vowels ranges from 350 Hz to 1200 Hz and F2 ranges from 1200 Hz to 3100 Hz (Hagiwara, 1997).

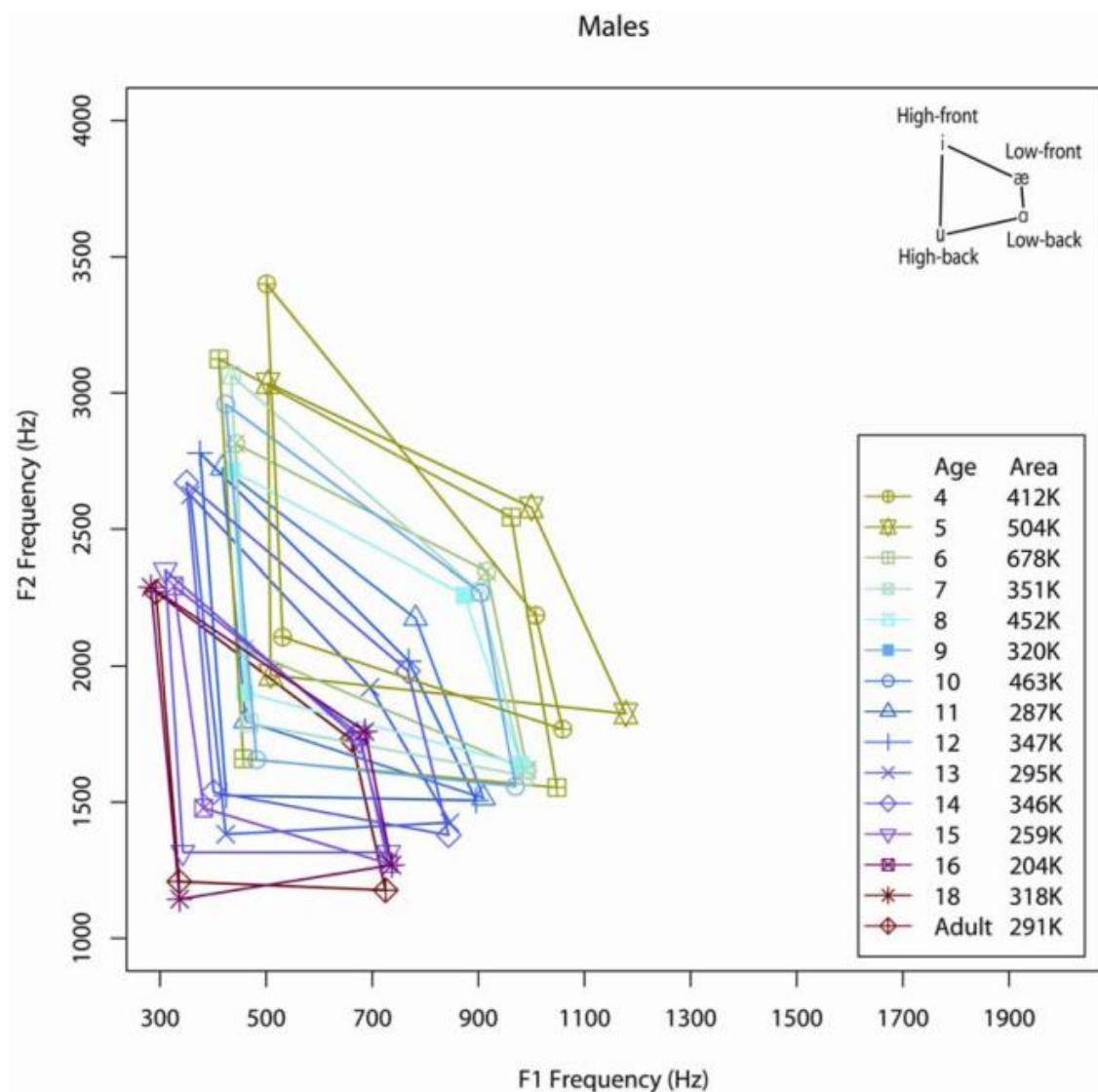


Figure 3 Average F1–F2 acoustic space for American English males aged 4 years through adulthood (Vorperian & Kent, 2007). Image reproduced with permission of the rights holder, Journal of speech, language, and hearing research : JSLHR.

### 1.3.3.2 Development stages

The biological development is accompanied by changes in vocal behaviours. The stages of infant vocal production are summarised in Figure 4. The infant starts



with producing sounds to express discomfort or anger including crying and vegetative sounds. Then, cooing occurs at 1-4 months of age, which is produced in the velar area where the tongue and the palate are in close contact (Vihman, 2014). At around 5-6 months after birth, the infant begins to produce consonant-vowel syllable trains, often referred to as reduplicated babbling (Stark, 1986) or canonical babbling (Oller & Eilers, 1988). The canonical syllables consist of fully resonant vowels and clear consonants with complete or nearly complete oral closure. The repetitive articulator movement is similar to other motor movements including the movements of the limbs, the torso and the fingers, which are also observed to be repeated at regular time intervals. It is suggested to be a general process for coordinating neuromuscular movements (Thelen, 1981). The infant seems to gradually acquire the control of laryngeal and articulatory movements at this stage. Prosodic features emerge along with frication noises, nasal murmurs, and bilabial and uvular trills. At 10-18 months of age, finally infants utter their first meaningful words. There are, however, overlaps between the development stages as well as individual differences (Stark et al., 1993).

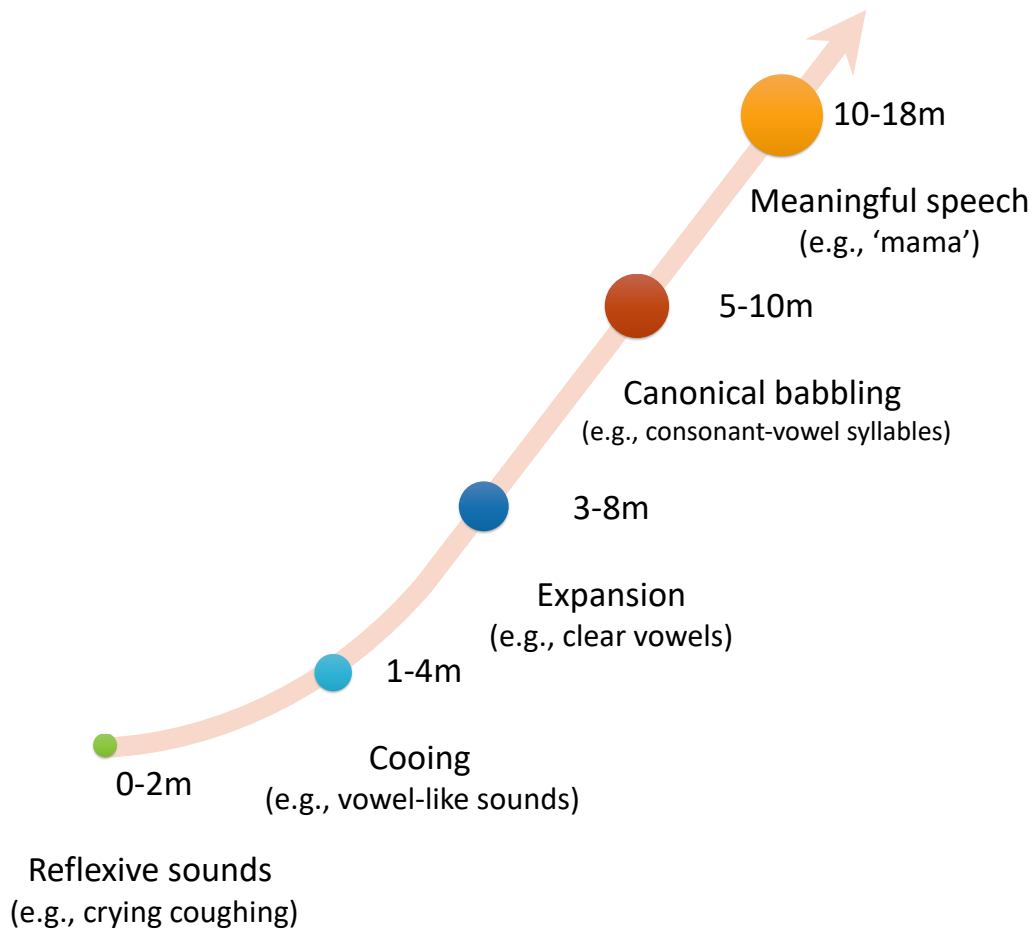


Figure 4 Child speech production development.

#### 1.3.3.3 Vowel development

The occurrence of corner vowels in English is relatively early, i.e., /a/ at 17 weeks, /i/ at 18 weeks and /u/ at 24 weeks (Buhr, 1980). This demonstrates how quickly a well-defined vowel triangle emerges. However, these vocalisations are not stabilised until 36 weeks for /a/ and /i/ and even more prolonged for /u/. Corner vowels such as /i, u, a/ seem easier for the children to acquire, while their mid vowels like /ɪ/, /e/, /ɛ/ and /ʊ/ are less accurate (Stoel-Gammon & Pollock, 2008). Substitutions of the difficult instances of vowels are commonly seen, showing a high variability in their production (Vihman, 1996). For example, /ɪ/ is often produced as /i/ or sometimes as /ɛ/. The usage of substitution declines over time. Between the age of 1 and 2, the accuracies of vowels increase rapidly with

uneven proficiency across vowel categories, that is, certain vowels are mastered earlier than others (Buhr, 1980; Hare, 1983; Paschall, 1983). By the age of 2, children's vowel production has extended to almost all the vowels except rhotic vowels, and by the age of 3, the mean vowel accuracy is reported to exceed 97% (Pollock & Berni, 2003). In contrast, the rhotic vowel, /ə/ is not acquired until 4 years old. However, the reported order of acquisition is controversial because of the different speech materials used in the previous experiments. More recent studies measure the identification of the vowels based on auditory transcriptions, but in some early studies the listeners could have relied on additional information such as a word list. Overall, the order of vowel acquisition in young English children can be summarised as follows: Corner vowels (except /æ/) > mid vowels > rhotic vowels (Stoel-Gammon & Pollock, 2008). The order of acquisition largely matches the pattern reported in Wellman et al. (1931) , as illustrated in Figure 5.



Figure 5 English vowel development in 204 children aged 2–6 years using correct production by 75% of children in a particular age group as the criterion (Wellman et al., 1931).

#### 1.3.3.4 Consonant development

A milestone in speech development, canonical babbling, is the benchmark for the occurrence of consonants. It has been found that the place of constriction of the consonant often assimilates with the following vowel in the babbling of infants: bilabial consonants precede central vowels (e.g., /bə/), alveolar consonants precede high vowels (e.g., /di/), and velar consonants are associated with back vowels (e.g., /ku/) (MacNeilage & Davis, 2000). Consonant substitutions in early words also follow such an assimilatory effect. For example, bilabial stops before a high vowel are produced as alveolar stops (Stoel-Gammon, 1983). The bilabial-central vowel and alveolar-front vowel association become weaker for older children during the 18-24 months (Tyler & Langsdale, 1996; Vihman, 1992). It is suggested that consonants are not fully acquired independently in the early babbling stage. Rather, the consonants and vowels are controlled as a single entity (B. L. Davis & Macneilage, 1995).

Sander (1972) notes that there is a distinction between the emergence of and the stabilisation of consonant production, referred to as ‘customary production’ and ‘mastery’ respectively. In the first stage, children can produce more than half of the sound in different positions (50% criteria) and they are able to produce consonants correctly at three word positions in the second stage (90% criteria). Figure 6 shows the development of consonant production by the 50% and 90%

threshold based on 15 studies on American English children (Crowe & McLeod, 2020). We can clearly see a long gap between the ‘customary’ and ‘mastery’ production. It is well documented that stops, nasals and glides appear in early vocalisations (Stoel-Gammon, 1985; Vihman et al., 1985), while other consonants were not fully mastered until more than 3 years old (Crowe & McLeod, 2020; Mcleod & Crowe, 2018). Children then acquire the affricates, the liquids and the fricatives, and the full set of consonants are not acquired until 5-6 years old (Crowe & McLeod, 2020). The mastery of consonant clusters is even more prolonged, which begins at as young as 2 years of age but not fully achieved until age 8-9 (Mcleod et al., 2001).

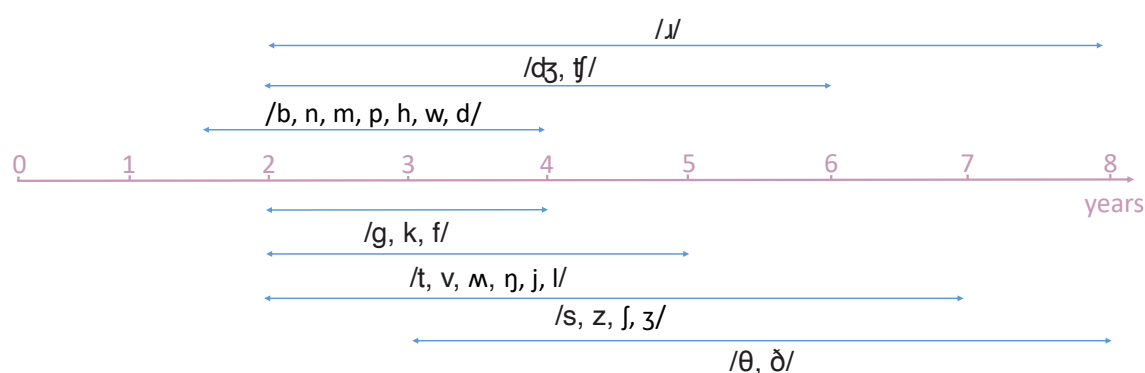


Figure 6 Average age of acquisition of American English consonants, adapted from Table 2 in Crowe & McLeod (2020). The arrow starts from 50% and ends with 90% criterion.

#### 1.3.4 SUMMARY

Studies on the development of speech perception show that infants are born with the ability to distinguish phonetic segments in all languages, and they gradually show perceptual attunement towards native languages (Werker & Tees, 1992). In other words, there is a transition from language-universal to language-specific perception during development (Kuhl, 2004; Werker & Lalonde, 1988). Difficulties arise, however, when an attempt is made to examine the impact of

perception inclination on production development by behaviour experiments alone, because of the overlap in the timeline of production and perception development (Figure 1). Moreover, the widely used method of manipulating auditory input in the animal studies (Section 1.2) is neither ethical nor practical on humans. Whether humans share the same mechanism with other vocal learners remains unclear.

## 1.4 SENSORIMOTOR INTERACTION

### 1.4.1 SENSORIMOTOR LEARNING

Humans are capable of learning complex motor actions using sensory information in various forms, known as sensorimotor learning. A broad definition of sensorimotor learning is the process of improving the performance of the motor behaviour with the assistance of sensory systems (Krakauer & Mazzoni, 2011; Makino et al., 2016; Wolpert et al., 2011). It includes sensory perceptual learning (i.e., detection of behaviourally relevant sensory information), sensorimotor associative learning (i.e., learning adaptive linkage between sensory and motor patterns) and motor skill learning (i.e., learning novel motor behaviours) (Makino et al., 2016). However, it is often used in the literature to refer to only speech adaptation learning, that is, the process of speech motor system being adjusted when the sensory feedback is altered (Parrell & Houde, 2019).

Based on substantial research into sensorimotor learning, three major mechanisms have been identified: 1) Error-based learning, 2) Reinforcement learning and 3) Use-dependent learning (Wolpert et al., 2011). Among them, error-based learning is the most well-studied. The idea is that when the motor system causes an error, it directionally modifies the command following the gradient of the error in the next movement. The process has been attested in motor adaptation experiments of hand movements such as gripping (Flanagan & Wing, 1997) and reaching (Krakauer et al., 2000). It has been long regarded as an essential form of learning motor control (Kawato et al., 1987) but it is

inadequate for the learning of motor movement with manifold solution, namely, many combinations of motor commands will produce a plausible solution (Wolpert et al., 2011). Reinforcement learning, in contrast, has the potential to further optimise the movement, because the system is adjusted based on reward signals rather than the motor error. To be more specific, providing reward that signals the utility of the motor command has been proven to facilitate the learning of motor skills (Abe et al., 2011). A possible neural underpinning of reinforcement learning is the projection of dopaminergic neurons that expedites the encoding of motor actions (Hosp et al., 2011; Luft & Schwarz, 2009). This learning mechanism has received increasing attention since the prevalent application of similar machine learning algorithms (Sutton & Barto, 1998). Finally, during use-dependent learning, the motor movements are biased by simple repetition without error estimation (Butefisch et al., 2000; Classen et al., 1998). For instance, in the case of reaching action, repetitive arm movements not only induce a tendency towards the neighbouring goal but also a reduction in the variability. This type of learning is sometimes modelled as Bayesian integration to simulate the adaptive changes (Verstynen & Sabes, 2011). Moreover, it has been found that use-dependent learning and error-based learning can simultaneously influence motor behaviours (Diedrichsen et al., 2010). However, these learning mechanisms are not well-attested in the case of speech motor learning.

#### 1.4.2 IMITATIVE LEARNING

Another line of studies has focused on sensorimotor learning by imitation. Imitation involves perceiving and reproducing motor actions. The process translates sensory signals into motor movements, which has been considered as an important method of acquiring novel actions (Heyes, 2001). Studies of imitative learning have seen an acceleration since the discovery of mirror neurons that signifies the remarkable interplay between sensory and motor systems (Kilner & Lemon, 2013). Mirror neurons are a group of neurons that are excited both when an individual performs a motor action and when he observes the same

or similar action executed by another individual (Fadiga et al., 1995; Gallese et al., 1996). Mirror neurons were first found in the ventral premotor cortex of the macaque monkey which responded both to actions such as grasping food and seeing others performing the same action (Gallese et al., 1996; Rizzolatti, Fadiga, Gallese, et al., 1996) and later in humans, homologous phenomenon was observed (Fadiga et al., 1995; Rizzolatti, Fadiga, Matelli, et al., 1996).

Regarding speech motor control, in transcranial magnetic stimulation (TMS) experiments, participants showed enhanced muscle activities in the tongue (Fadiga et al., 2002a, 2002b) and the lips (K. E. Watkins et al., 2003) when hearing speech. These findings are consistent with functional magnetic resonance imaging (fMRI) evidence showing that brain areas that are involved in speech production are activated when participants passively listen to speech (Pulvermuller et al., 2006; Wilson et al., 2004). These findings are often considered to be evidence of the linkage between speech production and perception (Hickok et al., 2011). Research on infant speech perception shows that the production-perception link in speech gradually emerges during vocal development (Imada et al., 2006). In this study, magnetoencephalography (MEG) was used to record the neural response of newborns, 6-month-olds and 12-month-olds while listening to speech sounds passively. The 6-month-old and 12-month-old infants showed activation in the inferior frontal cortex (Broca's area) which is involved in speech motor control, but no activation was found in the newborns. Interestingly, the advent of motor brain activation during speech perception is suggested to be due to canonical babbling which starts around 5 to 6 months after birth (Imada et al., 2006). This neurological study of infants indicates that the perception-production link is not innate but highly likely to be reliant on experience. However, more recent studies found that the perceptual sensitivity of preverbal infants can be influenced by the inhibition of oral movements, which suggests an early coupling between speech production and perception (Bruderer et al., 2015; Choi et al., 2019, 2021). One question that still needs to be asked, however, is what kind of sensorimotor experience can forge the link.



The crucial issue of how to link sensory experience with motor actions is known as the correspondence problem (Nehaniv & Dautenhahn, 2002). Heyes (2001) has proposed that associative sequence learning provides an essential basis for solving the problem. According to this view, the coupling of sensorimotor systems is the product of correlated motor actions and perceptual experience. The contingent and contiguous experience includes 1) self-observation (i.e., seeing our own actions), 2) synchronous actions (i.e., performing the same action in a social group) and 3) imitative partners (i.e., a caregiver imitating infant facial expressions). Experimental studies on infants have provided support for associative learning. For example, electroencephalogram (EEG) recording of infants at the age of 14 to 16 months showed that their motor systems were activated while watching a video of crawling and that the motor resonance was stronger for infants with more crawling experience (van Elk et al., 2008). Another EEG study adopted a training paradigm to investigate the development of sensorimotor coupling in pre-walking infants (de Klerk et al., 2015). The infants in the post-training session showed significantly more sensorimotor cortex activation during the observation of stepping. The account that sensorimotor coupling is formed by simultaneous excitation of sensory neurons and motor neurons shares similarities with Hebbian learning (Keysers & Perrett, 2004). Although Hebbian learning also recognises the critical role of sensorimotor experience, it further emphasises the value of experiential canalisation optimised by evolution (Giudice et al., 2009). Despite the existence of abundant theoretical works, one question that needs to be asked is how to apply associative learning to non-visual sensorimotor actions, such as speech.

#### 1.4.3 SUMMARY

Speech, a fine motor skill, is one of the most demanding cognitive challenges that humans can perform (Penfield & Roberts, 1959). Past research has suggested several learning mechanisms including error-based learning, reinforcement learning and use-dependent learning (Wolpert et al., 2011). Due to the many-to-

one mapping between articulation and acoustics (Xu et al., 2021), the possibility of driving the learning by error gradient is rather low. Reinforcement learning can cope with the manifold problem but has received much less attention when it comes to speech motor learning. Use-dependent learning may be involved during the phase of canonical babbling when infants repeat speech utterances but it seems unlikely to lead to the learning of intelligible words. On the other hand, imitative learning of sensorimotor systems has experienced a sharp surge especially following the discovery of mirror neurons that show activation during both observing and performing motor actions (Fadiga et al., 1995; Gallese et al., 1996). Even though it is widely known that there is an interaction between sensory and motor systems, it remains unclear how the linkage is forged. Some suggest that we are born with the imitative system and experience only strengthens the link (Lepage & Théoret, 2007), whilst others argue that correlated sensorimotor activities play a pivotal role in linking the two systems (Heyes, 2001). However, unlike most other sensorimotor tasks, there are no obvious visual cues for speech, as most of the articulators are hidden, which increases the difficulty of imitative learning. Vocal learning is therefore a special case in which the common sensorimotor learning strategies are largely inaccessible. The research on sensorimotor learning to date cannot fully explain how sensorimotor link in speech is acquired.

## 1.5 RESEARCH QUESTIONS AND THESIS OUTLINE

Throughout the animal kingdom, many species show vocal plasticity to a certain extent. Noticeable similarities can be seen in the vocal developmental pattern of songbirds and humans, in which a phase of auditory extraction paves the way for vocal practice (Doupe & Kuhl, 1999b). These observations have naturally led to the postulation that auditory experience guides vocal learning in humans and songbirds (Kuhl, 2003). It has been found that child speech perception changes from language-universal to language-specific perception (Kuhl, 2004; Werker & Lalonde, 1988) and production learning follows certain developmental patterns of

phoneme acquisition. Unlike birdsong learning, much uncertainty still exists concerning the nature of the auditory guidance for speech acquisition in humans, as the approaches of manipulating auditory feedback in animal studies are unethical. At the same time, studies on human sensorimotor learning have proposed several assumptions including error-based learning, reinforcement learning, use-dependent learning (Wolpert et al., 2011) and imitative learning (Cook et al., 2014). However, vocal learning is essentially dissimilar to other motor movements, because of the lack of visual information. What is yet unclear is whether these sensorimotor learning mechanisms likewise underlie vocal learning. The unsolved questions can be investigated using computational approaches, which allows hypothesis testing beyond observational studies. Although simulation studies have gained popularity in many scientific fields, it is rarely applied to the emulation of vocal development (ter Haar et al., 2021). With explicit computational models, it is possible to recreate the internal component of speech production and perception to probe into the black box of human vocal learning.

Inspired by the finding that in songbirds (Brainard & Doupe, 2002; Doupe & Kuhl, 1999b), mammals (Roupe et al., 2003; Stansbury & Janik, 2019, 2021) and humans (Brainard & Doupe, 2002; Kuhl, 2003), perception learning precedes production learning, it is speculated that successful simulation of vocal learning needs to a) use perception to guide vocal exploration, and b) emulate critical aspects of the articulation process. To test this idea, I developed a vocal learning model with these two components to mimic vocal learning. To the best of my knowledge, past work has not succeeded in simulating the learning of intelligible words with CV syllables (Appendix Table A Performance). The current study attempts a first step in this direction with the goal of learning intelligible English words.

The major aims of the thesis are as follows:

- 1) To investigate how language-specific perception and language-universal perception impact on production learning (chapter 4);

- 2) To investigate how articulatory dynamics during speech production can be modelled (chapter 4);
- 3) To investigate whether auditory-guided vocal learning in children and adults can be simulated (chapter 5);
- 4) To investigate whether anatomical development of the vocal tract influences learning (chapter 5);
- 5) To examine whether the vocal learning model shows resemblance to child speech development (chapter 5);

In chapter 2, I will first introduce previous vocal learning models with emphasis on the model architecture, followed by a review of previous theories on speech production and perception and its emulation in the vocal learning models. Finally, I will present the remaining issues and knowledge gap in the field of vocal learning simulations

In chapter 3, I present a simulation model of vocal learning. The model contains a state-of-the-art articulatory synthesiser with built-in articulatory dynamics, consisting of vocal tract models of an adult male, a 1-year-old child and a 3-year-old child. The vocal exploration scheme is guided by either acoustic features to simulate universal perception that detects phonetic differences in all languages (Kuhl, 2000; Werker & Lalonde, 1988), or an automatic phoneme recogniser to simulate language-specific perception that captures key phonetic properties that distinguish words in a language (Kuhl, 2000; Werker & Lalonde, 1988). The learning outcome of the simulation models by both a word recogniser and two types of listening experiments: an open-vocabulary transcription experiment and a close-set transcription experiment. The vocal learning model will be used to address the research questions in chapter 4 and 5.

In chapter 4, I investigate the speech sensory and motor control in detail. First of all, I examine the role of auditory and somatosensory feedback on vocal learning by comparing the performance of the learned speech in a series of controlled simulations. A word recogniser is used to assess the synthetic speech trained by acoustic features. Listening experiments are used to compare acoustic features

with an automatic phoneme recogniser to determine what kind of auditory feedback is more beneficial in guiding vocal learning. Secondly, I then explore how to model the dynamic articulatory control of CV coarticulation explicitly by the synchronised dimension-specific sequential target approximation model (Liu et al., 2022; Xu, 2020). The vocal tract parameters will be optimised through an analysis-by-synthesis approach, assisted by the embodiment constraints of the coarticulation model.

In chapter 5, I report the identification accuracy rate of the speech learned by the adult and the child vocal tract models. Whether the anatomical structure of the vocal tract model influences vocal learning is tested, based on the identification accuracy rate in the listening experiments. I additionally compare the child vocal tract model with the developmental changes during speech acquisition in real life. Additional factors that impact on identification accuracy rate such as syllable type and types of listening experiments are assessed.

In chapter 6, I will discuss all the findings and the research questions raised in the dissertation and present the limitations of this work and make suggestions for future research.

## **Chapter 2 REVIEW OF VOCAL LEARNING MODELS**

### **2.1 LEARNING STRATEGY**

Early vocal learning has long been a question of great interest and various computational models have been proposed (Pagliarini et al., 2021). Appendix Table A lists previous computational models of human vocal learning. Some models focus on simulating the developmental trajectories of learning stages (section 2.1.6 Self-organisation). Other models have centred on simulating different learning architectures (the rest of the section).

### 2.1.1 NEUROBIOLOGICAL MODELS

Several studies have probed the neural and cognitive mechanisms at play in early vocal learning by modelling the brain network. The earliest neurobiologically motivated computational model is the DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011), a neural-network-based model that simulates sensorimotor interactions during speech acquisition. An overview of the model is shown in Figure 7 (Tourville & Guenther, 2011). It consists of two main components: a) a feedforward control system that encodes the movement velocities of the articulators, and b) a feedback control system that encodes the time-varying sensory expectations. The feedforward articulator velocity map controls eight antagonistic pairs of cells that are in charge of the movement of the lips, the jaw, the tongue and the larynx. The feedback system incorporates auditory feedback based on the range of the first three formants and 22-dimensional somatosensory vectors that depict the expected tactile and proprioceptive signals. The model simulates sensorimotor interaction by finding appropriate synaptic weights for mapping the phonetic-to-orosensory space and orosensory-to-articulatory space. The synaptic weights that associate sensory error map and feedback control map are first tuned by co-occurring motor and sensory signals. The error signals are later used by the feedback control map to correct motor commands so that the trained model is able to adjust motor actions in the presence of sensory errors. Although the DIVA model has been widely applied to exemplify speech adaptation or compensation phenomena as a theoretical framework (Lane et al., 2007; Perkell et al., 2007), the computational implementation of speech acquisition has not led to intelligible speech.

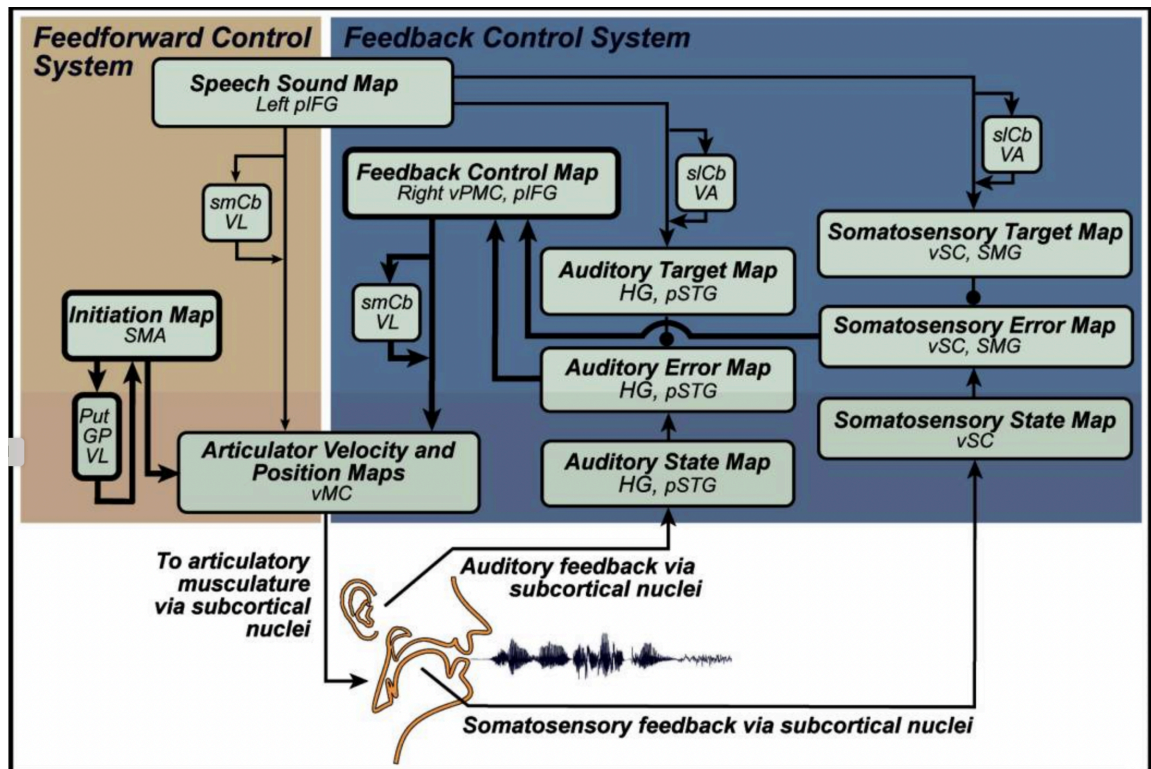


Figure 7 Illustration of the DIVA model (Tourville & Guenther, 2011) © copyright 2022, reprinted by permission of Informa UK Limited, trading as Taylor & Taylor & Francis Group, <http://www.tandfonline.com>.

As a complementary to the DIVA model, Kröger has proposed a neurocomputational model to establish a mapping between speech phonetics and sensory signals via self-organisation (Kröger et al., 2009). An artificial vocal tract model was trained to imitate speech sounds, which enabled the pairing of production and perception to be stored in self-organising maps (Kohonen, 1982). The learning was done by the adjustments of synaptic link weights between phonetic map and sensorimotor state map. The model contrasts with the DIVA model mainly in two respects: 1) it incorporates phonetic map and motor planning as intermediate levels, and 2) an error signal of predicted sensory feedback and actual feedback is not included. The model claimed to simulate the acquisition of CV sequences but no audio samples were provided.

Westerman & Miranda (2002, 2004) have proposed a sensorimotor learning model that simulates how mirror neurons are developed by imitation. The sensory

system and motor map were correlated through Gaussian activation of units in both receptive fields. Speech sounds evoked a response on the sensory map and the associated motor map through imitation. In short, the integration of the two maps was established by Hebbian connections (i.e., simultaneous activation) of the units on each map. Later, another similar attempt has been made to simulate the learning of point vowels by self-organising maps and Hebbian connections with additional dynamic components (Heintz et al., 2009). The framework has an articulatory vocal tract model and uses the first three formants as auditory input. Even though the model also comprises feedforward and feedback control systems, it puts more emphasis on the bidirectional association between the two systems. Heintz et al.'s model is only concerned with vowel acquisition and the articulatory maps do not contain consonant movements. Overall, the aforementioned neurobiological models have been inclined to compass neural processes of speech production and perception, whereas much less attention was paid to generating intelligible speech.

### 2.1.2 ACOUSTIC IMITATION

Vocal mimicry has long been regarded as a crucial mechanism for speech acquisition (Kuhl & Meltzoff, 1996). As a consequence, a great deal of research has been carried out to simulate vocal imitation by the distal learning framework. The distal learning provides a framework that describes how a dynamic system can learn actions to perform desired outcomes when supervised by a distal 'teacher' (Jordan & Rumelhart, 1992). The learning is divided into two phases: 1) the model learns a predictive forward model that transforms action space to sensation space; and 2) the model learns an inverse model to map desired sensation to actions by the utilisation of the forward model. The framework is suitable for addressing the correspondence problem in speech acquisition because it is applicable to non-convex many-to-one mapping relationship between articulation and acoustics (Xu et al., 2021). Moreover, speech acoustics can be the training data as a distal 'teacher' in an imitative learning process.



HABLAR model proposed by Bailly is perhaps the earliest model that aims to achieve audio-visual-to-articulatory inversion with the distal learning framework (Bailly, 1997). The model consists of an auditory system that detects static and dynamic status of the speech spectral information and a motor control system that converts phonological representations to articulatory movements. A crucial aspect of the model is that it explicitly simulates coarticulation by the consecutive activation of consonant and vowel goals. A forward model was built through polynomial interpolation of the articulatory parameters of X-ray data and auditory signals were represented by the first four formants. In line with the distal learning framework, an inverse model was derived by the Jacobian inversion of the forward model. However, one major drawback of this approach is that the process requires articulatory data for the sake of constructing the forward model.

Providing that learners do not have prior knowledge about articulation, another similar yet slightly different attempt was made by Howard and Huckvale (2005), which bypassed the utilisation of articulatory data. An inverse model between speech acoustics and speech motor control was trained by direct mapping and the distal learning. A babbling corpus was first constructed with the Maeda synthesiser (Maeda, 1990) to generate random CV sequences based on Hidden Markov Model Generator (HMM). The direct inverse model was then trained by a classical supervised regression algorithm. The distal learning model was supervised by the Euclidean similarity of formant frequencies. The synthesiser trained with its own speech had higher performance than the direct inverse model trained with human speech. The provided sound spectrograms and the supplementary audio samples showed that the learning of CV sequences driven by the distal learning was still unsatisfactory. The study further points out an essential difficulty in learning an inverse model, that is, the speaker normalisation problem.

More recently, a few studies have constructed vocal mimicry models with a structure similar to the distal learning. For example, Philippsen et al. (2014) simulated the learning of CV sequences via acoustic imitation. In the first stage, recurrent neural networks were trained to learn a forward and an inverse model

for CV syllables. In the second stage, the pre-trained models were refined by imitating auditory goals. The speech synthesiser used in the study was VocalTractLab (Birkholz, 2013), a highly sophisticated 3D articulatory synthesiser. Still, as reported in the paper, the model was not able to learn smooth articulatory trajectories even after fine-tuning, which reveals the difficulty in training an inverse model. In the same year, Prom-On et al. (2014a, 2014b) also trained VocalTractLab (Birkholz, 2013) to learn Thai vowels by acoustic imitation. Figure 8 shows an illustration of the model. A stochastic gradient descent algorithm was used to optimise the vocal tract parameters by minimising the Euclidean distance between the Mel-frequency cepstral coefficients (MFCCs) of the synthetic and the natural speech. The synthesis quality of the learned vowels was analysed based on both RMSEs of the mean formants and a listening experiment by native listeners. The results indicated that the synthetic vowels were close to natural speech in terms of formant values (F1, F2 and F3) and intelligibility, which indicates that the speaker normalisation problem (see Section 1.1 Background for details) can be resolved by acoustic imitation as far as vowel acquisition is concerned. However, so far, none of the simulation works has demonstrated successful learning of intelligible CV syllables.

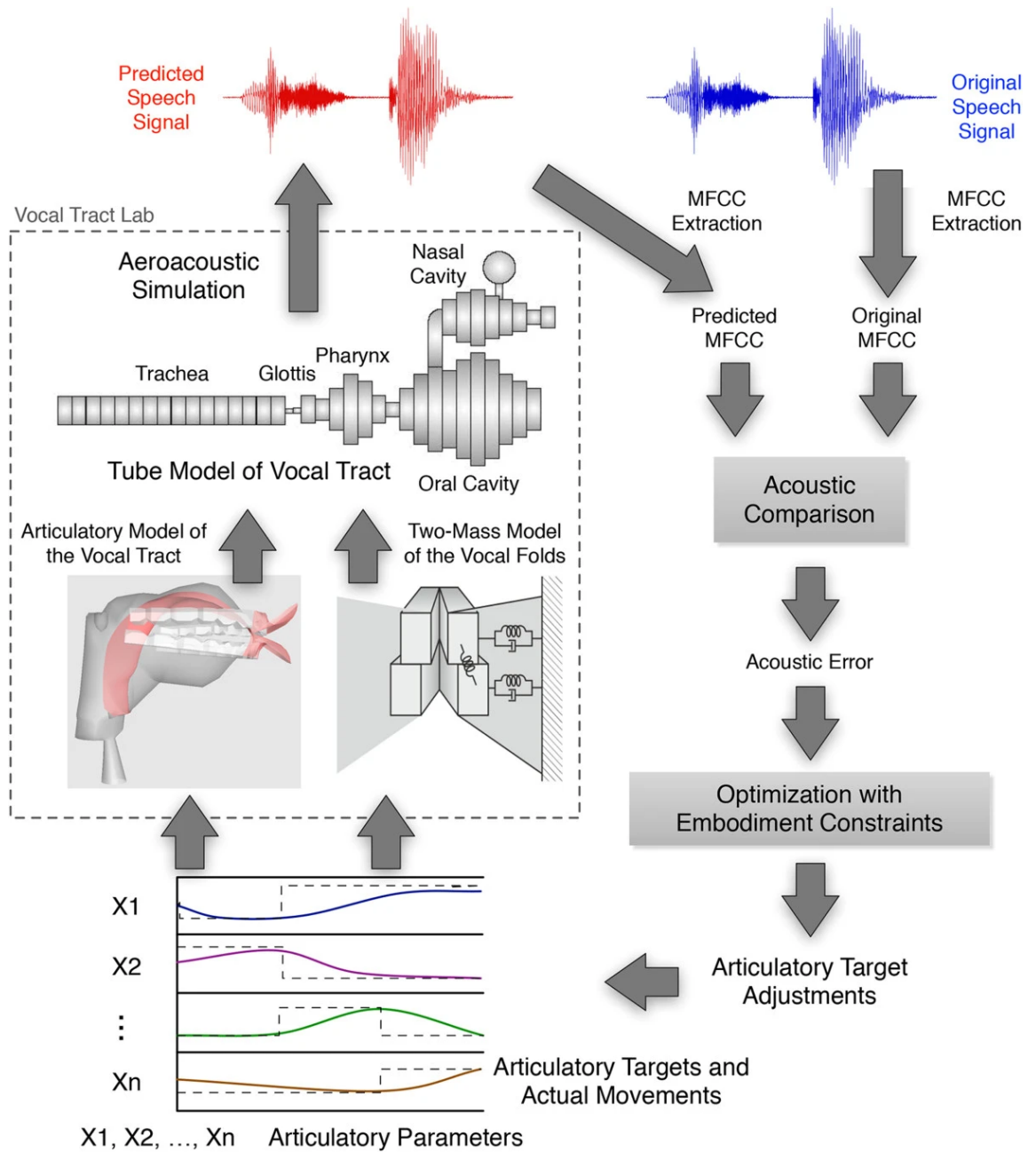


Figure 8 Architecture of an acoustic imitation model (Prom-On et al., 2014a, Figure 1)

### 2.1.3 INFANT-CAREGIVER INTERACTION

The speaker normalisation problem (i.e., the correspondence problem) has long been considered as a cumbersome burden for vocal learners, which has led to

an initiation of research into seeking assistance from caregivers (see Asada, 2016 for a review). Several attempts have been made to simulate social interactions between the infant and the caregiver to facilitate vocal learning. For example, Lyon and her colleagues trained a humanoid robot DeeChee with real-time reinforcement signals from a human teacher, who gave approving comments to the robot's appropriate words (Lyon et al., 2012). A similar social interaction paradigm was proposed by Cohen and Billard, whereby the agent produced vocalisations and the virtual caregiver gave reward or punishment (Cohen & Billard, 2018).

Other researchers, however, take the perspective that the caregiver's reformulation of the infant speech, instead of simple positive and negative feedback, assists speech acquisition (Asada, 2016). Huckvale, Howard and the others have built a virtual infant KLAIR to simulate interactive sensorimotor learning (Huckvale, 2011a, 2011b; Huckvale et al., 2009). The multimodal infant is able to produce and perceive real-time sounds and show facial expressions. KLAIR relies on caregiver's reformulation to reinforce the acquisition of speech to learn a mapping between adult speech and its motor pattern. Following the same principle, Howard and Messum proposed another interactive learning model, named Elija (Howard & Messum, 2007, 2014, 2011; Messum & Howard, 2015). As illustrated in Figure 9 (Howard & Messum, 2014), Elija is equipped with the Maeda synthesiser (Maeda, 1990) and the articulatory movements are calculated by the task dynamic model (Saltzman & Munhall, 1989b). Elija starts exploring speech sounds by unsupervised babbling. Elija then engages in imitative interactions with the caregiver iteratively. The caregiver first utters a word and Elija tries out different speech sounds that have been stored in the babbling repertoires. Elija keeps the vocal movements in the end when the caregiver is satisfied with his/her speech. In this way, the correspondence between his/her own vocal action and adult speech is established. Importantly, the central aspect of this learning architecture is the caregiver's judgement rather than the learner's own judgement. In the end, Elija has managed to learn some

vowels but the learned CV syllables did not sound intelligible and were not tested for intelligibility.

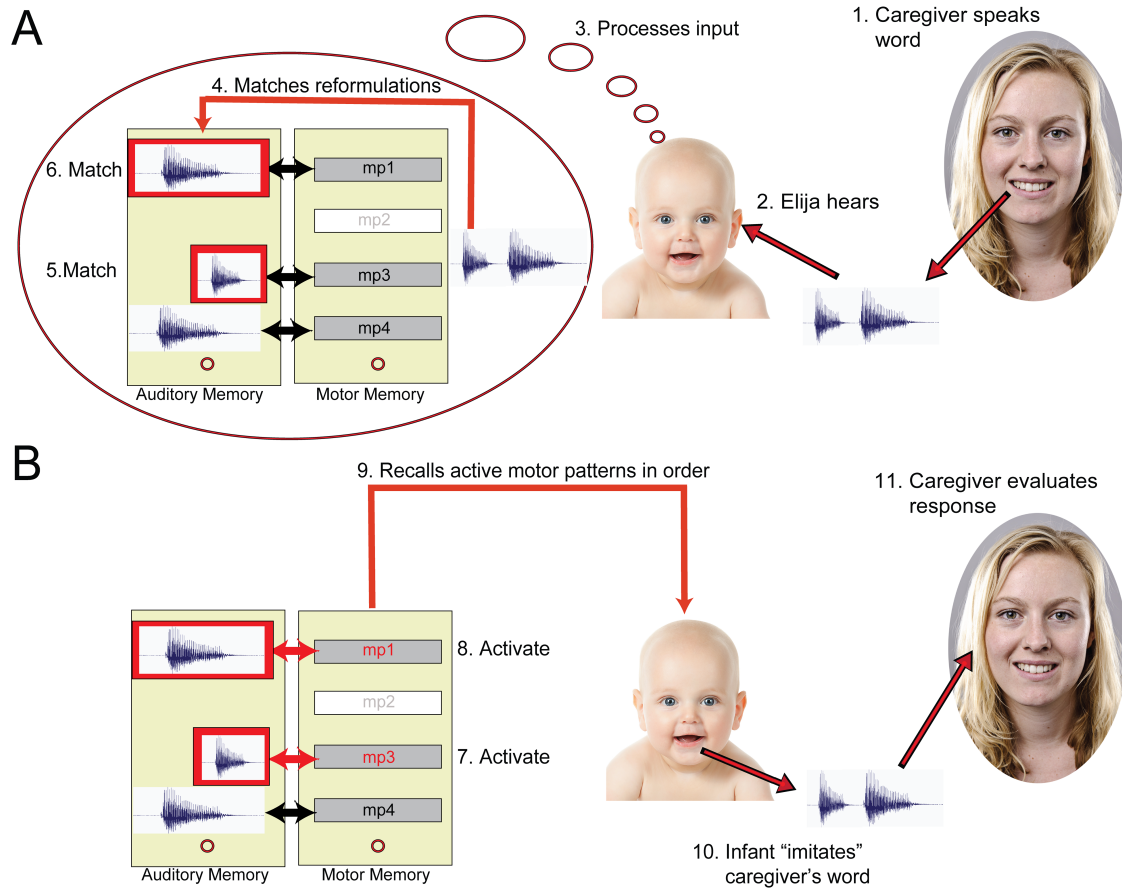


Figure 9 Illustration of imitative learning paradigm for Elija (Howard & Messum, 2014)

Another research group has pursued a similar path to solve the speaker normalisation problem independently. They also proposed a caregiver-robot interaction model for vowel acquisition (Ishihara et al., 2009; Miura et al., 2007; Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003), based on the finding that maternal vocal imitation elicits infant vocalisations (Pelaez et al., 2011). In this model, the robot produces random speech sounds and the caregiver imitates the robot's vowel production. More recently, the research group has revised their auto-mirroring model due to the developmental evidence

showing that only in 20% of cases the caregiver imitates the infant utterances (Gros-Louis et al., 2006). They show that with a less imitative caregiver, the model can still learn how to produce vowels (Miura et al., 2012). Rasilo and his colleagues have followed this line of research and built a more sophisticated model that can tackle natural caregiver-infant interactions with ambiguity (Rasilo et al., 2013; Rasilo & Räsänen, 2017). The model starts with uniformly sampled random articulations to explore vocal space. It is then guided by fully online social feedback from the caregiver (i.e., human participants) so that the infant can correspond its own production with vowel categories of the caregiver. In the end, the model has succeeded in learning eight vowels in Finnish. Taken together, the alternative approach of shifting the burden to the caregiver is again not effective enough in solving the speaker normalisation problem, as far as consonant acquisition is concerned.

#### 2.1.4 REINFORCEMENT LEARNING

Reinforcement learning is a mechanism in which an agent learns an action policy in a dynamic environment through trial and error (Kaelbling et al., 1996). The agent tries to find a balance between exploration of unknown regions and exploitation of available knowledge to maximise rewards. The algorithm requires neither explicit correcting actions, nor specifying how the task can be achieved. Actions simply get strengthened or weakened depending on the defined reward or penalty. On the one hand, the models of infant–caregiver interaction introduced in Section 2.1.3 use reinforcement learning based on extrinsic social rewards. On the other hand, a number of studies have explored the possibility of using intrinsic reinforcement. Warlaumont and her colleagues combined self-organisation models (Willshaw, 2006) with reinforcement learning to train speech motor learning (Warlaumont et al., 2013; Warlaumont & Finnegan, 2016). The model produces spontaneous speech with a self-organising map that controls the muscles of a speech synthesiser. The reinforcement signal comes from the auditory salience of these randomly generated sounds. Once it reaches the

auditory criteria of phonation or approximation to targeted vowels, the motor command will be reinforced. The auditory salience is calculated using Mel-transformed F0, F1 and F2 in Warlaumont et al. (2013) and a model of cochlear processing in Warlaumont & Finnegan (2016). Although the model has been trained to learn syllabic sounds, the spectrograms of the audio samples do not show any trace of consonants.

Reinforcement learning can be combined with vocal mimicry or parental-feedback-based learning. For instance, Murakami et al. (2015) has incorporated reinforcement learning with imitation to train VocalTractLab (Birkholz, 2013) to learn vowels. The agent adjusts the motor parameters in an iterative manner to maximise reward signals. The study shows that the supplementary visual reinforcement signal is advantageous in acquiring rounded vowels. Recent studies by Acevedo-Valle et al. (2017, 2018, 2020) integrate caregiver-interaction with self-motivated exploration. In this framework, the model is driven by social reinforcement from a simulated instructor with additional somatosensory feedback that predicts the motor action. The scope of the simulation is again restricted to vowel acquisition.

#### 2.1.5 GOAL BABBLING

Goal babbling is an approach for learning high-dimensional kinematics of robotics without prior knowledge (Rolf et al., 2010). The emphasis of this strategy is on trying to reach multiple goals from scratch. Through exploration, an internal model that describes the relation between motor commands and desired goals is established. The action-goal pairs get updated iteratively along a path towards the desired outcome. The method is different from feedback-error learning that demands prior knowledge of motor error (Kawato, 1990; Wolpert & Kawato, 1998). Forestier and Oudeyer (2017) brought together the learning of motor movements and speech production by goal babbling. The robot was trained to simultaneously learn arm movement, tool use as well as toy names, whereby the exploration was directed by the goal of retrieving objects by arms, tools or vocal requests.

Philippsen and her colleagues developed a vocal learning model with a goal space of vectors derived from acoustic features, as illustrated in Figure 10. An Echo state neural network (ESN) (Jaeger, 2001) was first trained to extract the time-series information of MFCC features. The obtained ESN representations were transformed to a two-dimensional sensory goal space by Principal Component Analysis (PCA) (Wold et al., 1987) and Linear Discriminant Analysis (LDA) (Fisher, 1936). The agent was equipped with an articulatory synthesiser (i.e., VocalTractLab (Birkholz, 2013)) and the articulatory trajectories were controlled by Dynamic Movement Primitives (DMP) (Schaal, 2006). The model managed to learn a mapping between sensory goal space and motor commands. However, the learned speech was limited to vowels and simple CV sequences including /maa/ and /baa/.

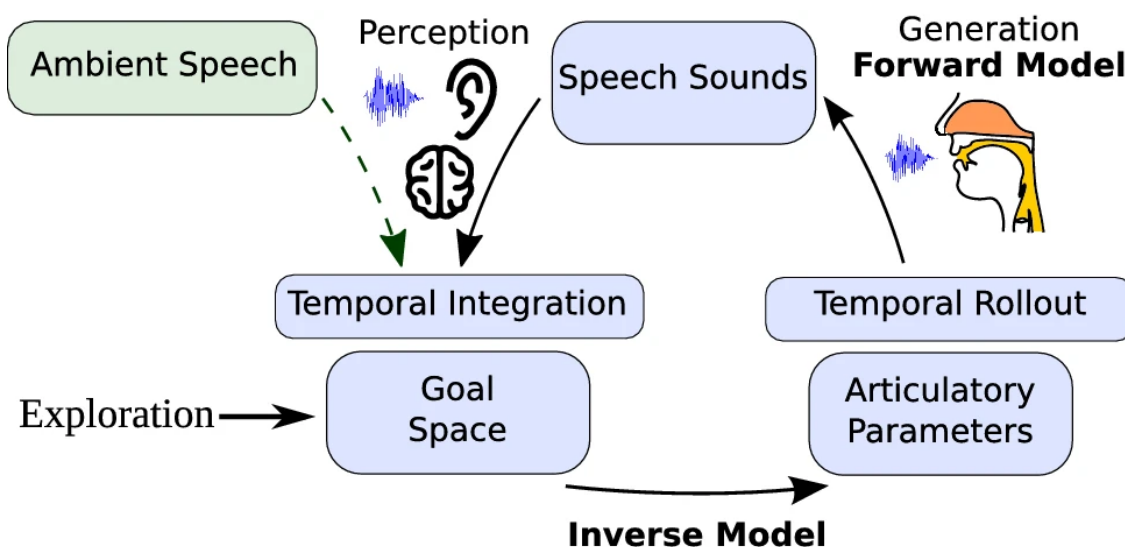


Figure 10 Illustration of goal babbling for speech acquisition (Philippsen, 2021, Figure 1)

## 2.1.6 SELF-ORGANISATION

Of particular concern is how to model the gradual learning behaviour during speech development. The progressive emergence of vocal sequences has been modelled as self-organised systems in some recent studies. Self-organisation,



grounded in evolutionary theory (Kauffman, 1992), refers to the process that a highly complex and dynamic system develops internal structure by interacting with its own distinct subsystems rather than through external instructions (Willshaw, 2006). The idea that spontaneous ordering arises among biological oscillators such as nervous systems has become a useful concept in computing (Watts & Strogatz, 1998). Self-organisation is suggested to play a role in shaping the sound inventory in world's languages (Lindblom et al., 1983). de Boer (2000) implemented the idea by training an agent with an imitation game to simulate how vowel systems emerge. Along this line, Oudeyer (2005) proposed a self-organisation model for speech acquisition, which is endowed with three virtual agents: 1) a vocal tract agent, 2) an ear agent, and 3) a brain agent that couples these two subsystems. The agent is able to discover vowel inventories based on its own subsystems without social interactions. A follow-up study revisited the model by adding a goal for babbling (Moulin-Frier & Oudeyer, 2012). They compared models of random exploration, random goal reaching and active goal reaching and found that active learning led to continuing exploration of auditory and acoustic space. They further provided the model with ambient language to test the possible influence of speech environment (Moulin-Frier et al., 2014). The model demonstrated developmental changes as a result of intrinsic motivations. Recently, the same research group proposed a model for the emergence of phonological systems, referred to as 'Communicating about Objects using Sensory–Motor Operations' (COSMO) (Barnaud et al., 2019; Moulin-Frier et al., 2015). Using a Bayesian modelling approach, motor and auditory systems were linked through linguistic objects to develop a mapping between articulation and acoustics. The framework showed how vowel systems and syllabic sounds could be self-organised during the early stages of learning.

Kanda et al. (2009) focused on how a model can learn to self-organise vowel articulation and speech segmentation. A recurrent neural network with parametric bias (RNNPB) (Tani, 2002) was trained to map time-series acoustic signals with articulatory movements of the Maeda synthesiser (Maeda, 1990). The parametric bias of the model was then manipulated to imitate vowel sequences by

approximating acoustic vectors derived from MFCCs. More recently, Najnin and Banerjee (2017) proposed a predictive coding framework to learn an internal model to predict sensory outcomes. The agent first learns from self-exploration and then imitates environmental speech driven by sensory prediction without reinforcement. Based on an intrinsically motivated architecture, the system learned to produce some vowels and syllables. The main focus of these studies is how an infant discovers sound systems at the very early developmental stage (i.e., babbling), but whether self-organisation is applicable to the learning of intelligible words remains unclear.

### 2.1.7 SUMMARY

On the one hand, some researchers have explored various possible computational models, including learning architectures based on neurobiologically motivated approaches (Kröger et al., 2009; Tourville & Guenther, 2011), acoustic imitation under the distal learning framework (Philippsen et al., 2014; Prom-On et al., 2014a), caregiver's feedback (Messum & Howard, 2015; Miura et al., 2012), reinforcement learning (Warlaumont & Finnegan, 2016) and goal babbling (Philippsen, 2021a; Philippsen et al., 2016). On the other hand, some other researchers are more interested in how children discover phonological systems. The developmental change has been modelled by self-organisation (Moulin-Frier et al., 2014) and Bayesian models (Barnaud et al., 2019; Moulin-Frier et al., 2015). However, even with the state-of-the-art machine learning algorithm and articulatory synthesisers, the speaker normalisation problem remains unsolved (see Appendix Table A). Although numerous learning strategies have been proposed as potential solutions of human vocal development, not enough attention was paid to articulatory dynamics and sensory feedback. More importantly, none of the studies have demonstrated successful learning of intelligible words containing CV syllables, not to mention that very few of them have even conducted systematic listening experiments to verify the perceptual quality of the vocal learning results.

## 2.2 SPEECH SENSORY SYSTEM

As far as vocal learning models are concerned, besides different learning strategies, a sophisticated sensory system is of great importance. Almost all of the previous simulations incorporate some auditory signals but only a few of them include somatosensory input (Appendix Table A Sensory system). In this section, I will first review observations and theories of speech perception and then discuss what kind of sensory representations are plausible training signals for the modelling of vocal learning.

### 2.2.1 BACKGROUND

A long-standing issue in speech perception is how the auditory system decodes linguistic categories despite variable acoustic information. First of all, the same phonetic units exhibit variable surface acoustic forms in different linguistic contexts. Liberman et al. (1954) was the first to report that drastically different acoustic signals can be perceived as the same consonant in different vowel environments. Take /d/ for example, the formant transition is distinctly different in /di/ and /du/, as shown in Figure 11. Secondly, not only does the linguistic context influence the acoustic signals, there are also cross-speaker differences. A seminal work by Peterson and Barney in 1952 demonstrates overlapping clusters of vowels produced by men, women and children in the acoustic space defined by the first two formants (F1 and F2). Interestingly, these vowels were still correctly identified by native listeners. For instance, spectrograms of /bad/ produced by male and female native British English speakers show that vowel formant frequencies are quite different for different individuals (Figure 12). The phenomenon that there is a considerable variation in speech acoustics across speakers while producing the same word, and that listeners still recognise them despite the variation, is known as ‘speaker normalisation’ (K. Johnson, 2005; K. Johnson & Sjerps, 2021). Together, transforming continuous acoustic space across linguistic context and speaker space to discrete perceptual space seems

especially critical for speech perception. It poses an important question: how do we perceive the highly variable sensory signals that are associated with linguistically invariant categories?

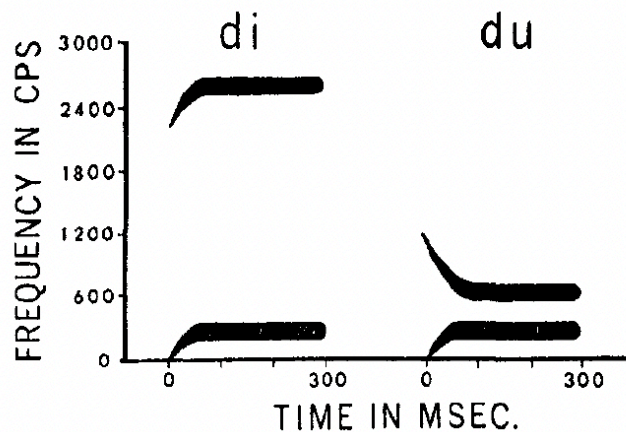


Figure 11 Simplified spectrograms of consonant /d/ followed by vowel /i/ and /u/, adapted from Liberman et al. (1954)

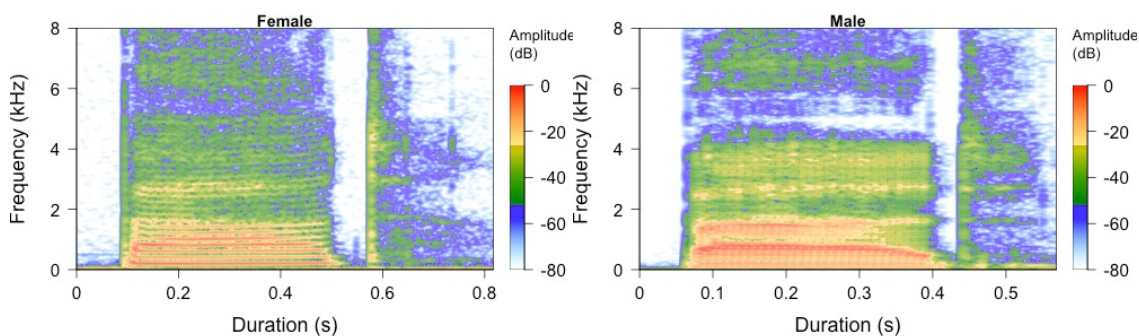


Figure 12 Narrow-band spectrograms of /bad/ produced by a female and a male native British speaker respectively.

## 2.2.2 MECHANISMS BEHIND SPEECH PERCEPTION

The motor theory of speech perception (Liberman et al., 1967; Liberman & Mattingly, 1985) is one of the earliest conceptual frameworks that tries to resolve the lack-of-invariance problem. Their proposals are 1) the motor system is recruited for speech perception, and 2) speech perception is perceiving gestures. The theory was first developed to explain the variable acoustic characteristics of

coarticulated syllables, as shown in Figure 11 (Liberman et al., 1952, 1954). Based on the experimental findings, Liberman and his colleagues suggest that the nature of speech perception is perceiving the intended phonetic gesture of the speaker. The assumption is shared by Fowler (1981, 1986, 1989), also from Haskins laboratories, who proposed direct-realist theory of speech perception, inspired by Gibson (1966). The theory resembles the motor theory, which also advocates that the object of speech perception is not acoustic but articulatory events. However, direct-realist theory is dissimilar from the motor theory in that the object of perception is not neuromotor commands and that the mechanism is not speech-specific.

The motor theory of speech perception has attracted much attention since the discovery of 'mirror neuron'. The accumulating research on mirror systems in speech perception seems to support the motor theory on the basis of the analogy in the description (Galantucci et al., 2006). However, neurological studies do not offer direct support for the presence of linguistically invariant gestures during speech perception. The neurological evidence assuring the existence of mirror neurons also demonstrated no statistical difference in the participants' motor-evoked potentials when real words and non-speech sounds (i. e., bitonal sounds) were played to them (Fadiga et al., 2002b). This indicates that the mirror system itself is not necessarily for the sake of perceiving articulation. Rather, it only shows a well-established fact that production and perception are linked. To accommodate new neurobiological evidence (Kohler et al., 2002), the advocates of the motor theory abandoned the original claim of speech-specific components (Galantucci et al., 2006), and the revised assumption largely aligns with the direct realist perspective (Fowler, 1981, 1986, 1989).

A counterexample to the motor theory is that speech perception can be intact even with deficits in speech articulatory control for patients with Broca's aphasia (Goodglass, 1993). As a consequence, there has been concern over the necessity of articulatory reference as an intermediate level between acoustics and linguistic categories (Lindblom, 1996; Ohala, 1986). A contrasting view to the motor theory and the direct realist is that the auditory characteristics of phonetic

segments are what the listeners are perceiving instead of articulatory events, known as the general approach (Diehl & Kluender, 1989). It is argued that the speech sounds are perceived in the same way as environmental sounds using general mechanisms of audition that have evolved to categorise complex acoustic signals. The accessibility to phonetic information is through auditory enhancement and contrast in spectral and durational features (Lotto & Kluender, 1998). The general auditory approach can account for the fact that some birds (Kluender et al., 1987) and chinchilla (Kuhl & Miller, 1978) can be trained to identify phonetic categories even though they cannot articulate human speech. Both the motor theory and the direct realist theory cannot fully explain why animals without human vocal apparatus are still able to detect phonetic cues in human languages after training.

In line with the general approach, some researchers posit that the context-dependent perception of linguistic units arise from listeners' accommodation of speaker's voice characteristics. There is considerable amount of behaviour research suggesting that the long-term and short-term frequency context of sounds facilitates the perception of phonemic contrast (Laing et al., 2012; A. J. Watkins, 1991). The idea of contrast enhancement postulates that the auditory system is tuned based on the statistical distribution of the acoustic input. The assumption is supported by experimental evidence revealing the neural underpinnings of speaker-normalised (Sjerps et al., 2019) and categorical (Chang et al., 2010) representations in the human auditory cortex. The neural responses to speech sounds show both context-dependent and context-independent normalisation (Sjerps et al., 2019).

The nature of speech perception has been controversial and much disputed in general. Much of the accounts up to now have been descriptive and lack specifications. An alternative way of investigating the underlying mechanisms is to model speech perception with realistic speech data. If we can simulate how the human perception system procedurally handles the statistical properties of ambient speech, it may lead us to a deeper understanding of the system.

### 2.2.3 SIMULATION OF SPEECH SENSORY SYSTEM

The existing body of research on computational models of vocal learning has explored different methods of emulating auditory feedback (see Appendix Table A Sensory system). Most of the research adopts vowel formants (i.e., F1, F2 and sometimes F3) to characterise vowel quantitatively (Acevedo-Valle et al., 2020; Bailly, 1997; Forestier & Oudeyer, 2017; Heintz et al., 2009; Howard & Huckvale, 2005; Miura et al., 2012; Rasilo & Räsänen, 2017). Vowel formants are the resonance frequencies of the vocal tract, which has been conventionally regarded as the most salient perceptual dimensions of vowels (K. Johnson, 2005). An alternative approach is to use scaled formants such as Mel-scale (Warlaumont et al., 2013) and Bark-scale (Barnaud et al., 2019; de Boer, 2000; Kröger et al., 2014; Moulin-Frier et al., 2014) formants to determine the perceptual space for vowel learning. Mel-scale reflects human perceptual distance of pitches (Stevens et al., 1937). In contrast, Bark scale denotes critical bands of frequency response of the human ear (Zwicker, 1961). A comparison of Mel-scale, Bark-scale and Hz is displayed in Figure 13. For both Mel-scale and Bark-scale, the sensitivity is high for low frequencies below 500 Hz, which is almost linear. Bark-scale is slightly more sensitive to low frequencies below 1000 Hz than Mel-scale. When the frequency is above 1000 Hz, the sensitivity decreases for both Mel-scale and Bark-scale. The two scales are in general similar in terms of the frequency sensitivity.

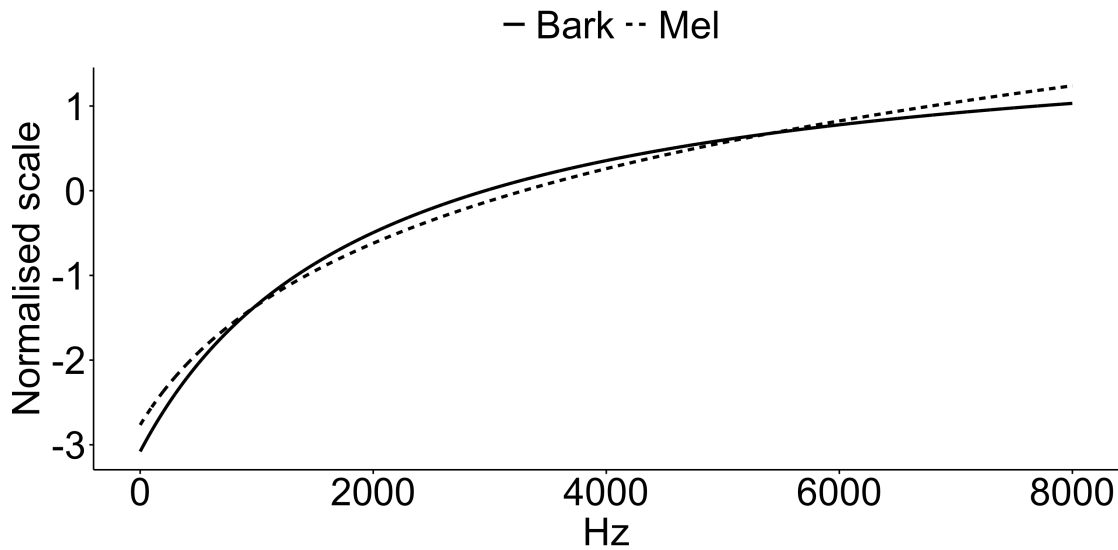


Figure 13 Bark and Mel scale as a function of frequency from 0 to 8000 Hz. X axis shows the physical frequency in hertz and Y axis shows the normalised scale.

However, formants might be less appropriate when the target of learning is beyond vowels and approximants. In addition to the lack of energy during the closure, the formant cues to consonants are limited to transitional movements. Recent research has sought for more sophisticated acoustic features. A few studies have made efforts to include more acoustic details using Bark-scale spectrograms (Kröger et al., 2014) or gammatone spectrograms (Howard & Messum, 2014; Messum & Howard, 2015). One of the most popular acoustic representations in the simulation studies is Mel-frequency cepstral coefficients (MFCCs) (Kanda et al., 2009; Prom-On et al., 2014a, 2014b; Rasilo et al., 2013). MFCCs are obtained from the log Mel-filterbank energy features. It is a robust parametric representation of speech acoustics (S. B. Davis & Mermelstein, 1980), widely used in speech recognition and Hidden Markov Model (HMM)-based synthesis. There are also some studies that have tried to combine MFCCs with formants as the auditory input (Najnin & Banerjee, 2017; Philippsen et al., 2014).

With development in speech signal processing, the approach towards realistic simulation of speech perception has gradually improved. Murakami et al. (2015) has proposed an auditory system that generates reward by BRIAN neural network simulator. The auditory system simulates the peripheral processing in



the cochlea by dual resonance nonlinear (DRNL) filtering of the sound signals (Fontaine et al., 2011; Lopez-Poveda & Meddis, 2001). It is implemented by Echo State Network (ESN), which is a recurrent neural network with connected hidden layers (Jaeger & Haas, 2004). ESN is suitable for modelling time-series data such as speech signals. Philippsen (2021) also adopts ESN to encode the temporal information of spectral changes. Unlike Murakami et al. (2015), Philippsen (2021) applied Principal Component Analysis (PCA) (Wold et al., 1987) and Linear Discriminant Analysis (LDA) (Fisher, 1936) to the ESN representations to reduce the dimensionality of the acoustic vectors. In addition to the tendency towards incorporating acoustic details and temporal characteristics, researchers have further explored the possibility of simulating categorical perception. DeeChee model proposed by Lyon et al. (2012) is equipped with an automatic phoneme recogniser (an adapted version of Microsoft SAPI 5.4) for capturing the statistical distribution of phonemes in speech perception. Although extensive research has been carried out, no research has systematically examined how various auditory representations influence vocal learning.

A wealth of evidence supports the idea that the auditory system plays a critical role in the learning of speech production (Kuhl, 2000, 2004). Somatosensory feedback has also been found to interact with auditory feedback and impact on the learning (Lametti et al., 2012). The somatosensory system integrates tactile and proprioceptive signals of articulator movements during speech production. By simultaneously manipulating two types of feedback during speech production, experimental studies found that auditory feedback and somatosensory feedback both influence speech learning but auditory feedback shows higher priority (Lametti et al., 2012). Studies on vocal learning has recognised the critical role played by somatosensory feedback in production learning. For example, the DIVA model simulates somatosensory representations by the temporal and spatial states of articulatory movements (Guenther, 1994; Kröger et al., 2009, 2014; Tourville & Guenther, 2011). Another alternative approach is to construct a somatosensory model that maps motor commands to internal somaesthetic representations, as in Acevedo-Valle et al. (2017, 2018, 2020). In this case, the

somatosensory system provides feedback on whether the articulation is fully executed.

Despite various methods of emulating auditory and somatosensory feedback, the simulations reviewed above did not systematically examine how it impacts on vocal learning. Therein lies the necessity of verifying the role of feedback by comparing the model performance in controlled simulation experiments. Previous vocal learning models have failed to generate intelligible words containing CV syllables, which makes it difficult to use them to evaluate the plausibility of different kinds of sensory feedback. In this study, the aim is to examine the role of auditory and somatosensory feedback by simulating the learning of real English words and quantitatively comparing the learning outcome.

## 2.3 SPEECH MOTOR CONTROL

*Part of this section has been previously been published in conference proceedings (ICPhS 2019) with the title “Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation” and have been made available at [assta.org/proceedings/ICPhS2019/papers/ICPhS\\_254.pdf](https://assta.org/proceedings/ICPhS2019/papers/ICPhS_254.pdf).*

Another key aspect of a vocal learning model is the plausibility of the motor control system. In contrast to the enormous amount of effort to explore learning strategies in previous works, surprisingly little attention has been paid to the articulatory dynamics during speech production (Appendix Table A Motor control and Synthesiser). In addition to a realistic model that can generate natural sounding speech, the control of the dynamic movements of the articulators is equally important. Of particular concern is the phenomenon of consonant-to-vowel (CV) coarticulation in speech production (i.e., the articulatory movement of the consonant varies with the following vowel), which is the main challenge faced by many previous models (Appendix Table A Learning target and Performance). In this section, I will first introduce the motor control systems used in past

simulations and then focus on discussing CV coarticulation and its implementation.

### 2.3.1 BACKGROUND

The early simulations modelled the vocal tract as a source-filter model (Miura et al., 2007; Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003) or a pipe model (Westerman & Miranda, 2002; Westermann & Miranda, 2004). A few studies adopted articulatory synthesis systems such as Praat synthesis (Warlaumont, 2012; Warlaumont et al., 2013; Warlaumont & Finnegan, 2016) and Rasilo's Articulatory model (Rasilo et al., 2013; Rasilo & Räsänen, 2017). The majority of the studies used the Maeda synthesiser (de Boer, 2000; Kanda et al., 2009) or its modified versions including VLAM (vocal linear articulatory model) (Barnaud et al., 2019; Heintz et al., 2009; Moulin-Frier & Oudeyer, 2012), VTCALCS (Acevedo-Valle et al., 2017, 2018, 2020; Guenther et al., 2006b; Howard & Messum, 2007, 2014; Messum & Howard, 2015; Moulin-Frier et al., 2014; Najnin & Banerjee, 2017; Tourville & Guenther, 2011) More recently, VocalTractLab, an articulatory synthesis with high-dimensional vocal tract parameter control, has been used in a number of studies (Murakami et al., 2015; Philippsen et al., 2014, 2016; Prom-On et al., 2014a, 2014b).

The various speech synthesisers are controlled by different motor systems to generate articulatory kinematics. Some studies use the task dynamic model (Fowler & Saltzman, 1993; Saltzman & Munhall, 1989) for controlling the trajectories of articulatory movements (Howard & Messum, 2007, 2014, 2011; Messum & Howard, 2015). The task dynamic model was initially a conceptual framework for explaining how linguistically contrastive units can be organised to generate contextually varying articulatory kinematics. It was later implemented mathematically as a second-order dynamical system (Saltzman & Munhall, 1989). As an alternative approach, Philippsen (2021) and Forestier and Oudeyer (2017) incorporated Dynamic Movement Primitives (DMPs) framework (Ijspeert et al., 2013; Schaal, 2006) to control VocalTractLab and DIVA synthesiser

(customised Maeda synthesiser), respectively. DMPs was developed to plan the trajectories for the motor movements of robots with discrete or rhythmic nonlinear dynamic primitives (Schaal, 2006). Whilst these computational models have been mostly restricted to vowel acquisition (Appendix Table A Learning target), up to now, no study has successfully simulated the learning of intelligible CV syllables (Appendix Table A Performance). This suggests that one of the main obstacles for the simulation of the motor control system is the modelling of CV coarticulation.

### 2.3.2 COARTICULATION

The term ‘coarticulation’ (‘Koartikulation’) was first proposed by Menzerath and De Lacerda (Menzerath & de Lacerda, 1933) to describe the phenomenon that the articulatory movement of the vowel in a CV sequence starts at the same time as the consonant (Kuhnert & Nolan, 1999). By now, however, it is mostly used to refer to any variation of a segment with adjacent or nearby segments. The contextual variability of segments has intrigued theoretical discussions on how linguistically invariant segments take various articulatory-acoustic manifestations in speech production.

Over the last century, researchers have shown enthusiasms for elucidating the mechanisms underlying CV coarticulation. Kozhevnikov & Chistovich (1965) proposed the concept of “Articulator syllable” supported by the evidence of co-onset of lip rounding and the movement of the first consonant in Russian complex syllables (i.e., CV, CCV and CCCV syllables). According to this theory, the motor command of the consonant and the vowel are set simultaneously at the syllable onset, resulting in coarticulation. In the task dynamic model (Fowler & Saltzman, 1993; Fowler, 1980; Saltzman & Munhall, 1989) and Articulatory Phonology (Browman & Goldstein, 1989, 1992; Ohala et al., 1986), it is assumed that there are temporal overlaps between linguistically relevant movements of the vocal tract, referred to as gestures. In a /VdV/ sequence, for example, the alveolar consonant and the vowel gesture compete for the control of the jaw, the tongue

tip and the tongue body. This coproduction process is called intergestural blending. In the window model of coarticulation, a segmental feature has a 'window' consisting of a maximum and a minimum physical value that reflects contextual sensitivity. Articulatory execution of segments is an interpolation process that paths through these windows (P. A. Keating, 1990).

Another line of study tended to focus more on how to quantify the coarticulatory movements. Bladon & Al-Bamerni (Bladon & Al-Bamerni, 1976) hypothesised that there was a specific 'coarticulation resistance' value associated with each segment, based on the observation that the allophones of English /l/ influenced the F2 of neighbouring vowels to varying extents. Later on, quantitative measurements of coarticulation resistance have developed with the advances in new articulatory imaging techniques. A degrees of articulatory constraint model of coarticulation (DAC model) has been proposed, which quantifies the coarticulation resistance of segments by means of acoustic, EPG and EMA data (Recasens, 1984; Recasens & Espinosa, 2009). Jackson-Singampalli's statistical model (Jackson & Singampalli, 2009) and Iskarous et al.'s Mutual Information Scale (Iskarous et al., 2013) both seek to identify the primary articulators for phones by measuring the distribution of their changes in physical positions in different linguistic environments. Specifically, the vertical anterior part of the tongue position was found to correlate with the articulation of alveolar stops; the vertical posterior part position of the tongue was critical for velar stops and the vertical lip position for bilabial stops. These measurements have revealed to what extent a specific articulator is involved with the presence of distinct surrounding segments.

In addition to coarticulation resistance, there have been some efforts that measure the coarticulation phenomenon quantitatively. Lindblom (1963) observed highly linear relation between formant frequencies at the centre of vowels and at the point right after the consonant release, and referred to this linear relation as locus equations. The linearity has been suggested to parallel findings in coarticulatory resistance (Brancazio & Fowler, 1998), degree of

coarticulation (Lindblom & Sussman, 2012; Löfqvist, 1999) and articulator synergy (Iskarous et al., 2010). On the other hand, some researchers failed to find the correlation between Locus Equation and the kinematic process during coarticulation (Tabain, 2000, 2002). As an extension of Locus Equation, Öhman (1967) proposed a mathematical function which treated the consonant gestures as constrictory gestures superimposed on a diphthongal movement of vowels. The results showed that the calculated vocal tract shapes almost resembled the empirical X-ray data. This work inspired a large body of literature that adopted different approaches to calculate vocal tract area functions for the modelling of coarticulation in speech synthesis (Birkholz, 2013; Carré & Chennoukh, 1995; Chennoukh et al., 1997; Story, 2005, 2009). For example, Story (2005, 2009)'s model implemented Öhman (1967)'s idea and simulated the superposition of the consonant on the vowel movements with the aid of MRI and X-ray data. Recently, Birkholz (2013) modelled the vocal tract shape of context-sensitive consonants based on weighted means of reference shapes of consonants following point vowels (i.e., /a/, /i/ and /u/), via acoustic optimization. These synthesis systems rely on articulatory data to pre-define the vocal tract shapes. Difficulties arise, however, when the coarticulation model is applied to the simulation of vocal learning, as learners would not have access to the knowledge of articulation behind the speech utterances they hear.

### 2.3.3 THE SYNCHRONISED DIMENSION-SPECIFIC SEQUENTIAL TARGET APPROXIMATION MODEL

Different from previous attempts, the synchronised dimension-specific sequential target approximation model offers a highly specific way of simulating the learning of coarticulation (Liu et al., 2022; Xu, 2020). The framework was proposed to explain how the articulators are coordinated during coarticulation, and how temporal coordination may benefit vocal learning. It is hypothesised that despite the co-onset of the underlying consonantal and vocalic targets, at the level of individual articulator dimension, the target approximation is sequential (Xu,

2020). Under this model, specific vocal tract parameters controlled by the consonant and the vowel asymptotically approach their targets simultaneously at the syllable onset, although ending at different time points. For example, in a /gV/ sequence, the tongue body vertically moves upward for a contact for the consonant, while also moving horizontally to achieve the tongue shape for the vowel, resulting in different velar contact locations depending on the vowel. The hypothesis is supported by experimental data from Electromagnetic Articulography (EMA) and acoustics that show synchronised CV coarticulatory movements in disyllabic words (Liu et al., 2022). So far, the synchronised dimension-specific sequential target approximation model has not been tested by computational modelling.

## **Chapter 3   SIMULATION OF VOCAL LEARNING**

In this chapter, I will describe a computational model of vocal learning. The model is distinct from the models reviewed in Chapter 2 in the following aspects.

1. Previous neurobiological models (Section 2.1.1) mainly concentrate on modelling the relationship between the sensory and motor systems, whereas the current approach aims at successful learning that reaches a high level of intelligibility.
2. The model simulates production learning guided by either universal perception or language-specific perception. The universal perception is emulated by acoustic features, in which case the learning process is in essence acoustic imitation (Section 2.1.2).
3. Unlike models of infant-caregiver interactions (Section 2.1.3), the model learns speech production guided by sensory feedback on its own without the assistance from the caregiver.

4. The framework is driven by the hypothesis of perception-guided learning in vocal learners, rather than based on a particular algorithm such as reinforcement learning (Section 2.1.4), goal babbling (Section 2.1.5) or self-organisation (Section 2.1.6).

5. The model simulates the phase of learning from babble to speech, in contrast with self-organised models of initial stages of speech acquisition (Section 2.1.6).

### 3.1 MODEL OVERVIEW

The model consists of a motor control and a sensory component, as illustrated in Figure 14. The motor control model begins with exploration of a set of articulatory targets within the available parameter range for the adult or the child vocal tract models (Figure 14A). The articulatory synthesiser is VocalTractLab 2.3 (Birkholz, 2013), a geometrical 3D vocal tract model. The adult vocal tract model is based on the volumetric MRI data of a German male speaker and the child vocal tract models are scaled versions of the adult model (Birkholz & Kröger, 2007; Goldstein, 1980). During the exploration, the vocal tract model adjusts 19 vocal tract parameters that determine the dynamics and physical locations of the active articulators. The kinematic trajectories that approach the articulatory targets are based on the timing relations specified by a coarticulation model – the synchronised dimension-specific sequential target approximation model (Xu, 2020). The model simulates context-sensitive realisation of consonants and vowels (Figure 14B). The time-varying vocal tract shapes are then converted to cross-sectional area functions for acoustic simulation (Figure 14C). The aero-acoustic simulation is based on the enhanced area function (Birkholz, 2014) of the time-varying vocal tract shapes to generate spoken words. The synthetic speech is evaluated either by acoustic features (Figure 14D) or by an automatic phoneme recogniser (Figure 14E). Mel-spectrograms of natural words and synthetic words were extracted to calculate the Mel-frequency cepstral coefficients (MFCCs) (S. B. Davis & Mermelstein, 1980). The automatic phoneme recogniser is a pre-trained deep learning model that maps acoustic feature



sequences to a contrastive auditory space. It evaluates the probability of the targeted onset consonants, vowels, and coda consonants, as represented by International Phonetic Alphabet (IPA) symbols around the circle. In addition to the two types of auditory feedback, somatosensory information of oral constriction sensing is also included (Figure 14F). The somatosensory information is provided by the cross-sectional areas to determine whether there is a closure in the vocal tract. The adult and the child vocal tract models were trained to learn English words guided by the sensory feedback options in Figure 14D to Figure 14F.

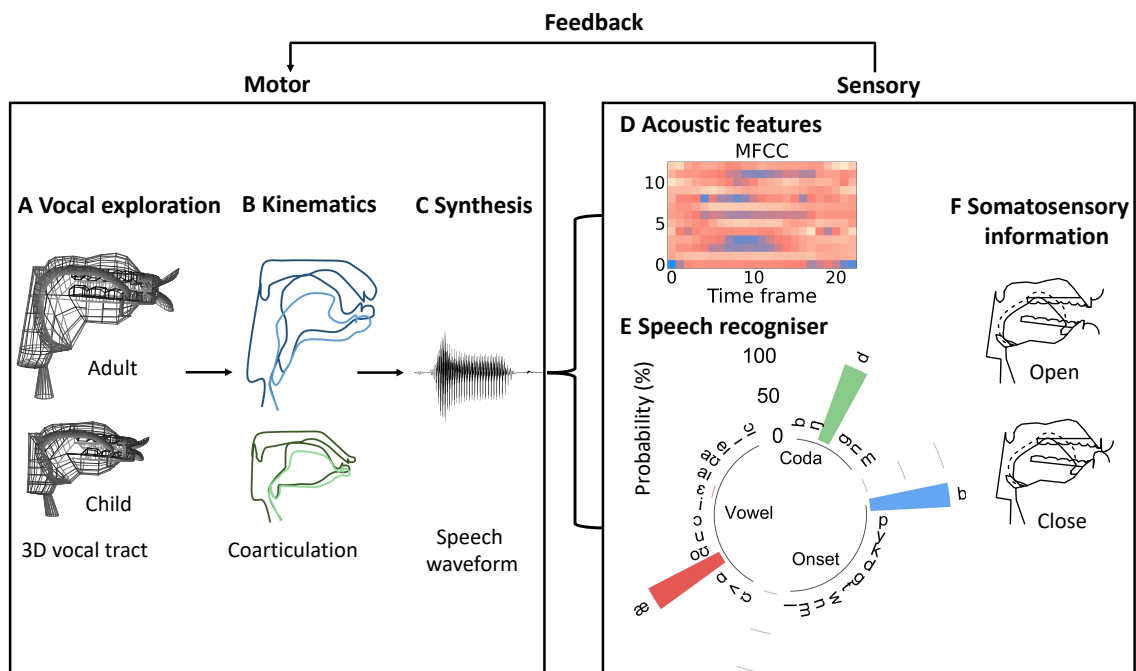


Figure 14 Overview of the vocal learning model

The learning targets of the vocal learning model are minimal pairs of real English words with consonant-vowel-consonant (CVC) and CVCV syllable structures, containing bilabial, alveolar, and velar stops, as shown in Table 1. Voiced stops were selected as the onset consonant to maximise the number of vowels available in English CVC words. The CVC words are first optimised during the simulation and the learned vocal tract parameters are reused to resynthesise CVCV words to test the generalisability of the learned parameters.

Table 1 Target words for the vocal learning model

	CVC		CVCV			
	/bVd/	/dVd/	/gVd/	/bVbi/	/dVbi/	/dVdi/
/i:/	bead	deed				
/ɪ/	bid	did				
/ɛ/	bed	dead			Debbie	
/æ/	bad	dad				daddy
/ɒ/	bod		god	body		
/u:/	booed					
/ʌ/	bud			buddy		
/ʊ/			good			

## 3.2 MOTOR CONTROL

A key feature of the vocal learning model is the explicit emulation of the motor control system in many respects. A biological-plausible articulatory synthesiser was adopted to simulate the child and the adult vocal systems. The articulatory dynamics of CV sequences were controlled by the synchronised dimension-specific sequential target approximation model.

### 3.2.1 ARTICULATORY SYNTHESIS

The articulatory synthesiser (Figure 14A) used is VocalTractLab 2.3 (vocaltractlab.de) (Birkholz, 2013), which calculates enhanced area functions (Birkholz, 2014) for an aero-acoustic simulation on the basis of a 3D vocal tract model and a geometrical glottis model. The adult vocal tract model is adapted from MRI data of a German male speaker. The static and the active articulators of the infant vocal tract models are scaled down based on the relative anatomy (Birkholz & Kröger, 2007; Goldstein, 1980). The vocal tract parameters define the

airway from the glottis to the lips, with 17 degrees of freedom, as shown in Table 2. The cross-sectional area of the oral cavity is converted to a transmission-line model for acoustic simulation in the time domain. The vocal tract parameters are sampled at 5 ms intervals to ensure the precision of the articulatory movement. The vocal fold model is a geometric glottis model which accounts for source-filter interaction during synthesis (Birkholz, 2014). Compared with the traditional glottal flow model, this voice source model is capable of generating asymmetric glottal area waveforms and diplophonic double pulsing. The vocal folds were set to be fully adducted with moderate longitudinal tension for the vowel targets, while the glottis parameters of the consonant targets including the distance between vocal cords, chink area and relative amplitude were free parameters. The intonation contours of the synthetic words were generated using pitch targets extracted from the natural recordings by PENTAtainer (Xu & Prom-on, 2014), an intonation modelling tool. The audio files were synthesised at a sampling rate of 44.1 kHz and a quantisation of 16 bit.

Table 2 Vocal tract parameters in the model.

Parameter	Description	Range
HX	Horizontal hyoid position	[0.0, 1.0] cm
HY	Vertical hyoid position	[−6.0, −3.0] cm
JX	Jaw position	[−0.5, 0.0] cm
JA	Jaw angle	[−7.0, 0.0] deg.
LP	Lip protrusion	[−1.0, 1.0] cm
LD	Vertical lip distance	[−2.0, 4.0] cm
VS	Velum shape	[0.0, 1.0]
VO	Velum opening	[−0.10, 1.0] cm <sup>2</sup>
TTX	Horizontal tongue tip position	[1.5, 5.5] cm
TTY	Vertical tongue tip position	[−3.0, 2.5] cm
TBX	Horizontal tongue blade position	[−3.0, 4.0] cm
TBY	Vertical tongue blade position	[−3.0, 5.0] cm
TCX	Horizontal tongue body centre position	[−3.0, 4.0] cm
TCY	Vertical tongue body centre position	[−3.0, 1.0] cm
TS1 – TS3	Tongue side elevation from the posterior to the anterior part of the tongue	

### 3.2.2 ARTICULATOR DYNAMICS

The temporal and spatial movements of the articulators were simulated by a motor control system that transforms articulatory targets to vocal tract parameter trajectories, as illustrated in Figure 15.

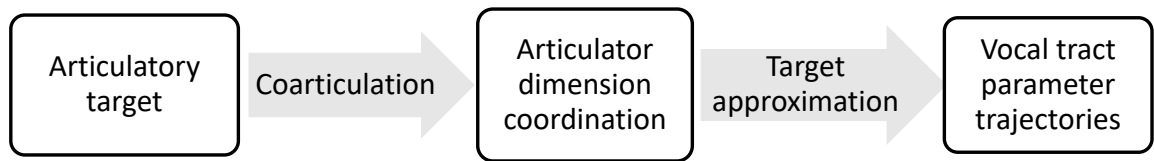


Figure 15 Workflow of the motor control system. The system takes articulatory targets such as consonant or vowel targets as input and returns 17-dimensional vocal tract parameter trajectories (Table 2) to be passed to the articulatory synthesiser.

First, during the approximation of the underlying articulatory targets, the coordination of multiple articulators is controlled by a coarticulation model, the synchronised dimension-specific sequential target approximation model (Liu et al., 2022; Xu, 2020). In this framework, the articulation of the phonetic segment originates from the execution of both the consonant and the vowel target from the syllable onset. Despite the co-onset of consonant and vowel targets, at the level of the individual articulator dimensions, the movement towards each articulatory target is sequential, so that each articulatory dimension is controlled either by the consonant or by the vowel at a particular moment in time. For example, in Figure 16, at the onset of a consonant-vowel (CV) syllable with a bilabial stop, the consonant target (dashed lines) controls the movement of jaw angle (JA), jaw horizontal position (JX) and lip distance (LD), while the vowel target (solid lines) governs the movement of the rest of the articulatory dimensions, such as the horizontal and vertical tongue tip positions (TTX & TTY). When the interval of consonant target approximation is over, JX, JA and LD start to move towards the vowel target. After the articulatory movements toward the syllable-initial consonant and vowel are terminated, all the articulator dimensions begin to approach the next set of articulatory targets. The final coda consonant was implemented as another hypothetical CV syllable with a voiceless schwa to ensure a closure in the oral cavity (coda target) and a release (into the voiceless

schwa) (Xu, 2020). The temporal domain of the motor control system is based on the timing alignment of the articulatory targets in the natural speech sample. The articulatory dimensions governed by the consonant are listed in Table 3, while the rest of the articulatory dimensions listed in Table 2 are controlled by the vowel target only.

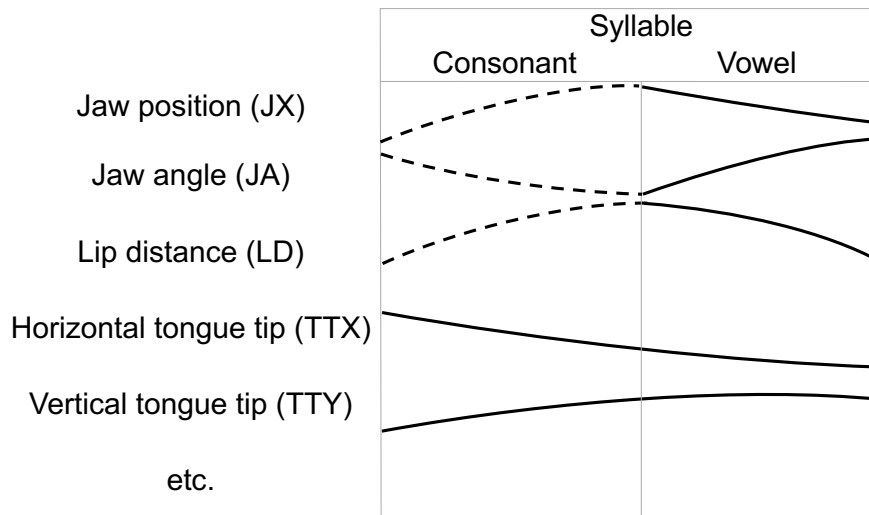


Figure 16 Illustration of the synchronised dimension-specific sequential target approximation model in the case of bilabial stop-vowel sequences. Dashed lines represent the articulatory trajectories of the consonant target and solid lines represent the articulatory trajectories of the vowel target.

Table 3 Vocal tract parameters controlled by the consonant

Consonant type	Vocal tract parameters
Bilabial stops	LD, JX, JA
Alveolar stops	TTY, TBY, TS3, JX, JA
Velar stops	TCY, TS2, JX, JA

Next, after the coarticulation model was applied, the dynamic trajectories of the 17 vocal tract parameters were calculated by the target approximation model. Quantitatively, each articulatory target is represented by the height (i.e., the position of the 17 vocal tract parameters), the slope and the strength (i.e., the

time constant). The slope was set to zero in the simulation because the speech materials include only monophthongs rather than diphthongs which involve dynamic articulatory movements (Xu, 2020). The movement of the articulators (vocal tract parameters) is modelled by a cascade of several identical first-order linear systems with the following transfer function:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{(1 + s\tau)^N}$$

where  $s$  and  $N$  denote the complex frequency and the order of the system respectively.  $\tau$  denotes the time constant, which determines how quickly the target is approached, hence the (inverse of the) strength of target approximation. Here,  $N$  equals 5, that is, a fifth-order system was used, which reproduces s-shaped asymptotic movement towards articulatory targets with bell-shaped velocity profiles. The time-domain representation of the previous equation can be derived using inverse Laplace Transform, which results in:

$$y(t) = (c_0 + c_1 t + \dots + c_{N-1} t^{N-1}) e^{-\frac{t}{\tau}} + x(t)$$

where  $x(t) = b$  is the position of the articulator target and  $t$  is the time from the beginning of the articulatory target. The coefficients are calculated based on the initial state of  $y$  and its derivatives of the articulator at the onset of the interval (which is equal to the final state of the previous target), as shown in the following equation (Birkholz et al., 2011):

$$c_i = \begin{cases} y(0) - b & n = 0 \\ \frac{y^{(n)}(0) - \sum_{i=0}^{n-1} c_i a^{n-1} \binom{n}{i} i!}{n!} & 0 < n < N \end{cases}$$

### 3.3 SENSORY SYSTEM

The sensory model contains two kinds of auditory feedback: 1) acoustic features for simulating universal perception of phonetic differences in all languages (Kuhl,

2000; Werker & Lalonde, 1988), and 2) an automatic phoneme recogniser for simulating language-specific perception of sound contrasts in native languages (Kuhl, 2000; Werker & Lalonde, 1988). Moreover, somatosensory feedback is provided by checking the closure areas within the oral cavity.

### 3.3.1 ACOUSTIC FEATURES

Acoustic features were used to simulate universal perception based on a Mel-filterbank: Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) and Log Mel spectrograms. The Mel-scale approximates human perception of frequency, which is more sensitive to low frequencies than high frequencies (Stevens et al., 1937). A Log Mel spectrogram is the log power output by each filter in the Mel filter bank for each temporal frame of the speech signal obtained through windowing. MFCCs are computed by applying the discrete cosine transform (DCT) to the Log Mel spectrum for each frame. MFCCs are widely used in speech recognition and Log Mel Spectrograms are used in machine learning based speech synthesis. Both are robust representations of phonetic contents of speech.

High-frequency emphasis was applied to the sound signals through pre-emphasis (coefficient = 0.97). Frames were then extracted using 25 ms Hamming windows with 5 ms overlap, to be consistent with the sampling rate of the vocal tract parameters during synthesis. 26 Mel filters<sup>3</sup> with a maximum frequency of 10 kHz were applied and the logpower of their output was calculated. 26 Mel filters The DCT of the Mel log power was calculated to obtain 22-dimensional MFCCs (including energy) with sinusoidal cepstral liftering (coefficient =  $2 \times$  number of MFCCs). 22-dimensional MFCCs<sup>4</sup> were selected to make the best use

---

<sup>3</sup> In a pilot experiment, 40 and 26 Mel filters were applied to obtain Log Mel Spectrograms and MFCCs. The number of filters did not significantly influence the intelligibility of the speech learned by the model.

<sup>4</sup> It has been reported that 12-dimensional and 22-dimensional MFCCs perform similarly in guiding vocal learning models for vowels (Gerazov et al., 2020).



of spectral information while excluding speaker information (Ryant et al., 2014). The acoustic error ( $E$ ) was calculated using the Euclidean distance between the 22-dimensional MFCCs or 26-dimensional Log Mel spectrograms of the target and the synthetic utterances, with the following equation:

$$E = \sum_{i=1}^n \sum_{j=1}^m (f_{ij} - \hat{f}_{ij})^2$$

where  $n$  is the number of time frames,  $m$  is the number of MFCC coefficients or the Log Mel spectrogram filters,  $f_{ij}$  is the  $j^{\text{th}}$  cepstral coefficient/ Mel frequencies of the  $i^{\text{th}}$  frame of the natural sound, and  $\hat{f}_{ij}$  is the  $j^{\text{th}}$  cepstral coefficient/ Mel frequencies of the  $i^{\text{th}}$  frame of the synthetic sound.

Audio recordings of natural speech were made by a female native speaker of American English (age: 27) in a sound-attenuated acoustic laboratory. The sound files were recorded with a studio-grade microphone and a professional audio interface at a sampling frequency of 44.1 kHz with 16-bit quantisation. The use of a female speaker was to address the speaker-normalisation problem (Section 2.2.1) by contrasting with the adult vocal tract model based on a German male speaker.

### 3.3.2 AUTOMATIC WORD RECOGNITION

A word recogniser developed in van Niekerk et al. (2022) and Xu et al. (2022) were used to evaluate the intelligibility of the synthetic words. The word recogniser was trained using the Kaldi Speech Recognition Toolkit and the LibriSpeech corpus (Panayotov et al., 2015). The corpus contains speech data extracted from audiobooks recorded by adult male and female speakers of varied ages. The model is based on Weighted Finite State Transducers (WFSTs) that use Gaussian mixture models (GMMs) to model the speech acoustics. The MFCC features were transformed with Linear Discriminant Analysis (LDA) and the Maximum Likelihood Linear Transform (MLLT) to reduce the dimensionality and the size of the acoustic model. The training data consists of 960-hour speech from LibriSpeech (Panayotov et al., 2015), normalised using Speaker Adaptive

Training (SAT). A small pretrained trigram language model was used in the decoding. The word recogniser was effective in evaluating speech sounds because it was tested to be robust in recognising natural speech, as shown in Figure 17.

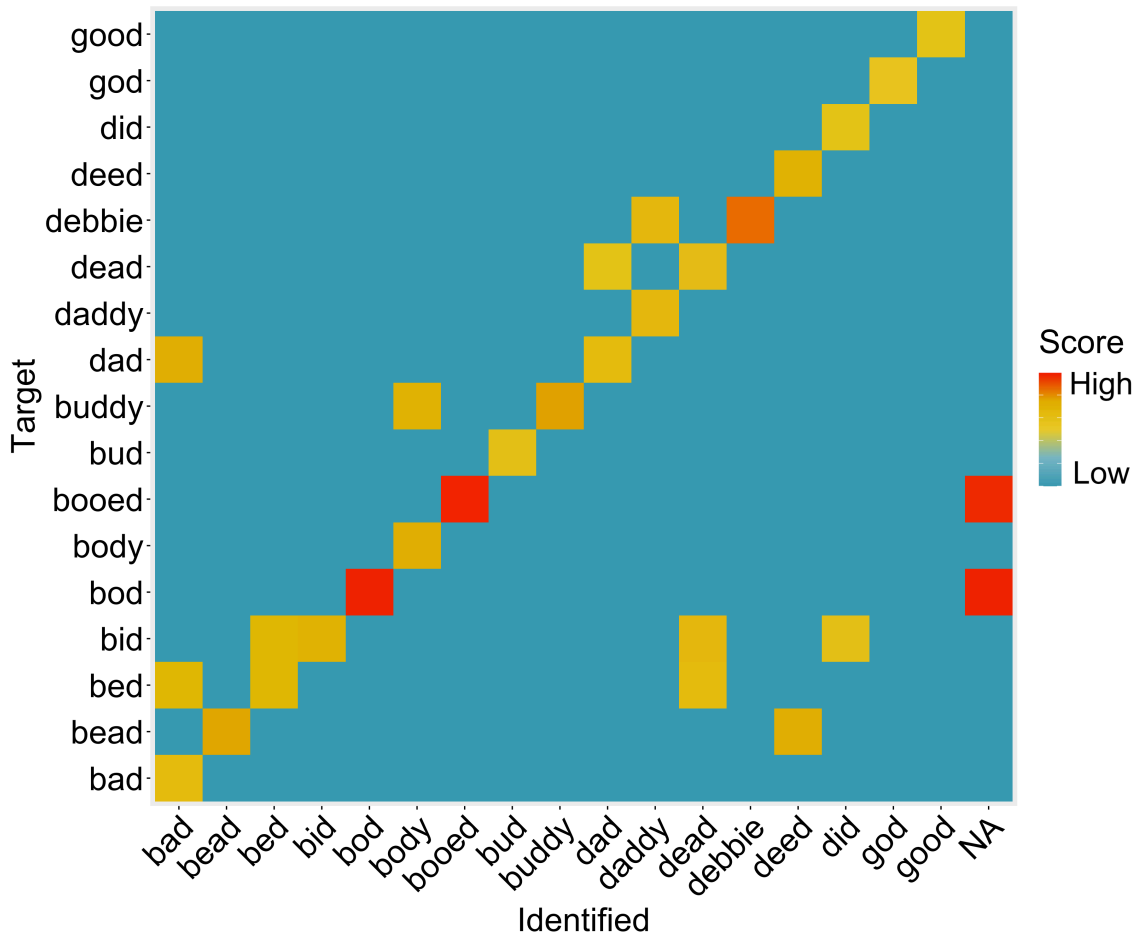


Figure 17 Confusion matrices of CVC words produced by a female native speaker, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

### 3.3.3 AUTOMATIC PHONEME RECOGNITION

An automatic phoneme recogniser developed in van Niekerc et al. (2022) and Xu et al. (2022) was used to simulate language-specific perception. The recognition system was trained using clean speech from LibriSpeech corpus (Panayotov et

al., 2015). To cover the phonemes in the target word list, speech segments were extracted with 11 onset consonants (/b/, /d/, /g/, /p/, /t/, /k/, /j/, /w/, /n/, /m/ and /l/) , 11 vowels and 6 diphthongs (stressed /aɪ/, /aʊ/, /eɪ/, /oʊ/, /ɔɪ/, /i/, /u/, /æ/, /ɑ/, /ɔ/, /ɛ/, /ɪ/, /ʊ/, /ʌ/ and unstressed /i/, /oʊ/, /ʌ/) and 6 coda consonants (/b/, /d/, /g/, /n/, /m/ and /ŋ/)<sup>5</sup>. 26-dimensional Log Mel spectrograms of the recordings were computed using the previously described parameters without preemphasis<sup>6</sup> and pre-padded to a length of 200 frames (spanning 1 s) to be used as the input for the training. A deep neural network was trained to learn a mapping from the Log Mel spectrograms to a 34-dimensional vector one-hot encoding the onset, the vowel and the coda phonemes, as illustrated in Figure 18. The network contains 8 convolutional layers (conv), 6 long short-term memory (LSTM) layers and 3 dense layers (Dense). Batch normalization layers after each conv, LSTM and Dense layers and dropout layers after each LSTM and Dense layers are not shown in the diagram. The architecture consists of 3 main parts: spectrotemporal feature processing, temporal feature processing and classification. The first convolutional layer module (in blue) was designed to learn the feature representations that may coexist or correlate in the spectral and temporal domains. The temporal feature processing module (in green) was designed to learn the temporal dependency and/or the state-based behaviour. Lastly, the classification module (in red) was used to learn the relationship between the features for classifying phonemes. Training proceeded with early stopping based on the validation set loss with a patience of 6 epochs.

---

<sup>5</sup> The speech data used to train the automatic phoneme recogniser included all the possible vowel categories in English. Although the learning targets of the vocal learning model only contained voiced bilabial, alveolar and velar stops, more consonant categories were included to ensure enough phonetic contrast. In a pilot experiment, the recogniser trained with only voiced bilabial, alveolar and velar stops performed badly in guiding the learning of the three stop consonants.

<sup>6</sup> Pre-emphasis was not applied to the speech sounds because the features would be passed to the deep learning model for dealing with the spectral information.

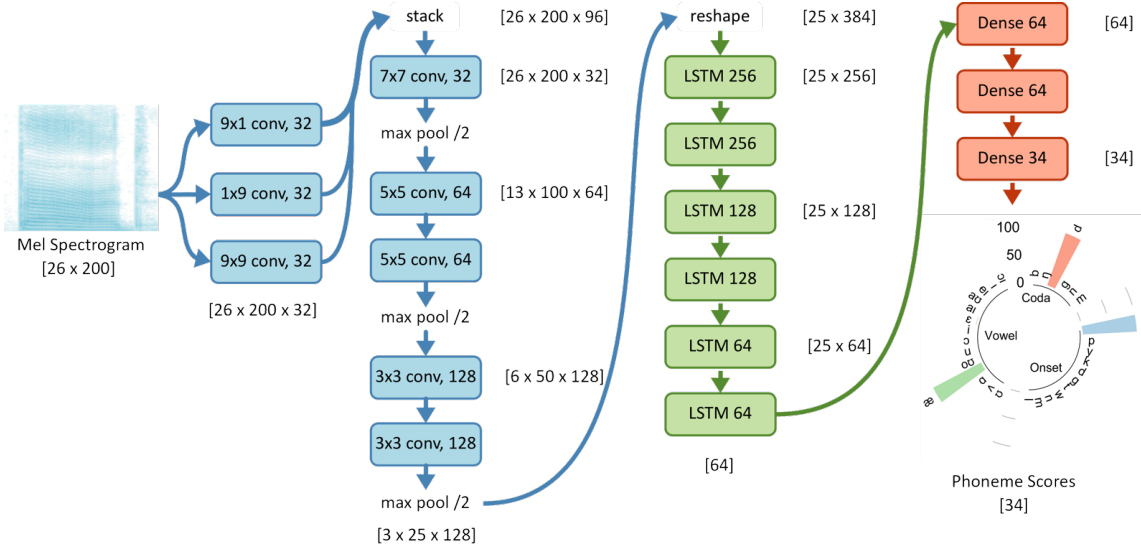


Figure 18 Schematic diagram of an automatic phoneme recogniser (Niekerk et al., 2022; Xu et al., 2022).

The details of the speech data used for the training are shown in Table 4. The speech data varied in syllable types, including 17 vowels, 187 CV syllables and 1122 CVC words. The trained recogniser had a 94% phoneme accuracy rate in the onset position, 88% in the vowel position and 98% in the coda position. A model trained using 22-dimensional MFCCs were also tested and the accuracy rate was 93%, 87% and 98% respectively. The Log Mel spectrograms had better overall performance (93%) than the 22-dimensional MFCCs (92%) and were thus adopted in the current simulations. The output vector of the automatic phoneme recogniser simulates a categorical perceptual space. The recognition loss of the CVC words is the Euclidean distance between the target vector and the recognised vector of the synthetic speech, by the following equation:

$$L = (p_i - q_i)^2, i = 1, \dots, N$$

where  $p_i$  represents the target phoneme vector and  $q_i$  represents the recognised phoneme vector. The  $p_i$  and  $q_i$  contain probability values between 0 and 1. The output vector is a 34-dimensional space for CVC syllables, i.e.,  $N = 34$  in the equation.

Table 4 Speech data extracted from LibriSpeech corpus (Panayotov et al., 2015) for training an automatic phoneme recogniser

	<b>Number of utterances</b>	<b>Size</b>	<b>Duration</b>
<b>Training</b>	2711615	21 G	116.7 h
<b>Validation</b>	337109	2.6 G	14.4 h
<b>Test</b>	345263	2.7 G	15 h

### 3.3.4 SOMATOSENSORY FEEDBACK

The somatosensory feedback was simulated by applying two kinds of constraints on the vocal tract parameters during vocal exploration. The vowel constraint is to ensure that the opening of the vocal tract is larger than a minimal cross-sectional area and the stop consonant constraint is to ensure a closure over a limited portion of the oral cavity. Two constraints were implemented by checking the tube area during the dynamic articulator movements. A tube area in the oral cavity larger than 0.25 cm<sup>2</sup> for the adult vocal tract model and 0.15 cm<sup>2</sup> for the child vocal tract models is considered as an open vocal tract, according to the tube area function in VocalTractLab (Birkholz, 2013). All the vowel targets that did not pass the check were filtered out. With regard to consonant target, the number of closed tube areas varied with the place of articulation of the target consonant. The total number of tube area sections is 40. A tube area less than 0.0001 cm<sup>2</sup> indicates a closed vocal tract. Up to 4 closed tube sections were allowed to ensure closed lips for bilabial stops. Due to the built-in interdependency between lip protrusion parameter and lip distance parameter in the articulatory synthesiser, the threshold of closed tube area was 0.15 cm<sup>2</sup> for bilabial stops preceding rounded vowels in /bood/. Moreover, the consonant constraint was implemented according to the uneven distribution of the sensory receptors on the tongue. Because the tongue tip is highly innervated compared to the tongue dorsum (Marlow et al., 1965), the closure tube length was set to be shorter in the anterior

tongue section and longer in the posterior tongue section. To be more specific, the number of closed tube area sections were set to be less than 3 for alveolar stops and less than 9 for velar stops, except for alveolar stops before high vowels. In English, alveolar stops preceding high front vowels are likely to be palatalized (Bateman, 2007), which suggests a larger area of contact during the consonant articulation. The number of closed tube area sections was therefore set to be less than 9 for /deed/ and 6 for /did/.

### 3.4 OPTIMISATION ALGORITHM

The articulatory targets were optimised for the speech sound by simulated annealing (Kirkpatrick et al., 1983), a stochastic optimisation algorithm that seeks an optimal solution through a coarse-to-fine criterion, suitable for non-linear, non-smooth and non-convex problem with many degrees of freedom, such as speech parameter estimation. The articulatory targets were iteratively adjusted and tested, and their acceptance is determined by a probability  $p$ .

$$p = \begin{cases} 1 & \text{if } \Delta E < 0 \\ e^{-\Delta E/T} & \text{otherwise} \end{cases}$$

where  $\Delta E$  is the change in the error of the objective function between the current and the previous attempt.  $T$  is the temperature that controls the annealing process. A uniformly distributed random number between 0 and 1 is generated as a criterion for deciding whether the current trial is accepted. If the error is lower than the current error, the current adjustment is accepted. The algorithm also keeps some changes that are not ideal. If the probability of acceptance  $p > r$ , the new attempt is still accepted. This allows a balance between exploration and exploitation of optimal parameters. The control temperature  $T$  gradually decreases throughout the process, which means that a new motor pattern in the earlier stages is likely to be accepted but only good trials with low errors are accepted in the later stages.

Due to the heuristic nature of the algorithm, there was a possibility that the final articulator targets were not optimal. Thus, the algorithm was implemented in two

stages to stabilise the learning outcome and to improve the chance of finding global optima rather than local optima, as shown in Figure 19. In the first stage, 10 processes were initiated in parallel, each with 2k iterations. Each process started with a neutral position (schwa) followed by random adjustments of the vocal tract parameters and gradual convergence to a solution. Next, all the trials were ordered by the loss and selected the best candidate of each of the 10 processes for a more localised optimisation. In the second stage, instead of a broad motor exploration, the 10 processes randomly walked around the selected set of articulatory targets for 200 iterations<sup>7</sup>. More specifically, the model generated a neighbour solution based on the previous trial as follows:

$$x'_i = x_i + RW_i, i = 1, \dots, N$$

in which  $x_i$  is the 20-dimensional articulatory target, including 17 vocal tract parameters and 1 time constant ( $N = 20$ ).  $W_i$  is added to adjust the relative step of the random walk, based on the range of the vocal tract parameters and the time constant.  $R$  is a uniformly sampled random number between -1 and 1.  $x'_i$  is further constrained by the range of the parameters, as shown in Figure 2.

---

<sup>7</sup> The refinement improved the intelligibility of difficult sound sequences such as /booed/.



Figure 19 Illustration of the two-step optimisation process. Step 1 Exploration: Uniform random parameter search; Step 2 Refinement: Random parameters search around good solutions.

### 3.5 MODEL EVALUATION

So far, previous studies have rarely conducted systematic listening experiments to evaluate the model outcome (see Appendix Table A Performance). The study initiates a new benchmark for vocal learning simulations, that is, making direct comparison with natural speech to see whether the model can achieve speech acquisition with high intelligibility. Quantitative measurements of intelligibility in listening experiments allow theoretical accounts of vocal learning to be linked to predictions.

#### 3.5.1 LISTENING EXPERIMENTS

Four listening experiments were conducted to evaluate the acoustic-feature-trained and phoneme-recogniser-trained models in a set of open-vocabulary experiments and a set of close-set transcription experiments. American English native speakers were recruited and screened on Prolific (prolific.co) and then directed to Gorilla (gorilla.sc) for the online experiments. Before the experiment,



the participants filled in a brief questionnaire for demographic and language background information (see Appendix Figure A). The listeners were all born and raised in the US, speaking American English as their first language. To verify their accents, participants were asked to read the first two sentences of the story “The North Wind and the Sun”, a well-established text recommended by the IPA for eliciting English phonetic contrast. Participants were asked to undertake the tasks on a computer in a quiet environment without noise or other distractions. A headphone screening was conducted to ensure that the participants were wearing headphones. The listeners were asked to choose the quietest sounds out of three pure tones with one of the tones presented 180° out of phase across the stereo channels. The listeners who were wearing headphones were more likely to discriminate the sounds because a loudspeaker would have resulted in phase cancellation (Woods et al., 2017). The participants who passed the screening were given five practice trials to get familiarised with the experiment. They were then randomly presented with the words produced by the female speaker and the synthetic sounds learned by the adult male, the 1-year-old, and the 3-year-old child vocal tract models. 3 unique tokens of the 17 target words (Table 1) were included in each condition. For the open-vocabulary transcription experiment, the participants were instructed to listen to the audio carefully and freely write down the word they had heard. For the close-set transcription task, the participants were asked to choose from the 17 target words the one they had heard. To make sure that each listening experiment can be finished within 30 minutes, the stimuli were divided into 4 listening experiments to assess the effect of auditory feedback (acoustic features vs. the automatic phoneme recogniser) and the effect of listening experiment type (open vs. close transcription). Participants were recruited separately for each listening experiment. The audio samples for the listening experiment can be found at [https://gitlab.com/Anqi\\_Xu/evoc\\_learn/-/tree/main/Stimuli](https://gitlab.com/Anqi_Xu/evoc_learn/-/tree/main/Stimuli).

### 3.5.2 PARTICIPANTS

173 monolingual American English native speakers between 18 to 50 years old participated in the experiment. The participants were born and raised in the US, without any self-reported speech or hearing disorders. Among them, 47 did not pass the headphone screening; 5 were excluded from the experiment because of apparently atypical American accents; and 1 was excluded because of noise in the submitted recordings that suggested a noisy listening environment. In the end, 30 participants were included in each intelligibility experiment (120 listeners in total). The procedure has been approved by the Department of Speech, Hearing and Phonetic Sciences, University College London and the experiments complied with all relevant ethical regulations. Informed consent from all the participants was obtained online via Gorilla.

### 3.5.3 ANALYSIS

The responses collected online were annotated with phone labels using the CMU pronunciation dictionary (Carnegie Mellon University, 2022) by *pronouncing* package (Parrish, 2022). Phone labels were then manually added for those responses without automatic annotation. In the case of phoneme insertion and deletion, recognised phonemes were aligned maximally as shown in Table 5. Responses recorded before the audio samples finished were excluded in the analysis. The recognition rate was calculated in terms of how many segments were correctly identified.

Non-parametric statistical tests were chosen for statistical analysis, in line with previous studies that involved evaluation of synthetic speech (Anumanchipalli et al., 2019). Moreover, outliers were found in the response data of the listening experiments and the distribution of the data was skewed (Hollander et al., 2014). In addition, the medians of the listening experiment are better estimates of the identification by native listeners rather than the means. Besides, the sample size is too small to determine a normal distribution for the experiment of comparing recognition errors of models with and without somatosensory feedback (Section 3.3.4). Kruskal-Wallis test, Wilcoxon signed rank test and Spearman correlation

were conducted to evaluate the reaction time and the phoneme accuracies. Multiple comparison correction was applied by Bonferroni correction.

Table 5 Transcript examples of phoneme insertion and deletion. Phonemes are labelled using CMU pronunciation dictionary (Carnegie Mellon University, 2022).

Target	B	AA	D
IPA	b	æ	d
Insertion	B	L AA	D
	Correct	Incorrect	Correct
Deletion		AA	DD
	Incorrect	Correct	Correct

## Chapter 4 SENSORY AND MOTOR SYSTEMS IN VOCAL LEARNING

### 4.1 SENSORY FEEDBACK

In this section, I will present the simulation results of the vocal learning model trained with different sensory options (Section 3.2 Sensory system). With regard to the auditory feedback, universal perception by MFCCs and Log Mel spectrograms were simulated, which are both parametric representations of speech acoustics. The effectiveness of the two types of acoustic features in guiding vocal learning was evaluated by the automatic word recogniser. Moreover, language-specific perception was simulated using an automatic phoneme recogniser which encodes sound contrast regardless of cross-speaker differences. Speech trained by acoustic features and by the recogniser were compared, using the automatic word recogniser and human perceptual

identification. Finally, the performance of models with and without somatosensory feedback were compared to examine its role in facilitating vocal learning.

#### 4.1.1 ACOUSTIC FEATURES: LOG MEL SPECTROGRAMS VS. MFCCs

The adult vocal tract model was trained to learn the 13 English CVC words with Log Mel spectrograms and MFCCs. The learned synthetic samples were then evaluated by the word recogniser. The two types of acoustic features resulted in very similar word error rates, as shown in Figure 20 and Figure 21. Only a few words were correctly identified in both cases.

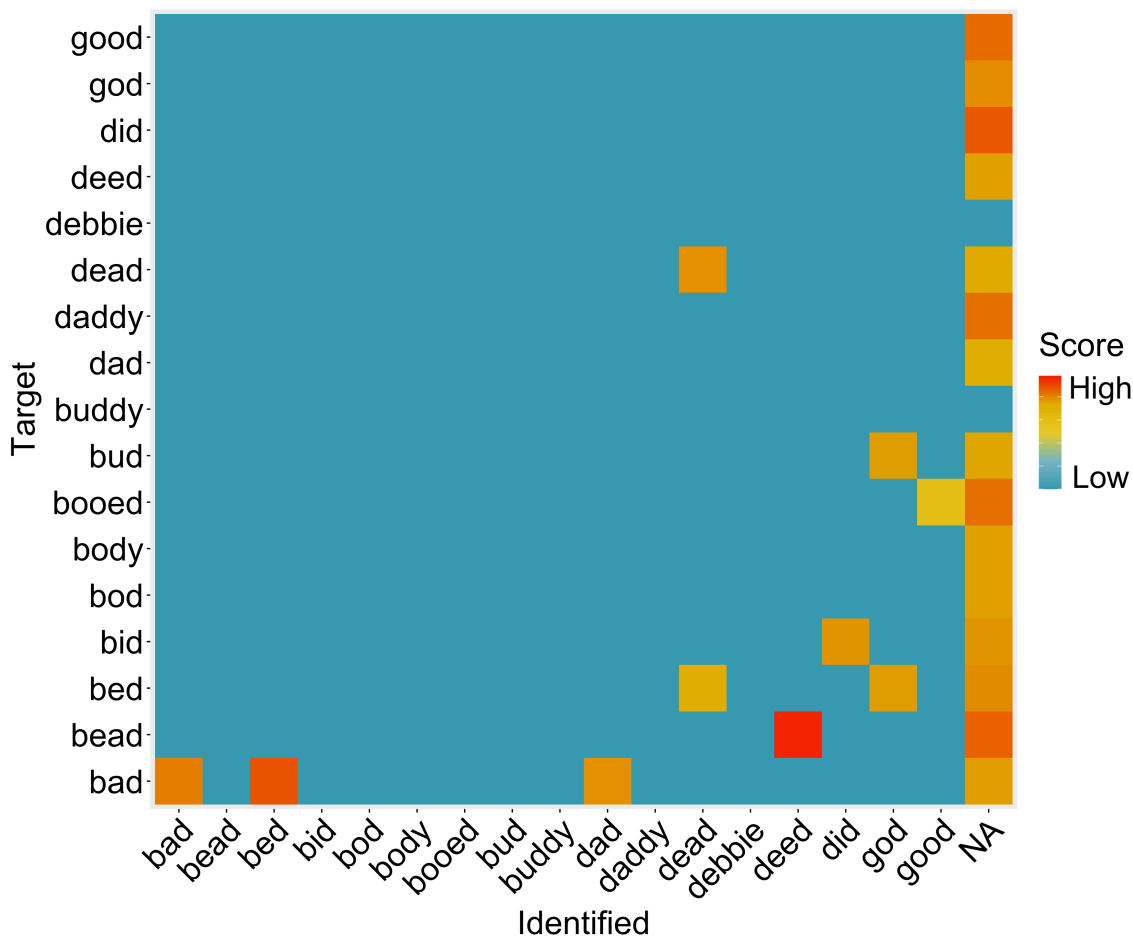


Figure 20 Confusion matrices of CVC words learned by adult vocal tract model when guided by Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

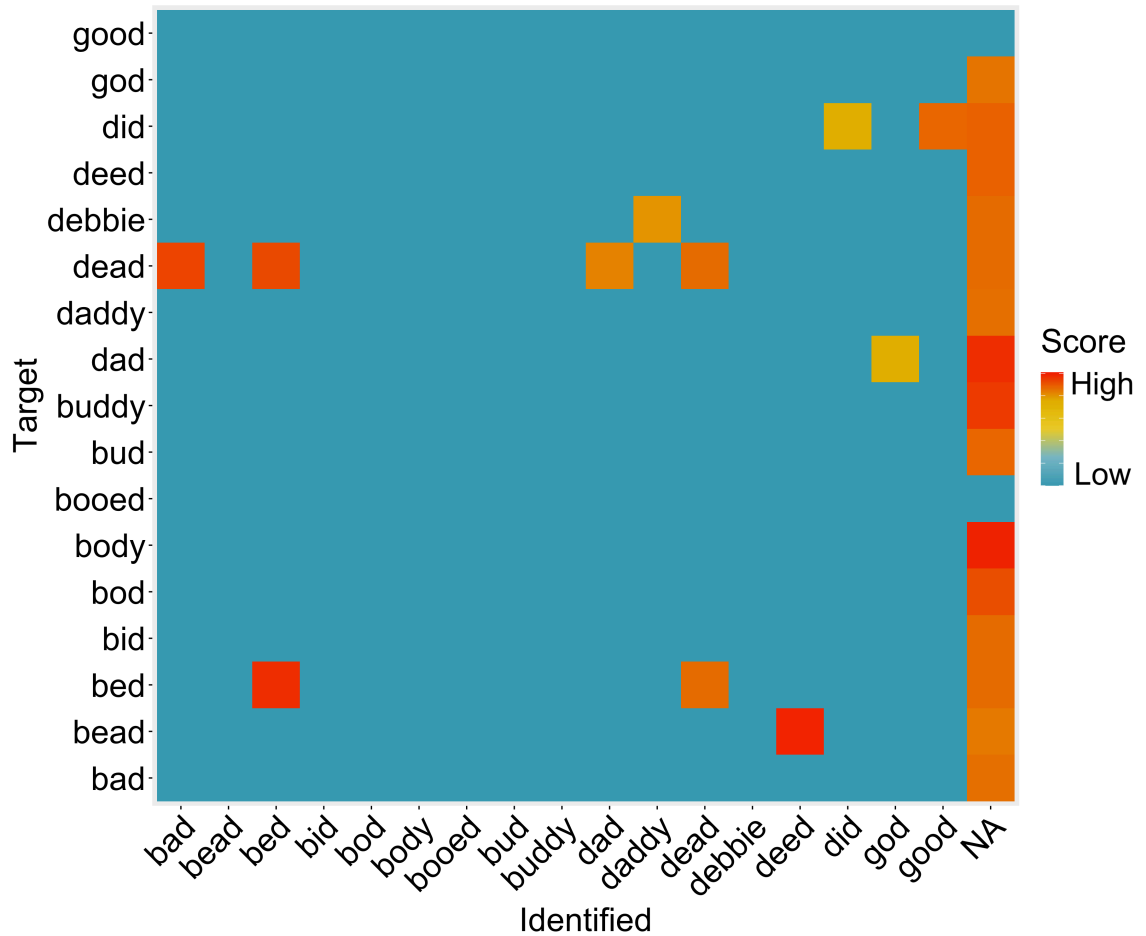


Figure 21 Confusion matrices of words learned by adult vocal tract model when guided by MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

Figure 22 shows the quality of the onset consonant trained by MFCCs and Log Mel spectrograms, evaluated by the word recogniser. The bilabial stops trained by Log Mel spectrograms had fairly high accuracies, while the alveolar stops trained by MFCCs had higher accuracies. However, the velar stops trained by Log Mel spectrograms were identified as /n/ or /w/. Overall, the speech trained by Log Mel spectrograms were slightly better identified than that trained by MFCCs.

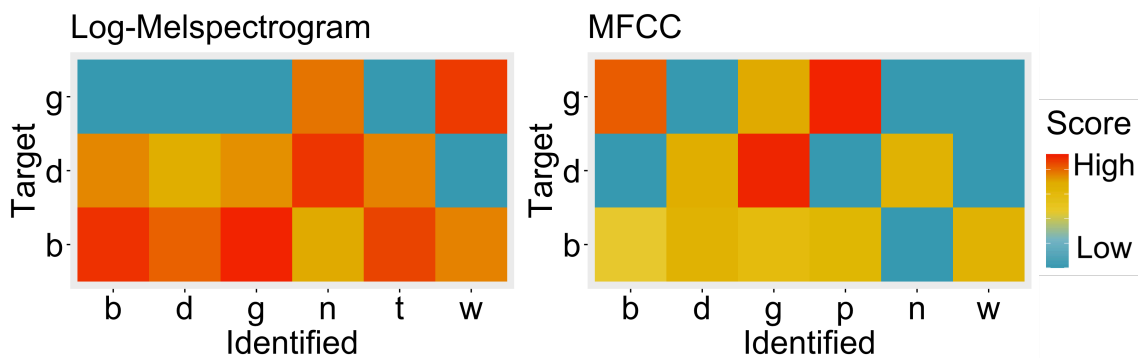


Figure 22 Confusion matrices of consonants trained by Log Mel spectrograms and MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

However, the performance of vowel quality showed entirely reversed patterns. As shown in Figure 23 and Figure 24, the recognition accuracies were higher for vowels trained by MFCCs than the ones trained by Log Mel spectrograms. Both acoustic features failed to guide the learning of intelligible vowels in ‘booed’ and ‘good’. For the rest of the vowel categories, those trained by MFCCs were identified more correctly than the ones trained by Log Mel spectrograms. Especially, the vowel in ‘bud’ was less successful when trained by Log Mel spectrograms. Overall, the two types of acoustic features had comparable performance in guiding vocal learning. The consonant quality was better when trained by Log Mel spectrograms, while MFCCs were more advantageous in training the vowels. The following analysis will be based on MFCCs, which are the most frequently used parametric representations in speech synthesis and recognition (Barry & van Dommelen, 2005; Davis & Mermelstein, 1980).

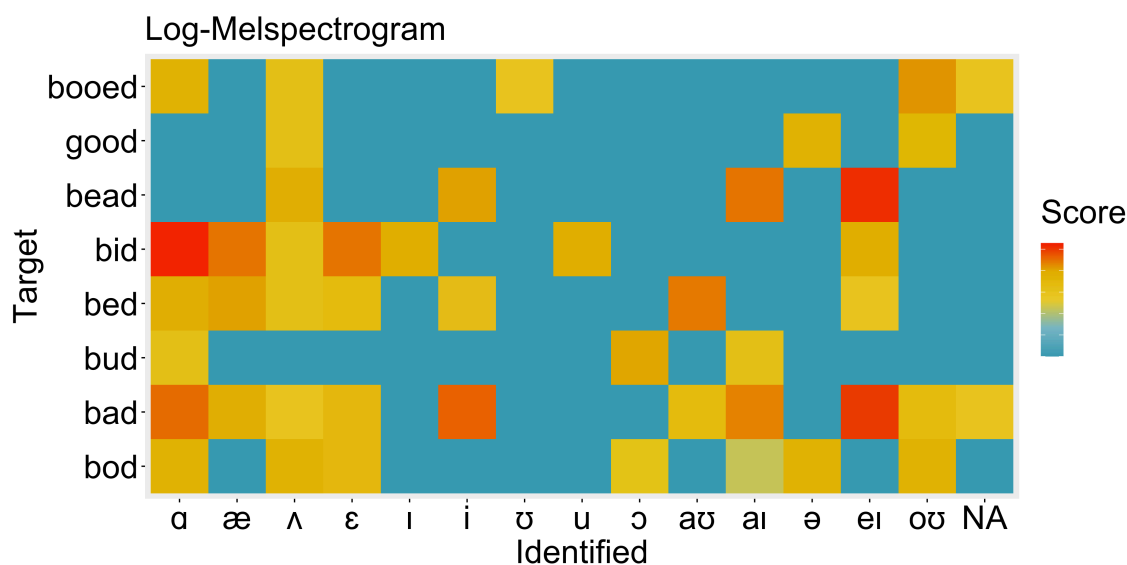


Figure 23 Confusion matrices of vowels trained by Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

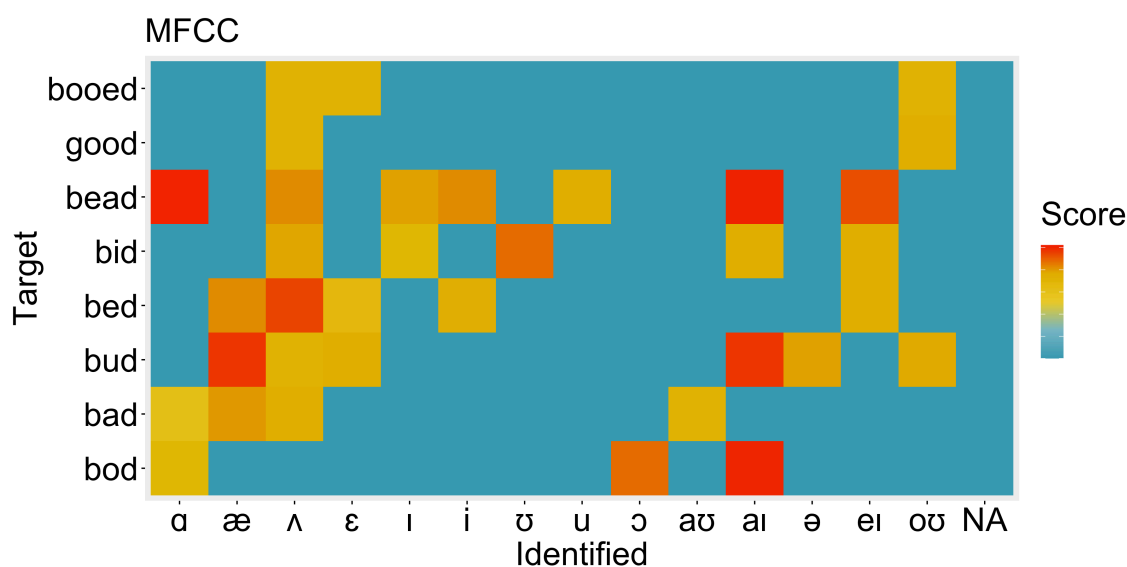


Figure 24 Confusion matrices of vowels trained by MFCCs, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.

#### 4.1.2 MFCCs VS. AUTOMATIC PHONEME RECOGNISER

With respect to auditory feedback, language-specific perception simulated by the automatic phoneme recogniser was more successful than language-universal perception simulated by MFCCs. The intelligibility of the speech trained by the recogniser and MFCCs by the automatic word recogniser were evaluated by the word recogniser. The results of the speech trained by the recogniser are shown in Figure 25. If we compare it with Figure 21, it is evident that the word recogniser performed better in recognising the CVC words trained by the recogniser than those trained by MFCCs, as more target words were correctly identified.

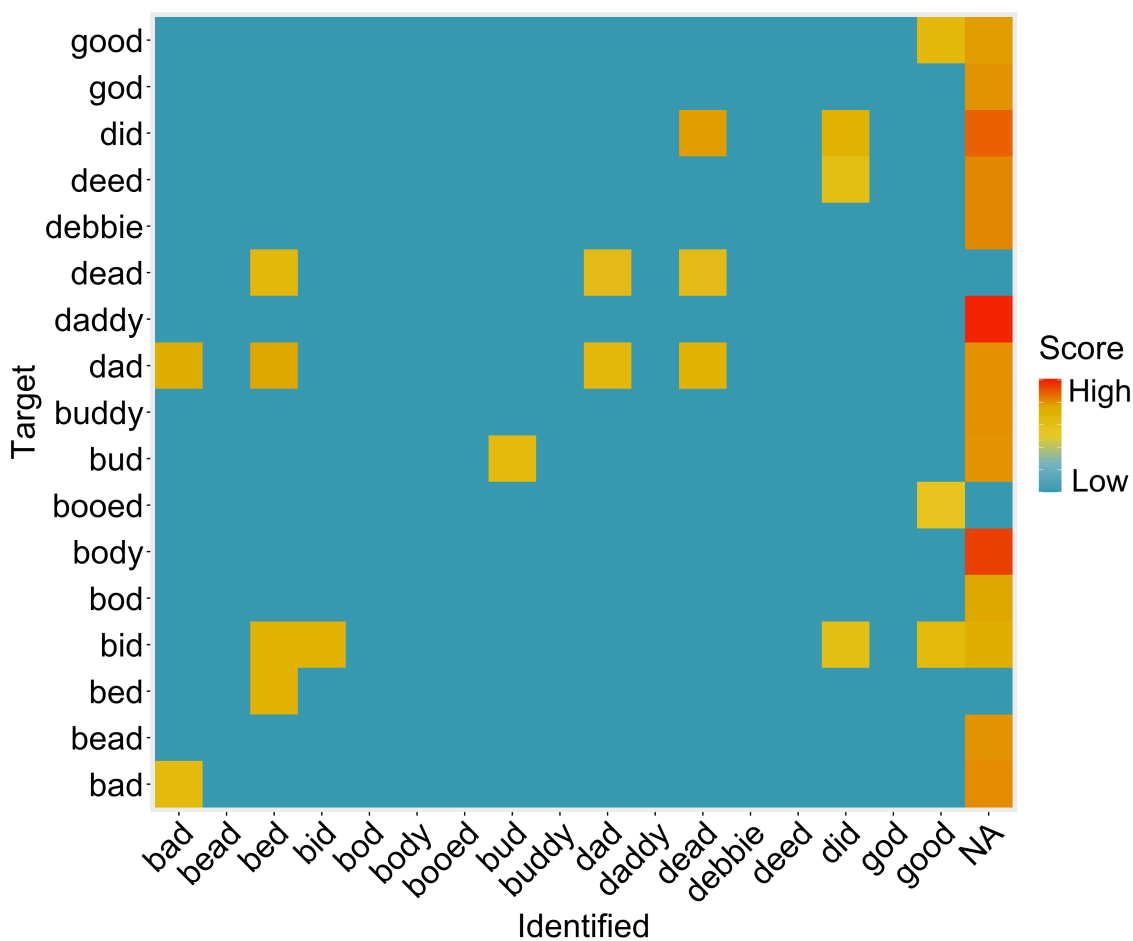


Figure 25 Confusion matrices of words learned by adult vocal tract model when guided by an automatic phoneme recogniser, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss.



To confirm the results, human perception experiments were run to examine the intelligibility of the learned speech. As shown in Figure 26, the synthetic words trained by the automatic phoneme recogniser were more intelligible than the ones trained by MFCCs in both the open-vocabulary transcription experiment and close-set transcription experiment (Kruskal-Wallis test:  $p < .001$ ). Wilcoxon Signed Rank test showed that the tendency was the same regardless of vocal tract models in both the open-vocabulary (1y:  $p < .001$ , 3y:  $p < .001$ , Adult:  $p < .001$ ) and close-set experiments (1y:  $p < .001$ , 3y:  $p < .001$ , Adult:  $p < .001$ ), as indicated by post-hoc comparisons.

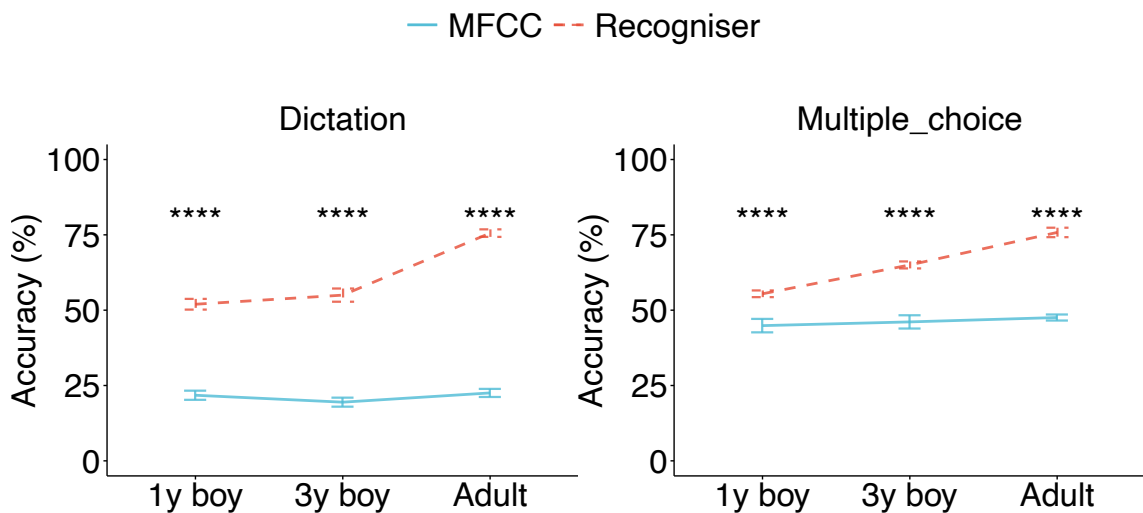


Figure 26 By-listener mean phoneme accuracy rates of CVC words learned by different vocal tract models in an open-vocabulary transcription experiment and a close-set transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

To test whether the type of auditory feedback influences the identification rate in different phoneme positions, the open-vocabulary transcription accuracy rate of the learned CVC words was compared. Figure 27 shows by-listener phoneme accuracy of vocal tract models of different ages. The onset consonant, the vowel and the coda consonant trained by the automatic phoneme recogniser were all more intelligible than the ones trained by MFCCs. The benefit is more evident in consonants than in vowels. As suggested by Wilcoxon Signed Rank tests, the

recogniser-trained words had higher accuracies in the onset ( $p < .001$ ), the vowel ( $p < .001$ ) and the coda ( $p < .001$ ) positions.

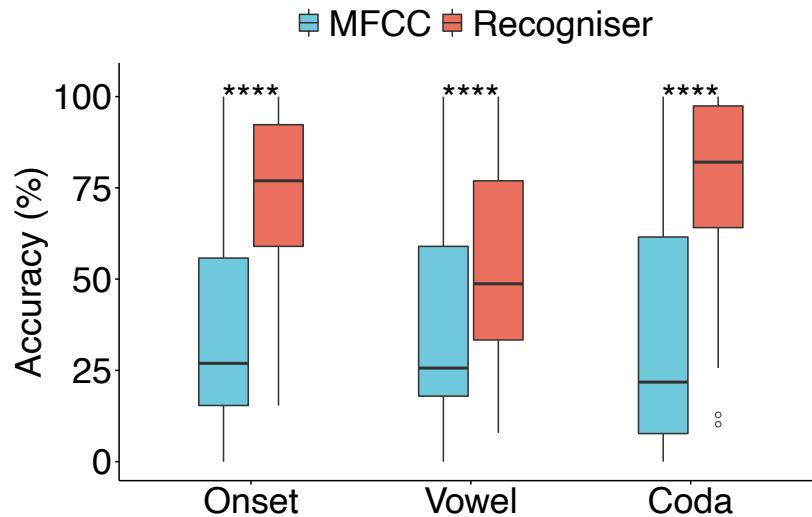


Figure 27 Distribution of by-listener phoneme accuracy rates for synthetic CVC words learned by vocal tract models of different ages in three syllable positions, evaluated by an open-vocabulary transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

In order to test the generalisability of the learned articulatory movements, CVCV words were regenerated based on the learned vocal tract parameters of the CVC words. The listening experiments show that the vocal tract parameters trained by the recogniser had better generalisability than the synthetic speech trained by the MFCCs. Figure 28 shows the identification rate of all the phoneme positions in the CVCV words trained by the two types of auditory feedback. The recogniser-trained CVCV words had higher accuracies than the MFCC-trained ones in all the phoneme positions (Wilcoxon Signed Rank test:  $p < .001$ ).

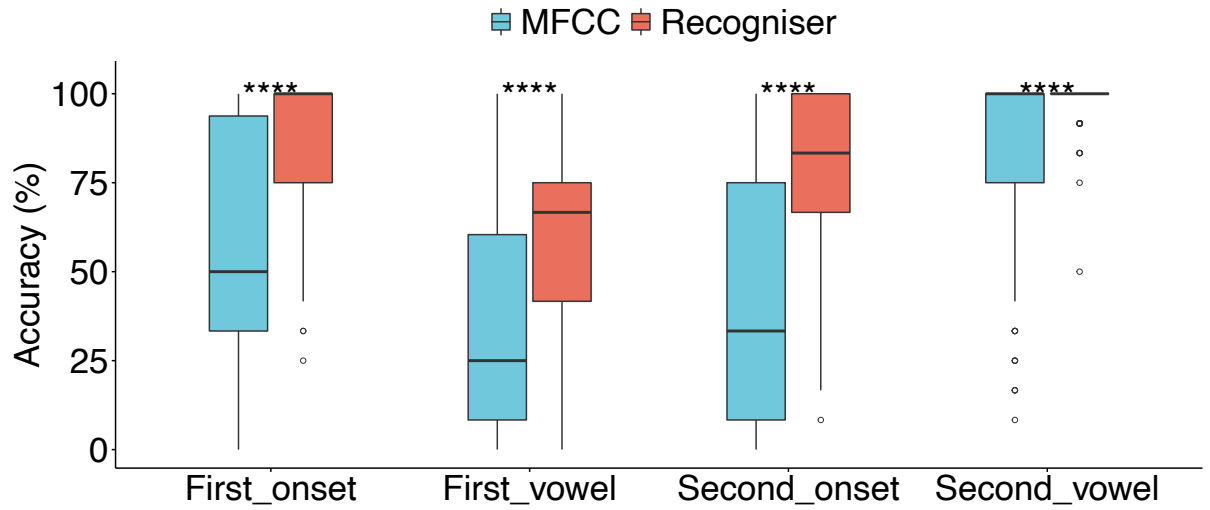


Figure 28 Distribution of by-listener phoneme accuracy rates for synthetic CVCV words learned by vocal tract models of different ages in four syllable positions, evaluated by an open-vocabulary transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

In addition, the reaction time of the listeners when identifying the CVC and CVCV words ( $n = 153$ ) learned by vocal tract models of different ages were analysed. Histograms and density plots of the reaction time in the open-vocabulary and close-set transcription experiments are shown in Figure 29. There were many more listeners showing hesitation while listening to the synthetic speech trained by MFCCs in both types of experiment. Wilcoxon Signed Rank test further confirmed that reaction time was significantly longer for MFCC-trained speech regardless of the task type (open-vocabulary:  $p < .001$ , close-set:  $p < .001$ ).

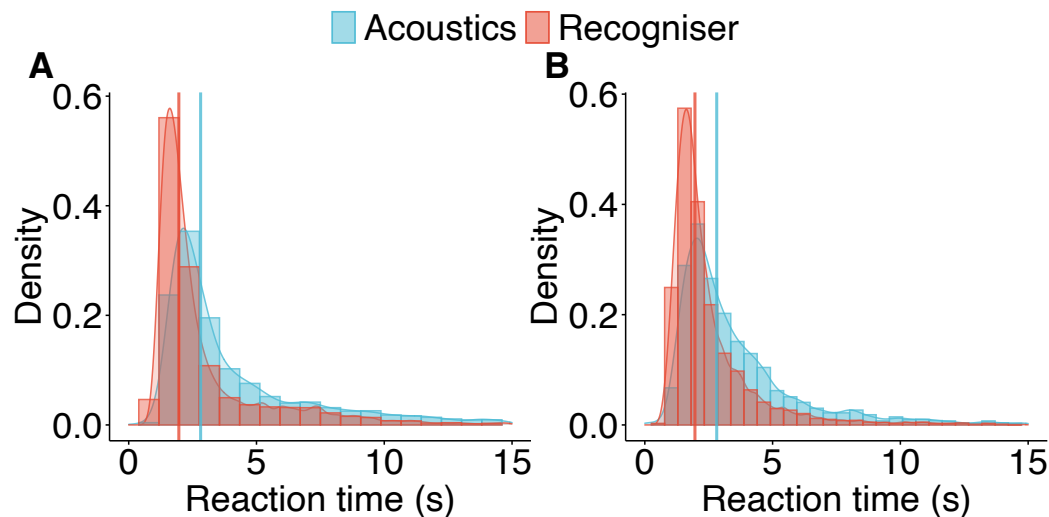


Figure 29 Reaction time of American English listeners in an open-vocabulary transcription (A) and in a close-set transcription experiment (B). The vertical lines represent the median reaction time for two types of auditory feedback.

To test which type of auditory feedback better reflects human speech perception, Spearman's rank correlation was conducted to assess the relationship between open-vocabulary transcription accuracies and type of auditory feedback. The scatter plots with correlation lines are shown in Figure 30. The correlation between identification accuracy and MFCC error was non-significant. In contrast, the recognition error returned by the recogniser significantly correlated with the phoneme identification accuracies in the open-vocabulary transcription experiment. The synthetic words with lower recognition error were more likely to be judged as having the correct phonemes. The results suggest that the recogniser emulates speech perception better than MFCCs.

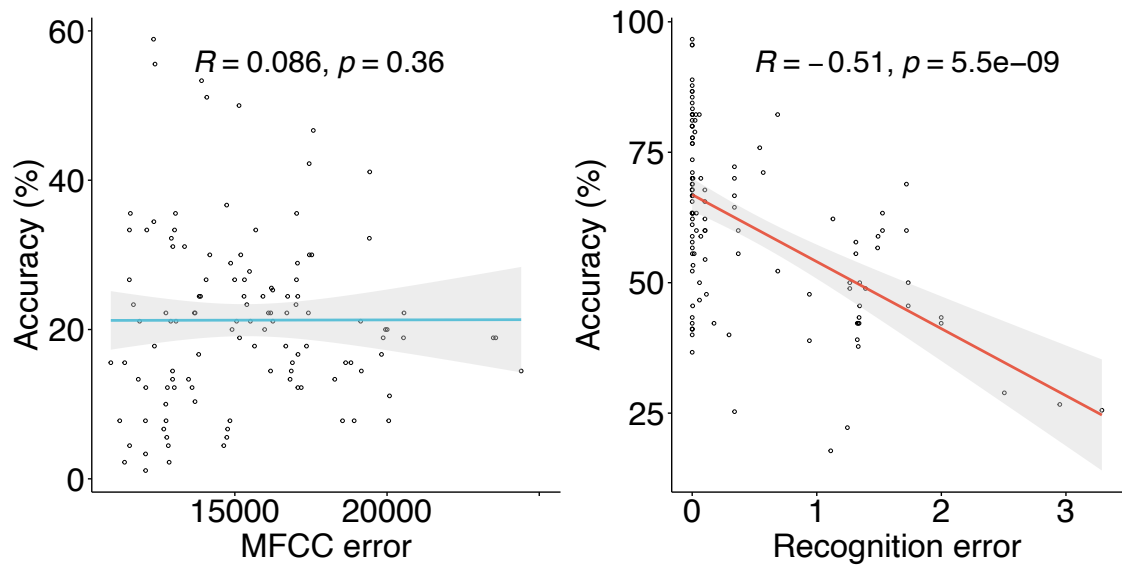


Figure 30 Relationship between phoneme identification rates and auditory feedback

In order to compare the automatic phoneme recogniser and MFCCs with human perception in detail, the phoneme accuracies judged by native listeners were normalised to values between 0 and 1. The phoneme errors judged by the recogniser and MFCC errors were also normalised to have the same range but in a reversed order. The normalised perceptual scores of target words evaluated by native listeners, the recogniser, MFCCs are shown in Figure 31. The distribution of normalised score of the recogniser and the listeners was almost symmetrical, whereas the distribution was asymmetrical for MFCCs and the listeners. The listener and the recogniser were relatively consistent with each other in identifying the learned synthetic speech. The disagreement was only found in the case of 'bid' and 'bud', whereby the native listeners might have been biased by the word frequency. It has been suggested that high-frequent words are usually preferred to phonetically similar low-frequent words in listening experiments (Savin, 1963). 'bud' had 57972 occurrences, while 'bid' had 70427114 occurrences in the Google Web Trillion Word Corpus (Tatman, 2017)<sup>8</sup>.

<sup>8</sup> 'bud' is one of the least frequent words, while 'bid' is one of the most frequent words among all the target CVC words starting with a bilabial stop.

As a consequence, 'bud' had fairly high recognition scores but the listeners failed to identify them in the open-vocabulary transcription experiment, probably due to the low word frequency. In contrast, the synthetic samples of 'bid' were not well identified by the recogniser, but the listeners were able to identify them. However, unlike the automatic phoneme recogniser, the synthetic speech with low MFCC errors was not correctly identified by the listeners.

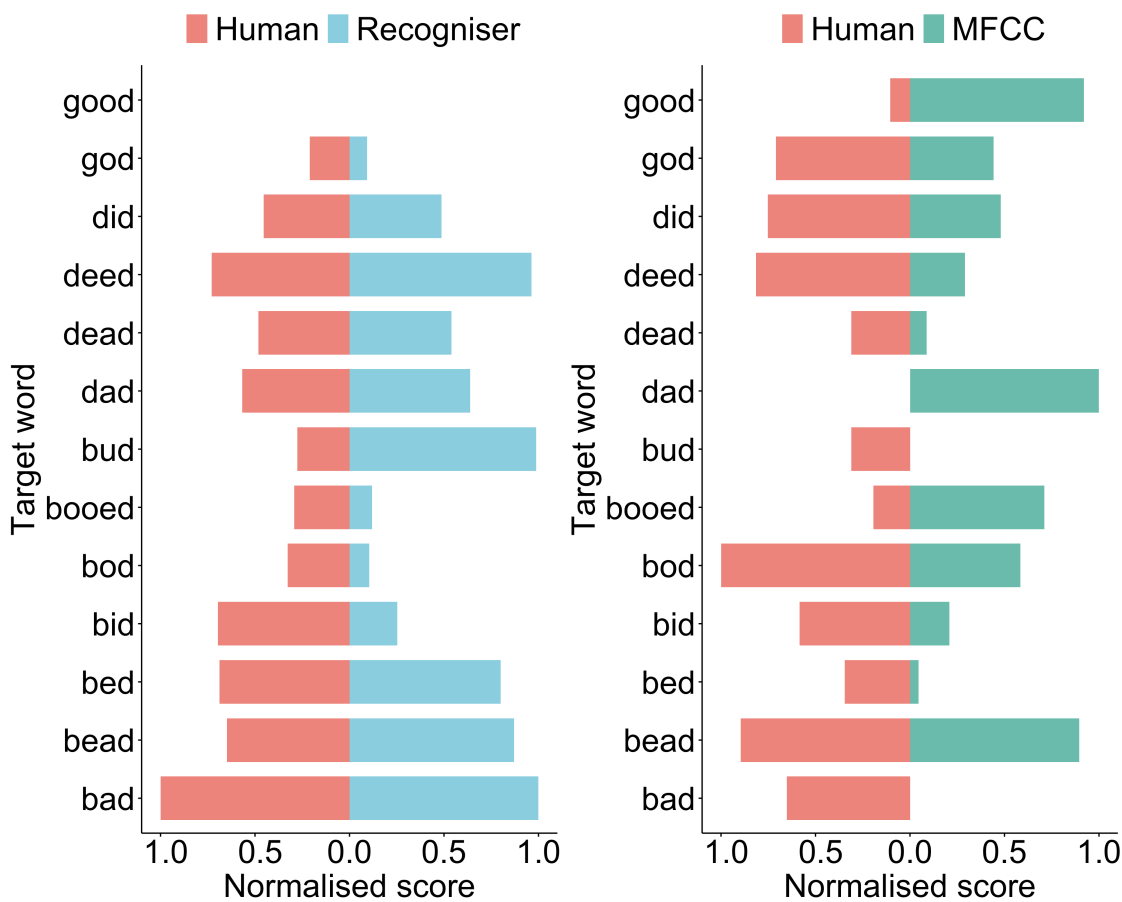


Figure 31 Comparison of human identification with an automatic phoneme recogniser and MFCCs by target words. Perceptual scores are normalised based on the phoneme accuracies judged by human participants, the recogniser and MFCCs respectively.

#### 4.1.3 SOMATOSENSORY FEEDBACK

Somatosensory feedback was simulated by a constraint on the degree of oral opening for each generated vocal tract configuration during vocal exploration. The constraint ensured an open vocal tract for vowels and a narrow vocal tract for consonants (see 3.2.4 Somatosensory feedback). I compared the recognition error of the best 10 instances<sup>9</sup> per target CVC words ( $n = 13$ ) trained with and without somatosensory feedback. As shown in Figure 32, with the same number of iterations, the model with somatosensory feedback learned more intelligible words than the baseline condition (Wilcoxon Signed Rank,  $p = .001$ ). This suggests that somatosensory feedback has effectively restricted the search space for the articulatory targets to facilitate learning.

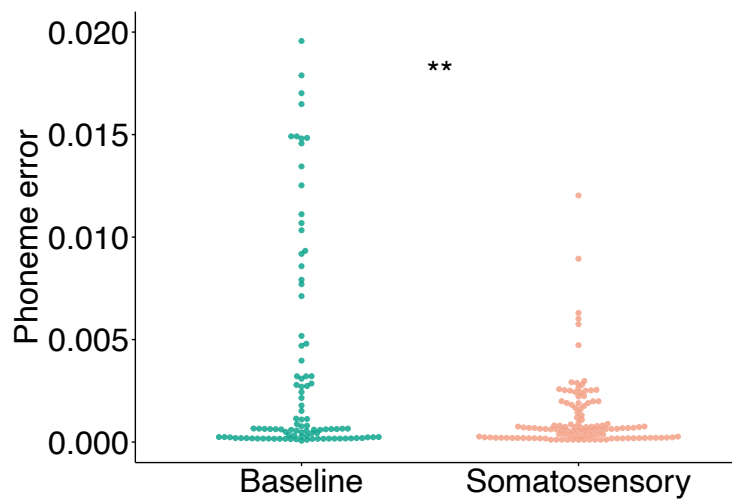


Figure 32 Effect of somatosensory feedback on the phoneme errors of synthetic CVC words learned by an adult vocal tract model, evaluated by an automatic phoneme recogniser. \*\*  $P \leq 10^{-2}$ .

#### 4.1.4 DISCUSSION

---

<sup>9</sup> 10 instances were chosen to ensure sample size for statistical analysis and to ensure intelligibility of the learned samples judged by the author.

Although various learning strategies have been explored to simulate vocal learning, not enough attention has been paid to proper modelling of the sensory motor system (see Appendix Table A Sensory system). In contrast with previous attempts, I systematically examined different types of sensory feedback by training the model to learn English words and compared the performance of learned speech in listening experiments. I simulated language-specific perception by an automatic phoneme recogniser and universal perception by acoustic features. I showed that language-specific perception is better than universal perception in guiding vocal learning. The recogniser simulates auditory experience that forms a distributional space shaped by all the speech sounds the learner has heard, so that the sound categories in it encompass varied forms by many speakers. This distributional space was powerful enough to even guide the child models although the recogniser was trained without child speech.

Past models have adopted various acoustic features to simulate the sensory system. A vast majority of the studies have used formants (Acevedo-Valle et al., 2020; Bailly, 1997; Forestier & Oudeyer, 2017; Heintz et al., 2009; Howard & Huckvale, 2005; Ishihara et al., 2009; Miura et al., 2007; Rasilo & Räsänen, 2017; Westermann & Miranda, 2004) or Bark-scaled formants (Barnaud et al., 2019; de Boer, 2000; Kröger et al., 2009; Moulin-Frier et al., 2014, 2015; Moulin-Frier & Oudeyer, 2012; Oudeyer, 2005) and Mel-scaled formants (Warlaumont, 2012; Warlaumont et al., 2013; Warlaumont & Finnegan, 2016). Other researchers have adopted Bark-scaled spectrograms (Kröger et al., 2014) or gammatone spectrograms (Howard & Messum, 2014; Messum & Howard, 2015), in order to keep more acoustic details of the speech sounds. More recently, quite a few studies have attempted to use MFCCs as the auditory feedback (Kanda et al., 2009; Najnin & Banerjee, 2017; Philippsen et al., 2014; Prom-On et al., 2014a, 2014b; Rasilo et al., 2013), which is the most popular parametric acoustic representation in speech synthesis and recognition (Barry & van Dommelen, 2005). MFCCs and Log Mel spectrograms were compared by training the model with these two acoustic features while keeping the rest of the settings the same. The learned English words had comparable identification rates while being



judged by the word recogniser. Similar to the previous attempts, the overall intelligibility of the synthetic words learned with acoustic guidance was relatively low.

One likely reason is that the acoustic features do not necessarily instantiate speech perception. If speech perception is, as has been observed, context-dependent (Liberman et al., 1954) and speaker-normalised (K. Johnson & Sjerps, 2021), a realistic model of speech perception should be equipped with knowledge of dynamic spectral changes that carry linguistic function as well as cross-speaker variations. There has been some effort to train recurrent neural networks to capture the temporal characteristics of speech acoustics, such as Echo State Network (ESN) (Murakami et al., 2015). The usage of recurrent neural networks, in effect, encompasses contextual information of time-series speech signals. What the system lacks is the representations of phoneme categories. Interestingly, Lyon et al. (2012) has built an on-line conversation robot with a phoneme recogniser (i.e., adapted version of Microsoft SAPI 4.5). In the same spirit, I simulated speech acquisition guided by an automatic phoneme recogniser. The combination of spectrotemporal feature processing, temporal feature processing and classification together in the recogniser used in this study may have led to the successful learning of English words. The recogniser is sensitive to the statistical distribution of acoustic features across context and speaker variances. It may therefore be analogous to the categorical representation (Chang et al., 2010) and the speaker-normalised representation (Sjerps et al., 2019) in the human auditory cortex.

A key innovation of the current study is that I developed a new way of testing different kinds of auditory feedback, on the basis of the intelligibility of the learned speech in listening experiments. For vocal tract models of different ages (Fig. 17), the synthetic speech was much more intelligible in all syllable positions when guided by the automatic phoneme recogniser than by MFCCs (Fig. 18 & Fig. 19), and listeners needed less time to identify words in both the open-vocabulary and close-set transcription experiments (Fig. 20). The benefit of the recogniser was most clearly seen in the correlation between the identification accuracies of the

human listeners and the evaluation of the recogniser. Statistical analysis showed no correlation between human identification accuracies and acoustic errors (Fig. 21A), whereas there was a significant negative relation between human identification accuracies and recognition errors (Fig. 21B). Furthermore, I found that native listeners and the recogniser had comparable judgement of synthetic words, while MFCCs did not realistically reflect human perception (Fig. 22).

Interpreted at the cognitive level, acoustic features are similar to the language-universal perception at the early developmental stage, whilst the speech recogniser is comparable to language-specific perception at the later stage (Kuhl, 2000; Werker & Lalonde, 1988). Newborns are capable of discriminating speech sounds in world's language universally (Eimas et al., 1971; Streeter, 1976). They then develop perceptual biases toward sound contrasts in native languages between 6-12 months after birth (Kuhl et al., 2006; Werker & Tees, 1984). The maturation of the perceptual system is the prerequisite for speech production acquisition (Kuhl, 2000). Acoustic features do not provide information about sound categories that distinguish words in a language. Rather, it merely reflects basic auditory processing without extracting information about linguistic contrasts (Chládková & Paillereau, 2020). A model of language-specific perception would, instead, be able to encode phonetic contrast, which is exactly what the automatic phoneme recogniser simulates. The results demonstrate that the ability of perceiving phonemes of a given language plays a pivotal role in guiding successful vocal learning. I have also shown that somatosensory feedback provides additional benefit for speech acquisition (Fig. 23).

## 4.2 COARTICULATORY CONTROL

Not only does the vocal learning model contain a sophisticated sensory system, but also the speech motor system is modelled explicitly. The synchronised dimension-specific sequential target approximation model was implemented to control the coarticulatory dynamics (Section 3.3.2). Here, I will report the learned

vocal tract parameters that yielded the highest intelligibility, which was trained by the automatic phoneme recogniser with somatosensory feedback.

#### 4.2.1 LEARNED ARTICULATORY KINEMATICS

The high intelligibility of the synthetic syllables is partially due to the articulatory model's ability to learn consonant configurations according to vowel contexts based on the coarticulation model. Figure 33 illustrates the learned vocal tract shapes of the bilabial stops at the moment of maximal constriction. The bilabial stop /b/, for example, is articulated with closed lips in all instances but the tongue shape of the consonant target is ready for the vowel. During the execution of the consonant target, the lips are closed, but at the same time the tongue is high and front in 'bead' and low and back in 'bod'. The learned lip distance parameters are all below zero<sup>10</sup> (Figure 34), which represent virtual articulatory targets of closed lips. This indicates that the model learned similar consonant targets despite different vowel contexts.

---

<sup>10</sup>The lip distance parameter can be below zero in VocalTractLab (Table 2), for the purpose of ensuring closed lips as a virtual target. When the parameter trajectories are passed to the synthesiser, the minimal tube area will be adjusted to 0.001cm<sup>2</sup> automatically.

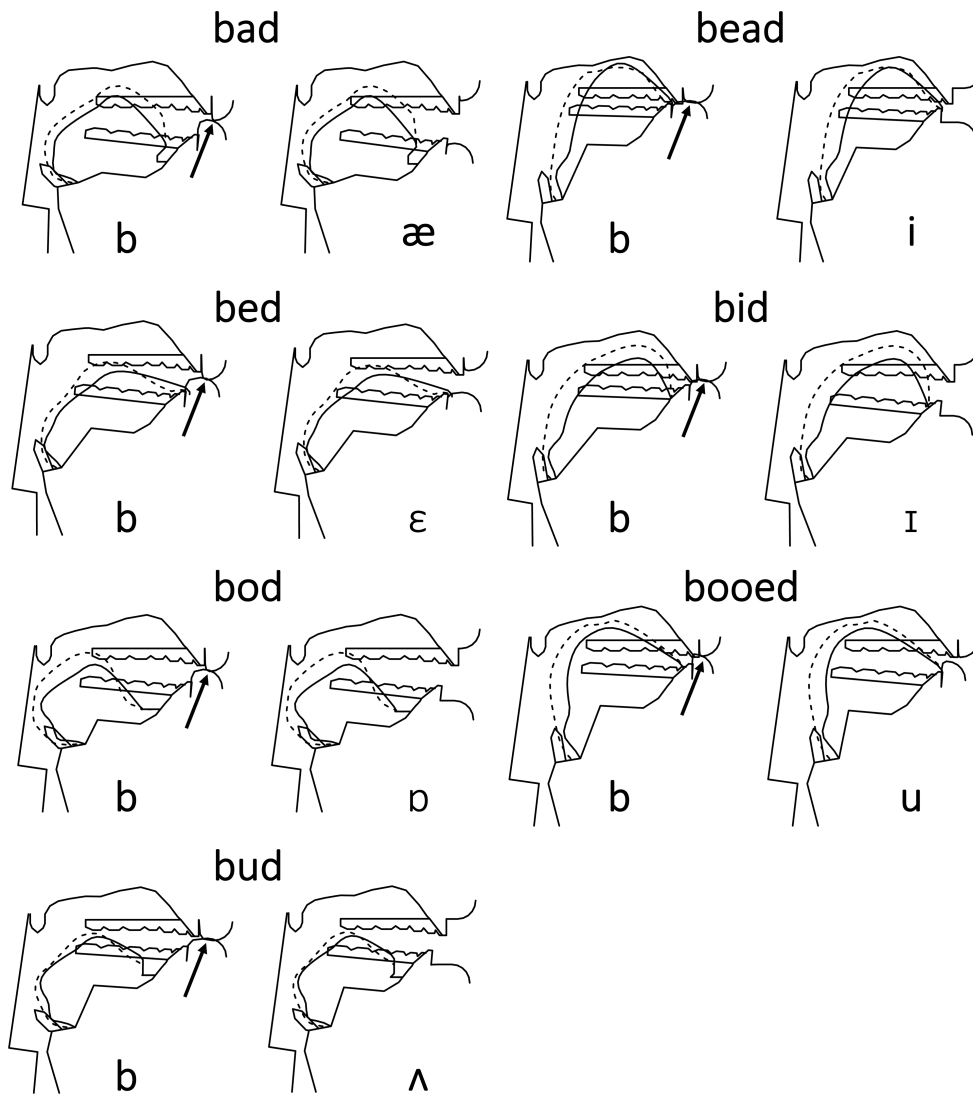


Figure 33 Midsagittal sections of the vocal tract shapes of bilabial stop-vowel sequences learned by an adult vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets.

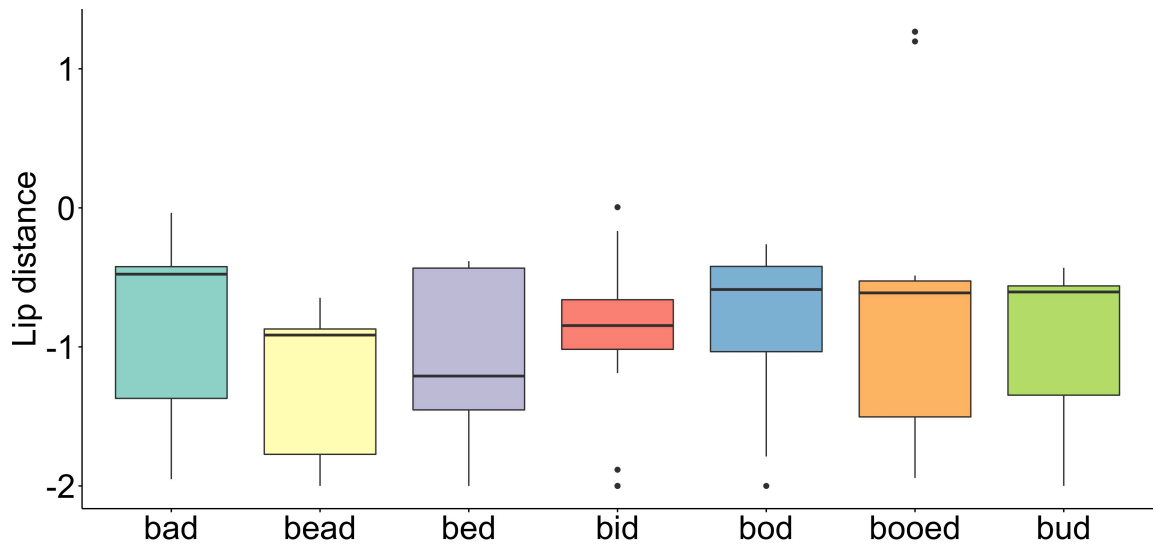


Figure 34 Boxplots of the learned lip distance parameters of all bilabial stop-vowel sequences learned by an adult vocal tract model.

As shown in Figure 35, in the case of /d/, an alveolar closure is formed for all of the learned consonant targets but the back of the tongue varies with the following vowel. If we compare ‘deed’ with ‘did’, it is not difficult to tell that the tongue body is higher in ‘deed’ than in ‘did’. The anterior part of the tongue is in a similar shape at the moment of the oral constriction, while the posterior part of the tongue is in a shape similar to the adjacent vowel. /ɪ/ and /i/ are very similar vowels, but /ɪ/ is slightly more open than /i/. The learned lip gestures and the jaw positions also suggest the trend of moving towards a more open vowel. Spearman’s correlation indicates that there is a significant negative correlation between the tongue tip parameters in the x and y-axis (Figure 36). This indicates that when there is a consonant constriction formed at the alveolar ridge, the fronter the tongue tip position, the lower the closure is formed. The tendency corresponds well with the shape of the alveolar ridge which is also low at the front.

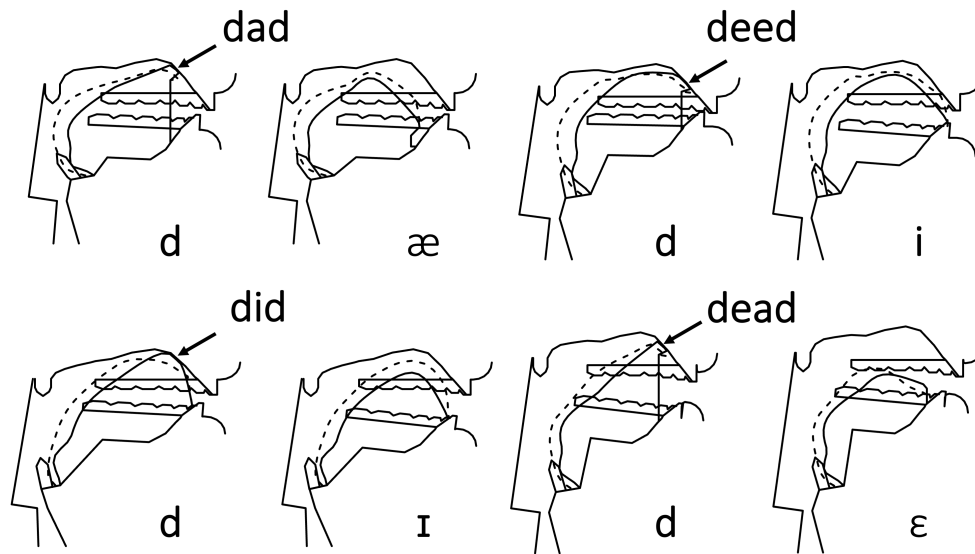


Figure 35 Midsagittal sections of the vocal tract shapes of alveolar stop-vowel sequences learned by an adult vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets.

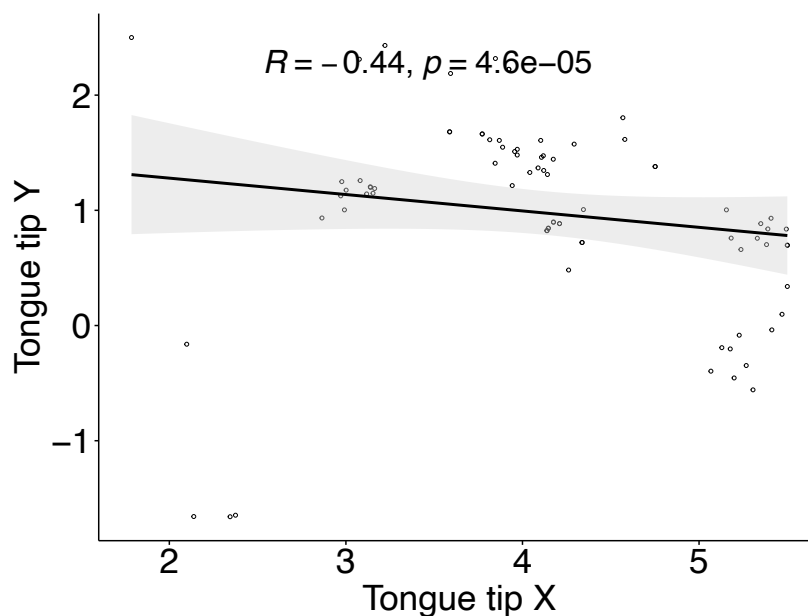


Figure 36 Correlation between the learned tongue tip parameters in the horizontal and vertical positions of an adult vocal tract model.

Figure 37 shows the learned vocal tract shapes of the alveolar stops at the moment of maximal constriction. For the velar stop /g/, the tongue body is more

advanced in 'good' than in 'god' because /ʊ/ is fronted in American English (Thomas, 2001). For horizontal tongue body centre position (TCX), the more positive the number, the more front the tongue position. The learned tongue body targets of velar stops are similar in the vertical dimension (TCY) but different in the horizontal dimension (TCX). As the target vowel changes from /ɒ/ to /ʊ/, the learned TCX becomes more positive, indicating more front tongue body positions. As the tongue body moves upwards to contact the soft palate, it also moves in the horizontal dimension toward the vowel. There was a significant positive correlation between the tongue body position in the x and y-axis, which corresponds with the shape of the palate (Figure 38). When there is a consonant constriction formed near the soft palate, the more anterior the tongue body position, the higher the closure is formed.

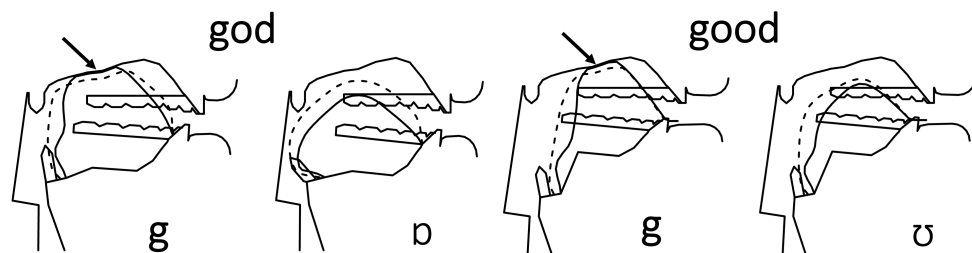


Figure 37 Midsagittal sections of the vocal tract shapes of velar stop-vowel sequences learned by an adult male vocal tract model. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the consonant targets.

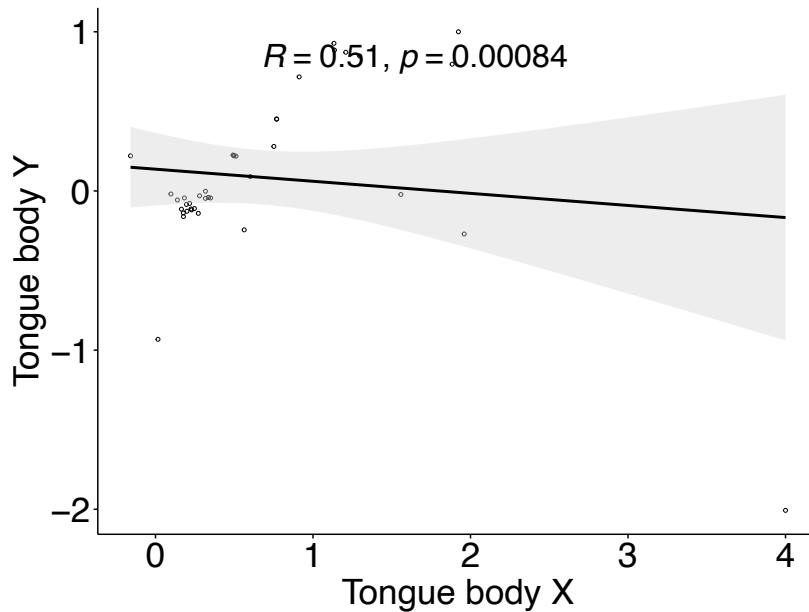


Figure 38 Correlation between the learned tongue body parameters in the horizontal and vertical positions of an adult vocal tract model.

#### 4.2.2 DISCUSSION

Past studies have not modelled the motor control system in much detail (see Appendix Table A Motor control and Synthesiser). The adopted vocal tract models are often simplistic, including only a few articulators (Barnaud et al., 2019; de Boer, 2000; Miura et al., 2012; Oudeyer, 2005; Warlaumont & Finnegan, 2016; Westermann & Miranda, 2004). Even when a sophisticated-vocal tract model (i.e., VocalTractLab) was used, the dynamic articulatory movements have been overlooked (Murakami et al., 2015; Philippsen et al., 2014). Some studies have applied Dynamic movement primitives (DMPs) framework (Forestier & Oudeyer, 2017; Philippsen, 2021a) and the task dynamic model (Howard & Messum, 2007, 2014, 2011; Messum & Howard, 2015) to model the time-varying articulatory kinematics. However, none of those has been able to generate intelligible CV syllables. A central advance in the current study was that I explicitly modelled CV coarticulation. Not only did the model learn context-sensitive consonant targets that yielded natural sounding syllables (Fig. 17) but also the learned articulatory



targets can be generalised to novel multisyllabic words (i.e., CVCV syllables in Fig. 19).

It has been well-established that different segments impact on the articulatory movement of the surrounding segments to a varying degree, known as ‘coarticulation resistance’. The contextual variations have been interpreted as originating from variable temporal overlap of gestures between the consonant and vowel in the task dynamic model and Articulatory Phonology, which involves gestural blending (Browman & Goldstein, 1989, 1992; Fowler & Saltzman, 1993; Fowler, 1980; Saltzman & Munhall, 1989). In the present study, I tested the alternative hypothesis that even though the consonant and the vowel are overlapped in time, the execution of the targets can be serially ordered for specific articulatory dimensions (Liu et al., 2022; Xu, 2020). Take the learned /bV/ sequences for example, the lip distance is controlled by the consonant, while the tongue moves toward the co-produced vowel at the same time. In the case of /dV/ and /gV/ sequences, the crucial articulator dimensions of the tongue move upwards to form constrictions in the vertical dimension, while the rest of the tongue moves backward or forward for the co-produced vowel. The simulated CV co-onset brings us back to the observation of vowel and consonant movement beginning at the same time upon which the term ‘coarticulation’ was originally proposed (Menzerath & de Lacerda, 1933).

The present study shows that the problem of massive contextual variability of the phonetic segment that has been a major irritant in concatenative speech synthesis can be resolved in a biologically realistic articulatory synthesis paradigm without excessive amounts of training data. Unlike previous articulatory synthesis that relies on articulatory data (Birkholz, 2013; Story, 2005, 2009), the current method tackles the acoustic-to-articulation problem by implementing analysis-by-articulatory-synthesis with internal synchronisation rules. Overall, the findings provide support for the hypothesis that CV coarticulation is realised by co-onset of multiple target approximation movements, each of which is sequential at the level of individual articulator dimensions. The model succeeded in simulating the learning of contextual articulatory variances with fairly high

intelligibility. The findings therefore offer new insight on the basic mechanisms of coarticulation and vocal learning, and may eventually have implications for high-quality articulatory synthesis. An issue that has not been addressed in this study is how the critical articulator dimensions for consonant targets are discovered by learners rather than being pre-set as was done in the present study. Another limitation of the current study is that coarticulation was only modelled for CV syllables. The co-onset of consonant and vowel was not only observed in CV syllables but also in more complex CCV and CCCV syllables (Kozhevnikov & Chistovich, 1965; Liu & Xu, 2021). Further research should be undertaken to test whether the synchronised dimension-specific target approximation also applies to consonant clusters.

## Chapter 5 HUMAN VOCAL LEARNING

*Part of this section has been previously submitted to a journal: Xu, A., Niekerk, D. R. v., Gerazov, Krug, P., Birkholz, P., Prom-on, S., Halliday, L., & Xu. Y. A computational simulation of human vocal learning. (Under review)*

In this chapter, I will compare the vocal learning performance of vocal tract models of different ages to test how the anatomical structure of vocal apparatus affects production learning. I will then compare the model performance with the developmental trajectories of phonetic acquisition to see how well the model can reflect the real-life learning scenario.

### 5.1 ADULT VOCAL LEARNING

In this section, I will first present the learning results of the adult vocal tract, so that they can serve as a reference for the child vocal tract results to be presented later.

### 5.1.1 CVC WORDS

I trained the adult vocal models to learn English consonant-vowel-consonant (CVC) words using the simulation model introduced in Chapter 3. The learned synthetic male speech shows a momentary burst of the onset consonant followed by clear vowel formants and high energy aspiration of the coda consonant, similar to the natural speech of a female speaker (Figure 39).

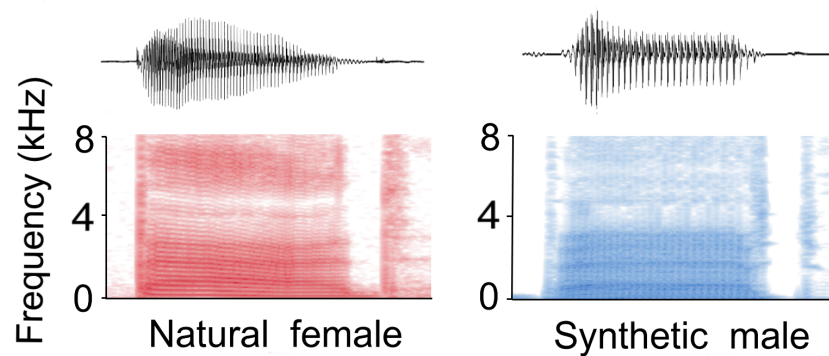


Figure 39 Waveforms and wide-band Mel-spectrograms of ‘bad’ produced by a native speaker and learned by an adult vocal tract model.

Given that the coda consonant remains the same in the word list, what was effectively evaluated in the close-set transcription experiment was the intelligibility of the initial CV portion of the words. For synthetic speech, the mean phoneme accuracy rate of CV syllables was 74% in the open-vocabulary experiment and 76% in the close-set transcription experiment. The mean phoneme accuracy rate of CVC words including the coda was 76% in the open-vocabulary experiment. With regard to natural female speech, the mean phoneme accuracy rate of CV syllables was 93% in the open-vocabulary experiment and 96% in the close-set experiment. The mean phoneme accuracy rate including the coda was 95% in the open-vocabulary transcription experiment. Figure 40 shows the overall phoneme accuracy rate of the natural female speech and the synthetic speech learned by the adult vocal tract model in the open-vocabulary and the close-set experiment. Although natural speech had overall higher phoneme accuracies, there was still a small proportion of native listeners

who were proficient at identifying synthetic speech in both types of listening experiments. I can also see that the phoneme accuracies were widely spread out for the synthetic speech, indicating a high variability in the listeners' identification. Meanwhile, the data points were slightly more concentrated in the close-set experiment than in the open-vocabulary experiment. The natural female speech was more intelligible than the synthetic male speech in both types of experiments (Wilcoxon signed-rank:  $p < .001$ ). There was a slight increase in the identification rate when the participants were provided with a limited vocabulary. However, Wilcoxon Signed Rank test showed that the overall phoneme accuracies did not differ across listening task types for either the natural speech ( $p = 0.111$ ) or the synthetic speech ( $p = .370$ ).

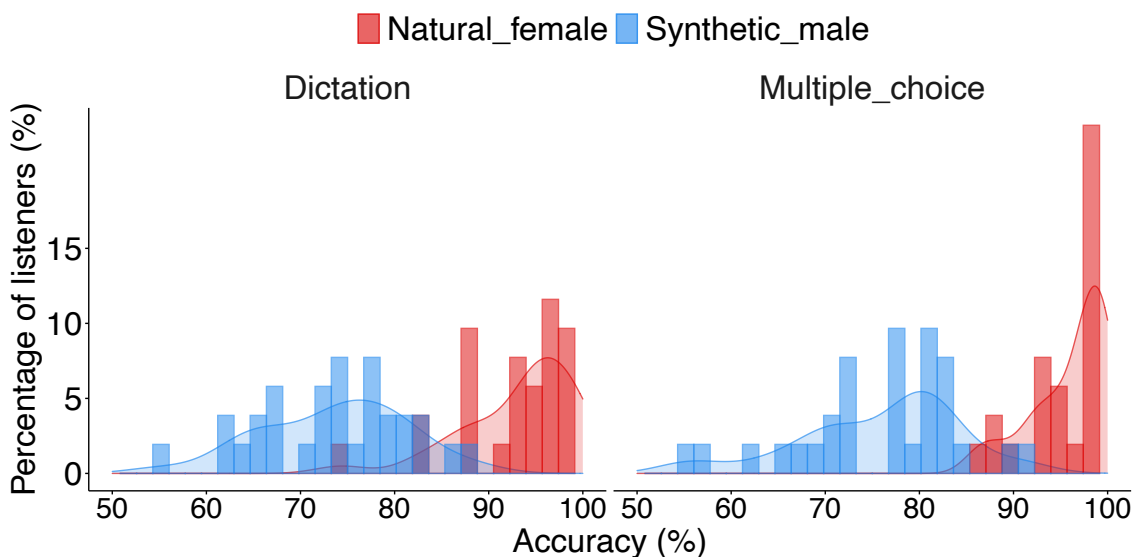


Figure 40 Histograms and Kernel density plots of by-listener mean phoneme identification accuracy rates of CV syllables in target CVC words produced by a female native speaker and by an adult male vocal tract model in listening experiments.

Figure 41 shows the phoneme accuracies of the natural and synthetic speech in each phoneme position. In the open-vocabulary transcription experiment, words learned by the adult vocal tract model were relatively intelligible, with a median phoneme accuracy rate of 87%, 60%, 82% in the onset, the vowel and the coda

position, respectively, although this was short of the natural female words which were almost flawlessly recognised (Onset: 95%, Vowel: 97%, Coda: 100%). Given that no context was provided in the experiment, the identification rate suggests that the model can learn to produce relatively intelligible speech. In the close-set transcription task, the listeners had a median of 100% accuracy rate and 96% accuracy rate in identifying the onset and the vowel respectively for the natural female speech, while the accuracy rate was 88% in the onset and 64% in the vowel for the synthetic male speech. Wilcoxon signed-rank test showed that the natural speech had higher phoneme accuracies in all the syllable positions in the open-vocabulary experiment (Onset:  $p < .001$ , Vowel:  $p < .001$ , Coda:  $p < .001$ ) and the close-set experiment (Onset:  $p < .001$ , Vowel:  $p < .001$ ). In addition, the type of listening experiment did not affect the identification rates of the natural female speech in either the onset ( $p = 0.056$ ) or the vowel position ( $p = 1.000$ ). Similar trend was found for the synthetic male speech in the onset ( $p = 1.000$ ) and vowel position ( $p = .554$ ).

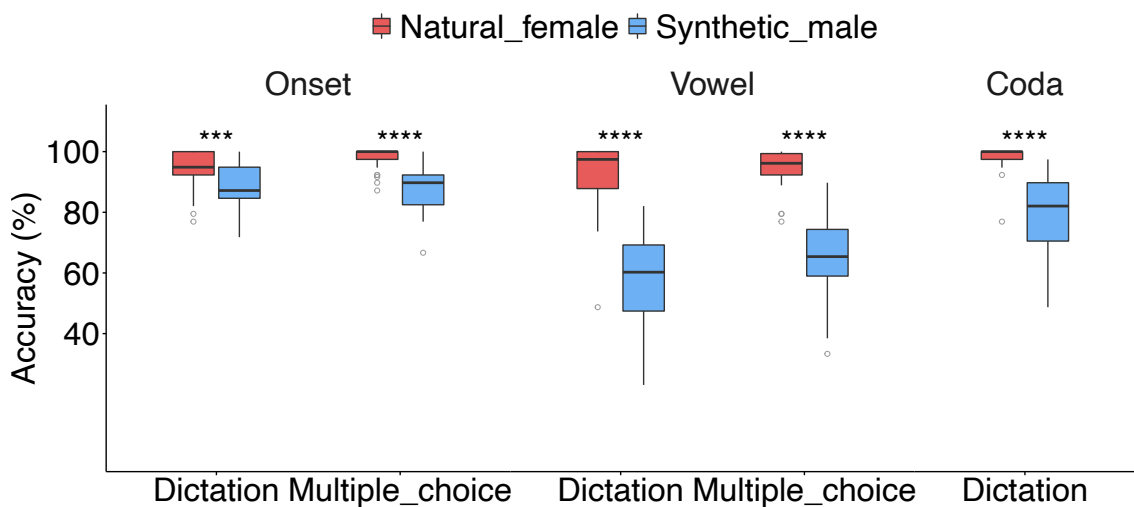


Figure 41 By-listener phoneme accuracy rates of natural female speech and synthetic male speech in different syllable positions, evaluated by an open open-vocabulary transcription experiment and a close-set transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

Figure 42 shows the learning performance of each target word in the onset, vowel and coda position in the open-vocabulary transcription experiment. Several synthetic onset consonants were perfectly identified, with accuracies greater than or equal to the natural words, such as the bilabial stops in ‘bed’, ‘bid’ and ‘bod’ and the alveolar stops in ‘deed’ and ‘did’. The accuracies of the onset consonant were relatively low in ‘good’, ‘dad’ and ‘booed’. The vowel learning was less successful, as shown in Figure 42. Compared with the high identification accuracy in ‘bad’ (i.e., 99%), only less than half of the mid vowels in ‘bed’ and ‘bid’ were correctly identified. In addition to mid vowels, there is room for improvement for the vowel in ‘booed’. It is worth noting that even for natural speech some vowels were sometimes misidentified. Finally, the coda accuracies were relatively high and stable across vowel contexts. There was only one exception in the word ‘bud’, which was frequently heard as ‘but’.

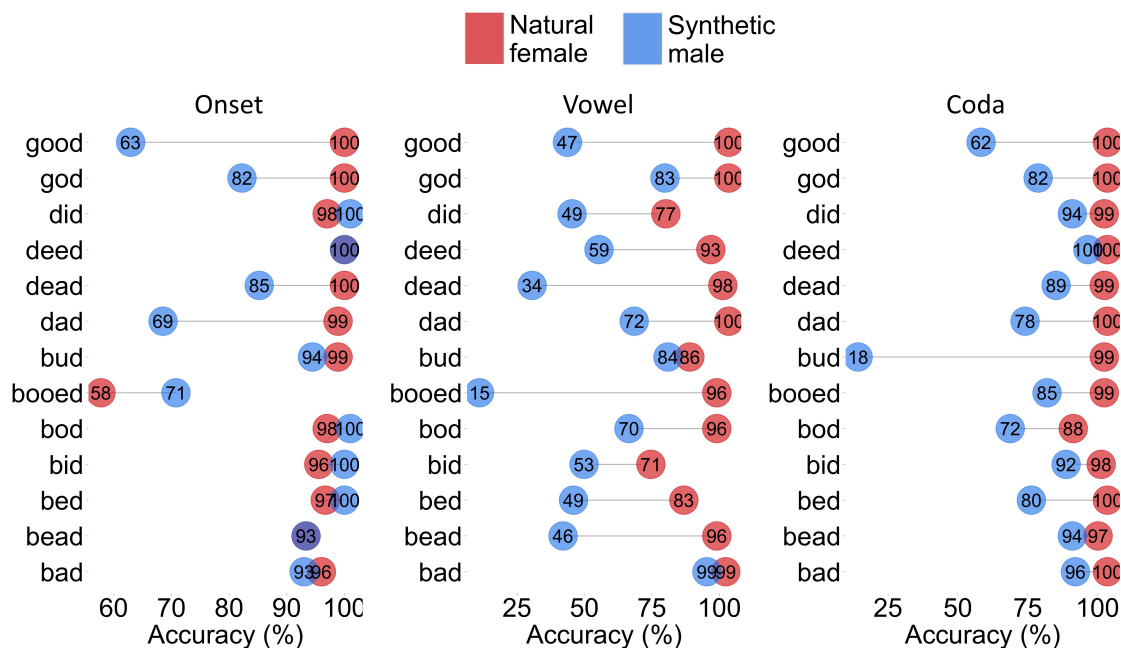


Figure 42 Mean identification rates of CVC words produced by a female native speaker and learned by an adult male vocal tract model in the onset position, vowel position and coda position in an open-vocabulary transcription experiment.

Figure 43 shows the mean identification of natural and synthetic utterances in the close-set transcription experiment. The synthetic male speech had identification

rates similar to the natural speech in the onset position for most of the CVC words. The identification rates were lower for the vowels in the synthetic CVC utterances compared to natural female speech. The CV combination in synthetic /booed/ had the lowest identification rates in both the onset and vowel positions.

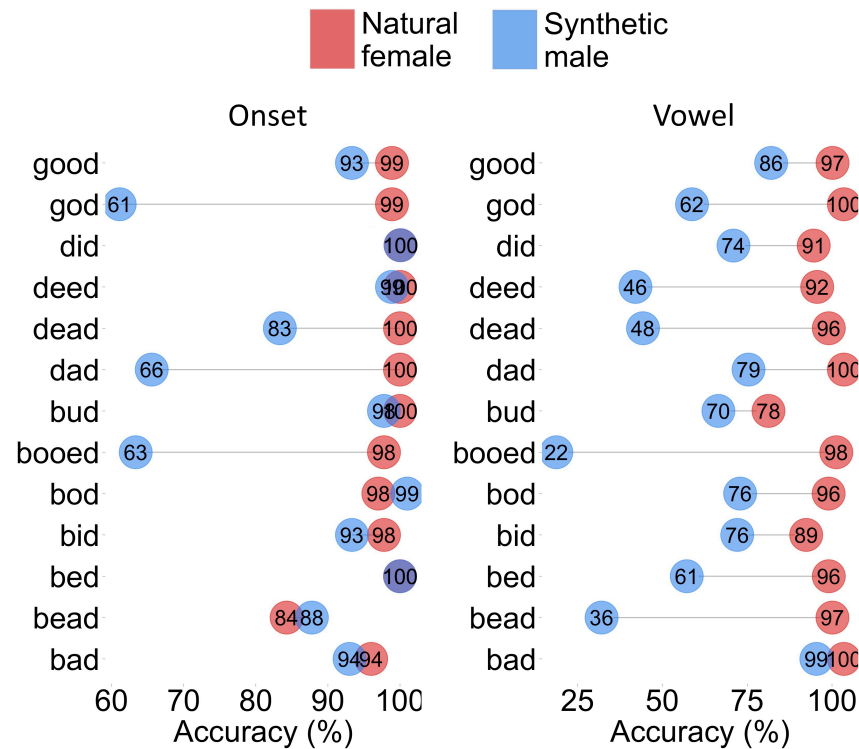


Figure 43 Mean identification rates of CVC words produced by a female native speaker and learned by an adult male vocal tract model in the onset position and vowel position in a close-set transcription experiment.

Figure 44 shows detailed confusion matrices of the natural and the synthetic speech in the close-set transcription experiment. Listeners identified natural female speech highly accurately when choosing from a word list, except in one instance where ‘bud’ was identified as ‘bod’. With regard to the synthetic male speech, there was more confusion on the vowels. For example, listeners identified ‘bead’ as ‘bid’, ‘dead’ as ‘dad’, and ‘deed’ as ‘did’. The low percentage of correctly identified ‘booed’ indicates that the vowel was confusing for the participants. ‘booed’ was frequently heard as ‘bud’ and ‘bid’. The synthetic onset consonants had fairly high accuracies but the place of articulation influenced the

learning performance. The listeners tended to have more difficulties in identifying the alveolar and velar stops in the synthetic speech, whereas bilabial stops were not easily mistaken. For example, the velar stop in 'god' was often identified as alveolar stops.



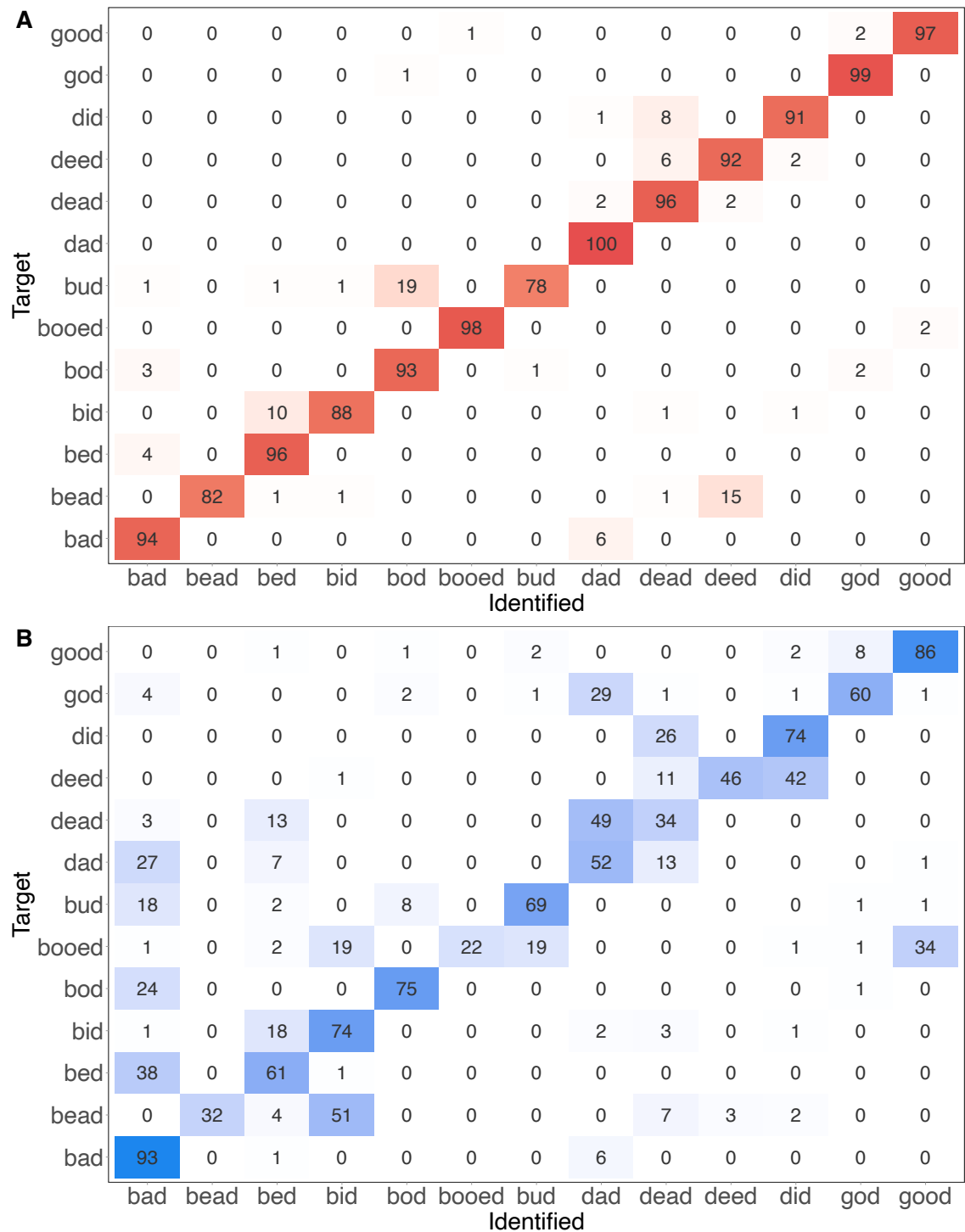


Figure 44 Comparison between natural female speech and synthetic male speech. Confusion matrix (%) of CVC words produced by a female native speaker

(a) and learned by an adult male vocal tract model (b), measured by a close-set transcription experiment.

### 5.1.2 CVCV WORDS

To test whether the articulatory targets learned from single syllables were generalisable to novel words that the model was never trained on, I resynthesised CVCV words based on the learned articulatory targets. The regenerated disyllabic words synthesised with targets from monosyllabic words achieved similar accuracies as natural words in the close-set transcription experiment, as shown in Figure 45. Especially in the close-set experiment, the distribution of the identification accuracies for the natural female speech and synthetic male speech was almost identical. In the open-vocabulary experiment, the mean identification accuracy rate was 88% and 95% for the synthetic male speech and natural female speech, respectively. The natural speech had significantly higher identification accuracies than the synthetic speech (Wilcoxon signed-rank,  $p < .001$ ). With respect to the close-set experiment, the mean identification accuracy rate was 96% for the synthetic words and 97% for the natural words. The synthetic speech and natural speech did not differ significantly (Wilcoxon signed-rank,  $p = .442$ ). The type of experiment did not significantly influence the identification accuracy of natural speech (Wilcoxon signed-rank,  $p = .066$ ). However, the mean identification accuracies were significantly higher in the close-set experiment for the synthetic male speech (Wilcoxon signed-rank,  $p < .001$ ). Overall, the results indicated that the regenerated CVCV words using learned vocal tract parameters were relatively intelligible.

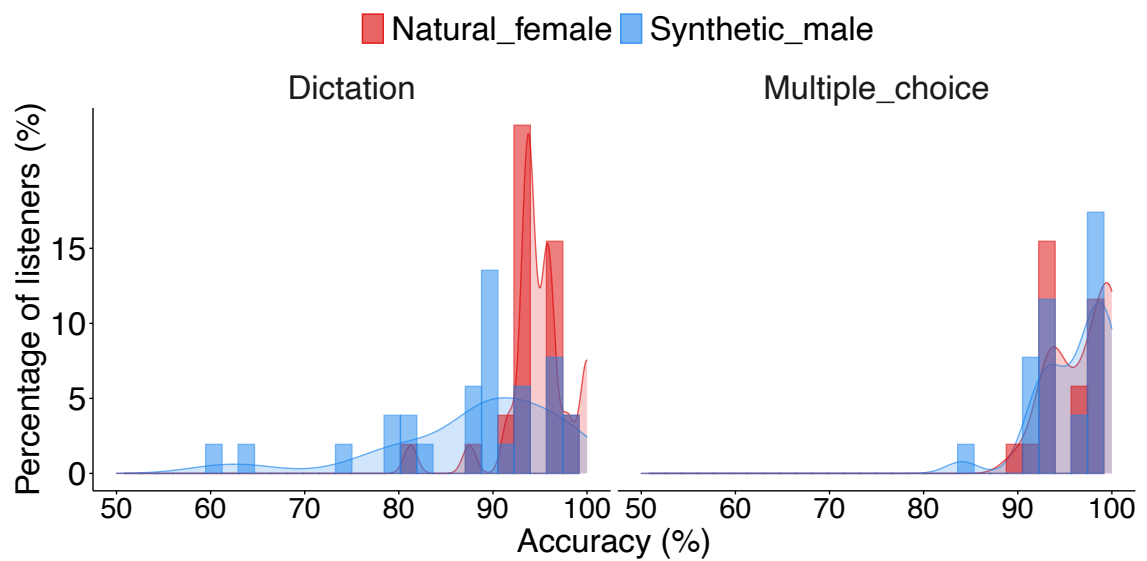


Figure 45 Distribution of by-listener mean phoneme identification accuracy rates of CVCV words produced by a female native speaker and learned by an adult male vocal tract model in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners.

I analysed the by-position phoneme accuracies of the natural and synthetic CVCV words in the open-vocabulary transcription experiment. The means and standard deviations of the phoneme accuracies are shown in Table 6 and the boxplots of the identification rate are shown in Figure 46. As we can see from the plot, the natural female speech had much higher identification rate mainly in the onset positions. The synthetic speech had a similar accuracy rate to the natural speech in the two vowel positions. The standard deviations of the identification rate were higher in the synthetic speech than the natural speech in all the phoneme positions (Table 6), indicating that listeners varied to a greater extent while identifying the synthetic speech. Wilcoxon signed-rank tests showed that the natural speech had higher accuracy rate than the synthetic speech in the first onset ( $p = .011$ ) and the second onset ( $p < .001$ ) position, but not in the first vowel ( $p = .104$ ) or the second vowel ( $p = 1.000$ ) position. This indicated that the adult vocal tract model learned close to natural vowels in CVCV words.

Table 6 Mean and standard deviation (in parentheses) of natural and synthetic CVCV words in each phoneme position, evaluated by an open-vocabulary transcription experiment.

	First onset	First vowel	Second onset	Second vowel
<b>Natural female</b>	99.72 (1.52)	80.56 (12.05)	98.61 (5.40)	99.72 (1.52)
<b>Synthetic male</b>	94.72 (10.83)	71.67(17.59)	86.94(14.79)	99.44(2.11)

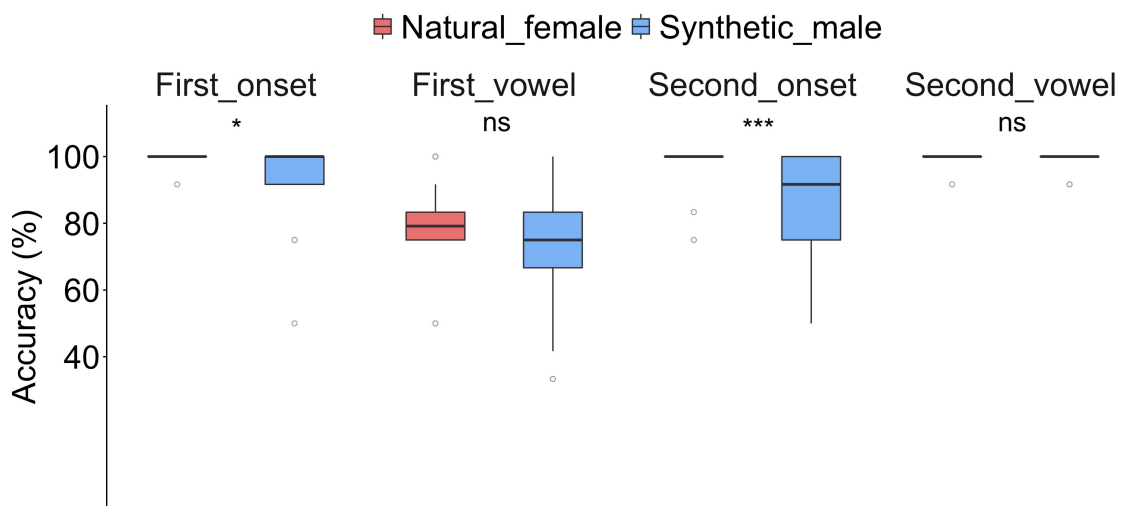


Figure 46 By-listener phoneme accuracy rates of CVCV words produced by a native female speaker and learned by an adult male vocal tract model in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns  $P > 0.05$ , \*  $P < 0.05$ , \*\*\*  $P \leq 10^{-3}$ .

The identification rates for each phoneme in the target words are shown in Figure 47. The accuracy rate of the first onset consonants in the synthetic speech was all over 90%. The accuracy of the first vowel of the synthetic speech was similar to the natural speech in most of the words except for the one in /daddy/. The second onset consonant in the learned /daddy/ was lower than the natural speech but the other three target words had identification rate close to the natural speech.

The final vowel in the synthetic CVCV words had 100% accuracy rate for /Debbie/ and /daddy/. Synthetic /buddy/ and /body/ also had nearly perfect identification rate in the final vowel position (99% and 98%, respectively). Wilcoxon signed-rank tests showed that the natural speech had higher identification rate in all the phoneme positions than the synthetic speech except the final vowel (First onset:  $p = .003$ , First vowel:  $p = .026$ , Second onset:  $p < .001$ , Second vowel:  $p = 0.570$ ).

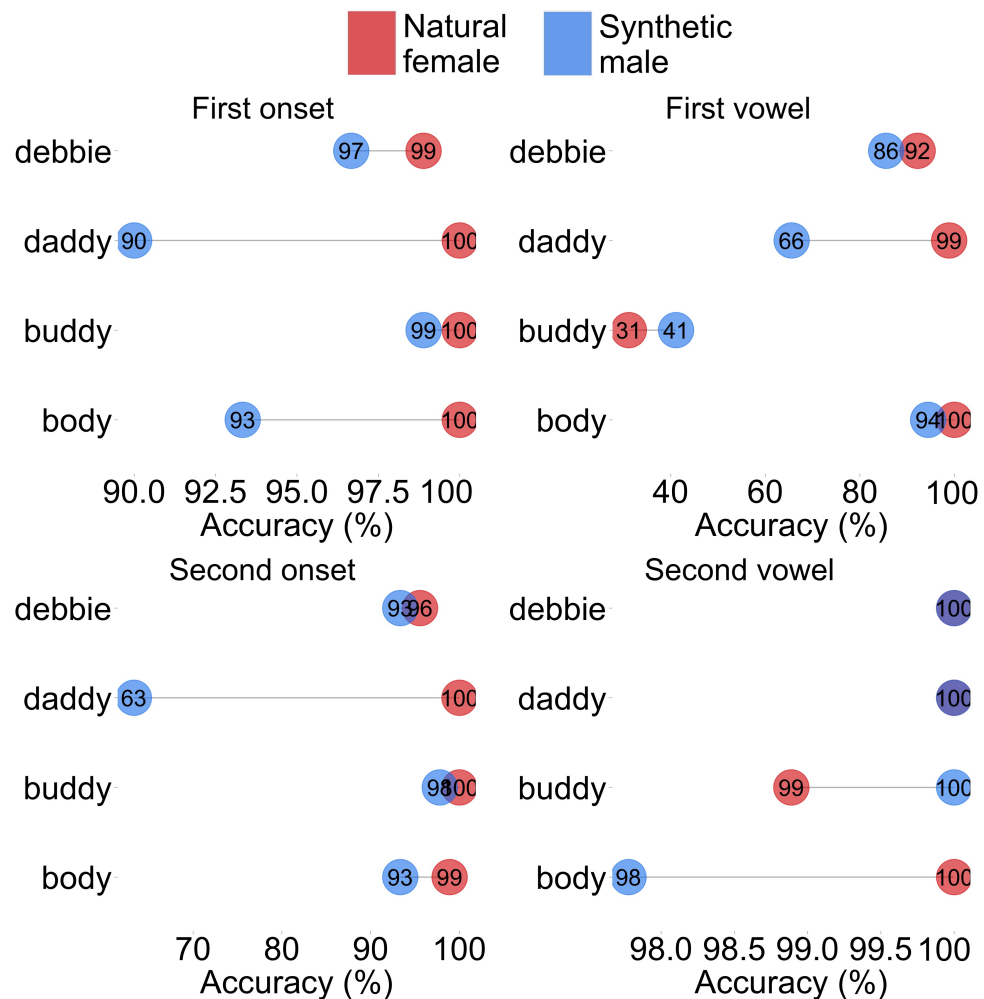


Figure 47 Mean identification rates of CVCV words produced by a female native speaker by an adult male vocal tract model in all the phoneme positions in an open-vocabulary transcription experiment.

Figure 48 shows the confusion matrices of the natural and the synthetic speech in the close-set transcription experiment. There were some confusions between

‘body’ and ‘buddy’ learned by the adult vocal tract model, but ‘Debbie’ and ‘daddy’ were almost perfectly identified. Again, vowels were the main source of confusion. ‘Debbie’ and ‘daddy’ had distinct consonant combinations and thus listeners could rely less on the vowel perception. Interestingly, ‘buddy’ learned by the adult male model even had a higher identification rate than the natural speech. Only 13% of ‘buddy’ learned by the vocal tract model was identified as ‘body’. In contrast, 40% of ‘buddy’ produced by the female native speaker was mistaken for ‘body’. In contrast, 40% of ‘buddy’ produced by the female native speaker was mistaken for ‘body’.

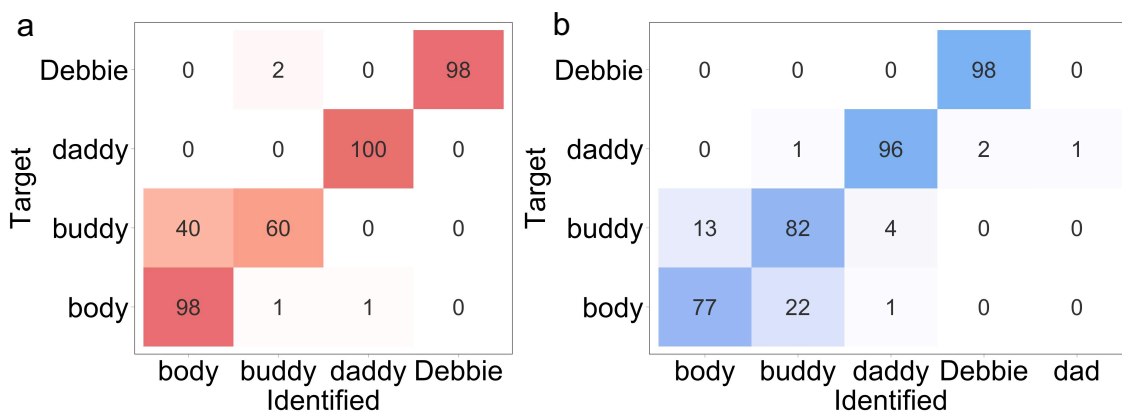


Figure 48 Comparison between natural female speech and synthetic male speech. Confusion matrix (%) of CVCV words produced by a female American English native speaker (a) and learned by an adult male vocal tract model (b), measured in a close-set transcription experiment.

Overall, the results showed that the learned vocal tract parameters could generalise to multisyllabic words that the model was never trained on.

## 5.2 CHILD VOCAL LEARNING

### 5.2.1 CVC WORDS

Fairly intelligible words were also learned by the child vocal tract models when trained by the recogniser. Figure 49 shows the mean phoneme accuracy rate of

the CV syllables in the CVC words learned by the two child vocal tract models. As we can see from the plot, the distribution of the two child vocal tract models overlaps greatly in the open-vocabulary experiment. Still, the 3-year-old vocal tract model had an overall higher mean accuracy rate than the 1-year-old model. The 3-year-old model had a mean phoneme accuracy rate of 49% for the target CV syllables in the open-vocabulary experiment, while the 1-year-old model had a mean accuracy rate of 44%. The mean phoneme accuracies in the close-set transcription experiment were 65% and 55% for the 3-year-old model and 1-year-old model, respectively. Wilcoxon signed-rank tests showed that the 3-year-old vocal tract model had higher phoneme accuracy rate than the 1-year-old model in the close-set experiment ( $p < .001$ ), but not in the open-vocabulary experiment ( $p = 0.120$ ). The type of experiment did not influence the phoneme accuracies for either of the child vocal tract models (Wilcoxon signed-rank:  $p = 1.000$ ).

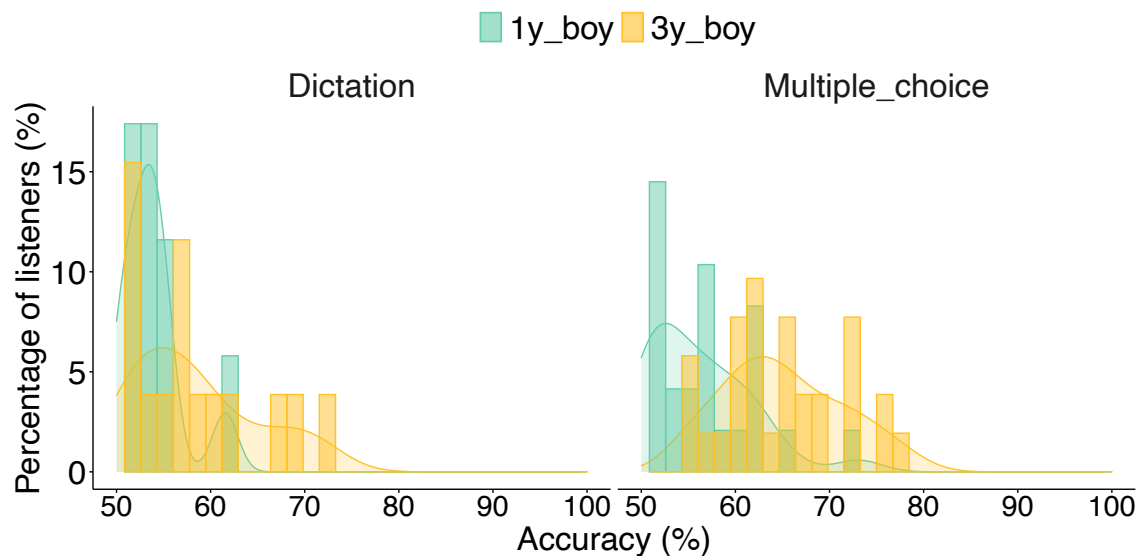


Figure 49 Distribution of by-listener mean phoneme identification accuracy rates of CVC words learned by a 1-year-old and a 3-year-old vocal tract model, tested in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners.

Furthermore, I analysed the by-phoneme position accuracy rate of the CVC words learned by the two child vocal tract models. In the open-vocabulary

transcription task, the 1-year-old model had a median of 56% phoneme accuracy rate in the onset position, compared with 63% for the 3-year-old model. Listeners correctly transcribed 32% of the vowels learned by the 1-year-old model and 38% with the 3-year-old model. For the coda position, the median identification was 68% for the 1-year-old model and 69% for the 3-year-old model. Both models had higher intelligibility in the consonant positions than the vowel position.

Figure 50 compares the identification accuracy rate of the phonemes in CVC words learned by the two child models. Each connected line represents the average phoneme accuracy rate of one listener. Solid lines indicate that the 3-year-old model has higher phoneme accuracies than the 1-year-old model and vice versa for the dashed lines. As shown in Figure 50, some listeners had higher identification rates when judging speech learned by the 1-year-old model than by the 3-year-old model in the open-vocabulary transcription task (dashed lines). In contrast, we can rarely see such cases in the close-set transcription task, that is, there were only a few cases where the words learned by the 1-year-old model were more intelligible than the 3-year-old model. Wilcoxon signed-rank tests showed that the 3-year-old model learned more intelligible speech in the onset ( $p = .019$ ) and the coda position ( $p = .019$ ), but not in the vowel position ( $p = .063$ ) in the open-vocabulary experiment. In the close-set experiment, similarly, the 3-year-old model had higher accuracies than the 1-year-old model in both the onset and the vowel position (Wilcoxon signed-rank:  $p < .001$ ). The increase in the perceptual accuracies suggests that the growing child vocal tract have enhanced capability to learn articulatory targets that yielded intelligible speech.



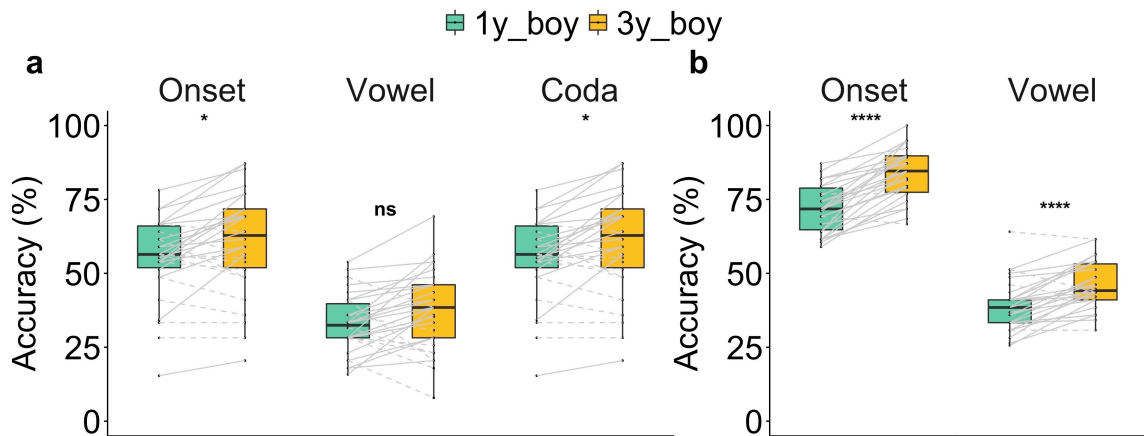


Figure 50 Boxplots of by-listener phoneme accuracy rates in different syllable positions of CVC words learned by a 1-year-old and a 3-year-old vocal tract model, measured in an open-vocabulary transcription experiment (a) and a close-set transcription experiment (b). ns  $P > 0.05$ , \*  $P < 0.05$ , \*\*\*\*  $P \leq 10^{-4}$ .

When the word list was given, listeners could identify the words learned by both models more easily. There were less variances in the phoneme accuracy rate of the close-set transcription task than the open-vocabulary transcription task. The median phoneme accuracies increased in the onset consonant position (Wilcoxon signed-rank: 1y:  $p < .001$ , 3y:  $p < .001$ ), which was 85% and 72% for the 1-year-old model and 3-year-old model respectively. There was improvement in the vowel accuracies as well (Wilcoxon signed-rank: 1y:  $p = 0.017$ , 3y:  $p = 0.003$ ). The median vowel accuracy rate was 44% and 38% for the 3-year-old model and 1-year-old model respectively.

Figure 51 shows the by-position phoneme accuracy rate of each target CVC word learned by the two child vocal tract models. The two child vocal tract models had similar accuracies for bilabial stops and alveolar stops. However, the 1-year-old vocal tract model learned poorer velar stops in 'god' and 'good', when compared with the 3-year-old model. With respect to the learning of vowels, the two models had comparable performance for most of the vowels. The 3-year-old model yet again showed better results in the case of /u/ in 'good' and /b/ in 'god' than the 1-year-old model. Both child models failed to learn intelligible /b/ in 'bod'.

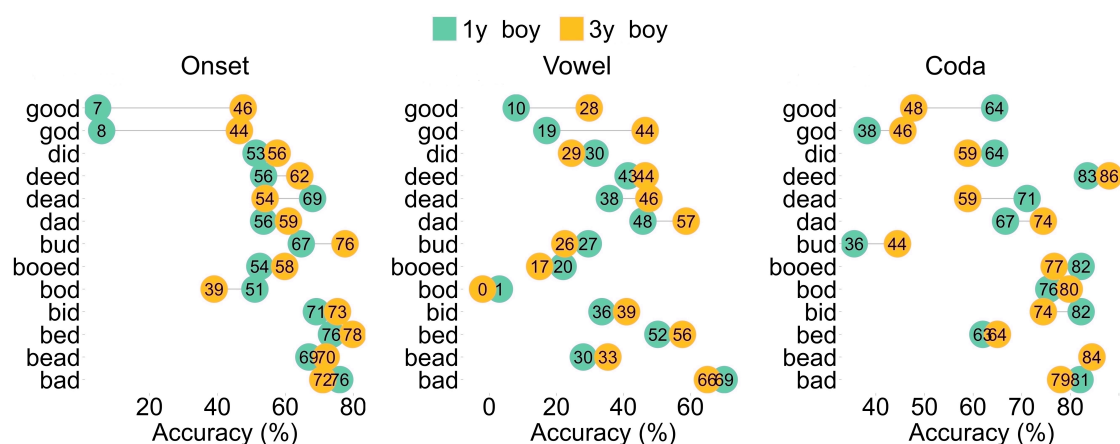


Figure 51 By-listener mean phoneme accuracy rates of utterances learned by a 1-year-old and a 3-year-old vocal tract models in the onset position, the vowel position, the coda position of the CVC words, measured by an open-vocabulary transcription task.

The confusion matrices of the CVC words of the two child vocal tract models in the close-set transcription experiment are shown in Figure 52. The child vocal tract model learned relatively intelligible vowels in ‘bad’ (1y: 71%; 3y: 68%), ‘bed’ (1y: 60%; 3y: 59%), and ‘bid’ (1y: 64%; 3y: 64%). The bilabial stops were rarely mistaken as other types of consonants except for the one in ‘bod’, which was sometimes mistaken as an alveolar stop. The place of articulation of alveolar stops was almost always correctly identified for both child vocal tract models. The learning of velar stops was relatively successful for the 3-year-old vocal tract model but not for the 1-year-old model. The velar stops learned by the 1-year-old model were often identified as alveolar stops and bilabial stops. Only a very small proportion of velar stops was correctly identified (4% in ‘good’ and 22% in ‘god’) for the 1-year-old model. In contrast, the velar stops learned by the 3-year-old model had a fairly high accuracy rate, which was 89% in ‘god’ and 93% in ‘good’.

Both models had difficulty in learning vowels with similar openness and tongue height. For instance, ‘bead’ was often mistaken as ‘bid’, and ‘bud’ as ‘bod’. The learning of the vowel /ɒ/ in ‘bod’ was unsuccessful for both child models, which was heard as /ɪ/ in ‘bid’. The rounded vowel /u/ was difficult for both child vocal tract models. Compared with the 3-year-old model, the 1-year-old model learned

much less intelligible vowels following velar stops in 'god' and 'good'. Only 7% was correctly identified for the 1-year-old model and 9% for the 3-year-old model.

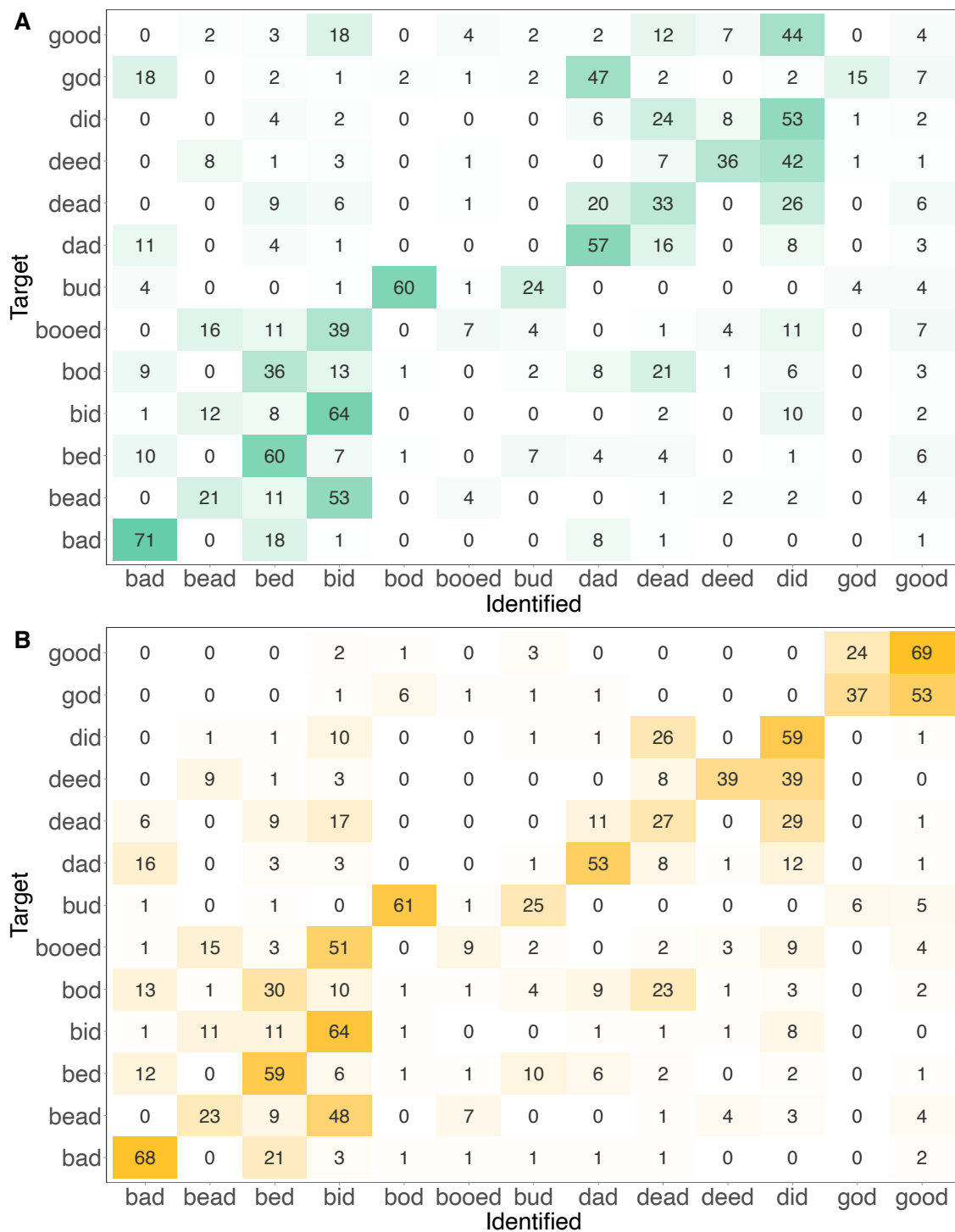


Figure 52 Confusion matrices (%) of CVC words learned by a 1-year-old (a) and a 3-year-old vocal tract model (b), measured by a close-set transcription experiment.

### 5.2.2 CVCV WORDS

The learned vocal tract parameters of the CVC words were used to regenerate CVCV words. The CVCV words of both child vocal tract models were fairly intelligible. Figure 53 shows the distribution of the mean identification rate of the CVCV words learned by the two child models. As shown in Figure 53, most of the listeners had higher accuracies while identifying the speech learned by the 3-year-old model than the 1-year-old model. A few listeners were able to identify all phonemes in both types of experiments, while the others had a difficult time in identifying the phonemes. The distribution of the mean identification rate was more concentrated in the close-set experiment than in the open-vocabulary transcription experiment. In the open-vocabulary transcription experiment, the mean identification accuracy rate of CVCV words was 64% for the 1-year-old vocal tract model and 78% for the 3-year-old model. The 3-year-old model achieved a significantly higher identification rate than the 1-year-old model in the open-vocabulary transcription experiment (Wilcoxon signed-rank:  $p < .001$ ). Furthermore, the mean accuracy rate of CVCV words was 76% for the 1-year-old model and 86% for the 3-year-old model in the close-set experiment. The learning performance of the two child vocal tract models was significantly different in the close-set experiment (Wilcoxon signed-rank:  $p < .001$ ). Wilcoxon signed-rank also showed that the mean accuracy rate was significantly higher in the open-vocabulary transcription experiment than in the close-set experiment for the 1-year-old model ( $p < .001$ ) and the 3-year-old model ( $p = .017$ ).

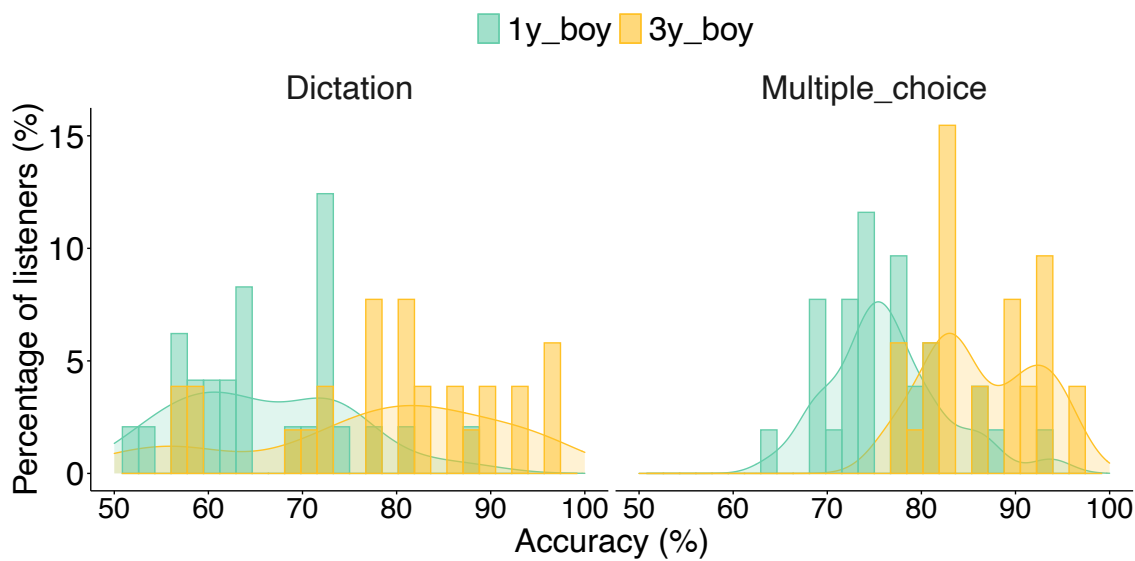


Figure 53 Distribution of by-listener mean phoneme identification accuracy rates of CVCV words learned by a 1-year-old and 3-year-old vocal tract model, tested in listening experiments. Kernel density estimate and histogram show the distribution of the performance of the listeners.

Figure 54 and Table 7 show the by-position phoneme identification accuracy rate of the CVCV words learned by the two child vocal tract models in the open-vocabulary transcription experiment. As shown in Table 7, the 3-year-old vocal tract model had higher accuracy rates than the 1-year-old model in all the phoneme positions. Both models had relatively high accuracies in the two onset consonant positions with the 3-year-old model being much more intelligible than the 1-year-old model. The final vowels learned by both child models were nearly perfectly identified, whereas the identification rate was lower for the first vowel. Figure 54 shows that for both child models there was much variability in the identification rate. The 3-year-old learned more intelligible phonemes in the first onset and the second vowel position (Wilcoxon signed-rank:  $p < .001$ ). However, the two child models had similar identification rates in the second onset (Wilcoxon signed-rank:  $p = 0.146$ ) and the second vowel (Wilcoxon signed-rank:  $p = 1$ ) position.

Table 7 Mean (%) and standard deviation (%) of CVCV words learned by the two child models in each phoneme position, evaluated by the open-vocabulary transcription experiment.

	First onset	First vowel	Second onset	Second vowel
<b>1-year-old child</b>	62.90(17.87)	34.58 (13.14)	62.37(20.73)	96.94 (9.66)
<b>3-year-old child</b>	86.11 (16.57)	53.81(20.42)	73.28 (21.24)	97.22 (6.32)

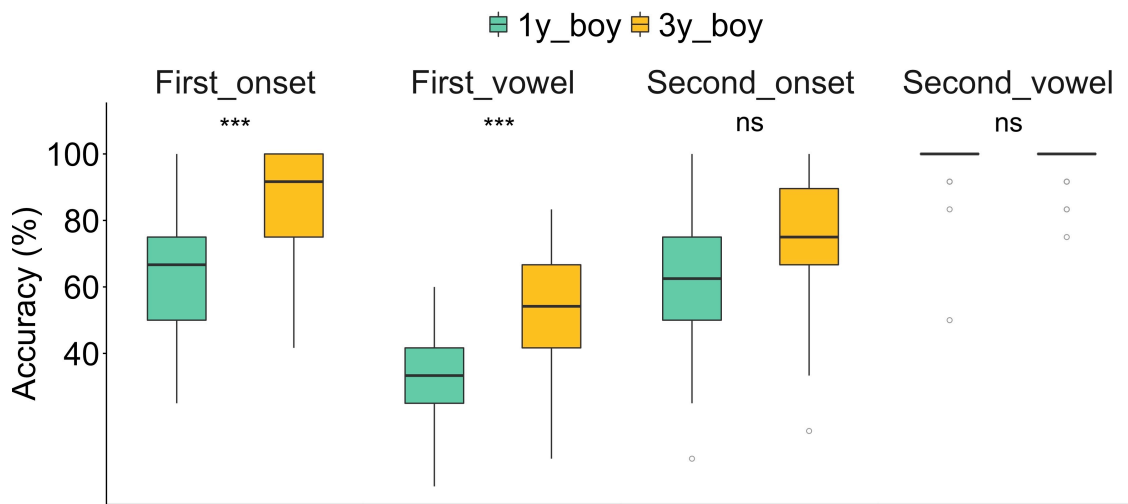


Figure 54 By-listener phoneme accuracy rates of learned CVCV words learned by two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns  $P > 0.05$ , \*\*\*  $P \leq 10^{-3}$ .

Figure 55 shows the phoneme accuracy rate of the four CVCV words learned by the two child vocal tract models. With regard to the first onset consonant, the 3-year-old model had fairly high intelligibility for all target words. Especially for 'Debbie' and 'body', the accuracy rate of the 3-year-old model was remarkably higher than the 1-year-old model. The onset consonants in 'daddy' and 'buddy' both had relatively high accuracy rate. The first vowel in 'daddy' and 'body' learned by the 3-year-old model were again more intelligible than the 1-year-old

model. Both models performed well in the learning of the first vowel in ‘daddy’, but had difficulties in learning the first vowel in ‘buddy’. As far as the second onset consonant is concerned, both child models learned intelligible alveolar stops in ‘daddy’, ‘buddy’ and ‘body’. However, the 3-year-old model again had much better performance in learning the bilabial stops in ‘Debbie’. Finally, the 3-year-old model outperformed the 1-year-old model for all the second vowels in the CVCV words.

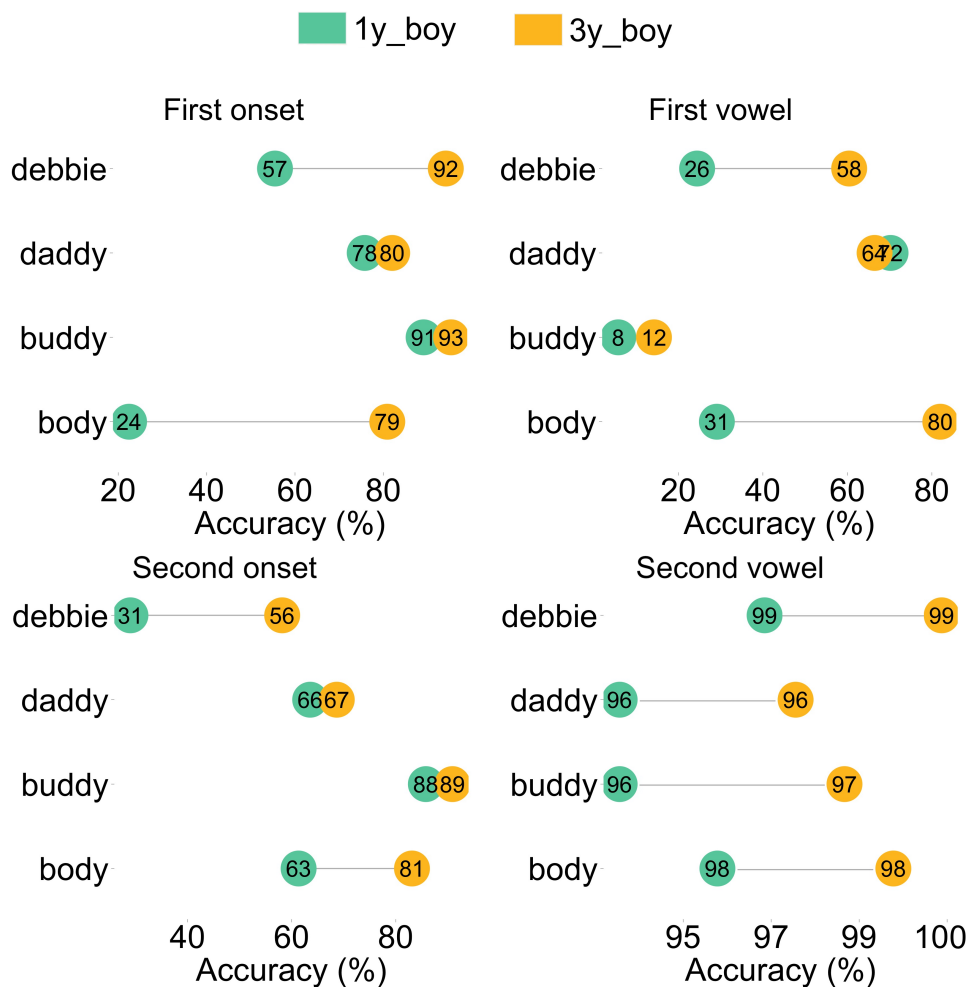


Figure 55 Mean identification rates of CVCV words learned by two child vocal tract models in all the phoneme positions in an open-vocabulary transcription experiment.

The confusion matrices of the close-set experiment for CVCV words are shown in Figure 56. Target word ‘body’ was often identified as ‘daddy’ for the 1-year-old



model, while the one learned by the 3-year-old model was quite intelligible. ‘buddy’ learned by both models was identified as ‘body’, but the 3-year-old model still had better performance. Both models learned a relatively intelligible ‘daddy’ with similar identification rate. ‘Debbie’ learned by both models was often identified as ‘daddy’, which indicated that the bilabial stops were the main source of confusion. Overall, the CVCV words regenerated by the learned vocal tract parameters of CVC words had fairly good performance. Moreover, the articulatory targets learned by the 3-year-old model had better generalisability than the 1-year-old model.

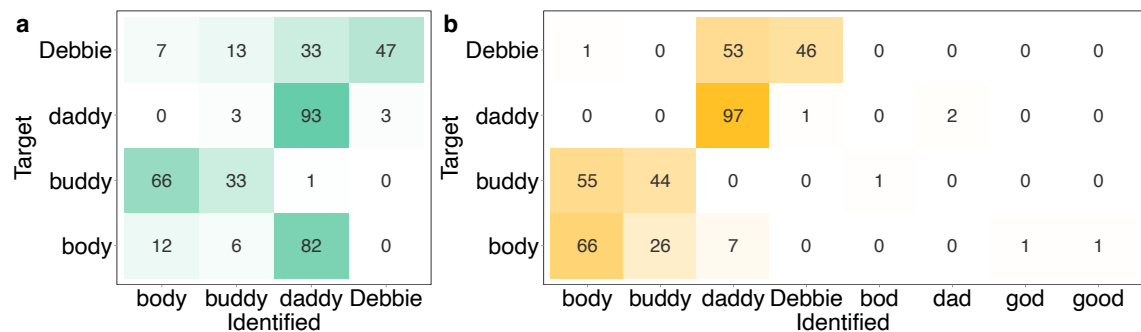


Figure 56 Comparison between two vocal tract models. Confusion matrices of CVCV words regenerated by learned vocal tract parameters of a 1-year-old (a) and a 3-year-old vocal tract model (b), measured by a close-set transcription experiment.

### 5.3 AGE-RELATED VOCAL TRACT DIFFERENCES

The results reported so far have demonstrated that learning performance of different vocal tract models vary with anatomical structure. If we compare the identification rates of the learned synthetic speech, the difference in perceptual quality is evident, as shown in Figure 57. The anatomical structure of the vocal tract model had a significant effect on accuracies in both the open-vocabulary and the close-set transcription experiments (Kruskal-Wallis test:  $p < .001$ ). Furthermore, the difference between the results for vocal tract models of each

age group was significant as well for all the comparisons (Wilcoxon signed-rank:  $p < .001$ ) except for the phoneme accuracies of two child models in the open-vocabulary transcription experiment (Wilcoxon signed-rank:  $p = .360$ ). The results indicated that the identification rate upsurged as the age of the vocal tract model increased<sup>11</sup>.

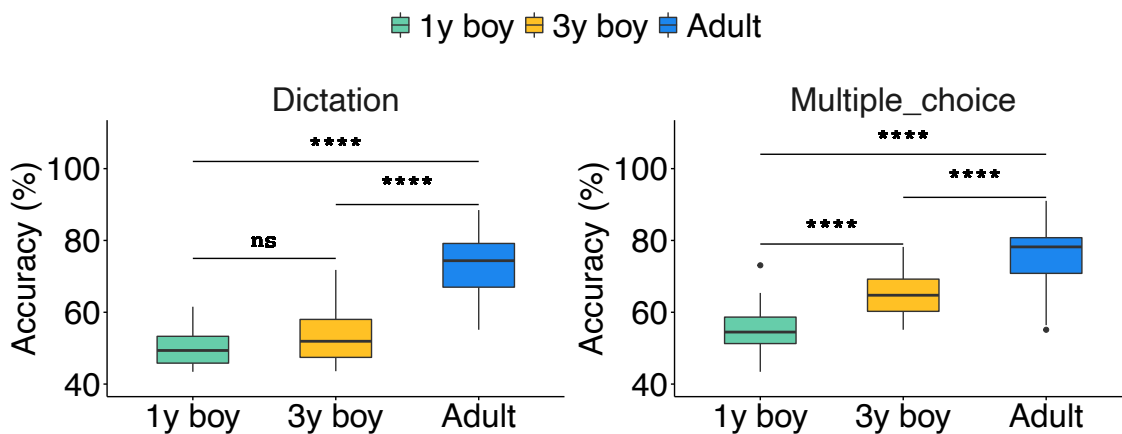


Figure 57 By-listener mean phoneme accuracy rates of CV syllables learned by a 1-year-old, a 3-year-old and an adult male vocal tract model, evaluated by an open-vocabulary transcription experiment and a close-set transcription experiment. Error bars show standard errors. ns  $P > 0.05$ , \*\*\*\*  $P \leq 10^{-4}$ .

Figure 58 shows the by-position phoneme accuracy rate of the CVC words learned by the vocal tract models. The overall tendency of the identification rate in each phoneme position was similar to the mean identification rate. The age of the model had a significant effect on the phoneme accuracy rate of the learned synthetic words in all the syllable positions (Kruskal-Wallis test:  $p < .001$ ). The adult vocal tract model learned more intelligible speech than the two vocal tract models in all phoneme positions (Wilcoxon signed-rank:  $p < .001$ ).

<sup>11</sup> See section 1.3.3.1 for the details of anatomical changes in the vocal tract during development.

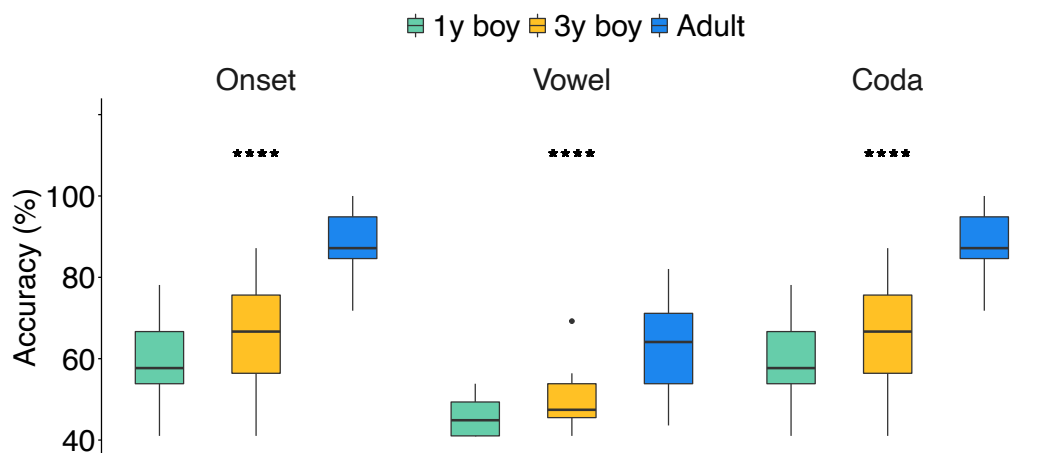


Figure 58 Phoneme accuracy rates of CVC words learned by an adult and two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

Figure 59 shows the phoneme accuracy rate of the regenerated CVCV words learned by the vocal tract models. Kruskal-Wallis test indicated that the anatomical structure of the vocal tract model had a significant effect on the phoneme accuracies of all the phoneme positions ( $p < .001$ ) except for the second vowel position ( $p = 1.000$ ). Wilcoxon signed-rank tests showed that the adult vocal tract had higher accuracies in the first onset position than the 1-year-old model ( $p < .001$ ), but not the 3-year-old model ( $p = .300$ ). The adult model learned more intelligible speech in the first vowel position than the two child vocal tract models (1y:  $p < .001$ , 3y:  $p = .010$ ). The adult model and the 3-year-old model had similar accuracies in the second onset position ( $p = .080$ ). The identification rate of the 1-year-old model was significantly lower than the adult model in the second onset position ( $p < .001$ ). Finally, all the vocal tract models learned intelligible final vowels and the performance was almost undistinguishable ( $p = 1.000$ ).

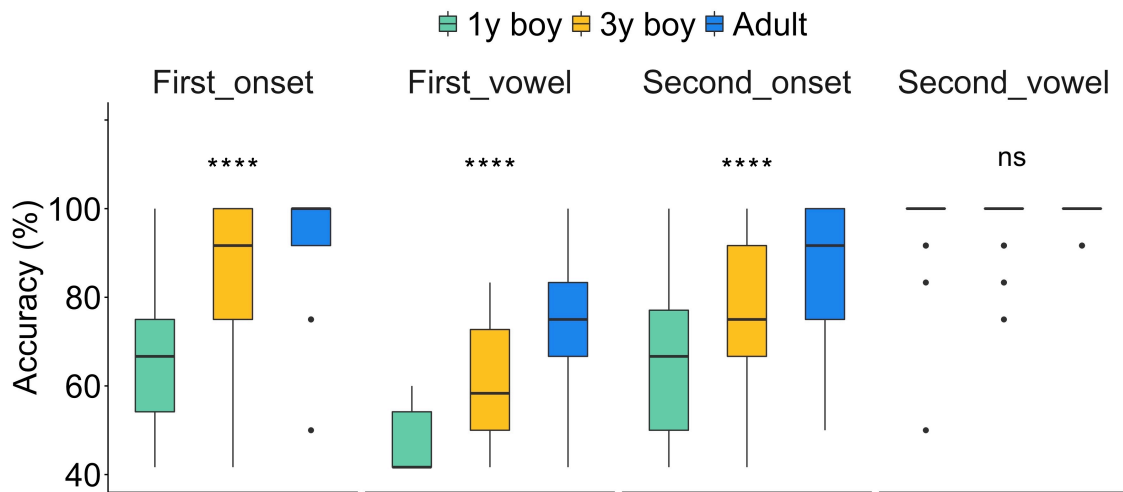


Figure 59 By-listener phoneme accuracy rates of CVCV words learned by an adult and two child vocal tract models in different syllable positions, evaluated by an open-vocabulary transcription experiment. ns  $P > 0.05$ , \*\*\*\*  $P \leq 10^{-4}$ .

The vocal tract models of different ages yielded divergent recognition errors with the same number of iterations, as shown in Figure 60. Consistent with the listening experiments, the models of different ages yielded significantly divergent recognition errors within the same number of iterations (Kruskal-Wallis test:  $p < .001$ ). Wilcoxon signed-rank tests showed that the adult male vocal tract model learned CVC words with lower recognition errors than the child models (1y:  $p < .001$  and 3y:  $p < .001$ ). The 3-year-old model had lower recognition errors compared with the 1-year-old model ( $p = .038$ ).

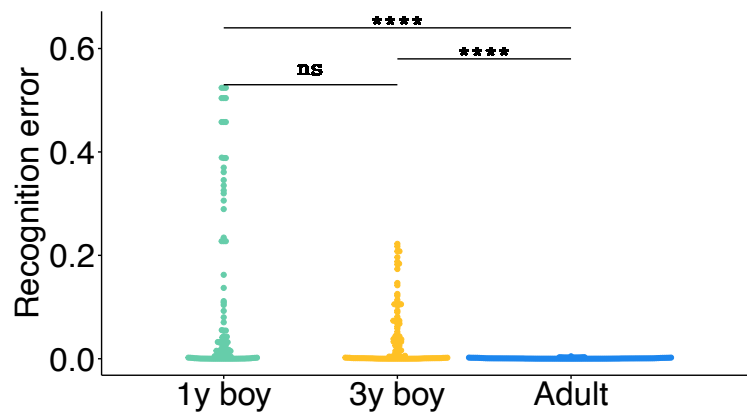


Figure 60 Recognition error distribution of 10-best CVC words evaluated by an automatic phoneme recogniser. ns  $P > 0.05$ , \*\*\*\*  $P \leq 10^{-4}$ .

The phoneme accuracy rate of the CV syllables in the learned CVC words ('bud', 'bod', 'bead', 'dead', 'deed', 'dad') and the regenerated CVCV words ('body', 'buddy', 'Debbie', 'daddy') were compared. The adult models outperformed the child models and that there was a marked facilitation effect of syllable type, as shown in Figure 61. Wilcoxon signed-rank tests showed that the adult vocal tract model had overall higher accuracies than the two child models ( $p < .001$ ), whereas the two child models did not differ significantly ( $p = .190$ ). As the number of syllables increased, word identification became easier for the listeners (Kruskal-Wallis test:  $p < .001$ ), regardless of the age of the model. Wilcoxon signed-rank tests indicated that the identification rate was higher for CVCV words than for CVC words for all the vocal tract models (1y:  $p < .001$ , 3y:  $p < .001$ , Adult:  $p < .001$ ).

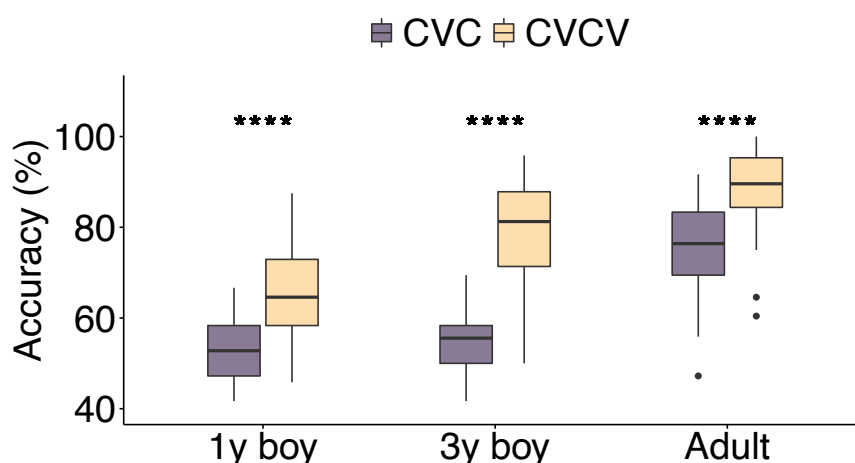


Figure 61 By-listener phoneme accuracy rates of the CV syllables in CVC words and CVCV words learned by adult and child vocal tract models in an open-vocabulary transcription experiment. \*\*\*\*  $P \leq 10^{-4}$ .

To compare the acoustic space of the learned vowels, I selected the best 20 instances per target word based on recognition error evaluated by the speech recogniser. I then compared the first formant (F1) and the second formant (F2) of the learned vowels in bilabial-vowel (bV) sequences by VocalTractLab (Birkholz, 2013), as shown in Figure 62. F1 and F2 of the vowels in the same category forms consistent clusters for the adult vocal tract model, which bears resemblance to the acoustic space of natural speech reported in previous studies (Clopper, Pisoni, & de Jong, 2005; Hagiwara, 1997; Peterson & Barney, 1952). However, there is no clear clustering of vowels for the child models. The acoustic space of the learned high vowels (i.e., /i/ and /u/) of the child vocal tract models heavily overlapped with one another. The high F2 values of vowel /u/ suggested that the learned tongue position was not retracted. The mid front vowels /ɛ/ and /æ/ were not well separated in acoustic space for both child models. The learning of the low back vowel /ɑ/ was also unsuccessful for both child model. To be more specific, the tongue position of the learned /ɑ/ by was not retracted for the 1-year-old model, while /ɑ/ learned by the 3-year-old model was indistinguishable from /ʌ/.

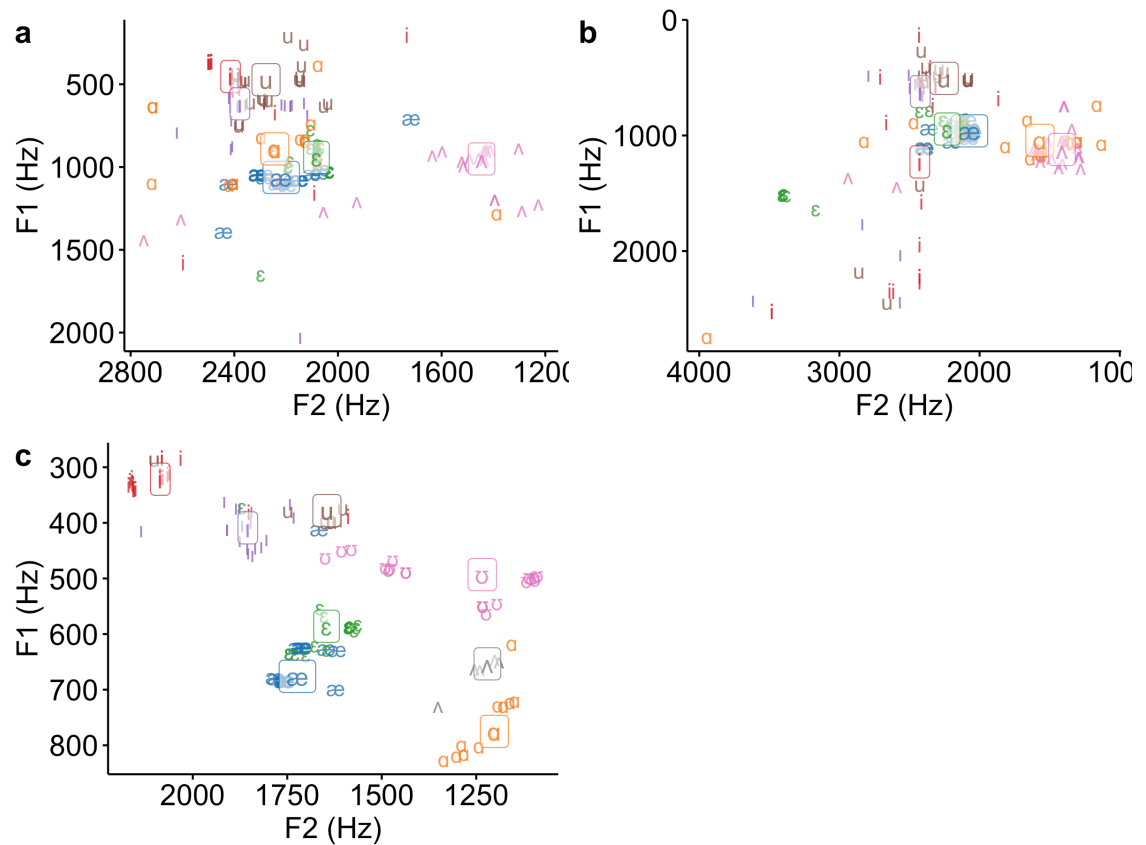


Figure 62 Comparison of vocal tract models of different ages. First formant (F1) and Second formant (F2) of vowels in CVC words with bilabial stops learned by a 1-year-old model (a), a 3-year-old model (b) and an adult male model (c). The squared IPA labels represent the median. The F1 and F2 were based on the vowel spectra calculated by VocalTractLab 2.3 (Birkholz, 2013).

## 5.4 DISCUSSION

Learning to speak requires the acquisition of sophisticated control of both sensory and motor systems. Previous computational models of vocal learning have centred on simulating learning architectures, based on neurobiologically motivated approaches (Kröger et al., 2009; Tourville & Guenther, 2011), acoustic imitation under the distal learning framework (Philippsen et al., 2014; Prom-On et al., 2014a), caregiver’s feedback (Messum & Howard, 2015; Miura et al., 2012), reinforcement learning (Warlaumont & Finnegan, 2016) and goal babbling

(Philippsen, 2021a; Philippsen et al., 2016). However, so far there has been no clear demonstration of successful learning of intelligible words containing CV syllables (see Appendix Table A Performance). In this study, I have demonstrated that vocal learning is achievable with two key components: 1) a speech perception system that encodes phonetic categories, and 2) coordinated articulatory movements that are context-dependent. I explicitly modelled language-specific perception and coarticulatory dynamics, which contrasts with previous attempts that focus more on the learning architecture. The model further shows learning progress that resembles the developmental patterns of child speech acquisition.

At the first glance, it is tremendously difficult for a child to acquire speech given the huge acoustic mismatch between the adult and the child speech (Vorperian & Kent, 2007) due to the large discrepancies between their vocal apparatuses (Fitch & Giedd, 1999), known as the speaker normalisation (K. Johnson, 2005)/correspondence problem (Brass & Heyes, 2005; Nehaniv & Dautenhahn, 2002). This problem has prompted a considerable amount of work on simulating caregiver-infant interactions (Acevedo-Valle et al., 2020; Cohen & Billard, 2018; Huckvale, 2011a; Messum & Howard, 2015; Miura et al., 2012; Rasilo & Räsänen, 2017). In these mirroring interaction paradigms, a child establishes the correspondence between his own motor repertoire and the mature speech by the caregiver's response to his vocalisations. The paradigm is based on the assumption that children are not equipped with the ability to judge the similarities between their own vocalisations and those of the adults (Messum & Howard, 2015). However, there is in fact evidence that children are able to learn speaker-normalised phonetic categories in the presence of variability in the auditory input (Rost & McMurray, 2009). Moreover, if we posit that being imitated by the caregiver is necessary for tackling the speaker normalisation problem, then vocal learning cannot be achieved without it. Yet, in the case of children raised in Gusii rural communities, verbal mother-child interactions are nearly absent but they manage to learn to speak (Lancy, 2014; LeVine, 2004; Mesman et al., 2021).



The fact that children can still learn to speak with limited verbal maternal interaction shows that children are capable of solving the speaker normalisation problem on their own to a great extent. Here, I use computation simulation to explore the possibility that a child can teach herself/himself to learn to speak when guided by auditory feedback. The approach is in line with the learning mechanism suggested for humans (Kuhl, 2000) and songbirds (Brainard & Doupe, 2002; Doupe & Kuhl, 1999b), that the motor manoeuvre can be learned by producing sounds to match the perceptual representations of ambient sounds. Songbirds first form internal song memory and then use the auditory feedback for song evaluation to drive the learning (Funabiki & Konishi, 2003; Keller & Hahnloser, 2009), while infants develop a perception system with a language-specific filter for phonetic categories which later guides production learning (Vihman, 1993; Kuhl, 1998). The absence of auditory feedback can lead to impaired vocalisations in songbirds (Marler & Tamura, 1964) and marmoset monkeys (Roupe et al., 2003). It is of no surprise that, in humans, the vocalisations of children with hearing difficulties are remarkably different from that of normal hearing children (Oller et al., 1985; Oller & Eilers, 1988). The onset of their babbling is delayed, and when it emerges, most of the sounds are visually prominent syllables such as /ma/ and /ba/ (Stoel-Gammon, 1988). It is also worth noting that the language-related genes such as FoxP2 show similar expression patterns in the auditory and motor systems in humans, songbirds (Teramitsu et al., 2004) and marmosets (Kato et al., 2014), who are all vocal learners. The notable resemblance may indicate that auditory-guided vocal learning can be a shared cognitive mechanism across species.

There are, of course, remarkable distinctions in the sound structure of child and adult vocalisations. What is not yet clear is the nature of the auditory representations in human vocal learning. Past models have attempted to use acoustics of individual utterances as the auditory feedback (Howard & Huckvale, 2005; Philippsen et al., 2014; Prom-On et al., 2014a, 2014b). It turned out that vowels can be acquired by simple acoustic matching (Prom-On et al., 2014a, 2014b), while consonant learning has been relatively unsuccessful (Philippsen et

al., 2014). The main obstacle for this method is the lack-of-invariance in the acoustic manifestation of CV coarticulation. It is well-known that while being perceived as the same phoneme, the acoustic signals can vary according to both linguistic context (Liberman et al., 1952, 1954a) and individual speakers (K. Johnson, 2005; K. Johnson & Sjerps, 2021). The present results show, however, that the lack of invariance problem is also solvable, provided that motor dynamics and perceptual guidance are both simulated, as done in the current study. These findings are still compatible with the idea that the vocal learning is guided by auditory representations (Vihman, 1993; Kuhl, 1998), but the nature of auditory representations in guiding vocal learning is not universal listening but language-specific perception.

Having successfully simulated the learning of intelligible English words, we can now evaluate the factors that may influence the learning. I compared the intelligibility of speech learned by vocal tract models of different ages. The two child models differ in terms of the phoneme accuracies of the learned speech. That the 1-year-old model had lower intelligibility than the 3-year-old model could be attributed to the anatomical differences in their vocal apparatuses. The child's vocal tract undergoes huge anatomical changes in the first three years of their life (Kent, 1992; Kent & Murray, 1982) and I have confirmed that the 3-year-old vocal tract model is more advantageous than the 1-year-old model while learning speech. In addition, the results show that although an infant vocal tract is speech ready, the learning performance is compromised compared to the adult vocal tract model. The findings need to be interpreted with caution because the automatic phoneme recognizer was trained without child speech data and thus it may be disadvantageous in evaluating child speech. Further studies should incorporate child data while training the recogniser and assess whether it would improve the child learning performance.

If we compare the model performance with speech development in real life, the difficult cases of speech sounds for the model seem to correspond well with the ones that are normally acquired later in real life. Children acquire corner vowels

(e.g., /i/, /a/, /ɒ/ and /u/) before mid vowels<sup>12</sup>, such as /ɪ/, /e/, /ɛ/ and /ʊ/ (Stoel-Gammon & Pollock, 2008). It has also been found that the adult vocal tract model had extremely high accuracies while learning corner vowels /æ/, /ɒ/ compared with mid vowels and both child models had the highest accuracy rate for vowels /æ/. One of the common confusions in children's production is /ɪ/ being mistaken as /i/ and /ɛ/ (Vihman, 1996). Interestingly, /ɪ/ learned by the child vocal tract models was also frequently mistaken as /i/ and /ɛ/ in the close-set transcription experiment as well. However, the corner vowel /u/, which is supposed to be an easy vowel for children to acquire, has been a difficult case for all the vocal tract models. The model learned unrounded vowels in 'booed', similar to the congenitally blind population (Ménard et al., 2014). It is therefore suspect that visual cues play a vital role in learning rounded vowels (Murakami et al., 2015). Further studies need to be conducted to investigate how visual signals influence the vocal learning process. Another possible reason for the learning difficulty is that the training data for the automatic phoneme recogniser is not optimal. The production of /u/ differs greatly across regional accents for American English native speakers (Clopper et al., 2005; Fridland et al., 2014) and thus the recogniser may not be efficient enough in discriminating the /u/ well from other vowels. This can explain why unlike the other vowels, the acoustic space of the learned /u/ varied to a great extent (Fig. 58c). More work will need to be done to trim the speech data to improve the performance of the recogniser.

With regard to the consonants, it has been reported that among the voiced stops, the production of bilabials occurs before alveolars and velars, which are also fully acquired the earliest (Crowe & McLeod, 2020). Similarly, both the adult and child models had higher identification accuracies for bilabials than for alveolars and velars. Most of the bilabial stops were correctly recognised but alveolar stops were sometimes identified as bilabial stops. The velar stops, on the other hand, were often identified as alveolar stops for vocal tract models of all the ages. The

---

<sup>12</sup> Mid vowels are normally acquired later in life probably because they require more precise control for tongue positions (Xu et al., 2021).

velar stops were extremely difficult for the 1-year-old vocal tract model with a phoneme accuracy rate of less than 10%, compared with around 45% for the 3-year-old vocal tract model. The growing anatomical structure of the child vocal tract and the maturation of muscle control is likely to accelerate the acquisition of velar stops. As far as syllables are concerned, it has been found that the CV combinations with consonants having the place of articulation similar to the following vowel was easier for the model to acquire, in line with the developmental patterns (MacNeilage & Davis, 2000). For example, CV sequences consisting of alveolar consonants followed by high vowels (e.g., 'deed', 'did') had higher identification rates than the other pairs (e.g., 'dad' and 'dead').

Moreover, I have also identified some external factors that impact on the identification accuracy rate. First of all, context information may interfere with phoneme perception. Child synthetic speech was better recognised in the close-set task than the open-vocabulary transcription task, while the same effect was not found for adult speech. Second, identification rates increased when the same articulatory targets for CVC words were used to regenerate CVCV words. As the number of syllables increased, the identification became easier for the listeners, regardless of the age of the model. The facilitation was more evident for the regenerated longer words learned by the two child models. In a word, it suggests that not only lexicon background but also syllable types can make up for the difficulty in perceiving child speech. These findings are consistent with previous studies showing that linguistic context supports word identification (Benichov et al., 2012). Given that the caregivers are very likely to know the limited vocabulary that children can produce, external factors may facilitate the perception to a great extent in daily life.

These findings raise intriguing questions regarding the interplay between speech production and perception through vocal imitation. Previous research on mirror neurons has established that the observation of actions activates the motor system that can perform the same action (Hari et al., 1998). The sensorimotor linkage has been found not only in motor actions such as hand movements (Fadiga et al., 1995; Gallese et al., 1996; Rizzolatti, Fadiga, Gallese, et al., 1996)

but also in speech production (Fadiga et al., 2002b, 2002a; Pulvermuller et al., 2006; K. E. Watkins et al., 2003; Wilson et al., 2004). Despite extensive research on verifying the existence of sensorimotor link, how it is forged remains dimly understood. Some advocates of mirror neurons believe that humans are born with perception-motor interaction and the experience of motor actions only enhances the existing coupling (Lepage & Théoret, 2007). Others argue that the linkage is forged through association of sensory and motor experience, such as self-observation, synchronous actions, and being imitated by social partners (Cook et al., 2014; Heyes, 2001; Keysers & Perrett, 2004). The associative learning theory has not been widely used to account for speech motor learning, as there are very limited visual cues. Establishing the link between speech production and perception have traditionally been challenging so far. Some researchers have resorted to shifting the burden to adults and posits that being imitated by the caregiver supports the establishment of the sensorimotor link (Messum & Howard, 2015). The simulation results instead indicate that it is the self-learning process that helps forge the link between speech perception and speech production. The model has solved the correspondence problem by self-guidance (i.e., self-observation) without any assistance from social partners (Brass & Heyes, 2005; Nehaniv & Dautenhahn, 2002).

As far as sensorimotor learning is concerned, previous computational approaches have modelled speech motor learning based on sensory feedback in a speech development scenario (Guenther, 1994; Kröger et al., 2009, 2014; Tourville & Guenther, 2011), but have seen little success in learning intelligible words. Even though the current model also makes use of auditory and somatosensory feedback for evaluation, it does not rely on either corrective motor movements as in the DIVA model (Guenther, 1994; Tourville & Guenther, 2011) or sensory predictions as in the SFC (Haith & Krakauer, 2013; Shadmehr et al., 2010) and the FACTS models (Parrell et al., 2019; Parrell & Houde, 2019). The high learning performance of the present model suggests that the learning of novel motor repertoires may not require a fully developed sensorimotor link. Still, it has been found that the link is forged gradually throughout speech

development. Toddlers are not capable of adjusting articulation when given altered auditory feedback (MacDonald et al., 2012), while older children and adults can compensate for the changes in auditory signals (Caudrelier et al., 2019; Shiller et al., 2010). So far, I have only modelled the learning progress from canonical babbling to first words (Fig. 32), while subsequent maturation of the sensorimotor systems is yet to be investigated by future simulations. The modelling approach opens a path toward resolving the mystery behind speech production and perception by implementations of the theoretical accounts.

## **Chapter 6    GENERAL DISCUSSION**

To the best of my knowledge, this dissertation reports the first ever successful simulation of vocal learning that is able to generate English words with a biologically plausible articulatory synthesiser. Although numerous models of vocal learning have been proposed previously (Pagliarini et al., 2021), none has modelled the speech motor and the sensory systems rigorously, and there has been only limited success in learning CV syllables (see Appendix Table A Performance). Here, I trained an articulatory synthesiser that emulated adult and child vocal systems with either acoustic features that simulates universal perception (Kuhl, 2000; Werker & Lalonde, 1988), or an automatic phoneme recogniser that simulates language-specific perception (Kuhl, 2000; Werker & Lalonde, 1988). I have demonstrated that perception-guided learning as suggested by the research on both birdsong learning (Brainard & Doupe, 2002; Phan et al., 2006; Zhao et al., 2019) and human vocal learning (Kuhl, 2000) is indeed feasible and probably necessary. The perception-guided vocal learning can resolve the long-standing problem of acoustic mismatch (Vorperian & Kent, 2007) due to anatomical differences between children and adults (Fitch & Giedd, 1999). This may further suggest that it is the vocal learning process that helps forge the link between perception and motor production (Fadiga et al., 2002a; Fadiga et al., 2002b; Liberman & Mattingly, 1985).

The following are the main novel aspects of the present work:

1. The study demonstrates that the hitherto considered unbreakable problem of cross-speaker variations in speech can be resolved by learners using language-specific perception to self-guide their vocal learning. This shows how children can learn to speak without formal instructions. More importantly, it shows a striking parallel to birdsong acquisition, which is also guided by auditory memory.
2. The study shows that the key to a successful learning of syllables is the articulatory dynamics and temporal coordination. Previous works have only been able to simulate the learning of isolated static vowels.
3. It shows a comprehensive use of human listening experiments to evaluate the intelligibility of synthetic speech in direct comparison with natural speech. The adult vocal tract model learned synthetic CVC and CVCV words with a mean accuracy rate of 82%, compared with 94% for natural speech in open-vocabulary transcription experiments. I pioneered the use of natural speech as the benchmark for quantitative evaluation of vocal learning models. This has set a new standard for assessing the quality of future simulation studies.

In chapter 3, I introduced a vocal learning model that can mimic vocal exploration at different ages. The model can be decomposed into two components: 1) a sensory system that provides auditory and somatosensory feedback, and 2) a motor system that controls the articulatory kinematics. The sensory system includes different kinds of sensory feedback including acoustic features, an automatic phoneme recogniser and somatosensory information. The speech motor system controls the coarticulatory dynamics of a state-of-the-art articulatory synthesis. I systematically evaluated the two types of auditory feedback by presenting the learned synthetic speech to native speakers in listening experiments.

In chapter 4, the results show that the recogniser trained speech had higher identification rates than the speech trained by acoustics. The effectiveness of

recogniser-guided exploration suggests that language-specific perception may be the critical link that has been missing from previous modelling work. In contrast, acoustic features, which reflect universal perception, correlated poorly with linguistic perception, and led to unsatisfactory learning outcomes. Moreover, the effectiveness of the motor control model plays a crucial role in vocal exploration, which has enabled the learning of syllables beyond static vowels and the extrapolation into new syllable types.

In chapter 5, I show that both the adult and the child vocal tract models learned intelligible CVC and CVCV words. It has been found that the child's vocal learning showed greater difficulty than the adult and the 1-year-old model had worse performance than the 3-year-old model partly because of the anatomical structure. The speech error patterns of the vocal tract models match empirical data on early speech acquisition. I have therefore demonstrated that perception-guided learning as suggested by research on both birdsong learning and human vocal learning is indeed feasible and probably necessary. The indispensable role of speech perception in human vocal learning highly resembles songbirds and mammals.

This work is far from complete. First of all, the auditory guidance for the child vocal learning model may not be ideal because the automatic phoneme recogniser used to simulate speech perception was trained by a corpus without child speech data. Consequently, the simulated auditory representations may not fully capture the phonetic properties of child speech. Further studies should incorporate child speech data and child-directed speech under even more ecological settings to investigate how auditory exposure and social interactions affect production learning. In addition, previous literature has found that visual input plays an important role in the learning of visually explicit sounds (Murakami et al., 2015). There is therefore abundant room for further progress in incorporating visual information to guide the movement of visible articulators such as the lips and the jaw. In addition, more research is also needed to determine the mechanisms behind speech adaptation (see section 1.3.1). Even though the current model integrates the speech motor control with sensory outcome, it is



unclear whether the model can adjust its articulation when the auditory feedback is altered. Finally, the scope of this study was limited in terms of the developmental stages in speech acquisition. The current model only simulated the learning process from vocal practice to producing intelligible words and thus there is a fruitful area for further work that elucidates early and later developmental stages. For example, it remains unclear whether the coordination of CV syllables is learned during canonical babbling. In future investigations, it is also necessary to explore how children can learn to segment continuous speech and discover the duration of the speech segments. Another natural progression of this work is to simulate the learning of complex syllable structures and connecting words to sentences.

The finding that vocal learning may be guided by auditory experience has implications on the long-standing debate over the link between speech perception and production, as it suggests that the process of vocal learning can help forge this intimate link. Moreover, this study has improved our understanding of the role of speech perception in production learning, which may carry clinical implications for the diagnosis and intervention of language disorders. The model is able to predict speech production development given perceptual ability. For children with hearing aids or cochlear implants, the model prediction of production accuracy can be used as a tool for assessing learning progress and designing individual interventions. All of these are fundamental issues about language, one of the most important attributes that make us human. This computational approach offers new possibilities of investigating shared cognitive mechanisms of vocal ontogeny across species. It raises the possibility that computational models can bridge between behaviour studies and theoretical work. The recent advance in artificial intelligence (AI) has demonstrated impressive achievement without prior knowledge of human intelligence but this work contributes to a growing literature on constructing computational systems to probe questions about human cognition. Therefore, it may also help to bridge the large gap between artificial intelligence and human intelligence.

## APPENDIX

Table A: Summary table of human vocal learning models

Model	Motor control	Synthesiser	Sensory system	Learning strategy	Learning target	Performance
HABLAR (Bailly, 1997)	8 articulator parameters (lip, jaw, tongue, apex positions);  Consonant goal imposed on the vowel goal to simulate coarticulation	Articulatory-to-acoustic model (Beautemps et al., 1996)	F1, F2, F3, F4+ lip area trajectories	Speech Maps: audio-visual-to-articulatory inversion (Abry et al., 1994)	Single vowel, vowel sequences, VCV	NA
(de Boer, 2000)	3 vocal tract parameters: tongue position, tongue height, and lip rounding	Maeda synthesiser (Maeda, 1990)	Bark-scale F1, F2, F3, F4	Self-organization, imitation	Vowels	NA
(Westerman & Miranda, 2002; Westerm	3 glottis parameters + 3 vocal tract parameters	Pipe synthesiser	Auditory map: F1 and F2;  Visual information	Sensorimotor integration, Hebbian connections	Vowels	NA

ann & Miranda, 2004)						
(Heintz et al., 2009)	12 articulatory parameters	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda, 1990)	Vectors derived from F1, F2, F3	Self-organising maps and Hebbian connections	Point vowels /a, i, u/	NA
(Kanda et al., 2009)	7 vocal tract parameters	Maeda synthesiser (Maeda, 1990)	5-dimensional vectors from low-third to low-seventh dimension out of 12-dimensional MFCCs; F0 analysis by STRAIGHT (Kawahara et al., 1999)	Self-organization, recurrent neural network with parametric bias (RNNPB) (Tani, 2002)	Vowels	NA
(Huckvale & Howard, 2005)	9 articulatory parameters	VTALCS (Maeda synthesiser (Maeda, 1990))	F1, F2	Distal supervised learning	Sentences containing vowels and consonants	The sound spectrogram shows that the vowel quality is

						good but the consonant quality is poor
KLAIR (Huckvale, 2011b, 2011a; Huckvale et al., 2009)	6 vocal tract parameters (Huckvale et al., 2009);  8 vocal tract parameters + 4 glottis parameters (Huckvale, 2011b, 2011a)	KLAIR's Synthesiser: infant-sized Maeda synthesiser (Maeda, 1990)	Adult reformulations	Online infant-caregiver interaction (caregiver imitate infant)	Words including vowels and consonants	NA
(Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)	8 articulators	modified Maeda synthesiser (Maeda, 1990)	Auditory feedback: F1, F2, F3;  Somatosensory feedback: 22-dimensional vector	Neurobiological modelling, neural networks	CVC (Guenther, 1994);  VV, CV, CVCV (Guenther et al., 2006)	NA
(Yoshikawa, Asada, et al., 2003;	5 motor controllers	Source-filter model	Formant vector	Caregiver's imitation of infant speech	Four vowels: /a, i, u, e/	NA

Yoshikawa, Koga, et al., 2003)						
(Ishihara et al., 2009; Miura et al., 2007)	6 vocal tract parameters (Miura et al., 2007);  NA (Ishihara et al., 2009);	Source-filter model (Miura et al., 2007);  NA (Ishihara et al., 2009)	Social feedback: F1 and F2 of caregiver's imitation of infant speech	Caregiver's imitation of infant speech	Five vowels: /a, i, u, e, o/ (Miura et al., 2007);  Vowels (Ishihara et al., 2009)	NA
(Miura et al., 2012)	NA	NA	Social feedback: F1 and F2 of caregiver's imitation of infant speech	Auto-mirroring bias (AMB): less imitative caregiver	Five vowels: /a, i, u, e, o/	NA
(Lyon et al., 2012)	NA	eSpeak synthesiser (Aslin et al., 1996)	Auditory feedback: Microsoft SAPI 5.4 (Phoneme recogniser);  Social feedback:	Human-robot interaction	V, CV, VC and CVC	NA

			Teacher's positive/negative feedback			
(Prom-On et al., 2014a, 2014b)	18 vocal tract parameters	VocalTractLab (Birkholz, 2013)	MFCCs	Distal learning, Gradient descent	Thai vowels	Good vowel quality
(Kröger et al., 2009)	270 proto-vocalic states	Articulatory vectors generated by VocalTractLab (Birkholz, 2013)	Auditory feedback: Bark-scale F1, F2, F3;  Somatosensory feedback: Vocal tract state	Neurobiological modelling	V, VC and CV	NA
(Kröger et al., 2014)	Motor plan states	Articulatory vectors generated by VocalTractLab (Birkholz, 2013)	Bark-scaled spectrogram  Somatosensory feedback: Vocal tract state	Self-organising maps and Hebbian connections	50 CV syllables	78% transcription by one phonetician
Elija (Howard & Messum, 2007,	2 vocal tract parameters for young infant, 7 vocal tract parameters for old	VTALCS (modified Maeda synthesiser (Maeda, 1990))	Sensory salience: spectral change and low frequency power	Caregiver's imitation of infant speech	NA (Howard & Messum, 2007)  CV, VC, or	NA (Howard & Messum, 2007, 2011);  Synthetic

2014, 2011; Messum & Howard, 2015)	<p>child + 2 glottis parameters (Howard &amp; Messum, 2007);</p> <p>7 vocal tract parameters + 2 glottis parameters (Howard &amp; Messum, 2014, 2011; Messum &amp; Howard, 2015);</p> <p>Task dynamics model (Fowler &amp; Saltzman, 1993; Saltzman &amp; Munhall, 1989)</p>		<p>(Howard &amp; Messum, 2007);</p> <p>Template-based dynamic time warping (Howard &amp; Messum, 2011);</p> <p>Gammatone spectrogram (Howard &amp; Messum, 2014; Messum &amp; Howard, 2015);</p> <p>Social feedback: Caregiver's reformulation of infant speech</p>		<p>CVV (Howard &amp; Messum, 2011);</p> <p>VV, CV, VC and CVV (Howard &amp; Messum, 2014; Messum &amp; Howard, 2015)</p>	<p>samples are provided. Good vowel quality;</p> <p>Consonants are unintelligible (no trace of consonant burst or frication) (Howard &amp; Messum, 2014; Messum &amp; Howard, 2015);</p>
(Murakami et al., 2015)	16 vocal tract parameters	VocalTractLab (Birkholz, 2013)	Auditory reservoir generated by BRIAN neural network	Reinforcement learning	Vowels	NA

			<p>simulator (Fontaine et al., 2011; Lopez-Poveda &amp; Meddis, 2001);</p> <p>Visual input</p>			
<p>(Warlaumont, 2012; Warlaumont et al., 2013; Warlaumont &amp; Finnegan, 2016)</p>	<p>Jaw and lips (Warlaumont, 2012);</p> <p>Lungs, trachea, larynx, pharynx, oral cavity, and nasal cavity (Warlaumont et al., 2013; Warlaumont &amp; Finnegan, 2016);</p>	<p>Praat synthesis of a female vocal tract</p> <p>Muscle activations controlled by a spiking neural network (Maass, 1997);</p>	<p>Caregiver's judgment as the reward (Warlaumont, 2012);</p> <p>Mel-scale F0, F1 and F2 (Warlaumont et al., 2013)</p> <p>Estimated auditory salience (Coath et al., 2009) (Warlaumont &amp; Finnegan, 2016);</p>	<p>Reinforcement learning</p>	<p>VCV sequences (Warlaumont, 2012);</p> <p>Vowels (Warlaumont et al., 2013);</p> <p>Single consonant and consonant clusters (Warlaumont &amp; Finnegan, 2016)</p>	<p>NA (Warlaumont, 2012);</p> <p>NA (Warlaumont et al., 2013);</p> <p>Synthetic samples are provided. No trace of consonants in the spectrogram (Warlaumont &amp; Finnegan, 2016)</p>



LeVI (Rasilo & Räsänen, 2017)	9 vocal tract parameters	Rasilo's Articulatory model	<p>Auditory feedback: 11 MFCCs without energy (Rasilo et al., 2013);</p> <p>F1, F2 (Rasilo &amp; Räsänen, 2017);</p> <p>Social feedback: Phase1: positive/negative feedback Phase 2: imitation of infants' babbles by caregivers</p>	Caregiver's imitation of infant speech	<p>VCVC sequences containing all 25 Finnish phonemes (Rasilo et al., 2013);</p> <p>CVCV sequences (Rasilo &amp; Räsänen, 2017);</p>	<p>Synthetic samples are provided (unintelligible) LeVI (Rasilo et al., 2013);</p> <p>Synthetic samples are provided (Vowels are not clear; Consonants are unintelligible and no trace of consonant burst/frication) (Rasilo &amp; Räsänen, 2017);</p>
----------------------------------	--------------------------	-----------------------------	--	--	---	--

(Najnin & Banerjee, 2017)	11 vocal tract parameters + 2 glottis parameters	DIVA model (Guenther, 1994a; Guenther et al., 2006a; Tourville & Guenther, 2011)/ modified Maeda synthesier (Maeda, 1990)	F1, F2, F3 + phonation level + 12 MFCCs	Self-organization	NN, CN, NC, VN, NV, VV, CV, VC, CC sequences	NA
(Forestier & Oudeyer, 2017)	7 vocal tract parameters	DIVA model (Guenther, 1994a; Guenther et al., 2006a; Tourville & Guenther, 2011)/ modified Maeda synthesier (Maeda, 1990);  Dynamic Movement	Auditory feedback: F1, F2  Social feedback: Simulated caregiver's guidance through objects  Sensory feedback: state of the environment	Goal-babbling	Vowel sequences including /o, u, i, e, y/	NA

		Primitives (DMPs) (Schaal, 2006) for controlling the articulatory trajectories	including the position of the caregiver, the stick and the toys			
(Cohen & Billard, 2018)	10 words	NA	PerAc (perception/action) architecture (Boucenna et al., 2010; Gaussier & Zrehen, 1995)	Caregiver-infant interaction through objects	CVCV sequences	NA
(Oudeyer, 2005)	3 vocal tract parameters	de Boer's synthesiser (Abstract linear articulatory synthesiser)	Perceptual representations based on Bark-scale F1, F2, F3, F4	Sensory motor interaction	Vowels	NA
(Moulin-Frier & Oudeyer, 2012)	7 vocal tract parameters	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda,	Bark-scale F1, weighted average of F2 and F3	Goal-babbling 1) Random motor exploration 2) Random goal selection with	Five vowels: /a, i, u, e, o/	NA


		1990)		reaching 3) Curiosity-driven active goal selection with reaching		
(Moulin-Frier et al., 2014)	7 parameters based on the PCA of the vocal tract shapes; over-damped spring-mass model for dynamic control	DIVA model (Guenther, 1994a; Guenther et al., 2006a; Tourville & Guenther, 2011)/ modified Maeda synthesier (Maeda, 1990)	Scaled F1, F2, intensity	1) Goal-babbling	VV, VC, CV, CC	NA
(Barnaud et al., 2019; Moulin-Frier et al., 2015)	Jaw, tongue body, tongue dorsum, lip protrusion, tongue tip, lip separation, larynx height	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda,	Bark-scale F1, F2, F3 (Moulin-Frier et al., 2015);  Bark-scale F1, F2 (Barnaud et al., 2019)	Bayesian modelling	VV, VC, CV, CC (Moulin-Frier et al., 2015);  Vowels (Barnaud et	NA


		1990)			al., 2019)	
(Acevedo-Valle et al., 2018)	NA	DIVA model (Guenther, 1994a; Guenther et al., 2006a; Tourville & Guenther, 2011)/ modified Maeda synthesier (Maeda, 1990)	Auditory feedback: F1, F2;  Somatosensory feedback: proprioceptive input	Reinforce learning through auditory and somatosensory feedback GMMs (Gaussian mixture models)	NA (Acevedo-Valle et al., 2018)	NA
(Acevedo-Valle et al., 2017, 2020)	7 vocal tract parameters + 2 glottis parameters (Acevedo-Valle et al., 2018);  10 vocal tract parameters + 3 glottis parameters (Acevedo-Valle et	DIVA model (Guenther, 1994a; Guenther et al., 2006a; Tourville & Guenther, 2011)/ modified Maeda	Auditory feedback: F1, F2;  Somatosensory feedback: proprioceptive input	Caregiver's imitation of infant speech: GMMs (Gaussian mixture models)	NA (Acevedo-Valle et al., 2018);  Vowel sequences containing 17 German vowels	NA


	al., 2020)	synthesier (Maeda, 1990)			(Acevedo- Valle et al., 2020)	
(Philipps en et al., 2014)	22 vocal tract parameters + 4 glottis parameters	VocalTractL ab (Birkholz, 2013)	39 MFCCs (energy, 12MFCCs, first and second derivatives)	Distal supervised learning by acoustic imitation (Echo State Network)	CV sequences containing 8 vowels and 8 consonants	Perceptual evaluation was conducted by the authors
(Philipps en et al., 2016)	20 vocal tract parameters	VocalTractL ab (Birkholz, 2013)	F1, F2, F3 + 39 MFCCs (energy, 12MFCCs, first and second derivatives) projected by Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to 10-D features	Goal babbling (exploration and adaptation)	6 vowels	NA
(Philipps en, 2021b)	18 vocal tract parameters + 3 glottis parameters	VocalTractL ab (Birkholz, 2013)	Echo State Network (ESN) 10-D vectors were based on	Goal babbling	6 vowels, /baa/ and /maa/	Good vowel and consonant quality

		Dynamic Movement Primitives (DMPs) (Schaal, 2006) for controlling the articulatory trajectories	F1, F2, F3 + 39 MFCCs (energy, 12MFCCs, first and second derivatives), then PCA and LDA were applied			
--	--	---	--	--	--	--





Figure A An introduction to the online repository




Evoc Learn   
Project ID: 30808788



Update README.md  
Anqi XU authored 1 month ago

Name	Last commit	Last update
 Code	<a href="#">Update Code/optimization.py</a>	2 months ago
 Demo	<a href="#">Upload New File</a>	1 month ago
 Stimuli	<a href="#">Delete_DS_Store</a>	10 months ago
 README.md	<a href="#">Update README.md</a>	1 month ago

 README.md

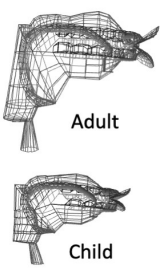
## Evoc-learn

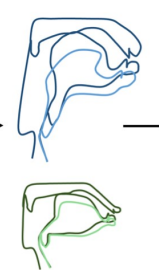
This project contains the code to reproduce the results in:


Anqi Xu, Daniel R. van Niekirk, Branislav Gerazov, Paul Konstantin Krug, Peter Birkholz, Santitham Prom-on, Lorna F. Halliday, Yi Xu. A computational model for human vocal learning.

The workflow of the vocal learning model is illustrated in the following plot:

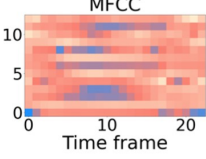
**Motor**

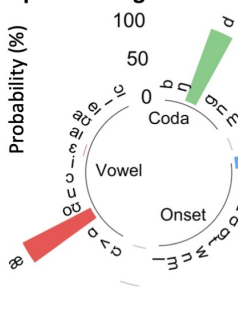
**A Vocal exploration**  
  
3D vocal tract

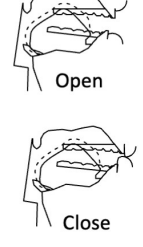
**B Kinematics**  
  
Coarticulation

**C Synthesis**  
  
Speech waveform

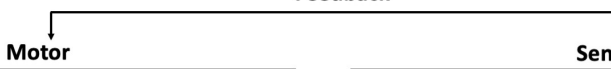
**Sensory**

**D Acoustic features**  
MFCC  
  
Time frame

**E Speech recogniser**  
Probability (%)  
  
Coda  
Vowel  
Onset

**F Somatosensory information**  
  
Open  
Close

**Feedback**



[Demonstration video](#) introduces the structure of the vocal learning model and shows the English words learned by the adult and the child vocal tract models with 3D articulatory movements.

[Stimuli](#) contains the synthetic samples used in the perception experiments.

[Code](#) contains the code to reproduce the results, please follow the instructions.



### Step 1: Installation

Create virtual environment to install the dependencies (recommended):

For Linux & Unix (Mac M1 is not supported at the moment)

```
python3 -m venv env
source env/bin/activate
pip install evoclearn-rec
```

For Windows

```
py -m venv env
.\env\Scripts\activate
pip install evoclearn-rec
```

### Step 2: Set up

Select and copy an initialisation file from [Initialization files](#)

Paste it in [Code](#)

\*See [Instructions](#) for further details of the initialization files

### Step 3: Run

For Linux & Unix

```
python3 optimization.py
```

For Windows

```
py optimization.py
```

The process will generate three synthetic samples with the lowest error. The auditory feedback errors of all the accepted trails will be listed in '\_costs.csv'. The learned articulatory parameters of accepted trails will be listed in '\_VTP1.csv'(consonant), '\_VTP2.csv'(vowel), '\_VTP\_taus.csv'(time constants of the consonant and the vowel). The learned glottis parameters of accepted trails will be listed in '\_GLP1.csv'(consonant), '\_GLP2.csv'(vowel), '\_GLP\_taus.csv'(time constants of the consonant and the vowel).

### Step 4: Exit


```
deactivate
```

Figure B Demographic and language background questionnaire on Gorilla

This is a questionnaire regarding your language background.

Age

Gender

Please Select... 


Country of origin (e.g. the US)

Country of residence (e.g. the US)

What is your native language or languages? (e.g. English)

If you have more than one native languages, please indicate the age of acquisition (e.g. English: 0 year old)

Have you ever had a speech or hearing impairment?

Please Select... 

Please specify (Skip this question if your answer was no)

## REFERENCES

- Abe, M., Schambra, H., Wassermann, E. M., Luckenbaugh, D., Schweighofer, N., & Cohen, L. G. (2011). Reward Improves Long-Term Retention of a Motor Memory through Induction of Offline Memory Gains. *Current Biology*, 21(7), 557–562. <https://doi.org/10.1016/j.cub.2011.02.030>
- Acevedo-Valle, J. M., Angulo, C., & Moulin-Frier, C. (2018). Autonomous Discovery of Motor Constraints in an Intrinsically Motivated Vocal Learner. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2), 314–325. <https://doi.org/10.1109/TCDS.2017.2699578>
- Acevedo-Valle, J. M., Hafner, V. v., & Angulo, C. (2017). Social reinforcement in intrinsically motivated sensorimotor exploration for embodied agents with constraint awareness. *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 255–262. <https://doi.org/10.1109/DEVLRN.2017.8329815>
- Acevedo-Valle, J. M., Hafner, V. v., & Angulo, C. (2020). Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2), 198–208. <https://doi.org/10.1109/TCDS.2018.2883249>
- Anumanchipalli, G. K., Chartier, J. & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 493–498 <https://doi.org/10.1038/s41586-019-1119-1>
- Asada, M. (2016). Modeling early vocal development through infant-caregiver interaction: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 8(2), 128–138. <https://doi.org/10.1109/TCDS.2016.2552493>

- Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22(2–3), 251–267. [https://doi.org/10.1016/S0167-6393\(97\)00025-3](https://doi.org/10.1016/S0167-6393(97)00025-3)
- Barnaud, M. lou, Schwartz, J. L., Bessière, P., & Diard, J. (2019). Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. *PLoS ONE*, 14(1), e0210302. <https://doi.org/10.1371/journal.pone.0210302>
- Barry, W. A., & van Dommelen, W. A. (2005). *The integration of phonetic knowledge in speech technology* (W. J. Barry & W. A. van Dommelen, Eds.; Vol. 25. Springer Netherlands. <https://doi.org/10.1007/1-4020-2637-4>
- Bateman, N. (2007). *A Crosslinguistic Investigation of Palatalization* [University of California, San Diego]. <https://escholarship.org/uc/item/13s331md>
- Benichov, J., Cox, L. C., Tun, P. A., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability, and cognitive function. *Ear and Hearing*, 33(2), 250–256. <https://doi.org/10.1097/AUD.0b013e31822f680f>
- Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2), 775–794. <https://doi.org/10.1121/1.1332378>
- Best, Catherine. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. C. Nusbaum (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words* (pp. 167–224). MIT Press.
- Best, Catherine. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

- Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4), e60603. <https://doi.org/10.1371/journal.pone.0060603>
- Birkholz, P. (2014). Enhanced area functions for noise source modeling in the vocal tract. *Proc. of the 10th International Seminar on Speech Production (ISSP 2014)*, 37–40.
- Birkholz, P., & Kröger, B. J. (2007). Simulation of vocal tract growth for articulatory speech synthesis. *Proc. of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, 377–380.
- Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech and Language Processing*, 19(5), 1422–1433. <https://doi.org/10.1109/TASL.2010.2091632>
- Bladon, R. A. W., & Al-Bamerni, A. (1976). Coarticulation resistance in English // *Journal of Phonetics*, 4(2), 137–150. [https://doi.org/10.1016/S0095-4470\(19\)31234-3](https://doi.org/10.1016/S0095-4470(19)31234-3)
- Boughman, J. W. (1998). Vocal learning by greater spear-nosed bats. *Proceedings: Biological Sciences*, 265, 227–233.
- Brainard, M. S., & Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, 1, 31–40. <https://doi.org/doi.org/10.1038/35036205>
- Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. *Nature*, 417, 351–358. <https://doi.org/doi.org/10.1038/417351a>
- Brancazio, L., & Fowler, C. A. (1998). On the relevance of locus equations for production and perception of stop consonants. *Perception & Psychophysics*, 60(1), 24–50. <https://doi.org/doi.org/10.3758/BF03211916>

- Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences*, 9(10), 489–495. <https://doi.org/10.1016/j.tics.2005.08.007>
- Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(2), 201–251. <https://doi.org/10.1017/S0952675700001019>
- Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, 49, 155–180. <https://doi.org/10.1159/000261913>
- Bruderer, A. G., Kyle Danielson, D., Kandhadai, P., Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, 112(44), 13531–13536
- Buhr, R. D. (1980). The emergence of vowels in an infant. *Journal of Speech and Hearing Research*, 23(1), 73–94. <https://doi.org/10.1044/jshr.2301.73>
- Butefisch, C. M., Davis, B. C., Wise, S. P., Sawaki, L., Kopylev, L., Classen, J., & Cohen, L. G. (2000). Mechanisms of use-dependent plasticity in the human motor cortex. *Proceedings of the National Academy of Sciences*, 97(7), 3661–3665. <https://doi.org/10.1073/pnas.97.7.3661>
- Carnegie Mellon University. (2022). *The CMU Pronouncing Dictionary*. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- Carré, R., & Chennoukh, S. (1995). Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. *Journal of Phonetics*, 7(3), 231–241. [https://doi.org/10.1016/S0095-4470\(95\)80045-X](https://doi.org/10.1016/S0095-4470(95)80045-X)
- Catchpole, C. K., & Slater, P. J. B. (2008). *Bird song: biological themes and variations* (2nd ed.). Cambridge University Press.
- Caudrelier, T., Ménard, L., Perrier, P., Schwartz, J. L., Gerber, S., Vidou, C., & Rochet-Capellan, A. (2019). Transfer of sensorimotor learning reveals

- phoneme representations in preliterate children. *Cognition*, 192, 103973.  
<https://doi.org/10.1016/j.cognition.2019.05.010>
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13, 1428–1432.  
<https://doi.org/doi.org/10.1038/nn.2641>
- Chennoukh, S., Carré, R., & Lindblom, B. (1997). Locus equations in the light of articulatory modeling. *The Journal of the Acoustical Society of America*, 102(4), 2380–2389. <https://doi.org/10.1121/1.419622>
- Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., & Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience*, 1(5), 351–353.
- Chládková, K., & Paillereau, N. (2020). The What and When of Universal Perception: A Review of Early Speech Sound Acquisition. *Language Learning*, 70, 1136–1182
- Choi, D., Dehaene-Lambertz, G., Peña, M., Werker, J. F. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proceedings of the National Academy of Sciences*, 118
- Classen, J., Liepert, J., Wise, S. P., Hallett, M., & Cohen, L. G. (1998). Rapid plasticity of human cortical movement representation induced by practice. *Journal of Neurophysiology*, 79(2), 1117–1123.  
<https://doi.org/10.1152/jn.1998.79.2.1117>
- Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, 118(3), 1661–1676.  
<https://doi.org/10.1121/1.2000774>

- Cohen, L., & Billard, A. (2018). Social babbling: The emergence of symbolic gestures and words. *Neural Networks*, 106, 194–204. <https://doi.org/10.1016/j.neunet.2018.06.016>
- Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, 37(2), 177–192. <https://doi.org/10.1017/S0140525X13000903>
- Cooper, W. E. (1974). Adaptation of phonetic feature analyzers for place of articulation. *Journal of the Acoustical Society of America*, 56(2), 617–627. <https://doi.org/10.1121/1.1903300>
- Crelin, E. S. (1973). *Functional Anatomy of the Newborn*. Yale University Press.
- Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the united states: A review. *American Journal of Speech-Language Pathology*, 29(4), 2155–2165. [https://doi.org/10.1044/2020\\_AJSLP-19-00168](https://doi.org/10.1044/2020_AJSLP-19-00168)
- Davis, B. L., & Macneilage, P. F. (1995). The Articulatory Basis of Babbling. *Journal of Speech and Hearing Research*, 38, 1199-1211. <https://doi.org/10.1044/jshr.3806.1199>
- Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>
- de Boer, Bart. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28, 441–465. <https://doi.org/10.006/jpho.2000.0125>
- de Klerk, C. C. J. M., Johnson, M. H., Heyes, C. M., & Southgate, V. (2015). Baby steps: Investigating the development of perceptual-motor couplings in infancy. *Developmental Science*, 18(2), 270–280. <https://doi.org/10.1111/desc.12226>



- Diedrichsen, J., White, O., Newman, D., & Lally, N. (2010). Use-dependent and error-based learning of motor behaviors. *Journal of Neuroscience*, 30(15), 5159–5166. <https://doi.org/10.1523/JNEUROSCI.5406-09.2010>
- Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, 1(2), 121–144. [https://doi.org/doi.org/10.1207/s15326969eco0102\\_2](https://doi.org/doi.org/10.1207/s15326969eco0102_2)
- Doupe, A. J., & Kuhl, P. K. (1998). Birdsong and human speech: Common Themes and Mechanisms. *Annual Review of Neuroscience*, 22, 567–631. <https://doi.org/10.1146/annurev.neuro.22.1.567>
- Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, 22, 567–631. <https://doi.org/10.1146/annurev.neuro.22.1.567>
- Eby, T. L., & Nadol, J. B. (1986). Postnatal Growth of the Human Temporal Bone. *Annals of Otology, Rhinology & Laryngology*, 95(4 Pt 1), 356–364. <https://doi.org/10.1177/000348948609500407>
- Eckel, H. E., Georg, M. S., Koebke, J., Pototschnig, C., Sittel, C., & Stennert, E. (1999). Morphology of the human larynx during the first five years of life studied on whole organ serial. *The Annals of Otology, Rhinology & Laryngology*, 108(3), 232–238. <https://doi.org/10.1177/000348949910800303>
- Egnor, S. E. R., & Hauser, M. D. (2004). A paradox in the evolution of primate vocal learning. *Trends in Neurosciences*, 27(11), 649–654. <https://doi.org/10.1016/j.tins.2004.08.009>
- Eilers, R. E., & Minifie, F. D. (1975). Fricative Discrimination in Early Infancy. *Journal of Speech and Hearing Research*, 18(1), 158–167. <https://doi.org/10.1044/jshr.1801.158>

- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, 16(3), 513–521. <https://doi.org/10.3758/BF03198580>
- Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics*, 18(5), 341–347. <https://doi.org/10.3758/BF03211210>
- Eimas, P. D., & Miller, J. L. (1980). Discrimination of information for manner of articulation. *Infant Behavior and Development*, 3, 367–375. [https://doi.org/10.1016/S0163-6383\(80\)80044-0](https://doi.org/10.1016/S0163-6383(80)80044-0)
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech Perception in Infants. *Science, New Series* 171(3968), 303–306.
- Elowson, A. M., Snowdon, C. T., & Lazaro-Perea, C. (1998a). Infant “Babbling” in a Nonhuman Primate: Complex Vocal Sequences with Repeated Call Types. *Behaviour*, 135(5), 643–664. [www.jstor.org/stable/4535550](http://www.jstor.org/stable/4535550)
- Elowson, A. Margaret., Snowdon, Charles. T., & Lazaro-Perea, C. (1998b). ‘Babbling’ and social context in infant monkeys: parallels to human infants. *Trends in Cognitive Sciences*, 2(1), 31–37. [https://doi.org/10.1016/s1364-6613\(97\)01115-7](https://doi.org/10.1016/s1364-6613(97)01115-7)
- Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002a). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, 15(2), 399–402. <https://doi.org/10.1046/j.0953-816x.2001.01874.x>
- Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002b). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, 15(2), 399–402. <https://doi.org/10.1046/j.0953-816x.2001.01874.x>
- Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of*

*Neurophysiology*, 73(6), 2608–2611.  
<https://doi.org/10.1152/jn.1995.73.6.2608>

Fernandez, A. A., Burchardt, L. S., Nagy, M., & Knörnschild, M. (2021). Babbling in a vocal learning bat resembles human infant babbling. *Science*, 373, 923–926.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Science*, 4(7), 258–267. [https://doi.org/10.1016/s1364-6613\(00\)01494-7](https://doi.org/10.1016/s1364-6613(00)01494-7)

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3), 1511–1522. <https://doi.org/10.1121/1.427148>

Flanagan, J. R., & Wing, A. M. (1997). The role of internal models in motion planning and control: Evidence from grip force adjustments during movements of hand-held loads. *The Journal of Neuroscience*, 17(4), 1519–1528. <https://doi.org/10.1523/JNEUROSCI.17-04-01519.1997>

Fontaine, B., Goodman, D. F. M., Benichoux, V., & Brette, R. (2011). Brian Hears: Online Auditory Processing Using Vectorization Over Channels. *Frontiers in Neuroinformatics*, 5(9). <https://doi.org/10.3389/fninf.2011.00009>

Forestier, S., & Oudeyer, P.-Y. (2017). A unified model of speech and tool use early development. *39th Annual Conference of the Cognitive Science Society (CogSci 2017)*. <https://hal.archives-ouvertes.fr/hal-01583301>

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech, Language, and Hearing Research*, 24(1), 127–139. <https://doi.org/10.1044/jshr.2401.127>

- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, 14(1), 3–28. [https://doi.org/10.1016/S0095-4470\(19\)30607-2](https://doi.org/10.1016/S0095-4470(19)30607-2)
- Fowler, C. A. (1989). Real objects of speech perception: A Commentary on Diehl and Kluender. *Ecological Psychology*, 1(2), 145–160. [https://doi.org/10.1207/s15326969eco0102\\_3](https://doi.org/10.1207/s15326969eco0102_3)
- Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36(2–3), 171–195. <https://doi.org/doi.org/10.1177/002383099303600304>
- Fowler, Carol. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetic*, 8, 113–133. [https://doi.org/doi.org/10.1016/S0095-4470\(19\)31446-9](https://doi.org/doi.org/10.1016/S0095-4470(19)31446-9)
- Fridland, V., Kendall, T., & Farrington, C. (2014). Durational and spectral differences in American English vowels: Dialect variation within and across regions. *The Journal of the Acoustical Society of America*, 136(1), 341–349. <https://doi.org/10.1121/1.4883599>
- Fripp, D., Owen, C., Quintana-Rizzo, E., Shapiro, A., Buckstaff, K., Jankowski, K., Wells, R., & Tyack, P. (2005). Bottlenose dolphin (*Tursiops truncatus*) calves appear to model their signature whistles on the signature whistles of community members. *Animal Cognition*, 8(1), 17–26. <https://doi.org/10.1007/s10071-004-0225-z>
- Funabiki, Y., & Konishi, M. (2003). Long memory in song learning by zebra finches. *The Journal of Neuroscience*, 23(17), 6928–6935. <https://doi.org/10.1523/JNEUROSCI.23-17-06928.2003>
- Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3), 361–377. <https://doi.org/10.3758/bf03193857>

- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609. <https://doi.org/doi.org/10.1093/brain/119.2.593>
- Gerazov, B. van Niekerk, D. R., Xu. A., Krug, P. K., Birkholz, P., Xu, Y. (2020). Evaluating Features and Metrics for High-Quality Simulation of Early Vocal Learning of Vowels. (Preprint)
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.
- Giudice, M. del, Manera, V., & Keysers, C. (2009). Programmed to learn? the ontogeny of mirror neurons. *Developmental Science*, 12(2), 350–363. <https://doi.org/10.1111/j.1467-7687.2008.00783.x>
- Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children* [Massachusetts Institute of Technology]. <https://dspace.mit.edu/handle/1721.1/22386>
- Gros-Louis, J., West, J. M., Goldstein, H. M., & King, P. A. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30(6), 509–516. <https://doi.org/10.1177/0165025406071914>
- Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent Speech production. *Biological Cybernetics*, 72, 43–53. <https://doi.org/doi.org/10.1007/BF00206237>
- Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3), 280–301. <https://doi.org/10.1016/j.bandl.2005.06.001>
- Gultekin, Y. B., & Hage, S. R. (2018). Limiting parental interaction during vocal development affects acoustic call structure in marmoset monkeys. *Science Advances*, 4(4), eaar4012. <https://doi.org/10.1126/sciadv.aar4012>

- Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America*, 102(1), 655–658. <https://doi.org/10.1121/1.419712>
- Haith, A. M., & Krakauer, J. W. (2013). Model-Based and Model-Free Mechanisms of Human Motor Learning. *Advances in Experimental Medicine and Biology*, 782, 1–21. [https://doi.org/10.1007/978-1-4614-5465-6\\_1](https://doi.org/10.1007/978-1-4614-5465-6_1)
- Hare, G. (1983). Development at two years. In J. v. Irwin & S. P. Wong (Eds.), *Phonological development in children: 18 to 72 months* (pp. 55–85). Southern Illinois Press.
- Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: A neuromagnetic study. *Proceedings of the National Academy of Sciences*, 95(25), 15061–15065. <https://doi.org/10.1073/pnas.95.25.15061>
- Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*, Cambridge University Press, New York.
- Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12, *Journal of Phonetics*, 28(4), 377–396, <https://doi.org/10.1006/jpho.2000.0121>.
- Heintz, I., Beckman, M., Fosler-Lussier, E., & Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 688–691. <https://doi.org/10.21437/interspeech.2009-238>
- Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, 5(6), 253–261. [https://doi.org/10.1016/S1364-6613\(00\)01661-2](https://doi.org/10.1016/S1364-6613(00)01661-2)
- Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization, *Neuron* 69(3), 407–422. <https://doi.org/10.1016/j.neuron.2011.01.019>

- Hollander, M., Wolfe, D. A., & Chicken, E. (2014). Nonparametric statistical methods. 3rd Edition, John Wiley & Sons, Inc. New York.
- Hosp, J. A., Pekanovic, A., Rioult-Pedotti, M. S., & Luft, A. R. (2011). Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. *Journal of Neuroscience*, 31(7), 2481–2487. <https://doi.org/10.1523/JNEUROSCI.5411-10.2011>
- Howard, I. S., & Huckvale, M. A. (2005). Training a vocal tract synthesizer to imitate speech using distal supervised learning. *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)*, 159–162.
- Howard, I. S., & Messum, P. (2007). A Computational Model of Infant Speech Development. *Proceedings of XII International Conference “Speech and Computer” (SPECOM’2007)*, 756–765.
- Howard, Ian. S., & Messum, P. (2011). Modeling the Development of Pronunciation in Infant Speech Acquisition. *Motor Control*, 15(1), 85-117. <https://doi.org/10.1123/mcj.15.1.85>
- Howard, Ian. S., & Messum, P. (2014). Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant. *PLoS ONE*, 9(10), e110334. <https://doi.org/10.1371/journal.pone.0110334>
- Huckvale, M. (2011a). Recording caregiver interactions for machine acquisition of spoken language using the KLAIR virtual infant. *Proceedings of Interspeech 2011*.
- Huckvale, M. (2011b). The KLAIR toolkit for recording interactive dialogues with a virtual infant. *Proceedings of Interspeech 2011*, 28–31.
- Huckvale, M., Howard, I. S., & Fagel, S. (2009). KLAIR: a Virtual Infant for Spoken Language Acquisition Research. *Proceedings of Interspeech 2009*.

- Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical Movement Primitives: Learning attractor models for motor behaviors. *Neural Computation*, 25(2), 328–373. [https://doi.org/10.1162/NECO\\_a\\_00393](https://doi.org/10.1162/NECO_a_00393)
- Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Neuroreport*, 17(10), 957–962. <https://doi.org/10.1097/01.wnr.0000223387.51704.89>
- Ishihara, H., Yoshikawa, Y., Miura, K., & Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *IEEE Transactions on Autonomous Mental Development*, 1(4), 217–225. <https://doi.org/10.1109/TAMD.2009.2038988>
- Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America*, 134(2), 1271–1282. <https://doi.org/10.1121/1.4812855>
- Iskarous, K., Nam, H., & Whalen, D. H. (2010). Perception of articulatory dynamics from acoustic signatures. *The Journal of the Acoustical Society of America*, 127(6), 3717–3728. <https://doi.org/10.1121/1.3409485>
- Jackson, P. J. B., & Singampalli, V. D. (2009). Statistical identification of articulation constraints in the production of speech. *Speech Communication*, 51(8), 695–710. <https://doi.org/10.1016/j.specom.2009.03.007>
- Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. *GMD Report - German National Research Institute for Computer Science*, 148.



- Jaeger, H., & Haas, H. (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667), 78–80. <https://doi.org/10.1126/science.1091277>
- Janik, V. M., & Knörnschild, M. (2021). Vocal production learning in mammals revisited. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1836). <https://doi.org/10.1098/rstb.2020.0244>
- Janik, V. M., & Sayigh, L. S. (2013). Communication in bottlenose dolphins: 50 years of signature whistle research. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 199(6), 479–489. <https://doi.org/10.1007/s00359-013-0817-7>
- Janik, V. M., Sayigh, L. S., & Wells, R. S. (2006). Signature whistle shape conveys identity information to bottlenose dolphins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21), 8293–8297. <https://doi.org/10.1073/pnas.0509918103>
- Janik, V. M., & Slater, P. J. B. (1997). Vocal Learning in Mammals. In P. J. B. Slater, J. S. Rosenblatt, C. T. Snowdon, & M. Milinski (Eds.), *Advances in the Study of Behavior* (Vol. 26). Academic Press. [https://doi.org/10.1016/S0065-3454\(08\)60377-0](https://doi.org/10.1016/S0065-3454(08)60377-0)
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60–99. [https://doi.org/doi.org/10.1016/0010-0285\(89\)90003-0](https://doi.org/doi.org/10.1016/0010-0285(89)90003-0)
- Johnson, K. (2005). Speaker normalization in speech perception. In D. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Blackwell.
- Johnson, K., & Sjerps, M. J. (2021). Speaker normalization in speech perception. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The*

*handbook of speech perception* (pp. 145–176). Wiley.  
<https://doi.org/10.1002/9781119184096.ch6>

Jordan, M. I., & Rumelhart, D. E. (1992). Forward Models: Supervised Learning with a Distal Teacher. *Cognitive Science*, 16(3), 307–354.  
[https://doi.org/doi.org/10.1016/0364-0213\(92\)90036-T](https://doi.org/doi.org/10.1016/0364-0213(92)90036-T)

Kaelbling, P. L., Littman, M. L., Moore, A. W., & Hall, S. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237–285.  
<https://doi.org/doi.org/10.1613/jair.301>

Kahane, J. C. (1978). A Morphological Study of the Human Prepubertal and Pubertal Larynx. *The American Journal of Anatomy*, 151(1), 11–19.  
<https://doi.org/doi.org/10.1002/aja.1001510103>

Kahane, J. C. (1982). Growth of the human prepubertal and pubertal larynx. *Journal of Speech and Hearing Research*, 25(3), 446–455.  
<https://doi.org/10.1044/jshr.2503.446>

Kanda, H., Ogata, T., Takahashi, T., Komatani, K., & Okuno, H. G. (2009). Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. *Proceedings - IEEE International Conference on Robotics and Automation*, 4438–4443. <https://doi.org/10.1109/ROBOT.2009.5152818>

Kauffman, S. A. (1992). The origins of order: Self-organization and selection in evolution. In F. J. Varela & JP. Dupuy (Eds.), *Understanding Origins. Boston Studies in the Philosophy and History of Science* (Vol. 130, pp. 153–181). Springer. [https://doi.org/doi.org/10.1007/978-94-015-8054-0\\_8](https://doi.org/doi.org/10.1007/978-94-015-8054-0_8)

Kawato, M. (1990). Feedback-Error-Learning Neural Network for Supervised Motor Learning. In R. Eckmiller (Ed.), *Advanced Neural Computers* (pp. 365–372). Elsevier. <https://doi.org/10.1016/B978-0-444-88400-8.50047-9>

Kawato, M., Furuka, K., & Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, 57, 169–185. <https://doi.org/doi.org/10.1007/BF00364149>

- Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology* (pp. 451–470). Cambridge University Press. <https://doi.org/10.1017/CBO9780511627736.026>
- Keating, P., & Buhr, R. (1978). Fundamental frequency in the speech of infants and children. *Journal of the Acoustical Society of America*, 63(2), 567–571. <https://doi.org/10.1121/1.381755>
- Keefe, D. H., Bulen, J. C., Arehart, K. H., & Burns, E. M. (1993). Ear-canal impedance and reflection coefficient in human infants and adults. *The Journal of the Acoustical Society of America*, 94(5), 2617–2638. <https://doi.org/10.1121/1.407347>
- Keller, G. B., & Hahnloser, R. H. R. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature*, 457(7226), 187–190. <https://doi.org/10.1038/nature07467>
- Kent, R. D. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications* (pp. 65–90). York Press.
- Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America*, 72(2), 353–365. <https://doi.org/10.1121/1.388089>
- Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences*, 8(11), 501–507. <https://doi.org/10.1016/j.tics.2004.09.005>
- Kilner, J. M., & Lemon, R. N. (2013). What We Know Currently about Mirror Neurons. *Current Biology*, 23(23), R1057–R1062. <https://doi.org/10.1016/j.cub.2013.10.051>

- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680. <https://doi.org/10.1126/science.220.4598.671>
- Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese Quail Can Learn Phonetic Categories. *Science*, 237(4819), 1195–1197. <https://doi.org/10.1126/science.3629235>
- Knörnschild, M., Behr, O., & von Helversen, O. (2006). Babbling behavior in the sac-winged bat (*Saccopteryx bilineata*). *Naturwissenschaften*, 93(9), 451–454. <https://doi.org/10.1007/s00114-006-0127-9>
- Knörnschild, M., Nagy, M., Metz, M., Mayer, F., & von Helversen, O. (2010). Complex vocal imitation during ontogeny in a bat. *Biology Letters*, 6(2), 156–159. <https://doi.org/10.1098/rsbl.2009.0685>
- Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons. *Science*, 297(5582), 846–848. <https://doi.org/10.1126/science.1070311>
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59–69. <https://doi.org/10.1007/BF00337288>
- Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift Für Tierpsychologie*, 22(7), 770–783. <https://doi.org/10.1111/j.1439-0310.1965.tb01688.x>
- Kozhevnikov, V. A., & Chistovich, L. A. (1965). *Speech: Articulation and perception* (Vol. 30). Joint Publications Research Service.
- Krakauer, J. W., & Mazzoni, P. (2011). Human sensorimotor learning: adaptation, skill, and beyond. *Current Opinion in Neurobiology*, 21(4), 636–644. <https://doi.org/10.1016/j.conb.2011.06.012>

- Krakauer, J. W., Pine, Z. M., Ghilardi, M.-F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *The Journal of Neuroscience*, 20(23), 8916–8924. <https://doi.org/10.1523/JNEUROSCI.20-23-08916.2000>
- Kröger, B. J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, 2(2). <http://www.epjnonlinearbiomedphys.com/content/2/1/2>
- Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9), 793–809. <https://doi.org/10.1016/j.specom.2008.08.002>
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93–107.
- Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: implications for theories of developmental speech perception. *Journal of Phonetics*, 21(1–2), 125–139. [https://doi.org/doi.org/10.1016/S0095-4470\(19\)31326-9](https://doi.org/doi.org/10.1016/S0095-4470(19)31326-9)
- Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, 4(6), 812–822. [https://doi.org/10.1016/0959-4388\(94\)90128-7](https://doi.org/10.1016/0959-4388(94)90128-7)
- Kuhl, P. K. (2000). A new view of language acquisition. *PNAS*, 97(22), 11850–11857. <https://doi.org/doi.org/10.1073/pnas.97.22.11850>
- Kuhl, P. K. (2003). Human speech and birdsong: Communication and the social brain. *PNAS*, 100(17), 9645–9646. [www.pnas.org/cgi/doi/10.1073/pnas.1733998100](http://www.pnas.org/cgi/doi/10.1073/pnas.1733998100)

- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843. <https://doi.org/10.1038/nrn1533>
- Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 979–1000. <https://doi.org/10.1098/rstb.2007.2154>
- Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100(4 Pt 1), 2425–2438. <https://doi.org/10.1121/1.417951>
- Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, 63(3), 905–917. <https://doi.org/10.1121/1.381770>
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2), F13–F21. <https://doi.org/10.1111/j.1467-7687.2006.00468.x>
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606–608. <https://doi.org/10.1126/science.1736364>
- Kuhnert, B., & Nolan, F. (1999). The origin of coarticulation. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, Data and Techniques (Cambridge Studies in Speech Science and Communication)* (pp. 7–30). Cambridge University Press. <https://doi.org/doi:10.1017/CBO9780511486395.002>

- Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00203>
- Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience*, 32(27), 9351–9358. <https://doi.org/10.1523/JNEUROSCI.0404-12.2012>
- Lancy, D. F. (2014). *The Anthropology of Childhood: Cherubs, Chattel, Changelings* (2nd ed). Cambridge University Press. <https://doi.org/10.1017/CBO9781139680530>
- Lane, H., Matthies, M. L., Guenther, F. H., Denny, M., Perkell, J. S., Stockmann, E., Tiede, M., Vick, J., & Zandipour, M. (2007). Effects of Short- and Long-Term Changes in Auditory Feedback on Vowel and Sibilant Contrasts. *Journal of Speech, Language, and Hearing Research*, 50(4), 913–927. [https://doi.org/10.1044/1092-4388\(2007/065\)](https://doi.org/10.1044/1092-4388(2007/065))
- Lecanuet, J. P., Granier-Deferre, C., DeCasper, A. J., Maugeais, R., Andrieu, A. J., & Busnel, M. C. (1987). Perception et discrimination foetales de stimuli langagiers; mise en évidence à partir de la réactivité cardiaque; résultats préliminaires [Fetal perception and discrimination of speech stimuli; demonstration by cardiac reactivity; preliminary results]. *Comptes Rendus de l'Academie Des Sciences. Serie III, Sciences de La Vie*, 305(5), 161–164.
- Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3), 1455–1468. <https://doi.org/10.1121/1.426686>
- Lenneberg, E. (1967). Biological foundations of language. *Hospital Practice*, 2(12), 59–67. <https://doi.org/10.1080/21548331.1967.11707799>

- Lepage, J. F., & Théoret, H. (2007). The mirror neuron system: Grasping others' actions from birth? *Developmental Science*, 10(5), 513–523. <https://doi.org/10.1111/j.1467-7687.2007.00631.x>
- LeVine, R. A. (2004). Challenging expert knowledge: Findings from an African study of infant care and development. In U. P. Gielen & J. L. Roopnarine (Eds.), *Childhood and adolescence: Cross-cultural perspectives and applications* (pp. 149–165). Praeger Publishers/Greenwood Publishing Group.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461. <https://doi.org/10.1037/h0020279>
- Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*. 54 (5), 358–368. <https://doi:10.1037/h0044417>
- Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954a). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8), 1–13. <https://doi.org/10.1037/h0093673>
- Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954b). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8), 1–13. <https://doi.org/10.1037/h0093673>
- Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 65(4), 497. <https://doi.org/10.2307/1418032>



- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36. [https://doi.org/doi.org/10.1016/0010-0277\(85\)90021-6](https://doi.org/doi.org/10.1016/0010-0277(85)90021-6)
- Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic Ability and Related Anatomy of the Newborn and Adult Human, Neanderthal Man, and the Chimpanzee. *American Anthropologist*, 74(3), 287–307.
- Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35(5), 783–783. <https://doi.org/10.1121/1.2142410>
- Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *The Journal of the Acoustical Society of America*, 99(3), 1683–1692. <https://doi.org/10.1121/1.414691>
- Lindblom, B., MacNeilage, P. F., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. *Linguistics*, 21(1), 181–204. <https://doi.org/10.1515/ling.1983.21.1.181>
- Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, 40(1), 1–19. <https://doi.org/10.1016/j.wocn.2011.09.005>
- Litovsky, R. (2015). Development of the auditory system. In *Handbook of Clinical Neurology* (Vol. 129, pp. 55–72). Elsevier B.V. <https://doi.org/10.1016/B978-0-444-62630-1.00003-2>
- Liu, Z., & Xu, Y. (2021). Segmental Alignment of English Syllables with Singleton and Cluster Onsets. *Interspeech 2021*, 3969–3973. <https://doi.org/10.21437/Interspeech.2021-187>
- Liu, Zirui., Xu, Yi., & Hsieh, F. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics*, 90, 101116. [doi.org/10.1016/j.wocn.2021.101116](https://doi.org/10.1016/j.wocn.2021.101116)

- Locher, H., Frijns, J. H., van Iperen, L., de Groot, J. C., Huisman, M. A., & Chuva de Sousa Lopes, S. M. (2013). Neurosensory development and cell fate determination in the human cochlea. *Neural Development*, 8(20). <https://doi.org/10.1186/1749-8104-8-20>
- Löfqvist, A. (1999). Interarticulator phasing, locus equations, and degree of coarticulation. *The Journal of the Acoustical Society of America*, 106(4), 2022–2030. <https://doi.org/doi.org/10.1121/1.427948>
- Lopez-Poveda, E. A., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, 110(6), 3107–3118. <https://doi.org/10.1121/1.1416197>
- Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, 60(4), 602–619. <https://doi.org/10.3758/BF03206049>
- Luft, A. R., & Schwarz, S. (2009). Dopaminergic signals in primary motor cortex. *International Journal of Developmental Neuroscience*, 27(5), 415–421. <https://doi.org/10.1016/j.ijdevneu.2009.05.004>
- Lyon, C., Nehaniv, C. L., & Saunders, J. (2012). Interactive language learning by robots: The transition from babbling to word forms. *PLoS ONE*, 7(6), e38236. <https://doi.org/10.1371/journal.pone.0038236>
- MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology*, 22(2), 113–117. <https://doi.org/10.1016/j.cub.2011.11.052>
- MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, 288(5465), 527–531.
- Maeda, S. (1990). Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model.

- In *Speech Production and Speech Modelling: Vol. NATO ASI Series*, 55 (pp. 131–149). Springer. [https://doi.org/10.1007/978-94-009-2037-8\\_6](https://doi.org/10.1007/978-94-009-2037-8_6)
- Makino, H., Hwang, E. J., Hedrick, N. G., & Komiyama, T. (2016). Circuit Mechanisms of Sensorimotor Learning. *Neuron*, 92(4), 705–721. <https://doi.org/10.1016/j.neuron.2016.10.029>
- Marler, P. (1970). Birdsong and speech development: could there be parallels? *American Scientist*, 58(6), 669–673. <http://www.jstor.org/stable/27829317>
- Marler, P., & Peters (1982). Developmental overproduction and selective attrition: new processes in the epigenesis of birdsong. *Developmental psychobiology*, 15(4), 369–378. <https://doi.org/10.1002/dev.420150409>
- Marler, P., & Tamura, M. (1964). Culturally transmitted patterns of vocal behavior in sparrows. *Science*, 11(3650), 1483–1486. <https://doi.org/10.1126/science.146.3650.1483>
- Marlow, C. D., Winkelmann, R. K., & Gibilisco, J. A. (1965). General sensory innervation of the human tongue. *The Anatomical Record*, 152, 503–511. <https://doi.org/doi.org/10.1002/ar.1091520410>
- Mcleod, S., & Crowe, K. (2018). Children's Consonant Acquisition in 27 Languages: A Cross-Linguistic Review. *American Journal of Speech-Language Pathology* 27, 1546-1571. <https://doi.org/10.23641/asha>
- Mcleod, S., van Doorn, J., & Reed, V. A. (2001). Normal Acquisition of Consonant Clusters. *American Journal of Speech-Language Pathology*, 10(2), 99-110.
- Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research*, 57(3), 793–804. [https://doi.org/10.1044/2014\\_JSLHR-S-12-0395](https://doi.org/10.1044/2014_JSLHR-S-12-0395)
- Menzel, E. W. Jr. (1964). Patterns of Responsiveness in Chimpanzees Reared Through Infancy Under Conditions of Environmental Restriction.

*Psychologische Forschung*, 27, 337–365.  
<https://doi.org/doi.org/10.1007/BF00421336>

Menzerath, P., & de Lacerda, A. (1933). Koartikulation, Steuerung und Lautabgrenzung: eine experimentelle Untersuchung. *Phonetische Studien*, 1.

Mesman, J., Basweti, N., & Misati, J. (2021). Sensitive infant caregiving among the rural Gusii in Kenya. *Attachment and Human Development*, 23(2), 124–133. <https://doi.org/10.1080/14616734.2020.1828512>

Messum, P., & Howard, Ian. S. (2015). Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics*, 53, 125–140. <https://doi.org/10.1016/j.wocn.2015.08.005>

Miksis, J. L., Tyack, P. L., & Buck, J. R. (2002). Captive dolphins, *Tursiops truncatus*, develop signature whistles that match acoustic features of human-made model sounds. *The Journal of the Acoustical Society of America*, 112(2), 728–739. <https://doi.org/10.1121/1.1496079>

Miura, K., Yoshikawa, Y., & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics*, 21(13), 1583–1600. <https://doi.org/10.1163/156855307782148596>

Miura, K., Yoshikawa, Y., & Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Advanced Robotics*, 26(1–2), 23–44. <https://doi.org/10.1163/016918611X607347>

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., & Jenkins, J. J. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, 18, 331–340. <https://doi.org/10.3758/BF03211209>

- Moulin-Frier, C., Diard, J., Schwartz, J. L., & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*, 53, 5–41. <https://doi.org/10.1016/j.wocn.2015.06.001>
- Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P. Y. (2014). Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*, 4, 1006. <https://doi.org/10.3389/fpsyg.2013.01006>
- Moulin-Frier, C., & Oudeyer, P.-Y. (2012). Curiosity-driven phonetic learning. *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 1–8. <https://doi.org/10.1109/DevLrn.2012.6400583>
- Murakami, M., Kröger, B., Birkholz, P., & Triesch, J. (2015). Seeing [u] aids vocal learning: babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. *Proceedings of 5th International Conference on Development and Learning and on Epigenetic Robotics*, 208–213. [https://doi.org/10.0/Linux-x86\\_64](https://doi.org/10.0/Linux-x86_64)
- Murayama, T., Iijima, S., Katsumata, H., & Arai, K. (2014). Vocal imitation of human speech, synthetic sounds and beluga Sounds, by a beluga (delphinapterus leucas). *International Journal of Comparative Psychology*, 27(3), 369–384. <https://escholarship.org/uc/item/51v1z12b>
- Najnin, S., & Banerjee, B. (2017). A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Communication*, 92, 24–41. <https://doi.org/10.1016/j.specom.2017.05.002>
- Nehaniv, C. L., & Dautenhahn, K. (2002). The correspondence problem. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in animals and artifacts* (pp. 41–61). Boston Review.

- Ohala, J. J. (1986). Against the direct realist view of speech perception. *Journal of Phonetics*, 14, 75–82.
- Ohala, J. J., Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219–252. <https://doi.org/10.1017/s0952675700000658>
- Öhman, S. E. G. (1967). Numerical Model of Coarticulation. *The Journal of the Acoustical Society of America*, 41(2), 310–320. <https://doi.org/10.1121/1.1910340>
- Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child Phonology* (pp. 93–112). Elsevier. <https://doi.org/10.1016/B978-0-12-770601-6.50011-5>
- Oller, D. K., & Eilers, E. R. (1988). The role of audition in infant babbling. *Child Development*, 59(2), 441–449.
- Oller, D. K., Eilers, R. E., Bull, D. H., & Carney, A. E. (1985). Prespeech vocalizations of a deaf infant. *Journal of Speech, Language, and Hearing Research*, 28(1), 47–63. <https://doi.org/10.1044/jshr.2801.47>
- Osberger, M. J., & McGarr, N. (1982). Speech production characteristics of the hearing impaired. In N. Lass (Eds.), *Speech and language: Advances in basic research and practice* (Vol. 8, pp. 221–283). New York: Academic Press.
- Oudeyer, P. Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3), 435–449. <https://doi.org/10.1016/j.jtbi.2004.10.025>
- Owren, M. J., Dieter, J. A., Seyfarth, R. M., & Cheney, D. L. (1992). “Food” Calls Produced by Adult Female Rhesus (*Macaca mulatta*) and Japanese (*M. fuscata*) Macaques, Their Normally-Raised Offspring, and Offspring Cross-

*Fostered between Species.* 120(3–4), 218–231.  
<https://doi.org/doi.org/10.1163/156853992X00615>

Pagliarini, S., Leblois, A., & Hinaut, X. (2021). Vocal Imitation in Sensorimotor Learning Models: A Comparative Review. *IEEE Transactions on Cognitive and Developmental Systems*, 13(2), 326–342.  
<https://doi.org/10.1109/TCDS.2020.3041179>

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Parrell, B., & Houde, J. (2019). Modeling the role of sensory feedback in speech motor control and learning. *Journal of Speech, Language, and Hearing Research*, 62(8S), 2963–2985. [https://doi.org/10.1044/2019\\_JSLHR-S-CSMC7-18-0127](https://doi.org/10.1044/2019_JSLHR-S-CSMC7-18-0127)

Parrish, A. (2022). *Pronouncing* (0.2.0).  
<https://pronouncing.readthedocs.io/en/latest/tutorial.html>.  
<https://pronouncing.readthedocs.org>

Paschall, L. (1983). Development at 18 months. In J. v. Irwin & S. P. Wong (Eds.), *Phonological Development in Children: 18 to 72 months*. Southern Illinois University Press.

Pelaez, M., Virués-Ortega, J., & Gewirtz, J. L. (2011). Contingent and Noncontingent Reinforcement With Maternal Vocal Imitation and Motherese Speech: Effects on Infant Vocalizations. *European Journal of Behavior Analysis*, 12(1), 277–287. <https://doi.org/10.1080/15021149.2011.11434370>

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton University Press. <http://www.jstor.org/stable/j.ctt7ztt6j>

Perkell, J. S., Denny, M., Lane, H., Guenther, F., Matthies, M. L., Tiede, M., Vick, J., Zandipour, M., & Burton, E. (2007). Effects of masking noise on vowel

and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users. *The Journal of the Acoustical Society of America*, 121(1), 505–518. <https://doi.org/10.1121/1.2384848>

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, 24, 175–184. <https://doi.org/doi.org/10.1121/1.1906875>

Phan, M. L., Pytte, C. L., & Vicario, D. S. (2006). Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. *PNAS*, 103(4), 1088–1093. <https://doi.org/doi.org/10.1073/pnas.0510136103>

Philippsen, A. K. (2021). Goal-Directed Exploration for Learning Vowels and Syllables: A Computational Model of Speech Acquisition. *KI - Kunstliche Intelligenz*, 35(1), 53–70. <https://doi.org/10.1007/s13218-021-00704-y>

Philippsen, A. K., Reinhart, R. F., & Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. *IEEE ICDL-EPIROB 2014 - 4th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 195–200. <https://doi.org/10.1109/DEVLRN.2014.6982981>

Philippsen, A. K., Reinhart, R. F., & Wrede, B. (2016). Goal Babbling of Acoustic-Articulatory Models with Adaptive Exploration Noise. *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 72–78. [https://doi.org/10.0/Linux-x86\\_64](https://doi.org/10.0/Linux-x86_64)

Pistorio, A. L., Vintch, B., & Wang, X. (2006). Acoustic analysis of vocal development in a New World primate, the common marmoset (*Callithrix jacchus*). *The Journal of the Acoustical Society of America*, 120(3), 1655–1670. <https://doi.org/10.1121/1.2225899>



- Polka, L., & Bohn, O. S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, 39(4), 467–478. <https://doi.org/10.1016/j.wocn.2010.08.007>
- Polka, L., & Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41, 221–231. [https://doi.org/10.1016/S0167-6393\(02\)00105-X](https://doi.org/10.1016/S0167-6393(02)00105-X)
- Polka, L., & Werker, Janet. F. (1994). Developmental Changes in Perception of Nonnative Vowel Contrasts. *Experimental Psychology: Human Perception and Performance*, 20(2), 421–435. <https://doi.org/10.1037//0096-1523.20.2.421>
- Pollock, K. E., & Berni, M. C. (2003). Incidence of non-rhotic vowel errors in children: Data from the memphis vowel project. *Clinical Linguistics and Phonetics*, 17(4–5), 393–401. <https://doi.org/10.1080/0269920031000079949>
- Prather, J. F., Nowicki, S., Anderson, R. C., Peters, S., & Mooney, R. (2009). Neural correlates of categorical perception in learned vocal communication. *Nature Neuroscience*, 12(2), 221–228. <https://doi.org/doi.org/10.1038/nn.2246>
- Prom-On, S., Birkholz, P., & Xu, Y. (2014a). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *Eurasip Journal on Audio, Speech, and Music Processing*, 2014, 23. <https://doi.org/10.1186/1687-4722-2014-23>
- Prom-On, S., Birkholz, P., & Xu, Y. (2014b). Estimating vocal tract shapes of Thai vowels from contextual vowel variation. *2014 17th Oriental Chapter of the International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, 1–6. <https://doi.org/10.1109/ICSDA.2014.7051442>

- Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20), 7865–7870. <https://doi.org/10.1073/pnas.0509989103>
- Ralls, K., Fiorelli, P., & Gish, S. (1985). Vocalizations and vocal mimicry in captive harbor seals, *Phoca vitulina*. *Canadian Journal of Zoology*, 63(5), 1050–1056. <https://doi.org/10.1139/z85-157>
- Rasilo, H., & Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication*, 86, 1–23. <https://doi.org/10.1016/j.specom.2016.10.010>
- Rasilo, H., Räsänen, O., & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, 55(9), 909–931. <https://doi.org/10.1016/j.specom.2013.05.002>
- Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *The Journal of the Acoustical Society of America*, 76(6), 1624–1635. <https://doi.org/doi.org/10.1121/1.391609>
- Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the Acoustical Society of America*, 125(4), 2288–2298. <https://doi.org/10.1121/1.3089222>
- Richards, D. G., Wolz, J. P., Herman, L. M., Bauer, G., Hoban, E., Jaeckel, D., Ketchum, J., & Mobley, M. (1984). Vocal mimicry of computer-generated sounds and vocal labeling of objects by a bottlenosed dolphin, *Tursiops truncatus*. *Journal of Comparative Psychology*, 98(1), 10–28. <https://doi.org/doi.org/10.1037/0735-7036.98.1.10>

- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), 131–141. [https://doi.org/10.1016/0926-6410\(95\)00038-0](https://doi.org/10.1016/0926-6410(95)00038-0)
- Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996). Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental Brain Research*, 111(2), 246–252. <https://doi.org/10.1007/BF00227301>
- Rogers, D. J., Setlur, J., Raol, N., Maurer, R., & Hartnick, C. J. (2014). Evaluation of true vocal fold growth as a function of age. *Otolaryngology - Head and Neck Surgery*, 151(4), 681–686. <https://doi.org/10.1177/0194599814547489>
- Rolf, M., Steil, J. J., & Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, 2(3), 216–229. <https://doi.org/10.1109/TAMD.2010.2062511>
- Roupe, S. L., Pistorio, A., & Wang, X. (2003). Vocal plasticity induced by auditory deprivation in the common marmoset. *Society for Neuroscience*.
- Ryant, N., Slaney, M., Liberman, M., Shriberg, E., & Yuan, J. (2014). Highly accurate mandarin tone classification in the absence of pitch information, *Proceedings of Speech Prosody*, 7.
- Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, 1(4), 333–382. [https://doi.org/10.1207/s15326969eco0104\\_2](https://doi.org/10.1207/s15326969eco0104_2)
- Sander, E. K. (1972). When are speech sounds learned? *The Journal of Speech and Hearing Disorders*, 37(1), 55–63. <https://doi.org/10.1044/jshd.3701.55>
- Sato, K., Hirano, M., & Nakashima, T. (2001). Fine structure of the human newborn and infant vocal fold mucosae. *The Annals of Otology, Rhinology & Laryngology*, 110(5), 417–424.

- Savin. (1963). Word-Frequency Effect and Errors in the Perception of Speech. *The Journal of the Acoustical Society of America*, 35(2), 200–206. <https://doi.org/10.1121/1.1918432>
- Schaal, S. (2006). Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics. In H. Kimura, K. Tsuchiya, A. Ishiguro, & H. Witte (Eds.), *Adaptive Motion of Animals and Machines* (pp. 261–280). Springer. [https://doi.org/10.1007/4-431-31381-8\\_23](https://doi.org/10.1007/4-431-31381-8_23)
- Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics*, 20, 213–223. <https://doi.org/10.1017/S0267190500200135>
- Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error Correction, Sensory Prediction, and Adaptation in Motor Control. *Annual Review of Neuroscience*, 33(1), 89–108. <https://doi.org/10.1146/annurev-neuro-060909-153135>
- Shiller, D. M., Gracco, V. L., & Rvachew, S. (2010). Auditory-motor learning during speech production in 9- 11-year-old children. *PLoS ONE*, 5(9). <https://doi.org/10.1371/journal.pone.0012975>
- Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, 10, 2465. <https://doi.org/10.1038/s41467-019-10365-z>
- Stansbury, A. L., & Janik, V. M. (2019). Formant Modification through Vocal Production Learning in Gray Seals. *Current Biology*, 29(13), 2244-2249.E4. <https://doi.org/10.1016/j.cub.2019.05.071>
- Stansbury, A. L., & Janik, V. M. (2021). The role of vocal learning in call acquisition of wild grey seal pups. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1836). <https://doi.org/10.1098/rstb.2020.0251>

- Stark, R. E. (1986). Vocal communication in the first 18 months of life. *Journal of Speech and Hearing Research*, 36(3), 548–558. <https://doi.org/doi.org/10.1044/jshr.3603.548>
- Stark, R. E., Bernstein, L. E., & Demorest, M. E. (1993). Vocal Communication in the First 18 Months of Life. *Journal of Speech and Hearing Research*, 36(3), 548–558. <https://doi.org/10.1044/jshr.3603.548>
- Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3), 185–190. <https://doi.org/10.1121/1.1915893>
- Stoel-Gammon, C. (1983). Constraints on consonant–vowel sequences in early words. *Journal of Child Language*, 10(2), 455–457. <https://doi.org/10.1017/S0305000900007893>
- Stoel-Gammon, C. (1985). Phonetic Inventories, 15–24 Months. *Journal of Speech, Language, and Hearing Research*, 28(4), 505–512. <https://doi.org/10.1044/jshr.2804.505>
- Stoel-Gammon, C. (1988). Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: A comparison of consonantal inventories. *Journal of Speech and Hearing Disorders*, 53(3), 302–315. <https://doi.org/10.1044/jshd.5303.302>
- Stoel-Gammon, C., & Pollock, K. (2008). Vowel Development and Disorders. In M. J. Ball, M. R. Perkins, & Müller Nicole (Eds.), *The Handbook of Clinical Linguistics* (pp. 525–548). Blackwell Publishing Ltd. <https://doi.org/10.1002/9781444301007.ch33>
- Story, B. H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation. *The Journal of the Acoustical Society of America*, 117(5), 3231–3254. <https://doi.org/10.1121/1.1869752>

- Story, B. H. (2009). Vowel and consonant contributions to vocal tract shape. *The Journal of the Acoustical Society of America*, 126(2), 825–836. <https://doi.org/10.1121/1.3158816>
- Strange, W., & Jenkins, J. J. (1978). Role of Linguistic Experience in the Perception of Speech. In R. D. Walk & H. L. Pick (Eds.), *Perception and Experience* (Vol. 1). Springer US. [https://doi.org/10.1007/978-1-4684-2619-9\\_5](https://doi.org/10.1007/978-1-4684-2619-9_5)
- Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature*, 259, 39–41. <https://doi.org/doi.org/10.1038/259039a0>
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. The MIT Press.
- Tabain, M. (2000). Coarticulation in CV syllables: a comparison of Locus Equation and EPG data. *Journal of Phonetics*, 28, 137–159. <https://doi.org/10.006/jpho.2000.0110>
- Tabain, M. (2002). Voiceless Consonants and Locus Equations: A Comparison with Electropalatographic Data on Coarticulation. *Phonetica*, 59(1), 20–37. <https://doi.org/10.1159/000056203>
- Takahashi, D. Y., Fenley, A. R., Teramoto, Y., Narayanan, D. Z., Borjon, J. I., Holmes, P., & Ghazanfar, A. A. (2015). The developmental dynamics of marmoset monkey vocal production. *Science*, 349, 734–738. <https://doi.org/10.1126/science.aaa7945>
- Takahashi, D. Y., Liao, D. A., & Ghazanfar, A. A. (2017). Vocal Learning via Social Reinforcement by Infant Marmoset Monkeys. *Current Biology*, 27(12), 1844–1852.e6. <https://doi.org/10.1016/j.cub.2017.05.004>
- Tani, J. (2002). Self-organization of behavioral primitives as multiple attractor dynamics: a robot experiment. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, 489–494. <https://doi.org/10.1109/IJCNN.2002.1005521>

- Tatman, R. (2017). English Word Frequency. *Kaggle*.  
<https://www.kaggle.com/datasets/rtatman/english-word-frequency?resource=download>
- ter Haar, S. M., Fernandez, A. A., Gratier, M., Knörnschild, M., Levelt, C., Moore, R. K., Vellema, M., Wang, X., & Oller, D. K. (2021). Cross-species parallels in babbling: animals and algorithms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1836), 20200239.  
<https://doi.org/10.1098/rstb.2020.0239>
- Thelen, E. (1981). Rhythmical Behavior in Infancy: An Ethological Perspective. *Developmental Psychology*, 17(3), 237–257.
- Thomas, E. R. (2001). *An acoustic analysis of vowel variation in New World English*. Duke University Press.
- Thorpe, D. W. H. (1954). The process of song learning in the chaffinch as studied by means of the sound spectrograph. *Nature*, 173, 465–469.  
<https://doi.org/doi.org/10.1038/173465a0>
- Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7), 952–981. <https://doi.org/10.1080/01690960903498424>
- Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America*, 120(4), 2285–2294.  
<https://doi.org/10.1121/1.2338290>
- Tyack, P. L. (1997). Development and social functions of signature whistles in bottlenose dolphins tursiops truncatus. *Bioacoustics*, 8(1–2), 21–46.  
<https://doi.org/10.1080/09524622.1997.9753352>
- Tyack, P. L., & Sayigh, L. S. (1997). Vocal learning in cetaceans. In C. T. Snowdon & M. Hausberger (Eds.), *Social influences on vocal development* (pp. 208–233). Cambridge University Press.

- Tyler, A. A., & Langsdale, T. E. (1996). Consonant-vowel interactions in early phonological development. *First Language*, 16(47), 159–191. <https://doi.org/10.1177/014272379601604702>
- van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P. & Xu. Y. (2022). Exploration strategies for articulatory synthesis of complex syllable onsets. <https://doi.org/10.48550/arXiv.2204.09381>
- van Elk, M., van Schie, H. T., Hunnius, S., Vesper, C., & Bekkering, H. (2008). You'll never crawl alone: Neurophysiological evidence for experience-dependent motor resonance in infancy. *NeuroImage*, 43(4), 808–814. <https://doi.org/10.1016/j.neuroimage.2008.07.057>
- Verstynen, T., & Sabes, P. N. (2011). How each movement changes the next: An experimental and theoretical study of fast adaptive priors in reaching. *Journal of Neuroscience*, 31(27), 10050–10059. <https://doi.org/10.1523/JNEUROSCI.6525-10.2011>
- Vihman, M. M. (1992). Early syllables and the construction of phonology. In C. A. Ferguson, L. Menn, & C. Stoel-Gamon (Eds.), *Phonological Development: Models, Research, Implications* (pp. 393–422). York Press.
- Vihman, M. M. (1996). *Phonological development: the origins of language in the child* [Book]. Blackwell.
- Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language*, 61(2), 397–445. <https://doi.org/10.2307/414151>
- Vihman, Marilyn. May. (2014). *Phonological Development: The First Two Years* (2nd ed.). Wiley-Blackwell.
- Vorperian, H. K., & Kent, R. D. (2007). Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data. *Journal of Speech, Language, and Hearing Research*, 50(6), 1510–1545. [https://doi.org/10.1044/1092-4388\(2007/104\)](https://doi.org/10.1044/1092-4388(2007/104))



- Vorperian, H. K., Kent, R. D., Gentry, L. R., & Yandell, B. S. (1999). Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. In *International Journal of Pediatric Otorhinolaryngology* (Vol. 49).
- Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 117(1), 338–350. <https://doi.org/10.1121/1.1835958>
- Warlaumont, A. S. (2012). A spiking neural network model of canonical babbling development. *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL 2012*, 1–6. <https://doi.org/10.1109/DevLrn.2012.6400842>
- Warlaumont, A. S., & Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLoS ONE*, 11(1), e0145096. <https://doi.org/10.1371/journal.pone.0145096>
- Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38, 64–75. <https://doi.org/10.1016/j.neunet.2012.11.012>
- Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, 90(6), 2942–2955. <https://doi.org/10.1121/1.401769>
- Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8), 989–994. [https://doi.org/10.1016/S0028-3932\(02\)00316-0](https://doi.org/10.1016/S0028-3932(02)00316-0)
- Watts, D., & Strogatz, S. (1998). Collective dynamics of “small-world” networks. *Nature*, 393, 440–442. <https://doi.org/10.1038/30918>

- Wellman, B. L., Case, I. M., Mengert, I. G., & Bradbury, D. E. (1931). *Speech sounds of young children* (2nd ed., Vol. 5). University of Iowa Press.
- Werker, J. F., & Lalonde, C. E. (1988). Cross-Language Speech Perception: Initial Capabilities and Developmental Change. *Developmental Psychology*, 24(5), 672–683. <https://doi.org/10.1037/0012-1649.24.5.672>
- Werker, J. F., & Tees, R. C. (1984). Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life. *Infant Behaviour and Development*, 7, 49–63.
- Werker, J. F., & Tees, R. C. (1992). The Organization and Reorganization of Human Speech Perception. *Annual Review of Neuroscience*, 15(1), 377–402. <https://doi.org/10.1146/annurev.ne.15.030192.002113>
- Westerman, G., & Miranda, E. R. (2002). Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research*, 31(4), 367–375. <https://doi.org/10.1076/jnmr.31.4.367.14166>
- Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, 89(2), 393–400. [https://doi.org/10.1016/S0093-934X\(03\)00345-6](https://doi.org/10.1016/S0093-934X(03)00345-6)
- Willshaw, D. (2006). Self-organization in the Nervous System. In R. Morris, L. Tarassenko, & M. Kenward (Eds.), *Cognitive Systems - Information Processing Meets Brain Science* (pp. 5–33). Academic Press. <https://doi.org/10.1016/B978-012088566-4/50004-0>
- Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7), 701–702. <https://doi.org/10.1038/nn1263>
- Winter, P. (1969). The variability of peep and twit calls in captive squirrel monkeys (*Saimiri sciureus*). *Folia Primatologica*, 10, 204–215. <https://doi.org/10.1159/000155200>

- Winter, P., Handley, P., Ploog, D., & Schott, D. (1973). Ontogeny of squirrel monkey calls under normal conditions and under acoustic isolation. *Behaviour*, 47(3–4), 230–239. <https://doi.org/doi.org/10.1163/156853973X00085>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, 12(12), 739–751. <https://doi.org/10.1038/nrn3112>
- Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8), 1317–1329. [https://doi.org/10.1016/S0893-6080\(98\)00066-5](https://doi.org/10.1016/S0893-6080(98)00066-5)
- Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, 79(7), 2064–2072. <https://doi.org/10.3758/s13414-017-1361-2>
- Xu, A., Niekerk, D. R. v., Gerazov, Krug, P., Birkholz, P., Prom-on, S., Halliday, L., & Xu. Y. A computational simulation of human vocal learning. (Under review)
- Xu, A., van Niekerk, D. R., Gerazov, B., Krug, P.K., Prom-on, S., Birkholz, P., Xu, Y. (2022). Modelling English diphthongs with dynamic articulatory targets. Preprint. <https://doi.org/10.31234/osf.io/532qj>
- Xu, A., van Niekerk, D., Gerazov, B., Krug, PK., Prom-On, S., Birkholz, P., Xu, Y., (2021). Model-based exploration of linking between vowel articulatory space and acoustic space. In: *Proceedings of the Annual Conference of the International Speech Communication Association: INTERSPEECH 2021*. (pp. 3191-3195). ISCA: Brno, Czechia. 10.21437/Interspeech.2021-1422

- Xu, A., Birkholz, P., & Xu, Y. (2019). Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation. *International Congress of Phonetic Sciences ICPhS 2019*. [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS\\_254.pdf](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_254.pdf)
- Xu, Y. (2020). Syllable as a synchronization mechanism that makes human speech possible. (Preprint). <https://doi.org/10.31234/osf.io/9v4hr>
- Xu, Y., & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication*, 57, 181–208. <https://doi.org/10.1016/j.specom.2013.09.013>
- Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, 15(4), 245–258. <https://doi.org/10.1080/09540090310001655075>
- Yoshikawa, Y., Koga, J., Asada, M., & Hosoda, K. (2003). Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching. *IEEE International Conference on Intelligent Robots and Systems*, 1, 149–154. <https://doi.org/10.1109/iros.2003.1250620>
- Zhao, W., Garcia-Oscos, F., Dinh, D., & Roberts, T. F. (2019). Inception of memories that guide vocal learning in the songbird. *Science*, 366(6461), 83–89. <https://doi.org/10.1126/science.aaw4226>
- Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2), 248–248. <https://doi.org/10.1121/1.1908630>