

## University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON

**Modelling the emergence of a basis for  
vocal communication between artificial  
agents**

by

Simon F. Worgan

A thesis submitted in partial fulfillment for the  
degree of Doctor of Philosophy

in the

Faculty of Engineering, Science and Mathematics  
School of Electronics and Computer Science

January 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS  
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Simon F. Worgan

Understanding the human faculty for speech presents a fundamental and complex problem. We do not know how humans decode the rapid speech signal and the origins and evolution of speech remain shrouded in mystery. Speakers generate a continuous stream of sounds apparently devoid of any specifying invariant features. Despite this absence, we can effortlessly decode this stream and comprehend the utterances of others. Moreover, the form of these utterances is shared and mutually understood by a large population of speakers. In this thesis, we present a multi-agent model that simulates the emergence of a system with shared auditory features and articulatory tokens. Based upon notions of intentionality and the absence of specifying invariants, each agent produces and perceives speech, learning to control an articulatory model of the vocal tract and perceiving the resulting signal through a biologically plausible artificial auditory system. By firmly establishing each aspect of our model in current phonetic theory, we are able to make useful claims and justify our inevitable abstractions. For example, Lindblom's theory of hyper- and hypo-articulation, where speakers seek maximum auditory distinction for minimal articulatory effort, justifies our choice of an articulatory vocal tract coupled with a direct measure of effort. By removing the abstractions of previous phonetic models we have been able to reconsider the current assumption that specifying invariants, in either the auditory or articulatory domain, must indicate the presence of auditory or articulatory symbolic tokens in the cognitive domain. Rather we consider speech perception to proceed through Gibsonian direct realism where the signal is manipulated by the speaker to enable the perception of the affordances within speech. We conclude that the speech signal is constrained by the intention of the speaker and the structure of the vocal tract and decoded through an interaction of the peripheral auditory system and complex pattern recognition of multiple acoustic cues. Far from passive 'variance mopping', this recognition proceeds through the constant refinement of an unbroken loop between production and perception.

# Contents

<b>Acknowledgements</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem statement . . . . .	3
1.2 Contributions . . . . .	3
1.2.1 Major claims . . . . .	3
1.2.2 Subsidiary claims . . . . .	4
1.3 Publications . . . . .	4
1.4 Outline of thesis . . . . .	5
<b>2 Grounding symbols in the physics of speech communication</b>	<b>7</b>
2.1 Overview . . . . .	7
2.2 Signal and symbol grounding . . . . .	11
2.3 Review of previous vowel system evolution models . . . . .	13
2.4 Basic agent architecture and its operation . . . . .	15
2.5 Contour spaces . . . . .	18
2.5.1 Introducing dispersive forces . . . . .	19
2.5.2 Attractive force: focalisation . . . . .	20
2.6 Results of simulations . . . . .	21
2.6.1 Reproduction of Oudeyer’s results . . . . .	22
2.6.2 Effect of the contour space . . . . .	23
2.6.3 Further comparison of the two systems . . . . .	23
2.6.4 Statistical hypothesis test . . . . .	26
2.7 Discussion and conclusions . . . . .	26
<b>3 Direct measure of articulatory effort</b>	<b>29</b>
3.1 Overview . . . . .	29
3.1.1 Weighting effort and distinctiveness . . . . .	31
3.1.2 Optimising quantal regions . . . . .	31
3.1.3 Finding local optima . . . . .	32
3.2 Capturing effort and dispersion . . . . .	32
3.2.1 Hill’s muscle model . . . . .	32
3.2.2 Guenther’s neural model . . . . .	34
3.3 Applying the model to H&H theory . . . . .	36
3.3.1 Developing emergent vowel systems . . . . .	36

3.3.2	Comparing effort and focalisation . . . . .	42
3.4	Summary . . . . .	43
<b>4</b>	<b>Interaction of place and voicing in the perception of initial stops</b>	<b>45</b>
4.1	Overview . . . . .	46
4.2	Modeling the categorisation of speech sounds . . . . .	48
4.2.1	Computational auditory/neural model . . . . .	51
4.2.2	Categorisation of synthetic Lisker and Abramson stimuli . . . . .	52
4.2.3	Categorisation of real speech . . . . .	55
4.3	Auditory-neural voicing identification . . . . .	59
4.3.1	Removing the peripheral auditory system component . . . . .	59
4.3.2	High-resolution FFT . . . . .	60
4.3.3	Modifying the peripheral auditory system . . . . .	62
4.3.4	Simplified model of perception of initial stops . . . . .	65
4.4	Discussion and summary . . . . .	68
<b>5</b>	<b>Modelling the cultural emergence of speech</b>	<b>72</b>
5.1	Overview . . . . .	73
5.2	True empiricism in language models . . . . .	74
5.3	Modelling the emergent perception of real speech . . . . .	75
5.4	Results and analysis . . . . .	77
5.4.1	Reproduction of CV utterances from limited auditory input . . . . .	78
5.4.2	Randomisation of auditory inputs . . . . .	79
5.4.3	Removal of the peripheral auditory system . . . . .	79
5.4.4	Sensitivity to SOM parameter settings . . . . .	81
5.4.5	Articulatory effort and auditory distinctiveness . . . . .	82
5.5	Language without strong constraints . . . . .	84
5.6	Summary . . . . .	86
<b>6</b>	<b>Speech perception as non-symbolic pattern recognition</b>	<b>90</b>
6.1	Overview . . . . .	91
6.2	Theories of speech perception and direct realism . . . . .	92
6.2.1	Nearey's framework: strong and weak relations . . . . .	92
6.2.2	Direct realism . . . . .	95
6.3	Phonetic and neuroscientific evidence . . . . .	98
6.3.1	Phonetic evidence . . . . .	98
6.3.2	Mirror neurons and perception as action . . . . .	99
6.4	Models and limiting abstractions . . . . .	101
<b>7</b>	<b>Conclusions and future work</b>	<b>104</b>
7.1	Limitations and future work . . . . .	105
7.2	Summary of work . . . . .	105
7.3	Wider implications and ASR . . . . .	107
7.4	Conclusion . . . . .	109

---

<b>A</b>	<b>Removing ‘mind-reading’ from the iterated learning model</b>	<b>111</b>
A.1	Overview . . . . .	111
A.2	Criticisms of the iterated learning approach . . . . .	112
A.3	Applying a self organising map to iterated learning . . . . .	113
A.4	Results . . . . .	116
A.5	Analysis . . . . .	118
A.6	Summary . . . . .	118
<b>B</b>	<b>Detailed models</b>	<b>120</b>
B.1	Oudeyer’s agent model . . . . .	120
B.1.1	Vocal tract model . . . . .	120
B.1.2	Cochlear model . . . . .	121
B.1.3	Neural model . . . . .	122
B.2	Our agent model . . . . .	124
B.3	Muscle model . . . . .	125

# List of Figures

2.1	The traditional view of symbol grounding links an <i>a priori</i> internal representation (cup) to its external referent cup. Reproduced from Pfeifer and Scheirer (1999, Fig. 3.4, p. 70).	11
2.2	(a) The ‘semiotic triangle’, reproduced from Vogt (2002, Fig. 1, p. 433). (b) A more complete picture of symbol grounding in which the FORM in (a) is grounded by interaction with the physical signal.	12
2.3	Illustration of signal grounding as a sub-problem of symbol grounding.	13
2.4	Architecture of the communicating multi-agent system, illustrated here for two agents. Redrawn from Oudeyer (2005c, Fig. 2, p. 439).	16
2.5	Convergence of Oudeyer’s model to a five-vowel system with 10 agents, $\sigma = 0.05$ and 2,000 iterations. Each cross represents a vector in auditory space; multiple vectors in the same region of space represent an equivalence class, or vowel. For a given equivalence class, individual vectors frequently overlay, giving the appearance of a single cross.	17
2.6	Typical plot of auditory dispersion versus number of iterations, showing convergence well before 2,000 steps.	21
2.7	Composite of typical results from our replication of Oudeyer’s simulation as $\sigma$ varies.	22
2.8	Composite of typical results from simulations of the new model with contour spaces with the same $\sigma$ values as in Figure 2.7. Realistic vowel systems emerge over a much wider range of $\sigma$ values.	24
2.9	Comparison of our replication of Oudeyer’s simulation with the new model based on DFT, illustrating the robustness to parameter variation resulting from inclusion of a dispersive force. Error bars are standard deviations over 500 runs.	25
2.10	Comparison of vowel systems observed in human languages and those produced by computer simulation with and without DFT (i.e., with and without dispersive forces).	25
3.1	Auditory spectrograms of a sample of vowels produced by a combination of the articulatory vocal tract and Guenther’s neural model, with CF corresponding to a particular ‘centre’ frequency. In each case the model has captured the appropriate formant values.	34
3.2	Plotted from data obtained from Schwartz et al., a wide number of candidates provide coverage of the space of possible vowels. Expanding on data from the IPA, candidate auditory prototypes were selected to provide adequate coverage of the vowel space.	36

3.3	Both figures show the effect of varying parameters ( $\alpha, \lambda, \beta$ ) on the resulting Euclidean distance within the vowel space – defined as the total distance between vowels. Clearly an increasing level of effort leads to a reduced total euclidean distance between vowels within the auditory space. . . . .	43
4.1	Two-stage (auditory/neural) computational model for the categorisation of speech sounds. . . . .	51
4.2	Time-frequency representations of Lisker and Abramson’s bilabial stimulus with VOT of 40 ms in the form of an auditory spectrogram produced by Lyon’s cochlear model: (a) is the full-data version produced with a sampling rate of 10 kHz; (b) is the reduced-data version produced by aggregating outputs into $12 \times 16$ time-frequency bins for input to the second-stage neural network model. Note that in (a), the centre frequency (CF) index reduces with increasing frequency (see text). . . .	53
4.3	Labellings of the Lisker and Abramson continuum produced by human and chinchilla subjects. From Kuhl and Miller (1978). . . . .	54
4.4	Labelling of the Lisker and Abramson stimuli by the computational auditory/neural model. Activations shown are averages across 10 presentations of the stimuli with the neural network trained from a different set of initial random weights on each presentation. Error bars are standard deviations. . . . .	55
4.5	Histograms showing the distribution of VOT(ms) for the (a) bilabial, (b) alveolar and (c) velar tokens in the Nossair and Zahorian speech database. Training data were formed from the means and standard deviations of voiced and unvoiced data, occasionally leaving overlapping test data, all three continua were then combined into two training sets (voiced/unvoiced) and one test set and presented to a single neural network. . . . .	57
4.6	(a) Auditory spectrogram produced by Lyon’s cochlear model for a real speech /bɔ/ (child speaker, VOT 42.1 ms) from the Nossair and Zahorian database; (b) reduced $12 \times 16$ time-frequency matrix presented to the second-stage neural network. The time bin width in (b) is 10.3 ms. . . .	58
4.7	Labelling curves for the real speech data in the child subset of the Nossair and Zahorian database. . . . .	59
4.8	(a) Fourier spectrogram for real speech /bɔ/ (VOT 42.1 ms) from the Nossair and Zahorian database; (b) Labelling curves for model with auditory preprocessor replaced by Fourier analysis. There is no obvious category boundary and, hence, correct movement of boundary with place of articulation is effectively abolished. . . . .	61
4.9	An illustration of the frequency enhancement present on the Bark scale. (Traunmüller 1990) . . . . .	62
4.10	(a) An example of the time-frequency representations of the Nossair and Zahorian data, reduced, from the original 128 frequency channels, to $68 \times 12$ network inputs; (b) Labelling curves for the high resolution FFT data generated from the Nossair and Zahorian database. . . . .	63



4.11	An idealised step function for comparison against the high-resolution FFT and peripheral auditory system. . . . .	64
4.12	(a) Auditory spectrogram for real speech /bɔ/ (VOT approximately 41.2 ms) from the Nossair and Zahorian database with hair cell simulation removed from Lyon's cochlear model; (b) An example of a reduced spectrogram for the real-speech data with hair cell simulation removed from Lyon's cochlear model. . . . .	64
4.13	Labelling curve for real speech /bɔ/ (VOT approximately 41.2 ms) from the child subset of the Nossair and Zahorian database with hair cell simulation removed from Lyon's cochlear model. There is no clear voiced/unvoiced boundary in this case; hence, the boundary movement effect is not evident. . . . .	65
4.14	Differences of unvoiced/voiced auditory spectrograms for Lisker and Abramson stimuli at 20 ms and 50 ms VOT: (a) bilabial, (b) velar (c) alveolar. . . . .	66
4.15	Labelling curves produced by new model with skeleton MLPs: (a) Lisker and Abramson synthetic stimuli with 4 inputs to MLP, (b) Nossair and Zahorian real-speech stimuli with 12 inputs to MLP. . . . .	67
5.1	Illustration of the artificial agents studied in this chapter. Signals are generated from the Lisker and Abramson speech database and from the utterances of the agent. These are processed by the agent's peripheral auditory system which activates the auditory and articulatory spaces, which are self-organising maps (SOMs), via weighted connections. Hebbian learning enables the agent to acquire a mapping from audition to articulation, while the SOMs adjust to represent the perceived signal and control the vocal tract. Once the validity of the individual agent has been established we replace the Lisker and Abramson database with the utterances of other agents in a population model. . . .	76
5.2	An example of auditory convergence for an agent trained on the four time-frequency bins derived from the Lisker and Abramson data set. After training there has been a significant reduction in the Euclidean distance between the 500 randomly initialised points. . . . .	78
5.3	When presented with random auditory input the auditory space fails to converge – the Euclidean distance between points is not reduced significantly. . . . .	80
5.4	An example agent equipped with the high resolution FFT detailed in Chapter 4. . . . .	81
5.5	The effect of a varying $\sigma$ value on the size of the final articulatory system. As can be seen realistic phonetic systems are only obtained under a narrow range of parameter values. . . . .	82
5.6	Sensitivity to parameter variation of a typical agent modified to trade articulatory effort against auditory distinctiveness according to Lindblom's H&H theory. This extension enables us to sustain emergence of plausible phonetic systems for a Gaussian width $\sigma$ over a wider range than hitherto. . . . .	83

5.7	By recording the Euclidean distance between the articulatory parameters of two randomly selected agents a measure of linguistic similarity is obtained. We can see how the articulatory parameters of each agent converge and diverge as they negotiate and settle upon a final shared set of articulatory parameters. The comparison between an agent within the population and an agent removed from it remains stable and distant.	84
5.8	The number of articulatory prototypes in a given agent's SOM over time	85
5.9	A comparison of the number of simulated articulatory prototypes and human phonetic systems. The human data were obtained from Tambovtsev and Martindale (2007) and the simulated data from 150 runs of the model under ideal conditions ( $\sigma = 0.21$ ).	86
6.1	Conflicting phonetic theories use evidence of strong constraints on articulation or audition to argue for different symbolic systems of perception. For example, the existence of invariant acoustic specifiers would be taken as evidence for the formation of abstract phonetic tokens during speech perception. See text for further explanation of (a), (b) and (c).	93
6.2	A comparison of Fowler's direct realism and double-weak direct realism. The phonetic evidence suggests a double-weak approach, while our own work proposes a direct realist cognitive theory. The arrow in the listener's head indicates implicit knowledge of the link between production and perception. See text for further details of (a) and (b).	96
6.3	The de Boer (2000a) model of phonetic emergence, explicitly capturing a variety of forces behind the change and evolution of phonetic tokens.	102
A.1	An adjustment of the previously implicit generalisation parameter showing the effect on the level of compression that each compositional meaning node can achieve.	113
A.2	Under the new IL model results similar to the ones obtained in previous implementations are maintained.	117
A.3	Previously unacknowledged generalisation parameter has a clear effect on the language, with a greater generalisation ability leading to a higher level of stability.	117
A.4	In a structured object space each meaning node can generalise over a greater number of objects, this will affect the stability of the language.	118

# List of Tables

3.1	The various muscles of the vocal tract and their role in speech production. $F_{max}$ represents the maximum exerted force, $\sigma$ the tension of the muscle and $m$ the mass. Values were obtained from (Sanguineti, Laboissiere, and Ostry 1998) and (Lucero and Munhall 1999). . . . .	35
3.2	Vowel acoustic prototypes. Formats $F_1$ , $F_2$ , $F_3$ , $F_4$ in Hz and in Bark, second perceptual formant $F'_2$ in Bark, and individual focalisation costs $E_F$ for $\alpha = 1$ . See Schwartz et al. (1997) for details of vowels not from the IPA. . . . .	38
3.3	A comparison of human and simulated vowel systems, obtained from the UCLA Phonological Segment Inventory Database (UPSID). Each of these simplified vowel spaces corresponds to the auditory prototypes presented in Figure 3.2. For clarity, central vowels are represented as white dots and vowels on the edge are black. . . . .	40
4.1	Mean and standard deviation of voice onset time (in ms) for the Nossair and Zahorian data used in this study. . . . .	57
B.1	The basic parameters for the muscle model. . . . .	126

## DECLARATION OF AUTHORSHIP

I, Simon F. Worgan declare that the thesis entitled “Modelling the emergence of a basis for vocal communication between artificial agents” and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- the work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as detailed in Section 1.3

**Signed:**

**Date:**

## **Acknowledgements**

Firstly, I would like to say thank you to my supervisors Prof. Robert I. Damper and Dr. Seth Bullock for their advice, guidance and insights over the last few years. Thanks also go to Dr. Richard Watson, Dr. Jason Noble and all of my colleagues from various research groups and seminars. Their constant stream of ideas has maintained my enthusiasm for my research and the wider field of AI. Most importantly of all I'd like to say thank you to my wife Melanie, whose love, understanding and support enabled me to start and finish my PhD. This work would not have been possible without her. Finally, I'd like to thank my friends and family for reminding me that there's more to life than the formation of symbolic phonetic systems.

# Chapter 1

## Introduction

Artificial Intelligence (AI) has followed two avenues of scientific research, the production of intelligent behaviour by artificial systems through any means possible and the discovery of underlying biological causes of intelligence through the implementation of theories found in the natural world. This thesis will take the second approach. We will investigate through computer simulation current theories of phonetic emergence and perception and in doing so we hope to contribute to our understanding of the evolution of language.

When faced with multiple competing theories, AI can make a valid contribution by testing their implications through simulation. Thus, we are able to apply a level of implementational rigour to existing, purely philosophical, theories. To exploit fully this advantage we must be cautious; every model contains abstractions and assumptions and a poor choice of either will invalidate the results and undermine the conclusions. By implementing these approaches carefully and properly, we can confidently progress from the underlying philosophy to specific phonetic questions. Accordingly, once we have tested the resulting models we will consider notions of cultural evolution and phonetic theory. These include H&H (hypo- and hyper-articulation) theory — the theory that speakers will maximise auditory distinctiveness and minimise articulatory effort (Lindblom 1990), and dispersion/focalisation theory — speakers not only seek auditory distinctiveness but they also seek acoustic stability (Schwartz, Boë, Vallee, and Abry 1997) and we will justify the embodiment of these theories through the adaptation of symbol grounding.

Within symbol grounding, atomic symbols are formed through the perception of and interaction with the surrounding environment so as to give meaning to the symbols and their combinations. Previously, abstract models faced a formidable problem,

famously articulated by Harnad (1990): “How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?” (p.335). Harnad characterises this problem as the symbol grounding problem. In light of its importance, various solutions have been proposed—see Belpaeme, Cowley, and MacDorman (2007) for some recent work—and within speech we propose that communicative systems can become grounded through ‘signal grounding’, where production and perception become embodied through the application of H&H theory.

According to H&H theory, speakers “tune their performance according to communicative and situation demands ... to vary their output along a continuum of *hyper-* and *hypospeech*” (Lindblom 1990, p.403). That is, in difficult communication conditions, speakers hyper-articulate in order to be understood, even though this requires additional energy be expended. In less demanding situations energy can be conserved and successful communication maintained by hypo-articulation. The ‘setting’ on the hyper-/hypo- continuum is determined by an on-line process in which the speaker continuously infers success of communication by monitoring linguistic and para-linguistic feedback from the listener. We assume that similar forces are at work in the process of vowel formation among a collection of communicating agents; that is, there is not only a drive towards distinctive sound categories (loosely corresponding to ‘hyper’), but also an inbuilt desire to minimise energy expended by the agent (loosely corresponding to ‘hypo’).

A full consideration of the above will equip us with the necessary models (cognitive, articulatory and auditory) to consider the role of real speech and its implications for current fundamental phonetic theories. Such theories have long been divided by the question: ‘What do we perceive when we perceive speech?’ Do we perceive symbolic phonetic tokens or do we perceive symbolic vocal tract gestures? Can we in fact do away with symbolic notions of perception altogether and take a more direct realist approach (Gibson 1979)? To guide our own work within this potential minefield, we will take the classification framework due to Nearey (1997), dividing phonetic theories according to the specifying invariance that each believes lies in the signal and gesture. This framework is useful because it highlights the assumed link between specifying invariance, and the hypothetical cognitive tokens of speech. By taking the modelling approach, this thesis will reconsider this link in a new light and question symbolic assumptions. Finally, we will speculate on the implications of applying this approach to the wider field of language and embodied cognition.

## 1.1 Problem statement

We are faced with two interrelated problems. Firstly, how does a shared understanding emerge within a population of speakers? Does a culture's language arise through a constrained, innate, biological process or does it follow from the cultural evolution of communicative intent? Secondly, how does the individual achieve phonetic perception and production? Do we perceive articulatory idealisations, symbolic phonetic tokens derived from the auditory signal or neither? These two questions are related because the emergence of a shared conversational system among a population of speakers is dependent upon the individual human capacity for phonetic production and perception. This crucial interdependence has been overlooked by previous modelling approaches and, frequently, phonetic challenges are abstracted away. We hope to correct this oversight in this thesis.

## 1.2 Contributions

We will now list the major and minor claims of the thesis giving brief references to supporting sections and results.

### 1.2.1 Major claims

- We have built a simulation of vowel system evolution among a population of intercommunicating agents that produces systems which are more like human vowel systems than have been produced by previous studies.
- We have shown how simulations of stop consonant perception, which exhibit the characteristic VOT boundary shift with place seen in human listeners, are reliant on particular filtering characteristics of the peripheral auditory system, a result which helps to explain the outcomes of earlier modeling studies.
- We have shown that an artificial agent, exposed to synthetic initial stops and equipped with physiologically-based auditory and articulatory models, is able to converge on a set of acoustic features and articulatory tokens, reproducing the relevant features of a set of initial stop consonants.



### 1.2.2 Subsidiary claims

- We have shown that a population of agents, producing and perceiving utterances without strong constraints, can converge to a shared set of auditory features coupled with articulatory tokens. By comparison, previous work has converged to a set of shared phonetic symbols.
- We have shown that the introduction of symbol grounding results in Gaussian models that are increasingly robust to parameter variation reducing the Euclidean distance, summed over an adjustment in Gaussian width from 0.02 to 0.15, from 21.8 vowels to 11.2 vowels, when compared to the human distribution of vowel systems, with a significance of  $\alpha = 0.043$ .
- We have shown that an abstract simulation of articulatory effort, coupled with the forces of dispersion focalisation theory, can produce vowel systems that are more plausible than those produced by models of dispersion focalisation theory alone.

Combining these claims we offer the opinion that speech should be viewed as the direct perception of interaction affordances, where the listener is aware of the speaker's intentionality and refines their perception appropriately. Current population models of speakers and listeners consider speech to be simply the transmission of abstract phonetic tokens. We argue that speech should be viewed, and modelled, as an unbroken loop of recursive perception and production.

## 1.3 Publications

This thesis presents a detailed body of work addressing current phonetic challenges and lending support to the cultural evolution of a shared system of communication. During the work a number of academic papers have been prepared for publication, listed below, and where appropriate these have been incorporated into the final thesis.

- Worgan, S. F. and Moore, R. K. (2009) *Spoken language processing as an aspect of human behaviour*. Distributed Language Group Symposium, Wenham, MA
- Worgan, S. F. and Damper, R. I. (2009) *Symbolism and enactivism: An experimental test of conflicting approaches to artificial intelligence* Journal of Experimental & Theoretical Artificial Intelligence, 21 (1). pp. 1–18.

- Worgan, S. F. and Damper, R. I. (2008) *Removing ‘mind-reading’ from the iterated learning model*. Evolang 2008, 7th Evolution of Language Conference, Barcelona, Spain. pp. 378–386.
- Worgan, S. F. and Mills R. (2008) *Initial modelling of the alternative phenotypes hypothesis* Proceedings of the Eleventh International Conference on Artificial Life. pp. 717–724
- Worgan, S. F. and Damper, R. I. (2007) *Grounding symbols in the physics of speech communication*. Interaction Studies, 8 (1). pp. 7–30.
- Worgan, S. F. and Damper, R. I. (2007) *Speech perception as non-symbolic pattern recognition*. Symposium on Language and Robotics, Aveira, Portugal. pp. 99–100
- Worgan, S. F. and Damper, R. I. (2006) *Grounding symbols in the physics of speech communication*. External Symbol Grounding Workshop 2006, Plymouth, UK.

## 1.4 Outline of thesis

The rest of this thesis is structured as follows.

First considering symbol grounding, in Section 2.2, we set out our conception of physical symbol grounding, which we call *signal grounding*, and relate this to more traditional views. Then, as a baseline for later discussion of our own work, we briefly describe Oudeyer’s (2005c) simulations of the emergence of vowel systems shared between a population of agents in Section 2.4. Section 2.5 introduces our extension to these simulations in the form of ‘contour spaces’. Section 2.6 illustrates the beneficial effects of this extension in terms of emergence of more realistic vowel systems. Section 2.7 discusses the implication of these findings and concludes by arguing for the use of more realistic articulatory/auditory modelling as necessary to move beyond production of static vowel systems and account for the dynamic consonant-vowel patterning of speech.

In Chapter 3, we move towards this more realistic articulatory/auditory modelling by justifying and constructing a muscle model to extend an existing vocal tract simulation (Cook 1993), providing us with a direct measure of articulatory effort. We then compare this model to existing approaches, previously detailed in Chapter 2, and analyse our

results in Section 3.4, concluding with an outline of how the production of real speech coupled with a direct measure of the effort of production will be used in the rest of the thesis.

In Chapter 4 we will then develop, in Section 4.2, a model of the peripheral auditory system. Detailed in Section 4.3, the importance of the peripheral auditory system in the perception of real speech will be highlighted. This model will be used to investigate the interaction of place and voice in the perception of initial stops, a long standing perceptual puzzle.

In Chapter 5, we will outline a model of phonetic emergence in Section 5.1. Then, we will construct and analyse the complete model in Section 5.3. This model will first focus upon the abilities of an individual agent before considering questions of phonetic emergence among a population of speakers. Section 5.4 develops this work's wider implications for grounding and speech.

After developing our final model we proceed, in Chapter 6, to speculate about the consequences of these results and we will conclude this chapter by highlighting the damaging abstractions present in existing phonetic models. These abstractions are considered as such due to their neglect of current challenges within the field of phonetics. Specifically, we will speculatively question the assumption that specifying invariants in either the acoustic signal or articulatory gesture provide evidence for the perception of symbolic articulatory or auditory idealisations. We then conclude in Chapter 7, considering the limitations and implications of this work.

## Chapter 2

# Grounding symbols in the physics of speech communication

We will first consider the role of symbol grounding in the modelling of speech. The traditional view of symbol grounding seeks to connect an *a priori* internal representation or ‘form’ to its external referent. But such a ‘form’ is usually itself systematically composed out of more primitive parts (i.e., it is ‘symbolic’), so this view ignores its grounding in the physics of the world. Some previous work simulating multiple talking/listening agents has effectively taken this stance, and shown how a shared discrete speech code (i.e., vowel system) can emerge. Taking the earlier work of Oudeyer (2005c), we have extended his model to include a dispersive force intended to account broadly for a speaker’s motivation to increase auditory distinctiveness. New simulations show that vowel systems result that are more representative of the range seen in human languages. These simulations make many profound abstractions and assumptions. Relaxing these by including more physically and physiologically realistic mechanisms for talking and listening is seen as the key to replicating more complex and dynamic aspects of speech, such as consonant-vowel patterning.

## 2.1 Overview

The computational metaphor that underpins cognitive science, and much of artificial intelligence and functionalist philosophy of mind also, sees intelligent behaviour as the product of the workings of a formal symbol manipulation system (e.g., Newell 1973; Minsky 1974; Fodor 1975; Newell and Simon 1976; Newell 1980; Newell 1990; Pylyshyn 1984; Dietrich 1990). But this view faces a formidable problem,

famously articulated by Harnad (1990) as: “How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?” Harnad calls this the *symbol grounding problem* (SGP) and comments: “The handicap has been noticed in various forms since the advent of computing”. The earliest reference that we know is that of Mays (1951), who writes “if we grant that these machines [*i.e.*, *digital computers*] are complex pieces of symbolism, ... it is clear that in order to acquire a significance the symbols need to be linked with a set of referents” (p.249). So if the computational metaphor is to offer any purchase in modelling and understanding cognition, the SGP poses a challenge that cannot be neglected (Cangelosi, Greco, and Harnad 2000). Since the long-term goal of our research is to understand via computer modelling and simulation how speech sound categories (broadly, ‘phonemes’) could have emerged during language evolution, and then how these could be combined systematically to lead to utterances with semantic content, we take this challenge seriously.

To some, the SGP is symptomatic of an incorrect view of AI and cognitive science, famously parodied as “good old-fashioned AI”, or GOF AI, by Haugeland (1985). For instance, as Pfeifer and Scheirer (1999, p.71) write: “... the symbol grounding problem is really an artifact of symbolic systems and ‘disappears’ if a different approach is used”. The different approach they have in mind is, of course, *embodied* or *nouvelle* AI as spearheaded by Brooks (1990, 1991, 1999), which seeks to replace the central role played by symbolic representation by non-symbolic interfacing to the physical world through cycles of perception and action, usually conceived as based on some connectionist or statistical machine learning principles. However, the complete banishment of symbolism from the scene is rather too radical for most AI scientists and cognitive psychologists, who continue to see a role for formal symbol systems, albeit in combination with some sort of connectionist component (e.g., Minsky 1990; Harnad 1990, 1993) in modelling and explaining the higher cognitive functions involved in using language, doing mathematics, decision making under uncertainty, etc., where *nouvelle* AI has arguably promised more than it has delivered.

A new view of the SGP has recently arisen in which the physics of the external world plays an important and simplifying role (Sun 2000; Vogt 2002). Vogt (2002) coins the term *physical symbol grounding problem* and writes, “It is based on the idea that symbols should be grounded (cf. Harnad 1990) and ... they should be grounded by physical agents that interact with the world (cf. Brooks 1990)” (p.435). Our thesis is broadly consonant with this view, treating the SGP (as does Vogt) as a technical problem by way of computer simulation.

Quite apart from the intrinsic scientific interest in studying the emergence of human speech and language for its own sake (Damper 2000), it makes an excellent context in which to consider the SGP. First and foremost, we believe human communication to be the clearest, certainly best-developed, example of externally-grounded cognition. As Vogt (2002, p. 431) writes, “language through its conventions offers a basis for invariant labeling of the real world.” Since human communication is a social phenomenon, we pursue an approach of multi-agent simulation, not unlike much previous work in ‘language games’ but with one important difference (see below).

In particular in this chapter, we argue that the emergence of speech sound categories can and should be grounded in the physics of speech communication between agents, recognising that the human’s contact with the external world of sound is via their articulatory and auditory systems. Important previous work along these lines is that of Steels (1997, 1998, 1999, 2003), de Boer (2000a, 2001, 2005), and Oudeyer (2005a, 2005b, 2005c), who have explored grounded speech-category formation by computer simulation of multi-agent systems, with agents equipped with rudimentary articulatory and auditory systems and associated ‘neural’ processing. Broadly speaking, this line of work had its beginnings in the early and influential efforts of Lindblom (1986a) and his colleagues to explain the origins of vowel systems in the world’s languages (Liljencrantz and Lindblom 1972; Lindblom, MacNeilage, and Studdert-Kennedy 1984; Lindblom 2000) based on “adaptive dispersion theory.” In their numerical simulations, the clustering of vowels in some metric space was predicted by minimising an energy function designed to reflect perceptual distinctiveness. An important question is exactly how realistic the simulations have to be (e.g., in terms of faithfully modelling the articulatory/auditory systems and brain mechanisms). A subsidiary goal of all work of the kind described in the thesis is to answer this question, although for this chapter we will restrict ourselves to relatively simple simulations such as have been used in other previous work.

Although Steels (1997) argues for a “limited rationality constraint” in multi-agent simulations (i.e., agents should not have access to each other’s internal states), this constraint is typically violated in language games where nonlinguistic feedback figures importantly. For instance, de Boer (2001) writes, “the initiator then communicates the success or failure to the imitator using nonlinguistic communication” (p. 52). In our view, this amounts to a form of ‘mind-reading,’ seriously undermining the credibility of the simulations. Hence, we wish to avoid this aspect of language games, and favour Oudeyer’s alternative approach where he dispenses with nonlinguistic feedback. As he writes, “it is crucial to note that agents *do not* imitate each other ... The only consequence of hearing a vocalization is that it increases the probability, for the agent

that hears it, of vocalizations ... similar to those of the heard vocalization” (Oudeyer 2005c, p.443). In spite of the absence of structured, coordinated interactions between agents, he achieves two results in his simulations which mirror important aspects of real language: “on the one hand discreteness and compositionality arise thanks to the coupling between perception and production within agents, on the other hand shared systems of phonemic categories arise thanks to the coupling across agents” (Oudeyer 2005c, p.445).

A related line of investigation is that of Kirby (2001) and Kirby and Hurford (2002) who describe the iterated learning model (ILM), see Appendix A for an elaboration of this approach. This, however, operates at the syntactic level, that is, learning agents receive from adult agents “meaning-signal pairs” (p. 103) that act as training data. Thus, the ILM already tacitly assumes the emergence of phonetic distinctiveness. Whereas the language-game style of simulations are concerned with language change once the basic mechanisms are in place, by contrast, Oudeyer (2005c) is concerned with the earliest origins of a phonemic sound system, as are we. Further, Oudeyer’s model is based on horizontal cultural interaction between agents of the same generation, following the works of Steels and colleagues, whereas the ILM is based on iterated learning among agents of one generation and agents of the previous generation (so this is more vertical learning).

However, Oudeyer’s work has its own drawback in that he ignores the tenets of dispersion theory. “There are no internal forces which act as a pressure to have a repertoire of different discrete sounds,” he writes (p.443). But to cite de Boer (2001, p.61), a successful vowel system has “its vowel clusters ... dispersed (for low energy) and compact (for high imitative success).” These ideas are broadly consistent with notions of H&H theory (Lindblom 1990), chapter 1, and the dispersion-focalisation theory (DFT) of Schwartz, Boë, Vallee, and Abry (1997). Although Oudeyer (2005c) tries to argue that the lack of a dispersion force is a virtue of his simulations (it is one less assumption), he also seems to recognise that it causes problems for the emergence of sound systems with realistically large numbers of vowels, writing, “Functional pressure to develop efficient communication systems might be necessary here” (p.447).

Accordingly, the principal purpose of the present chapter is to introduce ideas of H&H theory and DFT into Oudeyer-style simulations in the belief that more realistic vowel systems (i.e., more representative of those seen in a variety of human languages) will result, providing a first step towards the rigorous modelling of current phonetic theory. We will do this by extending the topological spaces in the neural maps used to couple auditory and articulatory processing as a vastly-simplified form of brain. We



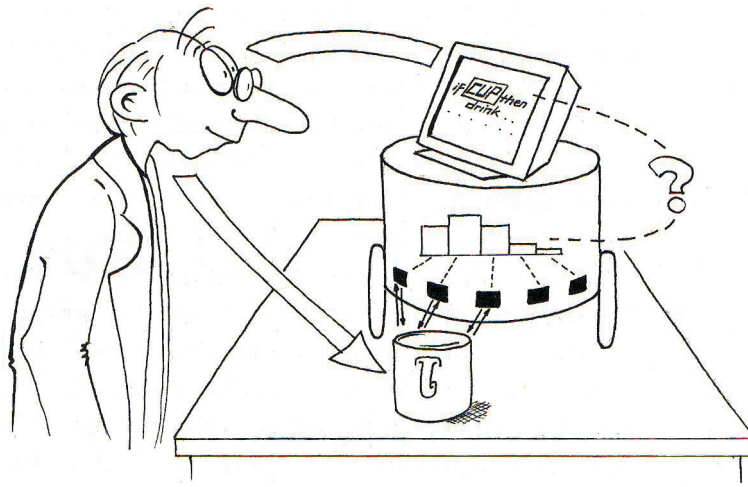


FIGURE 2.1: The traditional view of symbol grounding links an *a priori* internal representation (cup) to its external referent cup. Reproduced from Pfeifer and Scheirer (1999, Fig. 3.4, p. 70).

call these extensions *contour spaces*. The work is intended to form a baseline for the remainder of the thesis in which we will study the impact of increased realism of the agents' articulatory and auditory capabilities, as well as extending our simulations beyond prediction of static vowel systems to the emergence of connected speech sounds with appropriate consonant-vowel patterning.

## 2.2 Signal and symbol grounding

Before proceeding, it is necessary to discuss our relatively wide view of 'symbol grounding' and how it relates to the traditional, rather-narrower symbol grounding paradigm. Traditionally, the SGP has been seen as the problem of linking an internal symbolic representation like *cat* to the external (distal) object *cat*. For instance, Figure 2.1 (reproduced from the influential text of Pfeifer and Scheirer 1999) depicts a scenario linking the symbol *cup* with its external referent cup. But this traditional view already assumes the existence of some sort of internal representation, which is more or less symbolic (or at least compositional). In our view, any solution to the SGP must also explain how this internal representation gets composed from elementary parts, which we take to be close to the notion of 'icons' in the terminology of Harnad (1990) or 'perceptual symbols' in the terminology of Barsalou (1999). Because these elementary parts result from sensory-motor interaction, we cannot ignore the physics of the world. This leads us to the idea of signal grounding.

Symbol grounding is often discussed in the context of the semiotic triangle as in



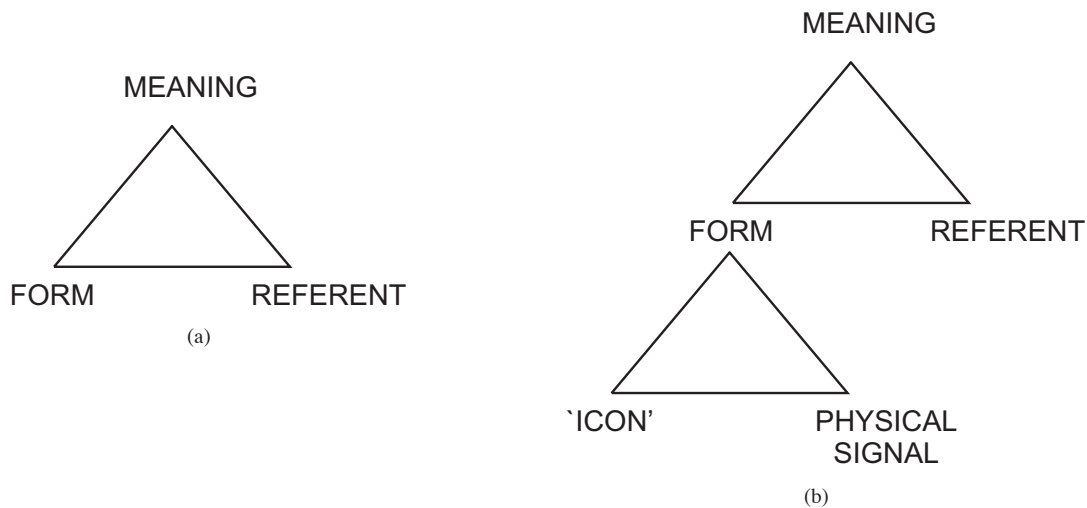


FIGURE 2.2: (a) The ‘semiotic triangle’, reproduced from Vogt (2002, Fig. 1, p. 433).  
 (b) A more complete picture of symbol grounding in which the FORM in (a) is grounded by interaction with the physical signal.

Figure 2.2(a), reproduced from Vogt (2002). But as just stated, we believe this picture to be incomplete, since the form is itself symbolic and ungrounded. A more complete view is depicted in Figure 2.2(b), where interaction with the physical world now grounds the form. In the case of interest here, this interaction is with the speech signal, hence the term ‘signal grounding,’ which can be seen either as a component part of symbol grounding, or as a specific instance of the SGP, albeit at a lower level than is usually considered. However it is viewed, we believe signal grounding is an indispensable part of symbol grounding.

For example, consider Figure 2.3. In this particular case of signal grounding, the distal object takes the form of an acoustic speech signal, produced by a vocal tract and perceived through the ear of a listener, linked to an arbitrary and iconic phoneme token (e.g., /æ/ using the notation of the International Phonetic Association 1999). The form *cat* (or, equivalently, /kæt/) is then composed in a way that is systematic, but nonetheless arbitrary, from these phonemic primitives. Signal grounding then presents numerous challenges when considering the practicalities of forming an equivalence class for the phoneme /æ/. We need to map a wide range of varied signals onto the same phoneme symbol; the system needs to adapt to linguistic change over time; and the grounding of these arbitrary tokens needs to be shared among a population of speakers. These challenges will be taken up in the remainder of the chapter.

To conclude this section, we remark that the ideas of signal and symbol grounding developed here are strongly related to notions of *double articulation*, stemming from the work of de Saussure (1983), which views a linguistic system as a series of differences of

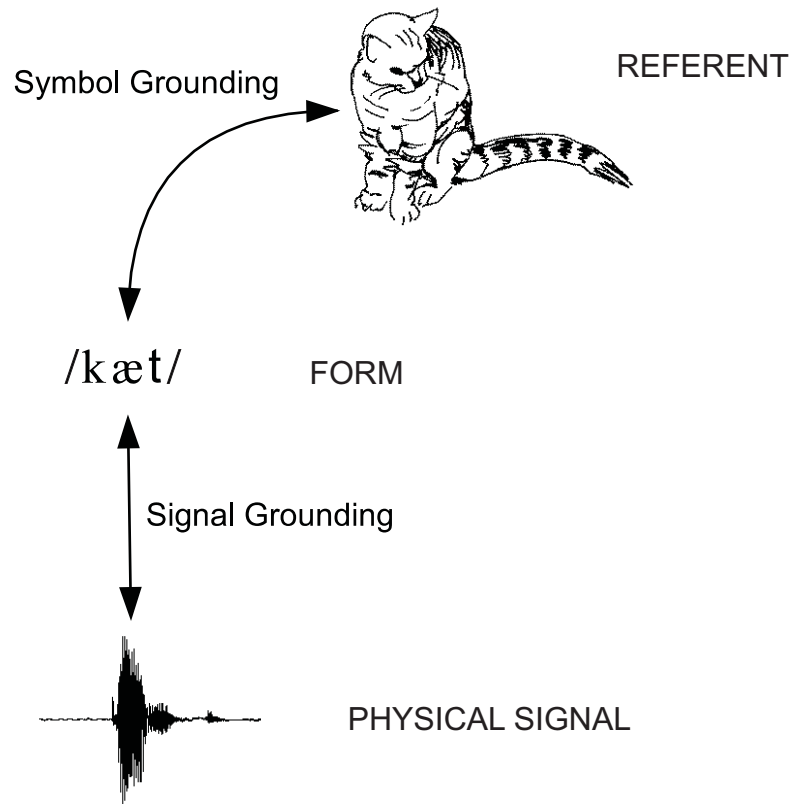


FIGURE 2.3: Illustration of signal grounding as a sub-problem of symbol grounding.

sound combined with a series of differences of ideas. At the level of the first articulation, meaningful units (morphemes, words) are combined syntactically to convey ideas. At the level of the second articulation, primitive or elementary sound units (phonemes) are combined to form the meaningful units of the first articulation. The level of the second articulation is vital to human language as a fully productive system, because it is the key (loosely quoting Wilhelm von Humboldt) to achieving infinite generativity from finite machinery. Yet this is the level that is typically ignored by the traditional view of the SGP as characterised in Figure 2.1.

## 2.3 Review of previous vowel system evolution models

Work by Vogt and de Boer (2009) summarises the current state of the art and highlights how computer modelling can help us understand the complex interactions between cultural and biological evolution. This approach is required as we cannot conduct our own empirical studies or historical review, as stated by Vogt and de Boer ‘The problem with both biological and cultural evolution is that they are historical processes, and therefore to a certain degree unpredictable and prone to loss of information overtime.’

Faced with such problems the importance of computational modelling is in providing a predictive formalism, ‘scientists carry out computer simulations and then need to assess whether the yielded predictions are in line with the empirical observations.’

Within computational modelling Vogt and de Boer (2009) identifies three frameworks:

- Analytical models (AM): ‘are mathematical models that typically describe an evolving system using a limited set of mathematical equations, so that the evolution can only be described at a meta-level.’
- Agent based analytical models (ABAM): ‘are models that define the dynamics and interactions within one or more populations of individuals (or agents) who are themselves defined by a mathematical equation.’
- Agent based cognitive models (ABCM): ‘are models that also describe the dynamic processes of a population, but where each individual is designed by a computer program that implements the production, interpretation and learning of linguistic elements.’

This chapter will take the Agent based cognitive models (ABCM) approach and ‘it is clear that the level of detail in ABCM is much higher than in AM. The level of detail regarding the empirical data that should be involved is likewise more complex for ABCM than for AM.’ As a result this chapter will consider the fundamental forces behind the production and perception of speech and test our model’s results against existing human vowel systems.

de Boer (2000b) further motivates the position of this chapter. Specifically, we need to test whether ‘innate predispositions are probably not necessary to explain the universal tendencies of human vowel systems.’ As an alternative it is proposed that that these structural regularities arise through self-organisation. Underlying this emergent behaviour it has been argued that human vowel systems seek to optimise maximum auditory distinctiveness and minimise articulatory effort.

In studying this approach de Boer (2000b) presents a population system where each agent possesses an articulatory and auditory model coupled with explicit imitation mechanisms and non-verbal feedback between participants. As de Boer himself acknowledges ‘direct nonverbal feedback might be considered unrealistic, as human children, when learning a language hardly get any direct feedback about the sounds they produce’.

However, this work does ‘shown that due to the interactions between the agents and due to self-organization, realistic vowel repertoires emerge. This happens under a large number of different parameter settings and therefore seems to be a very robust phenomenon. The emerged vowel systems show remarkable similarities with the vowel systems found in human languages’ Results suggest a similar, but not exact, match to the distribution of human vowel systems with a simulated frequency peak of 4 vowels instead of 5 for human systems.

In a development of this work Oudeyer (2005c) removes the ‘co-ordinated interactions’ of de Boer’s model. Instead, ‘The mechanism is based on a low-level model of sensory-motor interactions. We show that the integration of certain very simple and non-language-specific neural devices leads to the formation of a speech code that has properties similar to the human speech code.’ He presents an articulatory model that relies on the three major articulatory parameters: lip rounding, tongue height and tongue position. This is connected to the auditory system which is mapped to a two dimensional space consisting of the first formant and the second effective formant. Convergence within this neural model consists of two fully connected self organising maps which update according to a Gaussian activation function. While connections between auditory and articulatory spaces are updated according to hebbian learning. Perception of each agents own utterances allows for the formation of an accurate mapping from auditory to articulatory spaces and perception of other agents utterances causes the auditory space to converge towards the populations agreed vowel inventory. Results suggest that there is a very strong dependence on the width of the Gaussian function but the resulting systems show a strong similarity to human vowel systems with both peaking at 5 vowels. Unfortunately, for Oudeyer’s work detailed results and statistical tests were not presented, accordingly we are unable to judge the significance of his claims and results against our own.

## 2.4 Basic agent architecture and its operation

The kind of signal grounding just described, and argued to be fundamental to human speech and language as a fully generative system, is a feature of the multi-agent simulation work of Oudeyer (2005c). We will take his work as the basis for extensions aimed at producing more realistic sound systems, by defining a *contour space* which acts as an objective function embodying measures of both articulatory effort and phonetic distinctiveness, broadly in line with both H&H theory (Lindblom 1990) and dispersion-focalisation theory (Schwartz, Boë, Vallee, and Abry 1997).

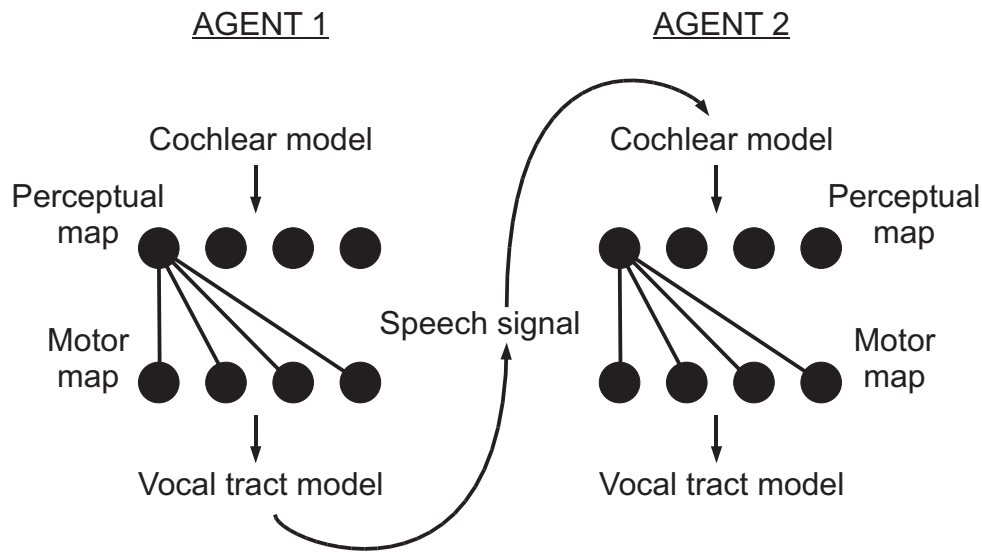


FIGURE 2.4: Architecture of the communicating multi-agent system, illustrated here for two agents. Redrawn from Oudeyer (2005c, Fig. 2, p. 439).

Figure 2.4 shows the basic agent architecture as used by Oudeyer (2005c) and in this work. Each agent has an artificial ear (cochlear model), an artificial vocal tract, and in Oudeyer’s words an artificial ‘brain’. Following Guenther and Gjaja (1996), the ‘brain’ features two coupled self-organising maps (SOMs, see Kohonen 1990) — a perceptual map taking input from the auditory system and a motor map driving the articulatory system. Each agent perceives sounds produced by other agents as well as by itself. Appendix B.1.2 sets out details of the cochlear, vocal tract and neural models used by Oudeyer (2005c), and in our replications of his work. Note that we have used the “realistic” non-linear articulatory/acoustic mapping (Oudeyer’s Section 6.2) rather than the “abstract” linear mapping (Oudeyer’s Section 6.1) throughout.

Our simulations use 10 agents (as compared to the 20 used by Oudeyer 2005c). But as he says of the number of agents, “This is a noncritical parameter of the simulations since nothing changes when we tune this parameter, except the speed of convergence of the system” (p. 443). Each ‘speaking’ agent is ‘heard’ by just one ‘listening’ agent picked at random. Oudeyer (2005c) states that “nothing changes” (p. 443) if a speaking agent is heard by more than one listener.

Initially, each agent produces utterances as dictated by its randomly-initialised ‘brain’ and also perceives the utterances of others. This, over some iterations, causes its SOMs to move from an unstable random configuration to a stable, converged, state of equilibrium. This process of convergence is driven by positive feedback (the basic self-organisation mechanism of the SOM), as each agent becomes increasingly likely to repeat the utterances that it has heard. Eventually, each SOM becomes partitioned into a

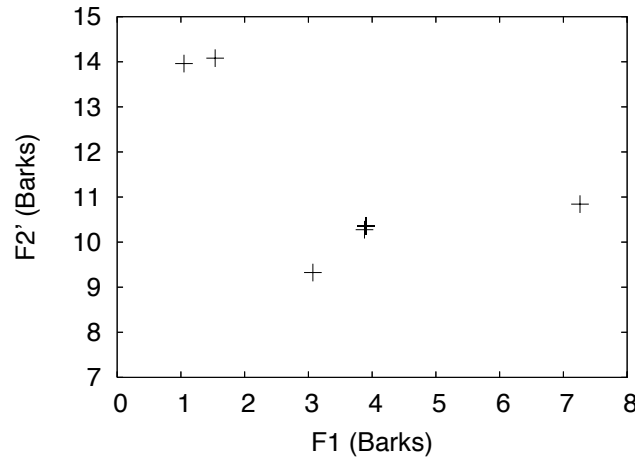


FIGURE 2.5: Convergence of Oudeyer’s model to a five-vowel system with 10 agents,  $\sigma = 0.05$  and 2,000 iterations. Each cross represents a vector in auditory space; multiple vectors in the same region of space represent an equivalence class, or vowel. For a given equivalence class, individual vectors frequently overlay, giving the appearance of a single cross.

variable number of basins of attraction as the nodes cluster around points of stability—determined by the utterances of the whole population. Any utterance which falls within the range of one of these basins of attraction is perceived by strong activation of the nodes around the centre point, so classifying a wide range of utterances.

The width of each SOM’s gaussian function ( $\sigma$  in equation B.1 of Appendix B.1.3) determines the size of the basin of attraction and, therefore, in the case of the auditory map, the variety of stimuli perceived as the ‘same’ utterance. In Oudeyer’s simulations, there is no dispersive force and, thus, as  $\sigma$  increases, convergence is to a single point. To quote Oudeyer (2005c, p.445), “if two neuron clusters ... get too close, then the summation of tuning functions in the iterative process of coding/decoding smooths their distribution locally and only one attractor appears.” This is not realistic behaviour within a language. However, it is clear that, with the right parameter settings, it is perfectly possible to cause the emergence of a feasible, shared, multi-vowel system. See for instance Figure 2.5, which depicts a typical result from our replication of Oudeyer’s simulation. Here, 500 points initially distributed randomly in  $F1$ - $F2'$  space, where  $F2'$  is defined as the second formant modified by the third formant, have converged to just five clusters. In fact, in the absence of a dispersive force, the ‘clusters’ have actually converged (almost) to overlay at the centre of their respective basin of attraction. In the remainder of this chapter, we will introduce a dispersive force and study its effect on convergence to linguistically-realistic vowel systems.

## 2.5 Contour spaces

In this section, we incorporate the basic ideas behind H&H theory (Lindblom 1990) and dispersion-focalisation theory (DFT) (Schwartz, Boë, Vallee, and Abry 1997) into our simulations. DFT encompasses more or less the same principles as H&H theory, but is formulated in the auditory (rather than articulatory) domain. This theory seeks to explain the formation of vowel inventories not so much in terms of energy expended by a speaker as via competing forces of “global dispersion based on inter-vowel distances; and local focalization, which is based on intra-vowel spectral salience” (Schwartz, Boë, Vallee, and Abry 1997, p.255). The dispersive force thus seeks to maintain distinctiveness between sound categories. The focalisation force in DFT is a little harder to visualise and justify. It is based on the ‘compactness’ of formant frequencies, formants being the resonant frequencies of the vocal tract that correspond to “concentration of acoustic energy, reflecting the way that air from the lungs vibrates in the vocal tract, as it changes its shape” (Crystal 1980, p. 150). These concentrations of energy are reflected in peaks in the frequency spectrum; the one occurring at the lowest frequency is called the first formant,  $F_1$ ; that occurring at the next highest frequency is called the second formant,  $F_2$ , and so on.

In the words of Schwartz, Boë, Vallee, and Abry (1997) (note the minor difference in notation for formant frequencies):

“a discrimination experiment involving stimuli with various  $F_2$ - $F_3$ - $F_4$  patterns ... demonstrated that patterns with the greatest formant convergence (namely with  $F_3$  close to either  $F_2$  or  $F_4$ ) were more stable in auditory memory ... while patterns with less convergence, namely with  $F_3$  at an equal distance from both  $F_2$  and  $F_4$ , were more difficult to memorize (Schwartz and Escudier 1989).” (p.259)

Schwartz, Boë, Vallee, and Abry (1997) further note, “the perceptual demonstration that formant convergence in the  $F_2$ - $F_3$ - $F_4$  pattern produced more stable patterns in discrimination experiments, led us to propose that formant convergence could result in an increased ‘perceptual value’ ... because of ‘acoustic salience’” (p.259). Hence, the focalisation force is designed to favour vowels in which the formants are close together in frequency.

### 2.5.1 Introducing dispersive forces

In the long term, we are seeking to minimise the articulatory effort of an utterance, at the same time maximising its perceptual distinctiveness to other agents. At this stage, however, we have no direct way to quantify articulatory effort; hence, we address the problem by using the established ideas of dispersion-focalisation theory (working in the auditory domain as opposed to the articulatory domain), as just discussed. In grounding terms, the drive for perceptual distinctiveness is important in shaping the coupled production-perceptual system. The higher the perceptual distinctiveness, the clearer the meaning of the utterance. When the topological space of our self-organising maps is augmented with dispersion based on inter-vowel differences (in addition to focalisation based on intra-vowel attraction), we refer to it as a *contour space*. By introducing the proposed contour spaces, we hope to achieve a greater robustness to parameter variation and a greater level of realism in the vowel systems that are produced.

We now describe how a repulsive force acting on the perceptual neurons of the agent is introduced. For each node  $i$  of the auditory map, at time  $t$ , we define an energy functional given by

$$E(v_i(t), v_j(t)) = \sum_{\substack{j=1 \\ j \neq i}}^N \frac{1}{d_{ij}^2} \quad (2.1)$$

$$\text{where } d_{ij} = \sqrt{(F1_i - F1_j)^2 + (F2'_i - F2'_j)^2}$$

In equation (2.1),  $j$  is an index over all  $N$  nodes in the auditory map,  $v_i = (F1_i, F2'_i)$  and similarly  $v_j = (F1_j, F2'_j)$ . (See Appendix B.1.2 for discussion of  $F2'$ .) This amounts to a measure of distance between the  $i$  and  $j$  vowels in the  $F1$ - $F2'$  auditory-map space.

Updating occurs as follows. At time  $t$ , for each neuron  $i$  in the auditory space, we generate eight ‘test positions’ around that neuron. These are spaced on a rectangular grid of side  $\sigma$  centred on  $i$ . The update equation is:

$$v_i(t+1) = v_i(t) + \gamma v_{\max} \quad (2.2)$$

where  $v_{\max}$  is the  $v_k(t)$  vector for which the energy  $E(v_i(t), v_k(t))$  is maximised, with



$k$  being an index over the eight neighbours of  $v_i(t)$ , and  $\gamma$  is a step size or learning rate. Thus, maximisation is performed by gradient ascent. In this way, we are moving the  $i$ th vowel in the direction that maximises the acoustic distinctiveness between it and all other vowels in the space.

### 2.5.2 Attractive force: focalisation

The articulatory space is three-dimensional, defined in terms of lip rounding  $r$ , tongue position  $p$  and tongue height  $h$ . As previously discussed, focalisation in our model follows Schwartz et al. (1997) in seeking to favour vowels with compact  $F2$ - $F3$ - $F4$  formant patterns by defining and minimising an energy functional.

The specific energy functional used is similar to that of Schwartz et al. (1997) (see their equations (4) to (7)) modified to fit our simulations using a self-organising map:

$$E(v_l(t) = (r_l, p_l, h_l)) = E_{12} + E_{23} + E_{34} \quad (2.3)$$

$$\begin{aligned} \text{where } E_{12} &= - \left( \frac{1}{(F2_l - F1_l)^2} \right) \\ E_{23} &= - \left( \frac{1}{(F3_l - F2_l)^2} \right) \\ E_{34} &= - \left( \frac{1}{(F4_l - F3_l)^2} \right) \end{aligned}$$

In (2.3), each neuron  $l$  has its associated  $(r_l, p_l, h_l)$  values, which allow computation of formant values via the vocal tract model (Appendix B.1.1). At time  $t$ , each such neuron has its vector  $v_l(t)$  updated according to:

$$v_l(t+1) = v_l(t) + \gamma v_{\min} \quad (2.4)$$

where  $v_{\min}$  is the  $v_m(t)$  vector for which  $E(v_m(t))$  is minimised,  $m$  is an index over the 26 neighbours of  $v_l(t)$  (on a grid of size  $\sigma$  in 3-D space), and  $\gamma$  is a step size or learning rate. Hence, we are minimising by gradient descent.

Note that although this mechanism of attraction is firmly based in perception, we are in fact minimising in  $(r, p, h)$  space. Hence, we view this as, effectively, a mechanism for

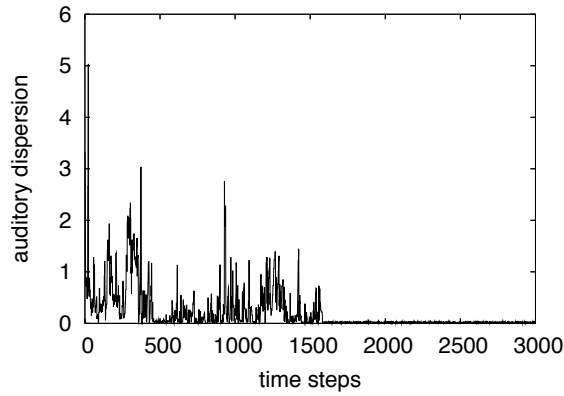


FIGURE 2.6: Typical plot of auditory dispersion versus number of iterations, showing convergence well before 2,000 steps.

reducing (if not actually minimising) articulatory effort in line with H&H theory.

## 2.6 Results of simulations

In this section, we first show some typical illustrative results obtained using Oudeyer's model to act as a benchmark before presenting typical results from the new model based on DFT. Thereafter, more thorough results (averaged over 500 runs) are given comparing the sensitivity of the two models to variation in the gaussian width parameter,  $\sigma$ . The two models are also compared with respect to the emergence of realistic vowel systems (i.e., their similarity to those observed in human languages). In all simulations, the nodes of the self-organising maps are initially randomised, that is, placed at uniformly-distributed positions in the appropriate space.

In these simulations, the optimisation step size,  $\gamma$  of equations (2.2) and (2.4), is set equal to the gaussian width,  $\sigma$  of equation (B.1) in Appendix B.1.3, enabling all three forces (i.e., dispersion, focalisation, self-organisation) to maintain their intended, relative level of influence. The gaussian width in the auditory space was scaled up to take account of the different range of the two maps  $[0, 1]^3$  for the motor map and 0..8 Bark, 0..15 Bark for the auditory map). All SOMs have 500 nodes, and simulations are stopped after 2,000 iterations of two-agent interaction. This stopping criterion was decided after examining how auditory dispersion (measured from the energy functional of eqn. (2.1)) varied during a few trials of the simulation. Figure 2.6 depicts a typical example. Although dispersion does not reduce monotonically, convergence is achieved well before 2,000 iterations.

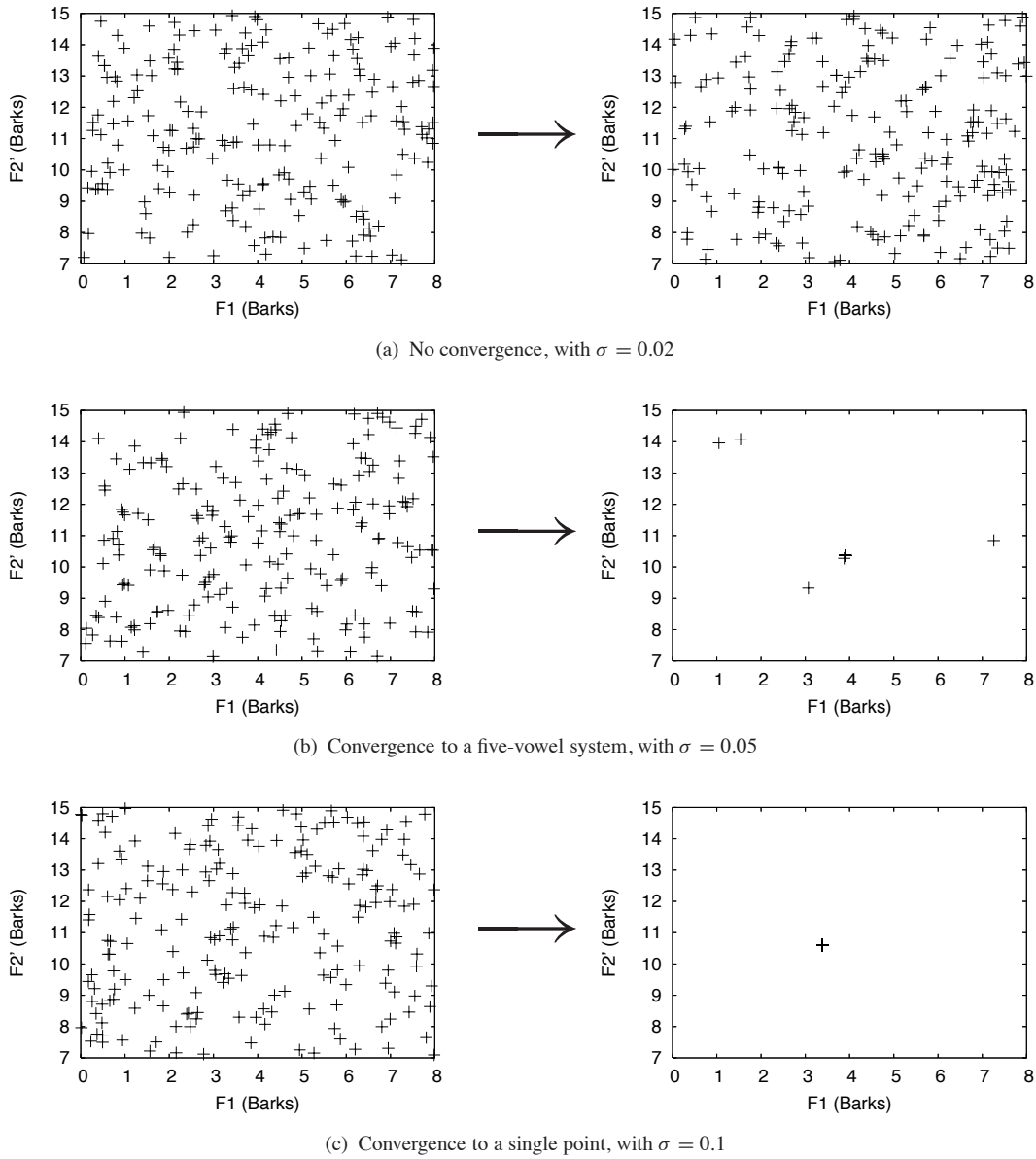


FIGURE 2.7: Composite of typical results from our replication of Oudeyer's simulation as  $\sigma$  varies.

### 2.6.1 Reproduction of Oudeyer's results

We have already shown an example of how the initial model can converge to a reasonable five-vowel system with  $\sigma = 0.05$  (Figure 2.5 earlier). Figure 2.7 shows a composite of typical results as  $\sigma$  varies. It is seen that realistic vowel systems emerge only for a restricted range of  $\sigma$  values.

## 2.6.2 Effect of the contour space

To assess the role of symbol grounding we will test the following null hypotheses, returning to them in Section 2.6.4:

*Hypothesis 2.6.1.* We hypothesise that the addition of dispersive and focalisation forces into the simulation will create vowel systems of a size more consistent with human vowel spaces. We will test this by comparing the distribution of vowel system size for many runs of the simulation against the distributions for human languages and for the baseline (Oudeyer 2005c) model.

*Hypothesis 2.6.2.* We state the null hypothesis that the introduction of signal grounding, a computational model of H&H theory, will have no effect on the observed accuracy and robustness of the self-organising vowel space simulations.

*Hypothesis 2.6.3.* The introduction of signal grounding, will increase the observed realism and robustness of the self-organising vowel space simulations. We will test this by varying the model parameters and measuring the realism of the resulting vowel systems.

Figure 2.8 shows a composite of typical results from simulations of the new model with contour spaces with the same  $\sigma$  values as in Figure 2.7. As can be clearly seen, realistic vowel systems emerge over a much wider range of  $\sigma$  values. There is also, we think, less tendency for the converged points to overlay exactly than in the original work (i.e., there is more of a ‘cluster’).

## 2.6.3 Further comparison of the two systems

To test further the assertion that the new system featuring dispersive forces (i.e., contour spaces) will possess a greater robustness to parameter variation than Oudeyer’s original, 500 repeated runs were made for different values of the gaussian width  $\sigma$ . The number of vowels present after convergence was then recorded for both systems. If convergence did not occur, results were discarded. Figure 2.9 shows the results averaged over the 500 runs; the error bars depict the standard deviation.

For the new system, a high level of variation in the number of vowels observed at convergence is seen across the whole range of  $\sigma$  values. We take this to be a positive feature of the new system, since human languages display a wide variety of vowel inventories (Maddieson 1984; Ladefoged and Maddieson 1996). By contrast, the Oudeyer (2005c) system (as replicated by us) shows unrealistic convergence to a single

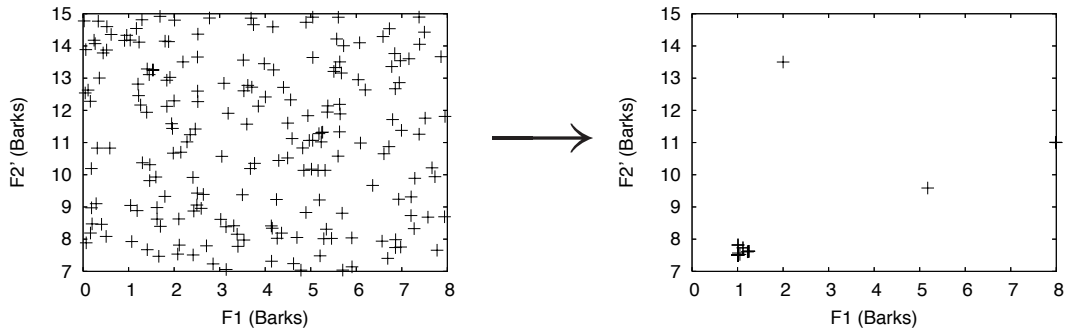
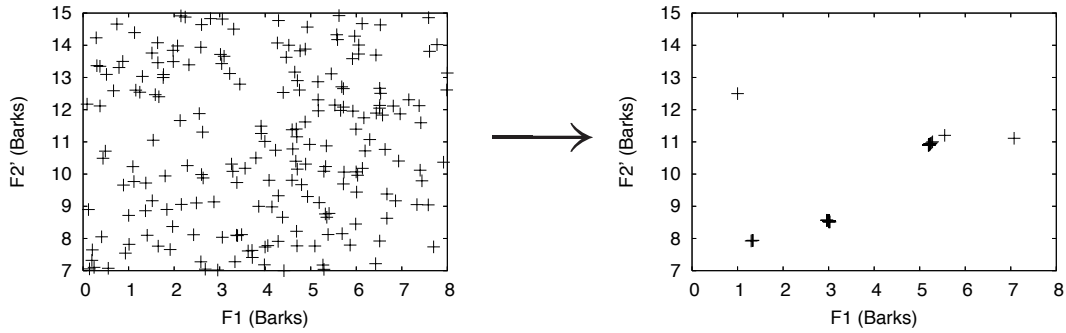
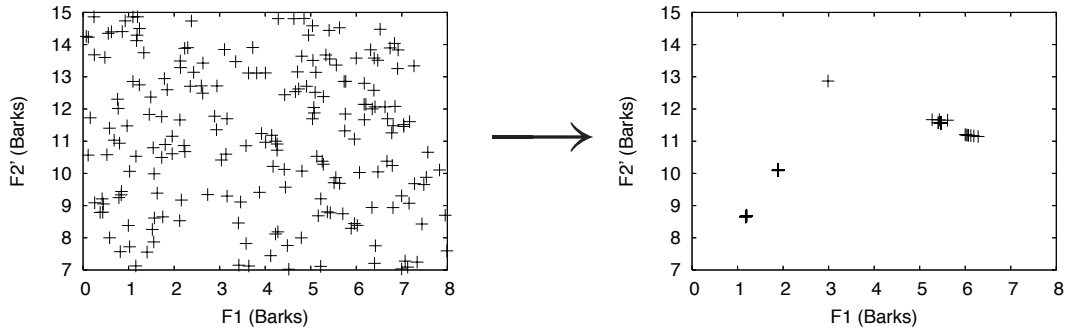
(a) Convergence to a four-vowel system, with  $\sigma = 0.02$ (b) Convergence to a five-vowel system, with  $\sigma = 0.05$ (c) Convergence to a five-vowel system, with  $\sigma = 0.1$ 

FIGURE 2.8: Composite of typical results from simulations of the new model with contour spaces with the same  $\sigma$  values as in Figure 2.7. Realistic vowel systems emerge over a much wider range of  $\sigma$  values.

‘vowel’ with zero variability for  $\sigma > 0.07$  and a total lack of convergence (to a sensibly small number of clusters) for  $\sigma < 0.05$ . Realistic convergence is maintained for the new system up to parameter values of 0.15. No simulations were performed for  $\sigma > 0.15$ .

Following Oudeyer (2005c, Figure 10, p.446), we have also compared the two systems with data for human languages, taking vowel frequencies from Ladefoged and Maddieson (1996). For the new computer model,  $\sigma$  was set to 0.05 and 500 simulations were run. Comparative data for Oudeyer’s system for the same value of  $\sigma$  and number of iterations were taken from his original paper, rather than the simulations being

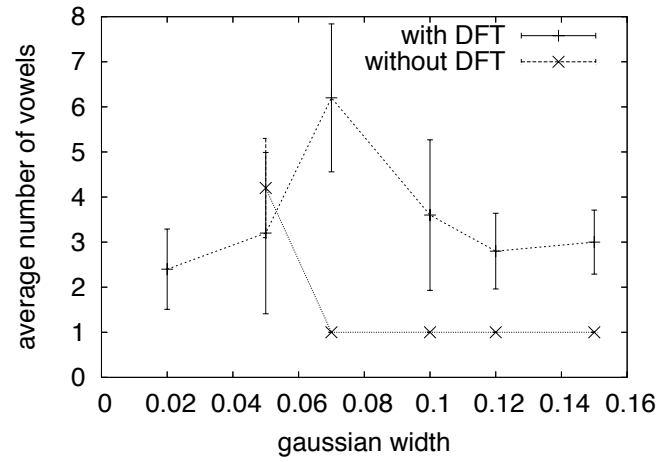


FIGURE 2.9: Comparison of our replication of Oudeyer’s simulation with the new model based on DFT, illustrating the robustness to parameter variation resulting from inclusion of a dispersive force. Error bars are standard deviations over 500 runs.

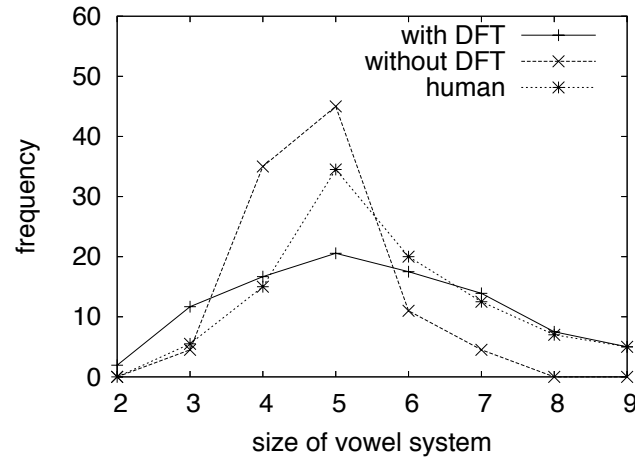


FIGURE 2.10: Comparison of vowel systems observed in human languages and those produced by computer simulation with and without DFT (i.e., with and without dispersive forces).

replicated here. Figure 2.10 shows the comparison, which reveals that the system with contour spaces has a slight preference for simpler vowel systems but is able to capture the emergence of the more complex systems, which is a problem for Oudeyer. Quantitatively, the mean square error (MSE) between the vowel frequency curve for Oudeyer’s data (labelled “without DFT”) and the human data is 91.28, whereas the corresponding MSE for our simulations (labelled “with DFT”) is 29.94. All three systems share a peak of five vowels. We emphasise that this comparison is made under conditions (namely,  $\sigma$  set at 0.05) which are maximally favourable to Oudeyer’s model. This is necessary because of the sensitivity of his model to the setting of  $\sigma$ .

## 2.6.4 Statistical hypothesis test

In hypothesis 2.6.1 we stated that: the introduction of signal grounding, will not reduce the Euclidean distance between simulated and real vowel space configurations. We find that the Euclidean distance is reduced from 60.5 to 28.39 when comparing the results of Oudeyer (2005c) (without signal grounding) with our own (with signal grounding). Unfortunately, the standard deviations were not provided with Oudeyer's results, consequently we were unable measure their significance.

In hypothesis 2.6.2 we stated that: the introduction of 'Signal Grounding', a computational model of H&H theory, will have no effect on the observed accuracy and robustness of the self-organising vowel space simulations. Defining accuracy as the total Euclidean distance between the simulations and human results we find that this distance is reduced from 21.8 to 11.2 summed over an adjustment in Gaussian width from 0.02 to 0.15 with a significance of  $\alpha = 0.043$  for a standard two tailed  $t$ -test, with mean and standard deviation values presented in Figure 2.9. We can therefore reject hypothesis 2.6.2 and support the alternative hypothesis 2.6.3.

## 2.7 Discussion and conclusions

The tension introduced by the addition of a dispersive force has clearly had a beneficial effect. This extension achieves an increased level of robustness to parameter variation and captures the emergence of some of the more complex vowel systems observed in human languages, in a way which Oudeyer (2005c) was unable to do. Despite a slight preference for the simpler vowel systems, the distribution is more representative of that seen in real languages, as confirmed by the much lower mean square error (see previous section).

How have these beneficial effects come about? Boë, Schwartz, and Vallée (1995) have already shown, although not in a multi-agent setting, how DFT can produce a range of vowel systems. (Rather, starting with a full set of vowel 'prototypes,' they show how DFT can be used to select realistic subsets typical of different languages.) In the present setting, the three forces of dispersion, focalisation and self-organisation act to produce convergence to attractors in the contour space. These attractors correspond to a physical grounding of the speech signals produced by the agents, as in Figure 2.2(b). The gradual, progressive nature of the convergence, over many interactions, ensures the final set of signal-grounded forms is shared among the population. So the physics

governing a population not only potentially accounts for a wide variety of human vowel systems but also allows for this set to become established within the population.

In our work, grounding of the external world is via these attractors in contour space. So, rather than connecting an arbitrary *a priori* abstraction (as when *cat* in the environment is miraculously labelled *cat* in one bound), we are connecting a more complete representation of the distal object, built on the physics of the situation. Through the formation of attractors, we have both a clear shared abstraction, its centre point, and a basin of attraction capturing the ambiguity and differences present in the real world. We feel that this view can answer some of the current criticisms of the symbol grounding paradigm (e.g., Lakoff 1993), just because the attractors capture the ambiguities and ‘shades of grey’ that challenge more traditional views of grounding (Davidsson 1993). This has similarities to previous work which has sought to explain grounding using connectionist models (e.g., Harnad 1993; Cangelosi, Greco, and Harnad 2000; Damper and Harnad 2000). These have been successful in displaying various aspects of human cognition. But, by considering grounding at the (sub-form) level of physical signals (Figures 2.2(b) and 2.3), we have developed a new framework in which this interplay between symbol grounding and connectionist systems can be further explored.

At present, agents do not exactly ‘hear’ sounds; rather, they have direct access to formant values. From  $F1$ ,  $F2$ ,  $F3$  and  $F4$  values specifying a vocalisation, they perceive  $F1$  directly and compute a perceived  $F2'$ . This is a very high level of abstraction, implicitly making many assumptions (e.g., about the role of formants in speech perception, and how the auditory system can extract them from the speech signal). First and foremost, therefore, this work must move to using actual sounds as the medium of interchange between agents. This move will make it necessary to use more physically realistic vocal tract and cochlear models and these will be detailed in Chapters 3 and 4 respectively. It is then a matter of some importance and interest to investigate how much increased realism/complexity impacts on the emergence of sound systems. We know from Oudeyer (2005c) and the present work that very simple, highly abstract models are adequate for the production of shared (static) vowel systems, but under rather strong assumptions. Furthermore, speech sounds do not consist entirely of vowels, but of dynamic consonant-vowel patterns forming syllables. Unfortunately, although there is general agreement among phoneticians and speech scientists that vowels can be reasonably well specified by formant values, there is no corresponding understanding of how consonant sounds can be similarly specified and distinguished.

Although Oudeyer (2005b) has extended his “abstract” linear model in the direction of “the formation of ... and patterns of sound combination” (p.328), this is done



without any acoustic, perceptual space, but with agents given direct access to the relevant parameters in what we believe to be an unsatisfactory (‘mind-reading’) manner. By moving to simulations in which actual, physical speech sounds are exchanged between agents, this thesis will now explore the emergence of speech as a dynamic phenomenon in a more realistic and satisfactory way by first developing a direct measure of articulatory effort.

# Chapter 3

## Direct measure of articulatory effort

Historically there have been a number of computational models that attempt to account for the formation of discrete vowel inventories in human speech. Often these models are loosely based upon H&H theory, as detailed in Chapter 1, where speakers seek to achieve an optimum between maximum auditory distinctiveness and minimum articulatory effort. Given the importance of energy conservation in this theory, it is surprising that no direct measure of vocal tract muscular effort has yet been modelled. Instead, current work attempts to capture the forces behind vowel formation within the auditory domain alone, e.g., dispersion-focalisation theory. We believe that without a direct measure of articulatory effort such theories will remain incomplete. This chapter seeks to rectify this oversight by adding a muscle model to an existing articulatory vocal tract simulation. Equipped with this system, we advance dispersion-focalisation theory by integrating articulatory effort into this existing approach. Our results show that the direct modelling of articulatory effort results in a wide range of plausible vowel systems. Furthermore, it illustrates two subtleties within H&H theory. First, auditory dispersion and articulatory effort are not necessarily in direct opposition to each other; certain key vowels can produce an optimisation of both forces and secondly that articulatory optimisation is possible within each vowel's corresponding auditory region. Having established the validity of this muscle model the way is now clear to study the role of hyper- and hypo-articulation in real speech.

### 3.1 Overview

Despite a highly flexible vocal tract, human vowel systems appear to be selected from only a limited range of configurations. To account for this, there have been a number

of investigations into the forces behind the formation and continuing cultural evolution of phonetic tokens. Accordingly, researchers have proposed a number of explanatory accounts, for example Stevens' quantal theory of speech, where articulatory stable acoustic regions are favoured by speakers, or Lindblom's H&H theory.

Comparing these two theories, we see that in quantal theory (Stevens 1978, 1981, 2002) speakers are driven towards regions of acoustic stability. Specifically, these regions ensure that small changes in articulatory configurations are not reflected in the speech signal, leading to a robustness in the final utterance. Theoretically, these regions correspond to areas of specifying invariance allowing for the perception of acoustic cues.

Historically, these theories have been tested in a variety of ways and means and one useful contribution has been from the field of computational modelling (Liljencrantz and Lindblom 1972; de Boer 2000a, 2003; Oudeyer 2005c). Using this approach, we can simulate the forces behind a theory and observe the resulting phonetic system, validating or refuting the initial account. In this chapter we will be continuing this tradition by constructing a complete model of H&H theory that captures both auditory distinctiveness and articulatory effort. The minimisation of effort is a crucial part of H&H theory. Accordingly, this chapter will address this oversight.

Previous work has neglected the articulatory domain in favour of the auditory domain. For example, Schwartz et al.'s (1997) dispersion-focalisation theory concentrates on the auditory dispersion and the acoustic stability (or focalisation) of the utterance to account for the formation of a variety of vowel systems. Focalisation captures the fact that "convergence in the  $F_2$ - $F_3$ - $F_4$  pattern produced more stable patterns in discrimination experiments" (p.259), this lead to the conclusion that utterances with high focalisation would have greater acoustic salience, and accordingly, robustness to interference. By varying the emphasis on dispersion and stability a variety of vowel configurations emerge. We argue that although this is a reasonable first step, the role of articulation cannot be discounted.

In the work of Schwartz et al., analysis of the speech signal is used as a replacement for Lindblom, MacNeilage, and Studdert-Kennedy's notion of articulatory effort. In addition to abstracting away from this fundamental motivating force, both quantal theory (Stevens 2002) and Lindblom's (1998) own work also hypothesise that there are regions of acoustic output that remain stable despite articulatory variation providing regions of articulatory optimisation distinct from the auditory signal. Therefore, to complement signal-based measurements of stability we require an accurate measure of articulatory effort to capture three key phenomena: the weighting between effort and

distinctiveness, articulatory optimisation within quantal regions and the discovery of local optima where these two forces are not necessarily in opposition to each other. We will now consider each of these points in turn.

### 3.1.1 Weighting effort and distinctiveness

Schwartz et al. contend that the variety of observed human vowel systems could have arisen through different weightings on dispersion and focalisation. Varying the emphasis on one or the other leads to either an increasingly dispersed or increasingly compact solution. We will take a similar approach and assess the effect of different weightings on effort and dispersion. In doing so, however, we are mindful that we are abstracting away from a complex interplay between cultural and individual timescales. Aside from the glossogenetic fixing of parameters, the individual themselves will place varying emphasis on effort and dispersion in their daily use of language, compensating for ambient noise and the perceptual abilities of the listener. Abstracting away from the cumulative effect of these day-to-day decisions is an unfortunate reality of this simulation.

### 3.1.2 Optimising quantal regions

In accordance with Stevens (2002), we agree that there are quantal, acoustically stable regions but we believe that the interplay between perceptual contrast and articulatory effort forms the main motivation for the creation of vowel inventories. This is in opposition to the tenet of quantal theory, which postulates acoustic stability as the main driving force. We propose that within these acoustically stable regions there remains a motivation to minimise articulatory effort and the presence of these quantal regions drives the theoretical need for an anatomical, articulatory effort function. By constructing a system that directly measures the effort of vocal tract motion, we hope that our new model of speech sound emergence will obtain a greater degree of realism as we are no longer inferring effort indirectly from the auditory signal.

Previous attempts to capture articulatory effort have been hampered by an inability to simulate the effort expended by human muscle. However, thanks to a combination of research (Cook 1993; Umberger, Karin, and Philip 2003; Lucero, Maciel, Johns, and Munhall 2005) developing muscle models and mapping out the structure of the vocal tract and facial musculature, the time is right to make an initial attempt at the simulation of articulatory effort. Once this approach has been established we can explore a number

of previously inaccessible subtleties within H&H theory, for example co-articulation and syllable formation.

### 3.1.3 Finding local optima

At the individual timescale, there is an implicit assumption that the two forces of H&H theory are in direct opposition to each other. In the daily use of a language, speakers compensate for ambient noise by over-articulating, requiring greater effort to enable the perception of vowels that have been already established by cultural evolution. However, when we consider the cultural timescale, the configuration of the vowel space becomes pliable. Freed from the necessity of over-articulating fixed vowels, the two forces of H&H theory are no longer in direct opposition to each other. The ability to reconfigure the vowel space enables the discovery of candidate vowels that satisfy the requirements of both effort and dispersion. We propose that by finding the same vowels across multiple configurations, we can identify the regions of local optima that satisfy these constraints.

## 3.2 Capturing effort and dispersion

This chapter's computational model consists of two parts: a configuration of muscles attached to the articulatory parameters of the vocal tract and a neural model to enable the vocal tract to produce the desired formant values for each candidate vowel. By moving the vocal tract from a neutral position into the required configuration, a measure of effort can be obtained.

### 3.2.1 Hill's muscle model

To develop a measure of articulatory effort, we will use the muscle model of Hill (1938). Widely considered as the most accurate model of energy expended by skeletal muscle, it captures a number of essential features. These include an accurate measure of the heat of shortening, the role of the fast twitch and slow twitch muscle fibres and, of course, the muscular composition of the vocal tract.

These muscle models have been used in a variety of work ranging from estimating dinosaur running speeds (Sellers and Manning 2007) to simulations of bipedal locomotion (Nagano, Umberger, Marzke, and Gerritsen 2005). There have also been

a number of adaptations and improvements to the Hill muscle model (Lichtwark and Wilson 2005) but we feel that our current level of abstraction, derived from Umberger, Karin, and Philip (2003), is sufficient for this initial work. For details of Umberger, Karin, and Philip's muscle model, see Appendix B.3.

To proceed we need to capture the 'stylised' physical constraints of the vocal tract. Any physically possible articulatory gesture should be possible in our model. At the same time, any physically impossible gesture needs to be disallowed.

One way of doing this is to create an anatomically accurate model of the vocal tract. When this model produces utterances, there are no preconceptions about their nature and any resulting sounds that occur through the manipulation of various vocal tract components. This gives us two advantages; first, there are no signal abstractions to undermine the model; and second, real speech-like utterances are produced, ensuring that the validity of any associated auditory system can be maintained.

Accordingly, a number of existing vocal tract models become worthy of consideration (Mermelstein 1973; Maeda 1982b; Maeda 1982a; Cook 1993; de Boer 1999). Cook's model is capable of producing a wide range of utterances; examples can be heard at <http://www.dcs.shef.ac.uk/~simon/samples/>. It forms a unified model of both the vocal tract and nasal tract idealised as an abstract set of filters. Eight control parameters then adjust the vocal tract shape to determine the acoustic output from a glottal pulse. Similar articulatory models have been used before in the study of emergent phonology, e.g., Oudeyer (2002) has used this approach to model the emergence of syllable systems within a multi-agent framework. We will be taking a similar agent-system approach in later chapters but will be focussing on specific phonetic theories instead.

We should not delude ourselves that these models represent a perfect reproduction of the human vocal tract. Faundez-Zanuy and McLaughlin (2002) illustrate a number of problems when defining an accurate model of the airflow within the vocal tract and this leads to a call for greater accuracy in synthesis. However, to be useful, a balance has to be struck between accuracy and computational efficiency.

Previous work (Sanguineti, Laboissiere, and Ostry 1998) details the various muscle groups and their role in speech production, see Table 3.1. These groups have been linked to the control parameters of the existing vocal tract model. Missing values that are required to calculate muscle energy expenditure have then been obtained from other sources (Sanguineti, Laboissiere, and Payan 1997), whereas the muscular properties of the lips were obtained from Lucero and Munhall (1999). By coupling this new muscle

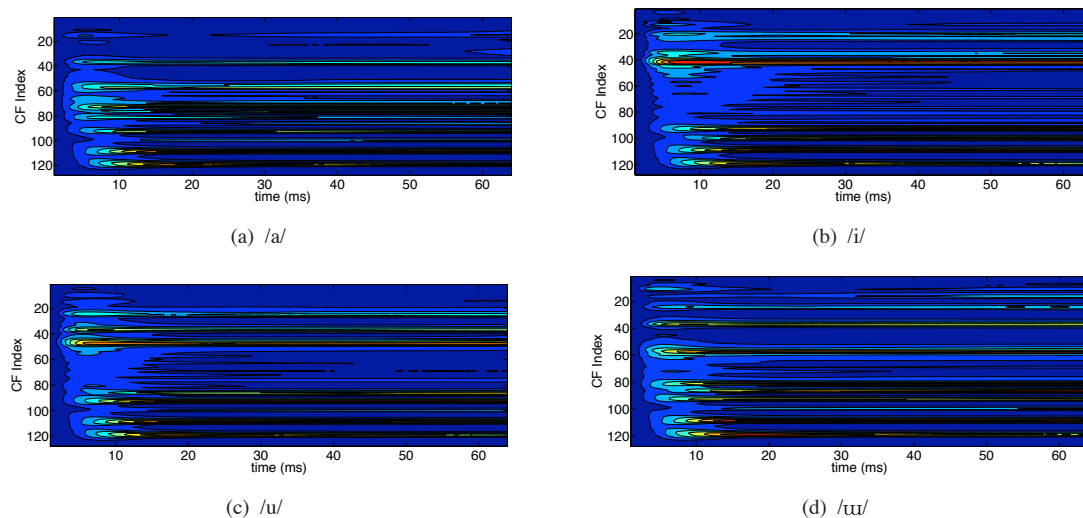


FIGURE 3.1: Auditory spectrograms of a sample of vowels produced by a combination of the articulatory vocal tract and Guenther’s neural model, with CF corresponding to a particular ‘centre’ frequency. In each case the model has captured the appropriate formant values.

model with our artificial vocal tract, we can determine the muscular effort required to generate each phoneme from a neutral initial position.

### 3.2.2 Guenther’s neural model

Reproduced from Guenther (1995), the abilities of the controlling neural model have been demonstrated in a number of publications (Guenther and Gjaja 1996; Guenther 2004). After a babbling phase in which a phonetic-to-orosensory mapping is acquired, the system then learns the appropriate vocal tract configuration for each candidate vowel in the auditory space. In Figure 3.1 we present a sample of the vowels produced by the vocal tract after training on the auditory prototypes presented in Figure 3.2. In training we followed the details in Guenther (1995). In accordance with Guenther’s work the formant values for each vowel were obtained successfully, forming valid articulatory prototypes for our effort calculations. Each of the vowels in Figure 3.1 has been plotted on a spectrogram produced by Lyon’s cochlear model, the details of which can be found in Chapter 4.

Control Parameter	Muscle	$F_{\max}$ (N)	$\sigma$ (N/m)	$m$ (kg)	Action
Tongue body	Genioglossus	67.8	0	0.1	Depresses the tongue
	Hyoglossus	65.1	0	0.1	Depresses the tongue
	Styloglossus	24.2	0	0.1	Retracts and elevates the tongue
Tongue tip	Superior longitudinalis	14.3	0	0	Raises the tip
	Inferior longitudinalis	19.4	0	0	Lowers the tip
	Verticalis	14.5	0	0	Flattens the tip
Jaw	Jaw opener	115	34.7	1	Lowers jaw
	Jaw closer	639	192	1	Raises jaw
Hyoid	Mylohyoid	40.9	32.2	0.1	Raises hyoid
Hyoid/larynx	Thyrohyoid	28.7	8.65	0.1	Lowers hyoid/elevates larynx
	Sternohyoid	28.7	8.65	0.1	Lowers hyoid and larynx
	Sternothyroid	28.7	8.65	0.1	Lowers larynx
Lip opening/protusion	Zygomatic major	0.000173	0.000173	0	Raises upper lip
	Levator labii superioris	0.00038925	0.0002595	0	Raises upper lip
	Depressor anguli oris	0.002768	0.0006920	0	Lowers bottom lip
	Depressor labii inferioris	0.00020933	0.0001903	0	Lowers bottom lip
	Mentalis	0.00008477	0.0001211	0	Protudes Lower lip
	Levator anguli oris	0.0001730	0.0001730	0	Raises corners of upper Lip
	Orbicularis oris superior	0.006228	0.0010380	0	Protudes lips
	Orbicularis oris inferior	0.006228	0.0010380	0	Protudes lips

TABLE 3.1: The various muscles of the vocal tract and their role in speech production.  $F_{max}$  represents the maximum exerted force,  $\sigma$  the tension of the muscle and  $m$  the mass. Values were obtained from (Sanguineti, Laboissiere, and Ostry 1998) and (Lucero and Munhall 1999).



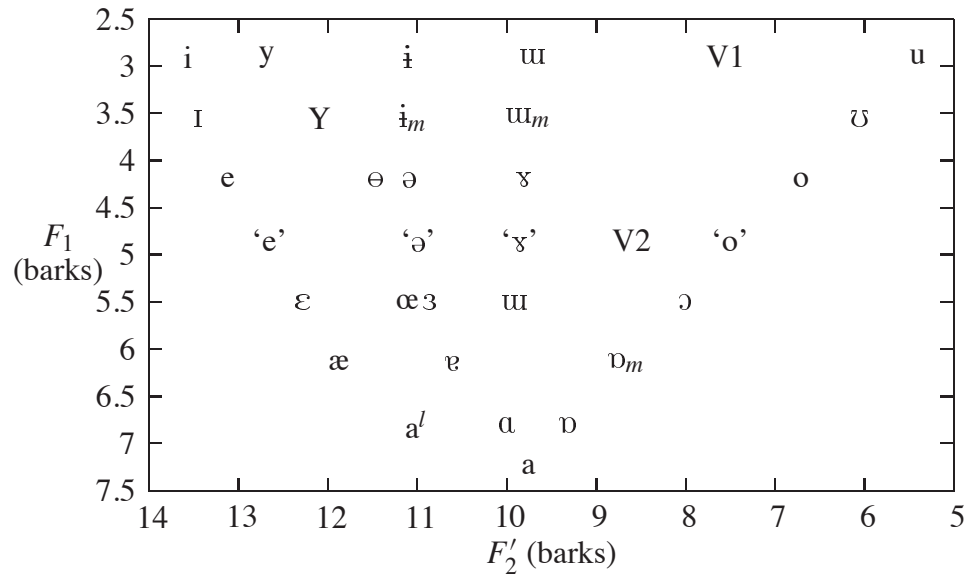


FIGURE 3.2: Plotted from data obtained from Schwartz et al., a wide number of candidates provide coverage of the space of possible vowels. Expanding on data from the IPA, candidate auditory prototypes were selected to provide adequate coverage of the vowel space.

### 3.3 Applying the model to H&H theory

Having constructed the system, we will now test whether it is suitable for inclusion in our final agent system. For this to be the case, our muscle model must capture the effort of production for each articulatory prototype. Due to the novelty of this approach we are unable to make a direct comparison to recorded effort measurements. Accordingly, we will conduct two indirect tests. In Section 3.3.2, we will compare the simulated vowel systems to recorded human data and in Section 3.3.1 we will measure the effect on the vowel space of an increasing emphasis on effort. Accordingly, we will first adjust the weighting on effort and dispersion and judge the extent to which plausible vowel systems emerge by judging them against the observed human vowel systems. Second, we will directly compare the effects of focalisation and effort by measuring the total Euclidean distance among all vowels within the vowel space under different parameter settings. If the Euclidean distance reduces with greater emphasis on effort it is clear that we have captured some aspect of the opposing forces within H&H theory.

#### 3.3.1 Developing emergent vowel systems

Having established the need for a new approach to the modelling of articulatory constraints and by constructing a muscle model to extend the selected vocal tract, we

can now seek to study directly the conflict between maximum auditory distinctiveness and minimum articulatory effort. In testing this model, we will establish optimal vowel systems, for a lexicon of a given size. This will allow for a direct comparison with human vowel systems and previous modelling approaches.

Defining each vowel according to its formant values, we can see in Figure 3.2 a number of auditory prototypes that provide complete coverage of the vowel triangle. These auditory prototypes then form the training data for the articulatory prototypes that we then use to calculate the effort of production. Having been reproduced from Schwartz et al., some of the notation is inconsistent with the IPA and certain vowels (V1, V2) cannot be found in normal speech but have been created for the sake of completeness. Then by generating the effort of production for each, we have a number of candidates that can be used to form a range of vowel systems. The auditory distance between each pair of these candidates  $V^i(F1^i, F2^i, F3^i)$  and  $V^j(F1^j, F2^j, F3^j)$  is determined, according to Schwartz et al., by:

$$d_{ij} = \left[ (F1^i - F1^j)^2 + \lambda^2 (F2'^i - F2'^j)^2 \right]^{\frac{1}{2}} \quad (3.1)$$

where  $\lambda$  represents an adjustable weight in the model and  $F2'$  is computed from the method put forward by Carlson, Fant, and Granström (1975). The final systems are then formed by altering the emphasis on auditory dispersion and articulatory effort by adjusting  $\lambda$  and  $\alpha$ , according to:

$$E_{DF} = E_D + \alpha E_E \quad (3.2)$$

with  $0 \leq \lambda \leq 1$  weighting the emphasis on the dispersion term and  $0 \leq \alpha \leq 1$  defining the effort term<sup>1</sup>. Dispersion is then calculated according to:

$$E_D = \sum_{i=1}^{n-1} \sum_{j=(i+1)}^n \left( \frac{1}{d_{ij}} \right)^2 \quad (3.3)$$

whereas articulatory effort is generated by the articulatory muscle model.

By directly measuring the effort of producing each phoneme from a neutral vocal tract position, we can see the relative effort of each. We observe that with the exception

---

<sup>1</sup> Although  $E_{DF} = (1 - \alpha)E_D + \alpha E_E$  might be more appropriate, we decided to maintain consistency with Schwartz et al.'s original work

Vowel	Formant frequencies in Bark					$E_F$	$E_E$
	$F_1$	$F'_2$	$F_3$	$F_4$	$F'_2$		
i	2.90	13.56	15.82	16.81	16.15	1225	10.5
y	2.90	12.68	13.63	16.81	13.00	1217	3.1
ɨ	2.90	11.10	13.86	16.81	11.85	261	7.7
ʉ	2.90	9.70	14.40	16.81	9.70	239	19.9
V1	2.90	7.55	14.29	16.81	7.55	226	10.4
u	2.90	5.40	14.18	16.81	5.40	318	2.7
ɪ	3.55	13.44	14.82	16.81	13.90	788	9.0
Y	3.55	12.09	13.63	16.81	12.60	534	9.2
ɪ <sub>m</sub>	3.55	11.04	14.09	16.81	11.60	261	9.0
ʉ <sub>m</sub>	3.55	9.75	14.40	16.81	9.75	239	19.9
ʊ	3.55	6.05	14.16	16.81	6.05	318	18.1
e	4.20	13.11	14.58	16.81	13.60	676	9.7
ə	4.20	11.46	13.69	16.81	12.20	323	3.8
ə	4.20	11.08	14.40	16.81	11.35	284	6.7
ɤ	4.20	9.80	14.40	16.81	9.80	251	13.4
o	4.20	6.70	14.14	16.81	6.70	318	13.9
‘e’	4.85	12.65	14.61	16.81	13.30	483	5.7
‘ə’	4.85	10.98	14.41	16.81	11.10	285	1.1
‘ɤ’	4.85	9.85	14.40	16.81	9.85	260	4.5
V2	4.85	8.60	14.26	16.81	8.60	256	7.2
‘o’	4.85	8.60	14.26	16.81	8.60	318	12.0
ɛ	5.50	12.27	14.47	16.81	13.00	411	7.0
œ	5.50	11.10	14.40	16.81	11.40	296	7.2
ɜ	5.50	10.85	14.40	16.81	10.85	286	12.0
ʉ	5.50	9.90	14.40	16.81	9.90	273	5.4
ɔ	5.50	8.00	14.04	16.81	8.00	318	9.9
æ	6.15	11.87	14.37	16.81	12.70	359	6.4
ɐ	6.15	10.60	14.40	16.81	10.60	292	8.2
ɐ <sub>m</sub>	6.15	8.65	14.40	16.81	8.65	318	5.0
a <sup>l</sup>	6.80	11.00	14.50	16.81	11.00	326	13.5
ɑ	6.80	10.00	14.40	16.81	10.00	321	8.6
ɒ	6.80	9.30	13.77	16.81	9.30	318	5.0
a	7.25	9.75	14.40	16.81	9.75	336	5.6

TABLE 3.2: Vowel acoustic prototypes. Formats  $F_1$ ,  $F_2$ ,  $F_3$ ,  $F_4$  in Hz and in Bark, second perceptual formant  $F'_2$  in Bark, and individual focalisation costs  $E_F$  for  $\alpha = 1$ .

See Schwartz et al. (1997) for details of vowels not from the IPA.

of ‘schwa’ and /o/, the central vowels require a greater degree of exertion to produce. Similarly, the vowels with higher  $F'_2$  require less effort than vowels with lower  $F'_2$ . Accordingly, we would expect the emergent vowel systems to favour peripheral vowels even without an emphasis on auditory dispersion. The one anomaly that would lead to a preference for a more central vowel is where /i/ requires slightly more effort than /y/.

Table 3.2 displays the formant values generated by the vocal tract, the individual focalisation costs,  $E_F$ , and the individual energy costs,  $E_E$ . Taken from Boë, Schwartz, and Vallée (1995), these formant values form the targets that the vocal tract needs to produce from a neutral initial vocal tract configuration. Using the neural network detailed in Guenter, Husain, Cohen, and Shin-Cunningham (1999), the model acquired the mapping between articulatory gesture and resulting formant frequencies. Having reproduced these vowel sounds using our vocal tract model, we were able to calculate the required effort and by changing the emphasis placed on articulatory effort and auditory distinctiveness we caused various vowel systems to emerge.

Systematically, we can use this variation to test the following null hypotheses.

*Hypothesis 3.3.1.* We hypothesise that the introduction of a model of articulatory effort will not improve the realism of the observed vowel spaces. Specifically, we will test this by comparing the configuration of the resulting vowel spaces with and without the addition of articulatory effort.

*Hypothesis 3.3.2.* To test the validity of H&H theory we hypothesis that the introduction of a model of articulatory effort will not reduce the level of auditory dispersion. This will allow for further study of the interactions between articulatory effort and auditory distinctiveness.

*Hypothesis 3.3.3.* The introduction of a model of articulatory effort will not introduce a number of persistent local optima across vowel configurations. The presence of local optima would suggest that articulatory effort and auditory dispersion are not necessarily in direct opposition to each other.

By varying the parameters governing the relative weight of effort and dispersion, in the range [0–1], we can compare the frequency of occurrence of artificial vowel spaces with human vowel spaces. Because of some extreme vowel system configurations, for example, extreme emphasis on effort instead of dispersion or *vice versa*, coupled with effort being measured without consideration of co-articulation within a syllable system, we did not expect our emergent vowel systems to match perfectly with observed human vowel systems. However, the model does display certain similarities to human vowel inventories. For example, the two most frequent artificial seven-vowel systems match the first and second most frequent human vowel systems.


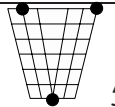
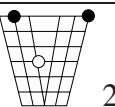
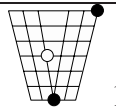

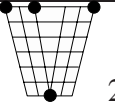

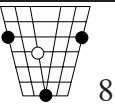
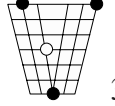
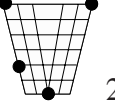
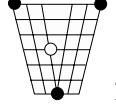
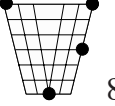
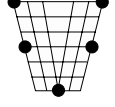
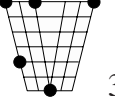
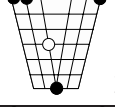
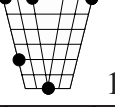
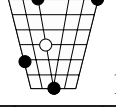
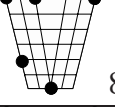
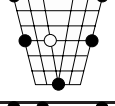
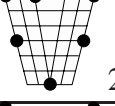
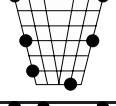
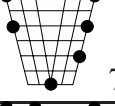
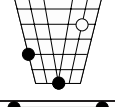
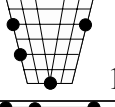
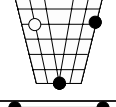
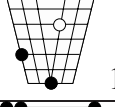
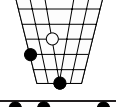
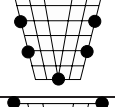
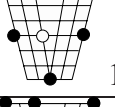
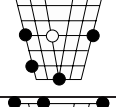
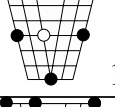
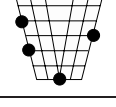
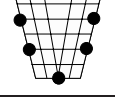
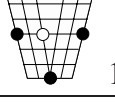
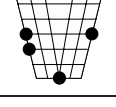
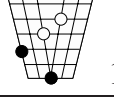
Number of Vowels		Common Vowel Systems and Frequency of Occurrences (%)				
3	Human	 100				
	Simulation	 57	 26	 17		
4	Human	 56	 20	 16	 8	
	Simulation	 32	 28	 24	 8	
5	Human	 97	 3			
	Simulation	 32	 12	 12	 8	
6	Human	 48	 22	 22	 7	
	Simulation	 36	 12	 12	 12	 12
7	Human	 53	 14	 12	 10	 7
	Simulation	 20	 16	 16	 12	

TABLE 3.3: A comparison of human and simulated vowel systems, obtained from the UCLA Phonological Segment Inventory Database (UPSID). Each of these simplified vowel spaces corresponds to the auditory prototypes presented in Figure 3.2. For clarity, central vowels are represented as white dots and vowels on the edge are black.

Clearly there are other factors that determine vowel formation and, as with any model, the difference between abstraction and reality is apparent. That is not to say that this model does not serve any purpose. By capturing the fundamentals of H&H theory, we have illustrated its plausibility as an account of vowel system formation. Furthermore, it is clear that even with a high emphasis on minimising articulatory effort, systems with a high level of auditory dispersion still form suggesting the presence of local optima where effort and dispersion are not in direct opposition. For example, certain vowels on the periphery of the auditory triangle require a relatively low energy to produce.

Looking in detail at Table 3.3 and Figure 3.2, certain features are worthy of note. We have to ask why in the fairly simple three-vowel system /i/ is consistently ignored in favour of /y/ in our model. As /y/ requires less effort to produce, as shown in Table 3.3, we would expect it to be favoured in human vowel systems. This could be caused by a number of factors that are removed in the abstraction of our model. For example, we have not considered the articulatory effort of our vowel systems under conditions of co-articulation. Additionally the notion of acoustic or articulatory robustness has not been taken into account. It could be that /i/ can be produced and perceived consistently under a variety of conditions, whereas /y/ is more likely to deteriorate when faced with a noisy environment.

Clearly certain features, such as the clustering in the top left of more complex vowel systems, are a result of effort optimisation taking precedence over auditory dispersion. It is, however, interesting to note the extent to which this happens in both human and simulated systems. It seems that the introduction of central vowels, for example in the five and six vowel system, can cause the peripheral vowels to be pushed into the corner to maximise the level of auditory distinction. This ‘central vowel pressure’ is absent in the simpler vowel systems causing the final configuration to drift from /i/ to /y/ as the considered optimum, suggesting that in many cases auditory dispersion takes precedence over articulatory effort.

In hypothesis 3.3.1 we stated the null position that: the introduction of a model of articulatory effort will not improve the realism of the observed vowel spaces. To test this statement we have defined realism as the Euclidean distance between the simulated and human vowel spaces. We find that the introduction of an effort factor to DFT reduces the Euclidean distance from 4.67 to 4.09, averaged across  $F_1$  and  $F'_2$  and all weightings of the various simulated forces.

In hypothesis 3.3.2 we stated that: the introduction of a model of articulatory effort will not reduce the level of auditory dispersion. We have defined auditory dispersion as the level of variance within the auditory space and found that the introduction of an effort

factor reduced the level of variance from 8.95 to 6.99.

In hypothesis 3.3.3 we stated that the: introduction of a model of articulatory effort will not introduce a number of persistent local optima across vowel configurations. Analysis of the data reveals that the total Euclidean distance, between vowel spaces across parameter variation, is reduced from 641.9 to 607.3 when an effort function is introduced. Additionally, subjective assessment of the data reveals that certain vowels, e.g., /y/ persist as the weightings on various parameters change.

In all cases no variance measures were taken and no significance tests performed as the model is deterministic with respect to parameter variation. In spite of this, we believe it is reasonable to reject hull hypotheses 3.3.1 and 3.3.2.

### 3.3.2 Comparing effort and focalisation

To compare effort and focalisation, we define focalisation as follows:

$$E_F = \beta(E_{12} + E_{23} + E_{34}) \quad (3.4)$$

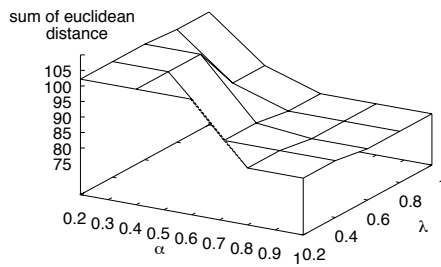
$$E_{12} = \sum_{i=1} \frac{1}{(F_2^i - F_1^i)^2} \quad (3.5)$$

$$E_{23} = \sum_{i=1} \frac{1}{(F_3^i - F_2^i)^2} \quad (3.6)$$

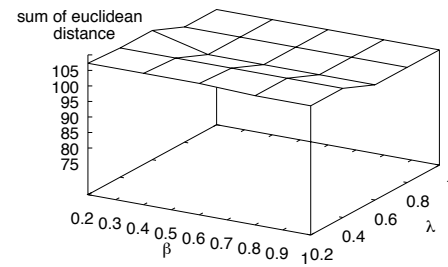
$$E_{34} = \sum_{i=1} \frac{1}{(F_4^i - F_3^i)^2} \quad (3.7)$$

where  $\beta$  determines the level of emphasis on focalisation and  $i$  represents each candidate vowel in the space.

In Figure 3.3 we present the resulting Euclidean distance within a vowel space under different parameter configurations in order to gain a clearer picture of the forces at work within H&H theory. For example, Figure 3.3(a) shows that an increased emphasis on articulatory effort causes a decrease in the dispersion of the final vowel systems. This effect seems to disappear when our effort function is replaced by Schwartz et al.'s focalisation function, i.e., Figure 3.3(b). Note that this effect is not dramatic, when all of the emphasis is placed on minimising effort, the level of dispersion does not drop to zero. This is because, within vowel systems, dispersion and effort are not necessarily directly opposed to each other and it is clearly possible to find an optimum



(a) The effect of effort and dispersion on vowel inventories ( $\alpha$  represents effort and  $\lambda$  represents dispersion).



(b) The effect of dispersion and focalisation ( $\beta$  represents focalisation and  $\lambda$  represents dispersion).

FIGURE 3.3: Both figures show the effect of varying parameters ( $\alpha, \lambda, \beta$ ) on the resulting Euclidean distance within the vowel space – defined as the total distance between vowels. Clearly an increasing level of effort leads to a reduced total euclidean distance between vowels within the auditory space.

point for both. This suggests that other forces are at play to account for the level of variation amongst human vowel systems. Potentially, the role of co-articulation and the interaction of an entire phonetic lexicon can account for the variation witnessed within vowel systems. Now that we are equipped with this plausible articulatory model, we are able to investigate these phenomena in future work.

A slight drop in distance as the level of dispersion increases can be seen with or without the focalisation parameter in Figure 3.3(b). However, in both cases it is not large enough to be significant. This suggests that although the focalisation of a vowel might have a role to play within the auditory system, it is not useful when attempting to infer the effort of an utterance, justifying the need for a direct measure of articulatory effort when constructing these models.

### 3.4 Summary

In this chapter, we have argued for the necessity of a direct simulation of articulatory effort when modelling phonetic theories like H&H. We have constructed a plausible vocal tract muscle model based upon established work from existing muscle models, and facial modelling. The utility of this model has then been underlined by the emergence of a number of plausible vowel systems. These have raised a number of interesting points within H&H theory and illustrated the fact that only by combining a direct measure of articulatory effort with a model of an entire phonetic inventory can a



complete account of H&H theory be rendered.

This complete account shows that articulatory effort and auditory dispersion are not necessarily in direct opposition to each other. Given that the resulting vowel systems do not entirely match existing human systems, we can also conclude that other factors are at work in their final formation.

We believe that the work reported in this chapter has captured two key features. First, we have demonstrated that this articulatory model is capable of producing speech and calculating the effort behind each utterance. This will become crucial in later chapters. Second, the persistence of certain key vowels across a variety of configurations suggests the presence of local optima within the auditory/articulatory space defined by H&H theory.

The decrease in the auditory distance of the vowel system given an increasing weighting of articulatory effort and the similarity to the DFT results (Schwartz et al. 1997) establishes that our method of an anatomical vocal tract coupled with a muscle model is a valid approach, suitable for further experimentation. We are constrained only by what is physically possible and only motivated by the desire to reduce articulatory effort in accordance with H&H theory.

An additional advantage of directly considering articulatory effort is the fact that we can now consider the selection of consonants and, ultimately, the formation of syllable structures. Work by Huckvale and Howard (2005) is encouraging in this regard as it demonstrates that an artificial vocal tract can learn to imitate stop consonants. However, this is only possible if we are able to obtain a measure of consonant distinctiveness and we will be approaching this challenge in future work. Having investigated the production of speech we will now turn our attention to the perception of speech by investigating the role and importance of the peripheral auditory system.

It is clear however that for both DFT and our own approach, it is not sufficient to simply project H&H, which has been found to be operating at the time scale of an individual, onto a cultural level. Rather, we propose that H&H theory does have an effect on the evolution of language but that this effect is expressed through a complex interaction between phylogeny, ontogeny and glossogeny.

By developing both production and perceptual models we can eventually return, in Chapter 5, to the population models proposed by Oudeyer. By marrying these systems to detailed phonetic considerations, we hope to gain new insights into the cultural evolution of speech.

## Chapter 4

# Interaction of place and voicing in the perception of initial stops

Having considered the role of articulatory effort in our population models we will now turn to the auditory system. We will argue that the mapping from signal to perception is far from trivial and illustrate this point by studying the role of the peripheral auditory system on the perception of initial stops. At this stage, we will consider a single individual agent as a step on the way to a final population model.

Extensive work on the perception of syllable-initial English stops over many years has established, in the words of Patricia Kuhl, an “unexplained perceptual interaction between ... voicing and place” (Kuhl 1988), in that the boundary between voiced and unvoiced token categories shifts systematically with place of articulation. In the same vein, Terry Nearey refers to stop consonant identification as a “long-standing perceptual puzzle” (Nearey 1997). Putative explanations of this phenomenon include an aeroacoustic account based on the inertia of the articulators involved, and the notion that the effect is an essentially uninteresting artifact of the particular (synthetic) stimuli used. In this chapter, we present a computational auditory/neural model of the perception of syllable-initial stops that replicates the observed interaction between place and voicing for two series of stimuli: one synthetic and the other real speech. Results with real speech are interpreted as denying that the effect is artifactual. One outstandingly advantageous aspect of computational modelling as an experimental paradigm in phonetics research is that the software model can be easily modified to explore and understand each stage of the perceptual process, in a way that is simply not possible with human or animal subjects. Analysis of the model developed here (by varying its component parts) leads to an explanation of voicing/place interaction in terms of localisation of auditory nerve activity at a particular low-frequency

region near the consonant-vowel border. This is argued to be more satisfying than an aeroacoustic ‘explanation’, since the latter must assume that ‘perception follows production’ in some sense.

## 4.1 Overview

In the middle part of the last century, researchers at Haskins Laboratories started to develop a highly distinctive, gesturally-oriented approach to the scientific study of speech perception to set against the more engineering-oriented work at Bell Telephones: see Liberman (1996) for personal, first-hand recollections of the history of this work. One particular—and, as it turned out, controversial—thread was the prominence given to so-called categorical perception (CP) and the ‘motor theory’ put forward to explain it. The experimental study of CP was especially influenced by the series of stimuli developed by Lisker and Abramson (1970), in which the voice onset time (VOT) of synthetic syllable-initial stops was varied. As stated by Lisker (1986), these stimuli have been extensively used in speech perception studies over many years (Lisker 1975; Kuhl and Miller 1978; Lisker 1986; Soli 1983) as well as in subsequent modelling work (Damper, Pont, and Elenius 1990; Darling, Huckvale, Rosen, and Faulkner 1992; Damper 1997; Damper, Gunn, and Gore 2000; Damper and Harnad 2000).

Early work at Haskins Labs. by Liberman and colleagues (Liberman, Harris, Hoffman, and Griffith 1957; Liberman, Delattre, and Cooper 1958; Liberman, Cooper, Shankweiler, and Studdert-Kennedy 1967) showed that English-speaking human listeners categorised synthetic initial stops varying in VOT into voiced and unvoiced classes. This categorisation was seen at the time as something of a ‘violation’ of known perceptual laws (cf. Weber’s law) calling for a special explanation. This was one of several lines of thought leading to the “speech is special” theoretical position with which the Haskins researchers have long been associated. An important finding of Liberman, Delattre, and Cooper (1958) was that the voicing boundary between the two classes depended systematically on place of articulation. In subsequent work, Lisker and Abramson (1970) found that, as the place of articulation moves back in the vocal tract from bilabial (for a /ba–pa/ VOT ‘continuum’) through alveolar (/da–ta/) to velar (/ga–ka/), so the boundary moves from about 25 ms VOT through about 35 ms to approximately 42 ms. Why this should happen is uncertain. For instance, Kuhl (1987, p.365) writes that “we simply do not know why the boundary ‘moves’” and refers to this as “an unexplained perceptual interaction between ... voicing and place” (Kuhl 1988, p.33). In the same vein, Nearey (1997, p.3251) refers to stop consonant

identification as a “long-standing perceptual puzzle”.

Putative explanations of this ‘puzzle’ include an aeroacoustic account based on the inertia of the articulators involved, and the notion that the effect is an essentially uninteresting artifact of the particular Lisker and Abramson stimuli used. However, as expanded in the remainder of chapter, we lean towards an account based on general auditory mechanisms, and *contra* the “speech is special” thesis. As MacWhinney (1998) writes: “It now appears that the ability to discriminate the sounds of language is grounded on raw perceptual abilities of the mammalian auditory system” (p.202). This position is consistent with the arguments of a number of influential and respected speech scientists (Kuhl and Miller 1978; Kuhl 1988; Diehl and Kluender 1989; Sussman, McCaffrey, and Matthews 1991; Kingston and Diehl 1994; Ohala 1996). Yet there is a difficulty inherent in attempting such an account. In the words of Summerfield (1982, p.51): “... the relationship between acoustical structure and perceived phonetic structure is complex and not obviously explained by known properties of the mammalian auditory system.”

How then are we to proceed? We believe that valuable insights into this complex relationship can be obtained by building computer models of the auditory system, and studying their ability to replicate perceptual data. Indeed, Kirby (2002, p.185) has written: “Computational simulation, informed by theoretical linguistics, is an appropriate response to the challenge of explaining real linguistic data in terms of processes that underpin human language”. As stated by Kuhl and Miller (1978, p.906): “Ideally, [*one would like*] experimental methods that somehow allow one to intervene at various stages of the processing of sound to observe the restructuring of information that has occurred at each stage.”

Traditional experimental methods based on human or animal psychophysics and neurophysiological investigations using animal models fall well short of this ideal. By contrast, the very nature of software means that a computational model based on general auditory principles, and which proves capable of replicating the data of interest, can be interrogated at “various stages of the processing of sound” to discover its operating principles. Additionally, or alternatively, we can simplify or exclude component parts of the model in order to discover the part that they play in “the restructuring of information”.

Hence, the approach taken in this chapter is to build a computational auditory/neural model of the categorical perception of syllable-initial stops. We show that this model is able to replicate the observed interaction between place and voicing (i.e., the voiced/unvoiced category-boundary movement effect) for the Lisker and Abramson

stimuli. It follows that an explanation of the natural phenomenon of boundary shift *need be no more complicated* than the model itself. It is important to emphasise that the model is *not* designed to replicate the data. Rather, it is designed to embody basic principles of auditory function; the replication of the boundary-shift phenomenon is consequential, following training on example stimuli from end-points of the voiced/unvoiced continuum. Further, we show that the shift occurs for a real speech ‘continuum’ (part of the dataset collected by Nossair and Zahorian 1991), so validating the model’s results for the synthetic data also.

Regarding ‘empiricist’ models of speech perception, Kluender and Lotto (1999, p.508) write: “The occasional demonstration of similar patterns of response data for speech and nonspeech analogs ... may no longer be sufficient. Instead, hypotheses about specific auditory processes must be generated”. In light of this concern, computational modelling has the distinct advantage of facilitating the formation of such hypotheses. We show that analysis and manipulation of our model then allows us to explain the boundary-shift effect in terms of the restructuring of information by the auditory system so as to emphasise a particular region of the time-frequency spectrum.

## 4.2 Modeling the categorisation of speech sounds

In previous work over many years (Damper, Pont, and Elenius 1990; Damper 1997; Damper, Gunn, and Gore 2000; Damper and Harnad 2000), researchers at the University of Southampton have studied extensively the emergence of phonetic categories in a variety of related computational models. We next consider a number of current models and describe the version used in this work.

In conducting this review we will highlight the advantages and pitfalls of the computational methodology and motivate our own approach. Previous work by Damper and Harnad (2000) presents a comprehensive overview of neural network models of categorical perception and argues ‘that a variety of neural mechanisms is capable of generating the characteristics of categorical perception. Hence, CP may not be a special mode of perception but an emergent property of any sufficiently powerful general learning system.’ In our own work this argument will be tested by our focus on the peripheral auditory system and selection of a general neural network architecture. As ‘an advantage of such computational models is that, unlike real subjects, they can be “systematically manipulated” to uncover their operational principles’. In considering previous work Damper and Harnad (2000) highlight a clear distinction between real and synthetic CP ‘the neural models of synthetic CP reviewed thus far have all taken their

inputs from artificial or novel dimensions, whereas the vast majority of real CP studies have used speech stimuli — most often stop consonants (or, more correctly, simplified analogs of such sounds).’ In an attempt to rectify this flaw, they have used the synthetic Lisker and Abramson stimuli from Haskins Laboratory. To expand upon this work we will be using not only this data but also real speech from Nossair and Zahorian.

Following this paper we will be focussing on a detailed computational model of the peripheral auditory system as ‘there has long been a view in the speech research literature that CP reflects some kind of “restructuring of information” by the auditory system in the form of processing non-linearities.’ In addition to this, certain methodological considerations will be taken into account in our own model. For example, the input to our neural network will be formed from 10ms time bins as ‘The 10ms width of the time bin corresponds approximately to one pitch period.’ Moving on to the cognitive model ‘there was a distinct net for each of the (bilabial, alveolar, velar) stimulus series.’ and initially, in their brain-state-in-a-box model, all of the training data was attracted to the same classification, indicating that the categorical perception of auditory training data is a non-trivial task. Classification was only successful when ‘the most important time-frequency cells were identified by averaging the endpoint responses and taking their difference.’ Accordingly, this chapter will produce novel research by attempting to reproduce CP by using a single network on the entire set of stimulus data before focussing on the most important regions of the signal. In Damper and Harnad’s own work they show that ‘Classical categorization is observed with a steep labelling curve and an ABX discrimination peak at the category boundary. Although the labeling curve is rather too steep and the actual boundary values obtained are slightly low (by about 5 or 10ms), the shift with place of articulation is qualitatively correct. Further work in this paper was developed through the use of a multi-layer perceptron and ‘it is apparent that the multi-layer perceptron is a rather better model of labeling behavior than is the brain-state-in-a-box.’ Accordingly, our own neural architecture will take the form of a multi-layer perceptron.

A key component of the Damper and Harnad’s approach is the Pont-Damper auditory system which we will now consider in more detail. This model consists of ‘a computational model of low-level neural processing which includes explicit details of dorsal cochlear nucleus (DCN) structure and function, based on recent anatomical and physiological findings.’ Consisting of two separate stages ‘simulating afferent activity in:

- the cochlea and auditory nerve
- the dorsal cochlear nucleus’

In total the model consists of an array of 512 neurons organised into 128 columns with each neuron acting ‘as a ‘leaky integrator’, summing its synaptic inputs at each time instant and adding these to a decayed version of the potential from the previous instant’ firing according to a probability value over some threshold function. Their results on bilabial, alveolar and velar stops show that ‘At auditory nerve (AN) level, the model’s spatio-temporal patterns of neural firing correspond well with available physiological data.’ Furthermore, in response to varying VOT from the Lisker and Abramson stimuli it was shown that there was no ‘obvious visible evidence that some feature (or features) changes abruptly at the boundary value of 30–40ms in a way that could underlie the observed labelling behaviour.’ Suggesting a clear motivation for an abstract neural component, one that is able to exploit large classification features. Similar to the work by Damper and Harnad this paper shows ‘that a composite model, consisting of the computational auditory model feeding an artificial neural network (ANN) trained by back-propagation, is able to reproduce the non-uniform identification and non-monotonic discrimination behaviour of listeners with these stimuli.’

Further consideration of the neural correlates behind categorical perception is presented by Damper, Gunn, and Gore (2000) who studied three learning systems: ‘single-layer perceptions, support vector machines and Fisher linear discriminants.’ Of particular significance to our choice of peripheral auditory system is the result ‘that the phonetic percept of voicing is easily and directly recoverable from auditory (but not acoustic) representations.’ In greater detail ‘This work has revealed very clearly that any reasonably general learning system (i.e. ‘neural network’) – acting as a ‘synthetic listener’ – is able to categorize the patterns of simulated auditory nerve activation in a way which mimics the psychophysical behaviour of real listeners.’ Given this strong conclusion we can proceed with confidence and consider the acoustic properties of real speech in this chapter. Accordingly we will expand upon the methodology of this work specifically ‘we seek (1) to see if a learning system is capable of mimicking the behaviour of real listeners and then (2) to analyse the system to see what it has learned.’

However, in following this method the work of Darling, Huckvale, Rosen, and Faulkner (1992) sounds a note of caution, as it is clear from their comparisons that not all models of the peripheral auditory system are functionally equivalent. Certain changes in the choice of representation for the peripheral auditory system lead to the removal of the boundary shift effect under consideration. However, we still feel that it remains the most flexible tool for analysing the perception of initial stops. As stated by Damper (1998) “While this intervention is difficult or impossible to achieve in experiments using human or animal listeners, it is immeasurably easier ‘to observe the restructuring of information’ in a software model of auditory processing.”



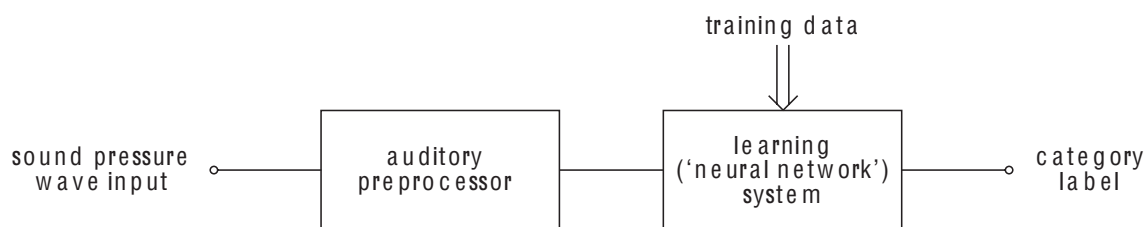


FIGURE 4.1: Two-stage (auditory/neural) computational model for the categorisation of speech sounds.

### 4.2.1 Computational auditory/neural model

Previous work has used computational models having the general, two-stage form shown in Figure 4.1 in which an auditory preprocessor mimics restructuring of information by the listener’s peripheral auditory system, and a ‘neural network’ component simulates processing by the central auditory system. Since the operating principles of the central auditory system are very poorly understood, if at all, it is convenient (perhaps mandatory) to train this second stage to categorise clear exemplars of the voiced and unvoiced distinction, for which we use end-point stimuli from the continuum. This mimics the operant training used in animal studies by, e.g., Kuhl and Miller (1978) in which chinchillas were trained on end-points from the Lisker and Abramson stimuli and tested on the entire continuum, including generalisation testing on unseen, intermediate points.

In this work, for the auditory preprocessor we use Lyon’s ‘passive long-wave cochlear model’ described by Slaney (1998). By virtue of its availability in MATLAB’s Auditory Toolbox, this model has been widely used in the work of others (Aleksandrovsky et al. 1996). It strikes a good balance between abstraction and biological accuracy. It captures important properties of the auditory system, for example the role of outer hair cells in loudness recruitment, but through necessity abstracts away from “factors such as displacement of the stereocilia due to the influence of Brownian motion and stochastic resonance” (Araújo, Magalhaes, Souza, Yehia, and Loureiro 2005).

Lyon’s cochlear model represents the probability of auditory nerve firing at a particular time. For comparability with most of our earlier work using a different auditory preprocessor (e.g., Damper, Gunn, and Gore 2000; Damper and Harnad 2000), and because it effects a good compromise between computational efficiency and neurophysiological realism, we have employed 128 cochlear filter channels. In this study, we use both synthetic and real speech stimuli: those of Lisker and Abramson (1970) and the child part of the dataset collected by Nossair and Zahorian (1991), respectively. The sampling rate was 10 kHz for the Lisker and Abramson synthetic



stimuli, and 16 kHz for the Nossair and Zahorian real-speech data.

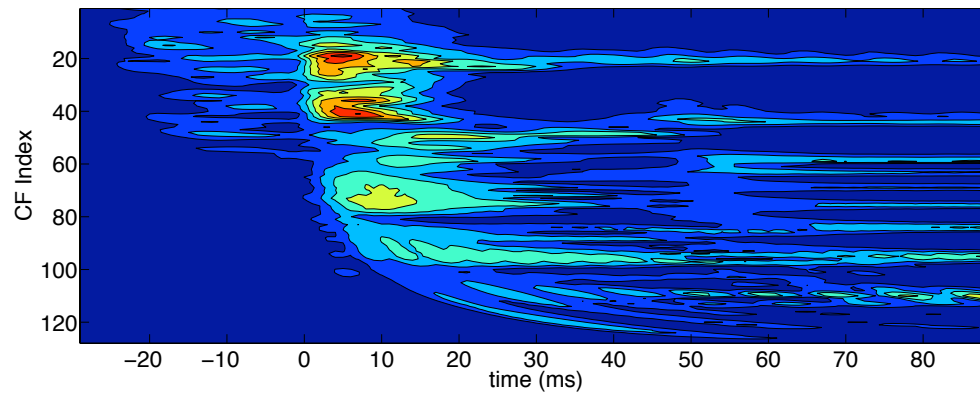
It is not very explicitly stated in Slaney (1998) how the centre frequencies (CFs) of the auditory filters are spaced. It appears that “the center frequency of each channel decreases by `EarStepFactor` of the bandwidth at the previous stage” (p. 19) where `EarStepFactor` was 0.25 in this work. This means that frequency channel indices *reduce* as frequency increases, so that a CF index of 128 corresponds to the Nyquist frequency (sampling frequency divided by 2), i.e., 5 kHz for the Lisker and Abramson stimuli and 8 kHz for the Nossair and Zahorian data, and a CF index of 1 corresponds to a frequency close to zero hertz.

Figure 4.2(a) shows the time-frequency spectrogram produced by this auditory preprocessor for the Lisker and Abramson bilabial stimulus with VOT of 40 ms. For the second-stage neural network, trained to convert the auditory time-frequency patterns of firing into a category label, it would not be possible to train this component successfully (i.e., estimate statistically stable connection weights) on such a rich representation as this without dramatic data reduction. As in earlier work, this is done by aggregating outputs over the time range 25 ms before onset to 95 ms after onset in a small number of time-frequency bins. Accordingly, a  $12 \times 16 = 192$ -element matrix of inputs is produced, as shown in Figure 4.2(b), and is then presented as input to the second-stage neural network.

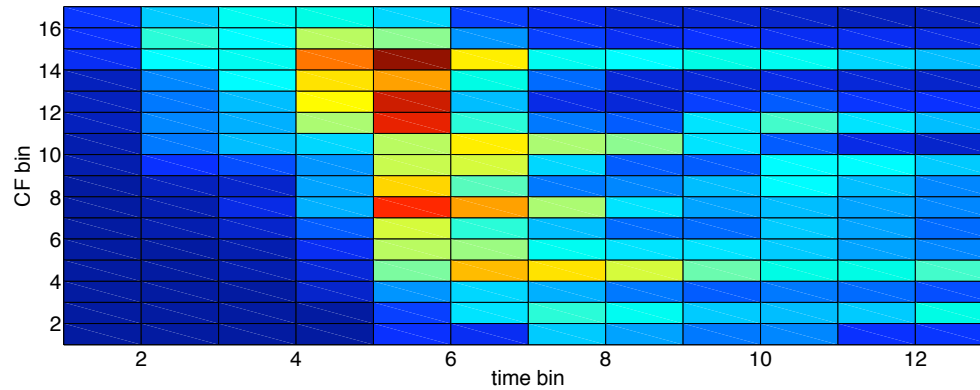
The specific network used in this work was a multilayer perceptron (MLP) with 5 hidden nodes and a single sigmoidal output node, trained to output an activation of 0 on the unvoiced endpoint stimuli and 1 on the voiced endpoint stimuli. Training was from initial random weight settings over 2000 training epochs using Levenberg-Marquardt back-propagation (Marquardt 1963; Hagan and Menhaj 1994) using MATLAB function `trainlm`. This function adjusts the training rate (starting from initial value 0.001) according to the mean squared error over all inputs; there is no momentum term. Categorisation of these stimuli (at least, the Lisker and Abramson synthetic tokens) by neural network is known not to be especially sensitive to network architecture and/or parameter settings (Damper and Harnad 2000).

#### 4.2.2 Categorisation of synthetic Lisker and Abramson stimuli

To validate the auditory/neural model, we compare our labelling results obtained under the same conditions with those for human and chinchilla listeners from Kuhl and Miller (1978) using the Lisker and Abramson stimuli, consisting of three synthetic series—bilabial (/ba/–/pa/), alveolar (/da/–/ta/) and velar (/ga/–/ta/)—varying in place



(a) Auditory spectrogram



(b) Reduced auditory spectrogram

FIGURE 4.2: Time-frequency representations of Lisker and Abramson’s bilabial stimulus with VOT of 40 ms in the form of an auditory spectrogram produced by Lyon’s cochlear model: (a) is the full-data version produced with a sampling rate of 10 kHz; (b) is the reduced-data version produced by aggregating outputs into  $12 \times 16$  time-frequency bins for input to the second-stage neural network model. Note that in (a), the centre frequency (CF) index reduces with increasing frequency (see text).

of articulation of the stop consonant. Although we would like to be able to give a quantitative measure of agreement between the labelling results from the auditory/neural model and the human and nonhuman animal results, unfortunately we cannot do this, because the Kuhl and Miller raw data points (as opposed to the smoothed, fitted data that they present in their paper) are not available.

Figure 4.3 shows Kuhl and Miller’s labelling curves for the Lisker and Abramson stimuli. A key characteristic of these results is the movement of category boundary with place of articulation: as the constriction in the vocal tract moves back from bilabial through alveolar to velar, so the boundary moves to longer VOTs. Kuhl (1988, p.33) describes this as “an unexplained perceptual interaction between ... voicing and place”. Broadly similar results on the ability to form phoneme-like equivalence classes have been found in studies with other animal species, including macaques

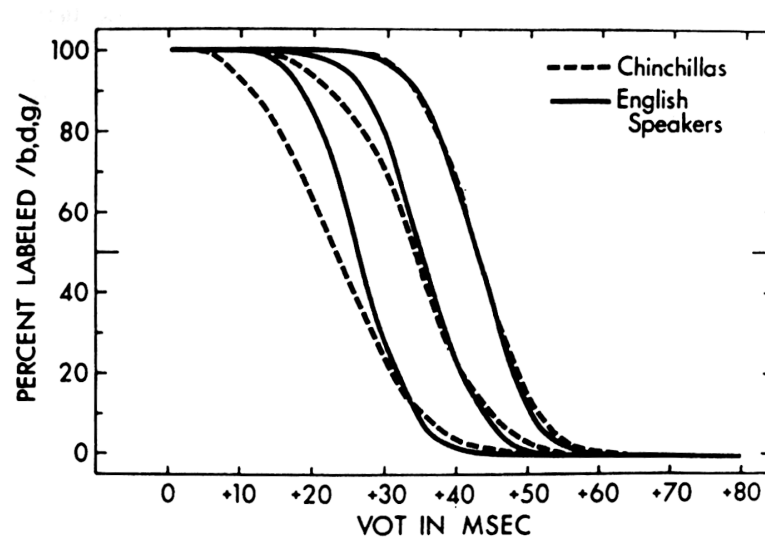


FIGURE 4.3: Labellings of the Lisker and Abramson continuum produced by human and chinchilla subjects. From Kuhl and Miller (1978).

(Kuhl and Padden 1982, 1983), Japanese quail (Kluender, Diehl, and Killeen 1987) and budgerigars (Dooling, Soli, Kline, Park, Hue, and Bunnell 1987; Dooling, Okanoya, and Brown 1989). However, differences in the auditory physiology of birds and mammals apparently lead to differences in the exact boundary values obtained. Of course, none of these animal listeners is likely to possess a specialisation for speech perception.

Figure 4.4 shows corresponding labelling results obtained by the computational model, in the form of average activation against VOT. In this plot, presentation of the stimuli to the neural network has been repeated 10 times, with the network retrained each time from a different, random initialisation of the connection weights. Error bars are shown in the form of plus and minus the standard deviation across the 10 repetitions, but note that the actual activation of the output node could not go above 1 or below 0. As in Kuhl and Miller (1978), smooth curves have been fitted to the raw data for presentational purposes. In this work, we have used least squares to fit sigmoids to raw data points. The labelling curves produced by the model clearly replicate the movement of the category boundary with place of articulation seen in Kuhl and Miller's data, although the separation of the bilabial and alveolar curves is perhaps less convincing. It is worth noting, however, that the results of Fig. 4.4 are obtained on really quite sparse data. The Lisker and Abramson stimuli consist of 3 series (bilabial, alveolar, velar) each of 9 points (0 to 80 ms VOT), 27 tokens in all. Of these, 6 endpoint tokens are used for training and the remaining 21 are used for generalisation testing. The paucity of data may be one reason why the bilabial and velar labelling curves for the computational model are less well separated than those of Kuhl and Miller for

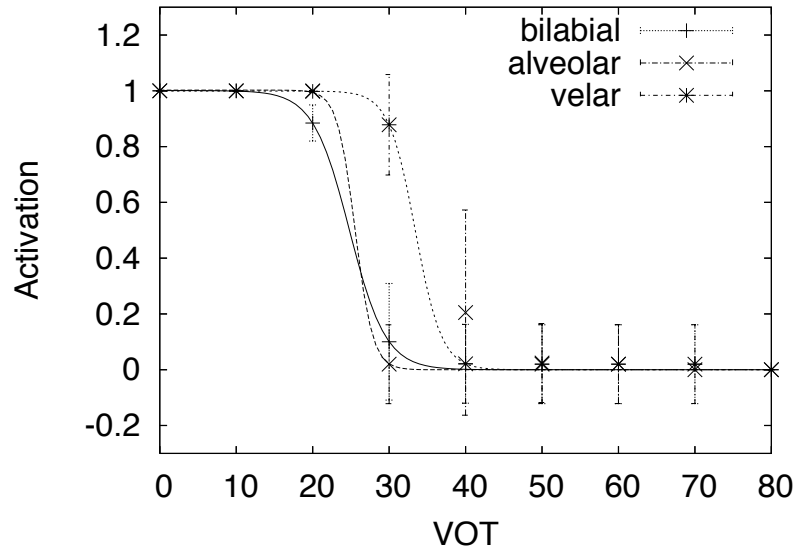


FIGURE 4.4: Labelling of the Lisker and Abramson stimuli by the computational auditory/neural model. Activations shown are averages across 10 presentations of the stimuli with the neural network trained from a different set of initial random weights on each presentation. Error bars are standard deviations.

real (chinchilla and human) listeners.

### 4.2.3 Categorisation of real speech

One concern in the work described thus far is that the (single set of) stimuli used are synthetic, and are very few in number. Hence, we cannot be sure that they are truly representative of real speech and its variability. In particular, other researchers (e.g., Soli 1983) have reported discontinuities in the acoustics of these stimuli or (e.g., Kluender 1991) have reported different effects for other voicing series. Hence, it is important to examine the model's performance on other VOT series, and preferably on large numbers of real tokens rather than small numbers of synthetic tokens. This should eliminate a long-standing concern in the field that the Lisker and Abramson stimuli may not be perfectly acoustically-equated across series steps from blemishing the interpretation of the model's capacity to predict category boundaries. To this end, we have chosen to use syllable initial stop tokens (/b/, /d/, /g/, /p/, /t/, /k/) collected by Nossair and Zahorian (1991), and hand-labelled with their VOT, for the purpose of studying the automatic recognition of initial stops. With these data we can test the following hypothesis:

*Hypothesis 4.2.1.* A computational model combining an artificial peripheral auditory system, with a simple neural abstraction will be sufficient to reproduce the observed boundary shift effect in the perception of initial stops.

Initial exploratory work with the database supplied to us by Nossair and Zahorian (1991) revealed that the single neural network was unable to replicate the boundary shift across all speakers and continua. We feel we were possibly asking too much of the model to learn to classify such extensive and variable real speech data. Accordingly, to make the problem more tractable we formed our data set from the 9 child speakers only, reducing the data set from 2481 stop consonants to 634 and making the set of speakers the model has to cope with more homogeneous. However, even this subset of the complete database is a vast improvement on the sparse Lisker and Abramson stimuli, consisting of just 27 tokens.

A very obvious difference between Lisker and Abramson's synthetic tokens and these ecological data is that there is no external control of the VOT in the latter case. One of the strong reasons for using synthetic data in our previous work has been the serious concern that real speakers will avoid productions in the 'ambiguous', boundary region, so leading to a paucity of data for generalisation testing. In the event, the distribution of VOT across the productions was as shown in Figure 4.5. Although there are relatively fewer productions in the ambiguous zone, we are confident that these are sufficient for our purposes. Another very obvious and welcome difference is that the real speech has variability typical of actual intra- and inter-speaker differences, albeit within the subset of child speakers.

Figure 4.5 also depicts the breakdown of the Nossair and Zahorian data into training and test tokens. The criteria for selection of tokens for training was as follows. For each of the three places of articulation, the means and standard deviations of the productions labelled as unvoiced by Nossair and Zahorian were computed, similarly the means and standard deviations for the voiced tokens. These are shown in Table 4.1, where the figures are consistent with Nossair and Zahorian's Table I (p.2980). For the voiced case, those tokens with VOT less than the relevant mean plus the relevant standard deviation were taken as 'endpoint' (voiced) training data. For the unvoiced case, those tokens with VOT greater than the relevant mean minus the relevant standard deviation were similarly taken as training examples of the unvoiced class. All the remaining tokens were considered to lie in the ambiguous zone and were taken as test data. The labellings given by Nossair and Zahorian were confirmed by listening.

The speech data were 'trimmed' for presentation to the model as follows. The maximum VOT for any of the tokens was 137.9 ms, so the data were cut to encompass the range  $-25$  ms (i.e., 25 ms before initial burst) to 140 ms. This gave a time bin width of  $165/16 = 10.3$  ms, close to the 10 ms used for the Lisker and Abramson stimuli.

Figure 4.6(a) shows the auditory spectrogram produced for a /bɔ/ token (child speaker,

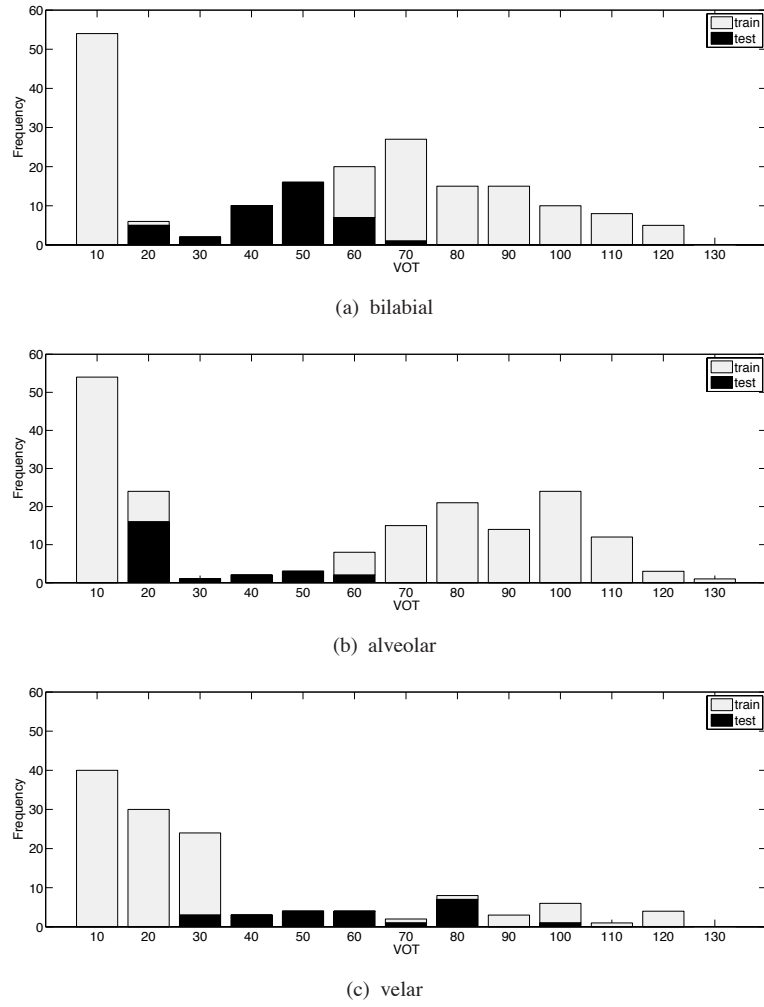
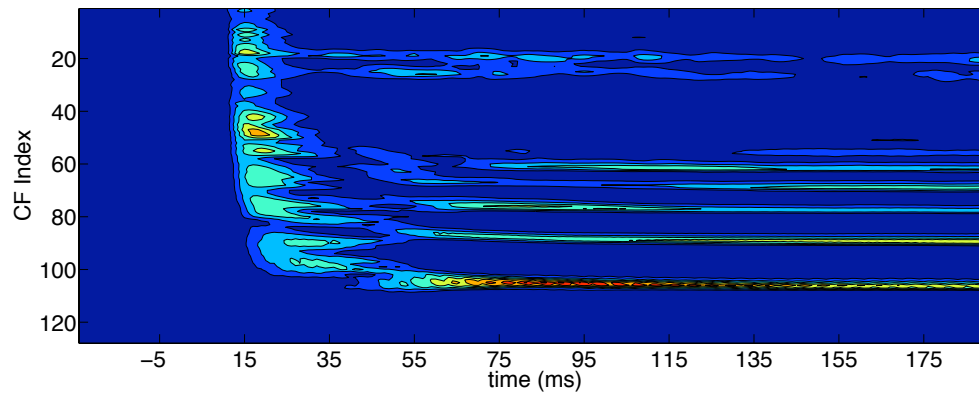


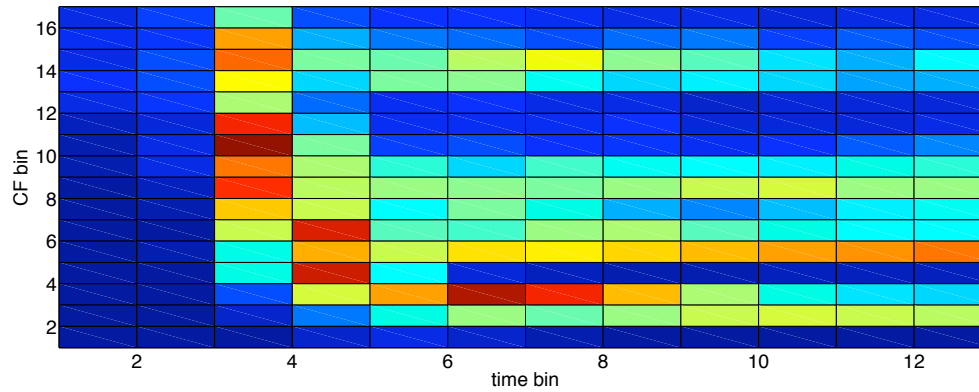
FIGURE 4.5: Histograms showing the distribution of VOT(ms) for the (a) bilabial, (b) alveolar and (c) velar tokens in the Nossair and Zahorian speech database. Training data were formed from the means and standard deviations of voiced and unvoiced data, occasionally leaving overlapping test data, all three continua were then combined into two training sets (voiced/unvoiced) and one test set and presented to a single neural network.

(a) voiced				(b) voiceless			
		mean	SD			mean	SD
bilabial	/b/	10.8	12.3	bilabial	/p/	81.5	25.9
alveolar	/d/	16.6	6.3	alveolar	/t/	91.7	27.5
velar	/g/	25.1	13.5	velar	/k/	101.9	30.2

TABLE 4.1: Mean and standard deviation of voice onset time (in ms) for the Nossair and Zahorian data used in this study.



(a) Auditory spectrogram



(b) Reduced auditory spectrogram

FIGURE 4.6: (a) Auditory spectrogram produced by Lyon's cochlear model for a real speech /bɔ/ (child speaker, VOT 42.1 ms) from the Nossair and Zahorian database; (b) reduced  $12 \times 16$  time-frequency matrix presented to the second-stage neural network. The time bin width in (b) is 10.3 ms.

VOT 42.1 ms) by the computational model, and Figure 4.6(b) illustrates the reduced spectrogram obtained by averaging nerve firing probabilities into  $12 \times 16$  time-frequency bins. The results of labelling the reduced auditory spectrograms for both the training and test data are given in Figure 4.7, again for 10 repetitions of stimulus presentation and random initial weight settings. The training parameters were as in Section 4.2.2 for the Lisker and Abramson stimuli.

In these results with real speech, the movement of the boundary with place of articulation is clearly and appropriately replicated. One intriguing feature of the results is that the network seems to have greater difficulty labelling the voiced tokens as such than it does labelling the unvoiced tokens, especially for the velar sounds. This latter observation may be due to a relative lack of unvoiced velar training examples (see the histogram in Fig. 4.5(c)).

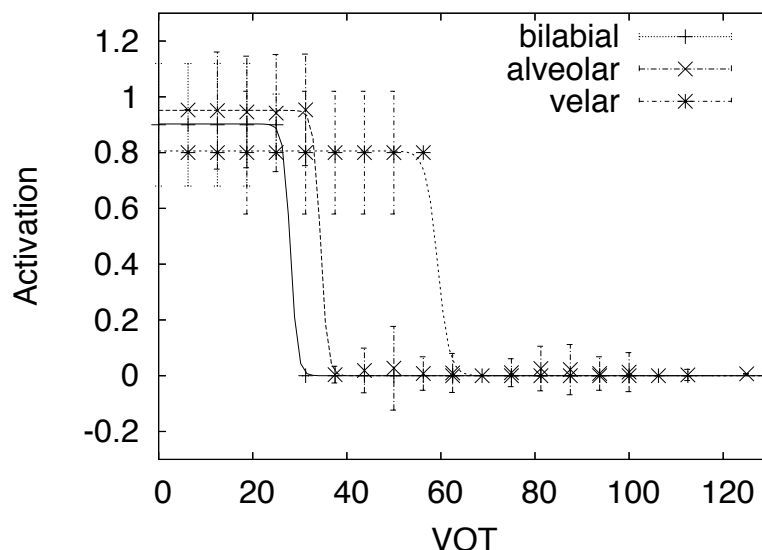


FIGURE 4.7: Labelling curves for the real speech data in the child subset of the Nossair and Zahorian database.

### 4.3 Auditory-neural voicing identification

At this point, with respect to the boundary shift effect for initial stops, we have demonstrated the equivalence of our auditory/neural model with human and animal listeners for both synthetic (Lisker and Abramson stimulus continuum) and real speech (9 child speakers from the Nossair and Zahorian database). Although we believe that this is in itself an exciting and thought-provoking finding, we are mindful of the opinion expressed by Kluender and Lotto (1999, p. 508) that: “The occasional demonstration of similar patterns of response data for speech and nonspeech analogs . . . may no longer be sufficient. Instead, hypotheses about specific auditory processes must be generated”. In light of this consideration, we now attempt to identify the precise auditory and/or neural processes within our model that underlie the observed behaviour. In this way, we can hope to generate further testable hypotheses of the kind envisaged by Kluender and Lotto. As emphasised throughout this chapter, one outstanding feature of a software, computational model as used in this work is the ease with which it can be modified to explore its operating principles.

#### 4.3.1 Removing the peripheral auditory system component

As we have a two-stage auditory/neural model, as depicted in Fig. 4.1, the first and most obvious question to ask is: what role does the auditory preprocessor play? A direct way to address this question is (following Damper 1998) simply to replace the



auditory preprocessor, which attempts to mimic known details of auditory biophysics and neurophysiology, by the simplest conceivable time-frequency analysis, namely Fourier analysis using the fast Fourier transform (FFT).

Figure 4.8(a) shows the Fourier spectrogram for the 42.1 ms VOT /bɔ/ token whose auditory spectrogram was shown earlier in Fig. 4.6(a). This was produced using MATLAB function `spectrogram()` with parameters `segment` set to 128 and `ntrans` set to 1, i.e., 128-point FFT. Such a parameter choice gives a reasonable trade-off between time and frequency resolution. Figure 4.8(b) shows the labelling behaviour of the neural network for the case where the auditory preprocessor is replaced by Fourier analysis. It is very obvious that correct simulation of the boundary movement with place of articulation has been entirely abolished. Indeed, there is not really any voicing boundary *to* move!

Why has this happened? We need to further consider the effect of the auditory processor on the correct modelling of the boundary shift effect. The most striking, obvious difference between the auditory and Fourier spectrograms is that the non-linear ‘cochlear’ frequency scale utilised by the former gives far greater prominence to the low frequency content (below about 2 kHz) than does the linear (hertz) frequency scale employed by the FFT, Figure 4.9. We take this as strong circumstantial evidence that the correct modelling of the ear’s frequency sensitivity (possibly including hair-cell function) is important not just for mimicking the boundary movement effect, but for there being any voicing boundary to move in the first place.

Formally, we can challenge these assertions, and investigate the argument that the resolution of the FFT data is too low, with the following two null hypotheses:

*Hypothesis 4.3.1.* The low frequency enhancement abilities of the peripheral auditory system will have no effect on the model’s categorisation of initial stops.

*Hypothesis 4.3.2.* The low frequency enhancement abilities of the peripheral auditory system will not enable the boundary shift effect.

To test these we will now present the results of a high-resolution FFT control experiment.

### 4.3.2 High-resolution FFT

It is very possible that the failure to replicate the boundary shift effect with the above FFT analysis could be due to loss of frequency resolution in the region emphasised

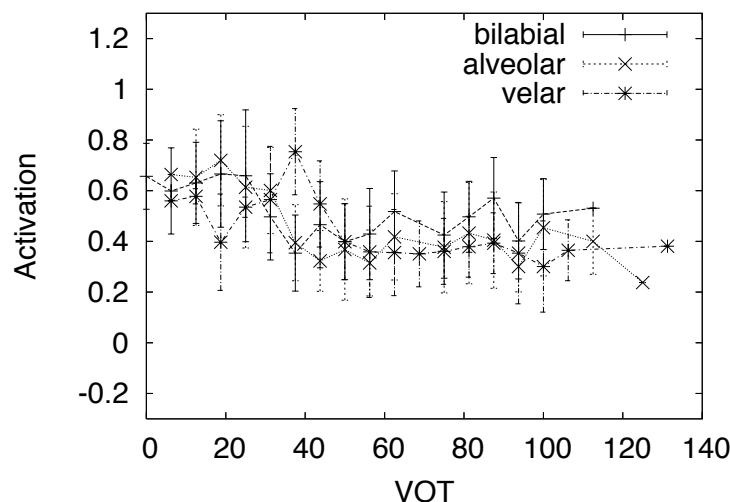
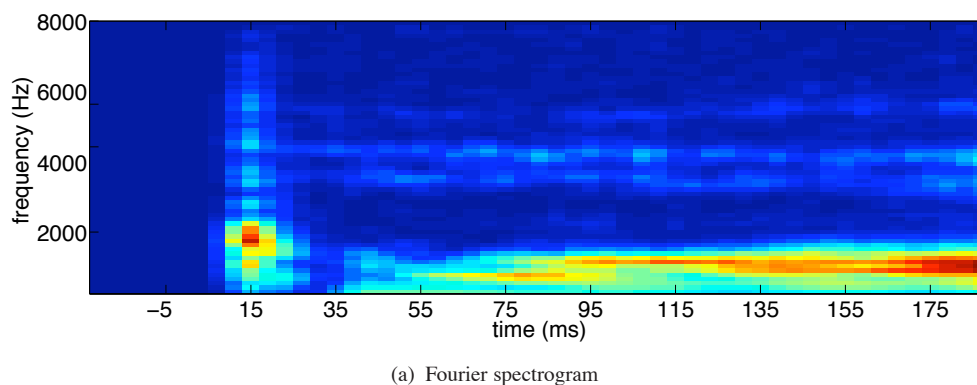


FIGURE 4.8: (a) Fourier spectrogram for real speech /bɔ/ (VOT 42.1 ms) from the Nossair and Zahorian database; (b) Labelling curves for model with auditory preprocessor replaced by Fourier analysis. There is no obvious category boundary and, hence, correct movement of boundary with place of articulation is effectively abolished.

by the auditory model (i.e., below about 2 kHz). To test this possibility, as shown in Figure 4.10(a), the number of CF bins was increased to 68 at each time step and the number of network inputs was increased accordingly. This allowed for a far higher resolution in the crucial low frequency region even without the enhancements provided by the peripheral auditory system. Consequently, in Figure 4.10(b), the classification accuracy of the network, trained on the FFT data, increases but not to the point where the boundary could be said to exist, except possibly in the case of the alveolar tokens. Clearly then the enhancement of the region below 2 kHz remains a key part of the observed boundary shift effect.

We first stated in Hypothesis 4.3.1 that the low frequency enhancement abilities of the peripheral auditory system will have no effect on the model's categorisation of

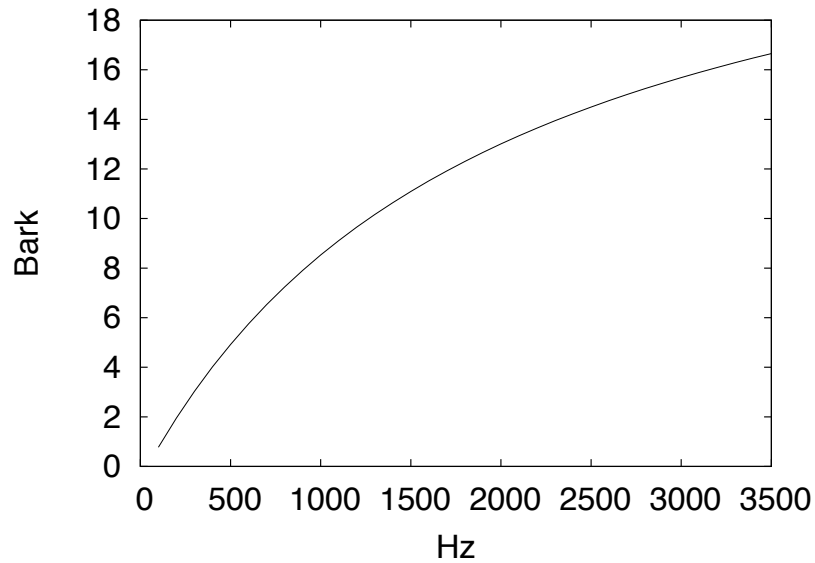
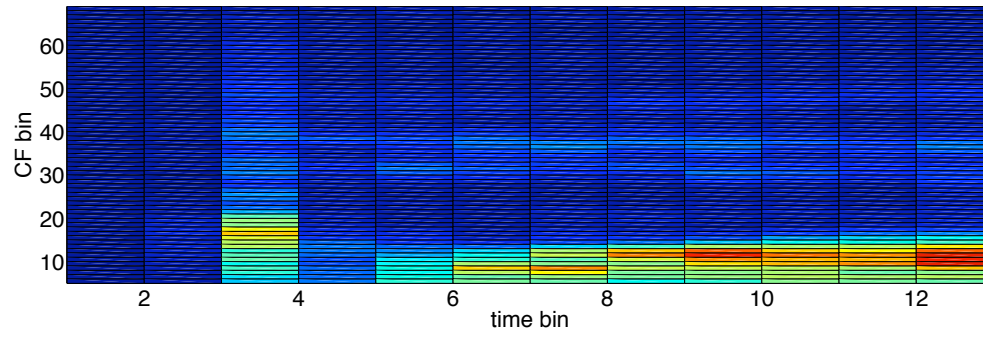


FIGURE 4.9: An illustration of the frequency enhancement present on the Bark scale.  
(Traunmüller 1990)

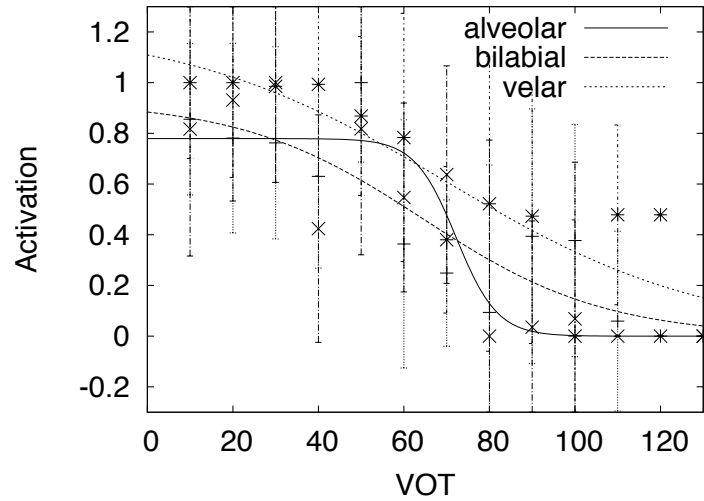
initial stops. Although the model was only trained on idealised end points we have also selected idealised categorisation boundaries for each data set (Figure 4.11), as a simple benchmark with directly plotted labelling curves, and we find that the Euclidean distance between the simulation and idealisations has been reduced from an average of 2.94, for the high resolution FFT, to 1.40 for the peripheral auditory system. This reduction in Euclidean distance can be attributed to the introduction of the boundary shift effect, allowing us to reject null Hypothesis 4.3.1 and 4.3.2.

### 4.3.3 Modifying the peripheral auditory system

Although the precise function of the outer hair cells in the mammalian cochlea is unknown, it is generally believed that mechanical amplification by hair cells is necessary to enhance the sensitivity and frequency-selectivity of hearing (Jia and He 2005, p.1028). Thus, according to Moore and Oxenham (1998, p.109): “... many experiments have shown that damage to the outer hair cells causes dramatic changes in the mechanical responses of the basilar membrane to low-level sounds; the tuning is broadened, and the sensitivity decreased”. In Lyon’s cochlear model, the role of the outer hair cells is defined by a set of automatic gain controls. By removing these, we can investigate their role in the perception of initial stops. Because of the consequent loss of tuning, we would expect onsets, e.g., initial burst and switch-on of voicing, to be poorly represented.



(a) Fourier spectrogram



(b) Labelling curve for model with high-resolution Fourier spectrograms

FIGURE 4.10: (a) An example of the time-frequency representations of the Nossair and Zahorian data, reduced, from the original 128 frequency channels, to  $68 \times 12$  network inputs; (b) Labelling curves for the high resolution FFT data generated from the Nossair and Zahorian database.

Figure 4.12(a) shows the auditory spectrogram of the /bɔ/ token featured earlier in Fig. 4.6(a) when hair cells are removed from the model. Figure 4.12(b) shows the reduced auditory spectrogram in this case, which should be compared with Fig. 4.6(b). Comparing the two figures, the loss of detail for the model lacking hair cells is obvious, as is the effect of reduced onset enhancement. Figure 4.13 shows the resulting labelling behaviour of the model without hair cells, whereupon it is clear that there is no real voiced/unvoiced boundary in evidence, and hence the boundary shift effect is not correctly simulated. Thus, it seems that correct modelling of the ear's onset-enhancement ability is important for correct behaviour on the part of the model.

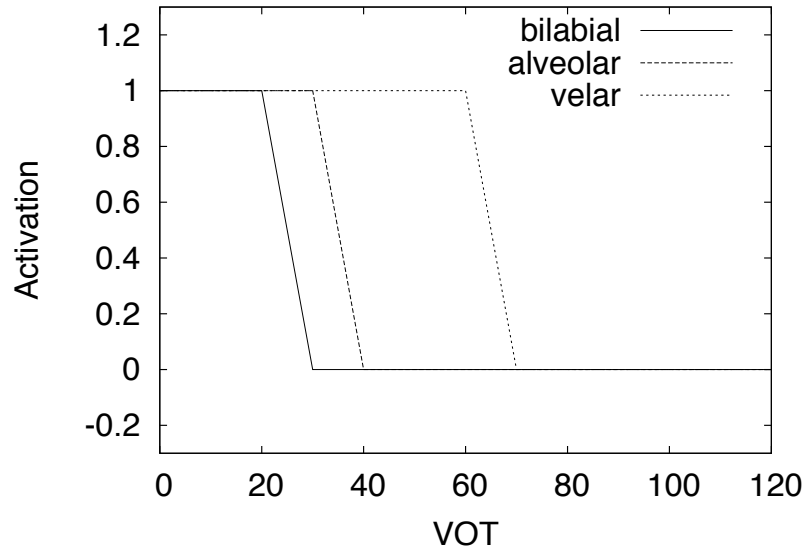
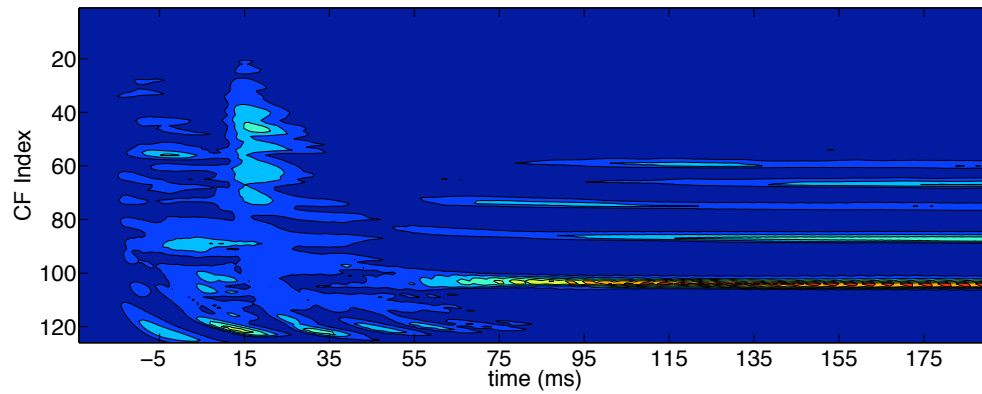
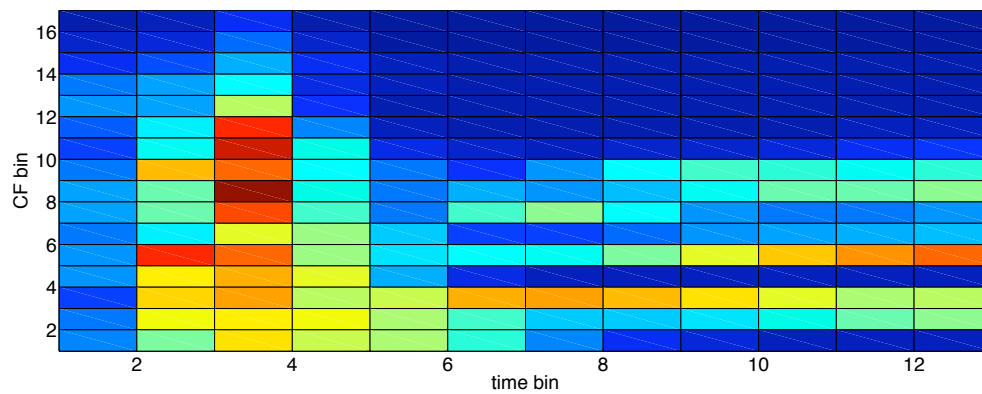


FIGURE 4.11: An idealised step function for comparison against the high-resolution FFT and peripheral auditory system.



(a) Auditory spectrogram, no hair cells



(b) Reduced auditory spectrogram, no hair cells

FIGURE 4.12: (a) Auditory spectrogram for real speech /bɔ/ (VOT approximately 41.2 ms) from the Nossair and Zahorian database with hair cell simulation removed from Lyon's cochlear model; (b) An example of a reduced spectrogram for the real-speech data with hair cell simulation removed from Lyon's cochlear model.

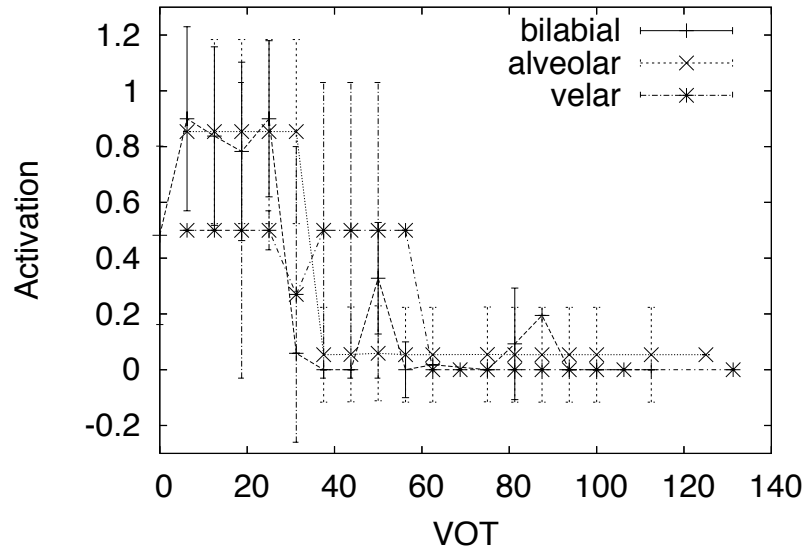


FIGURE 4.13: Labelling curve for real speech /bɔ/ (VOT approximately 41.2 ms) from the child subset of the Nossair and Zahorian database with hair cell simulation removed from Lyon's cochlear model. There is no clear voiced/unvoiced boundary in this case; hence, the boundary movement effect is not evident.

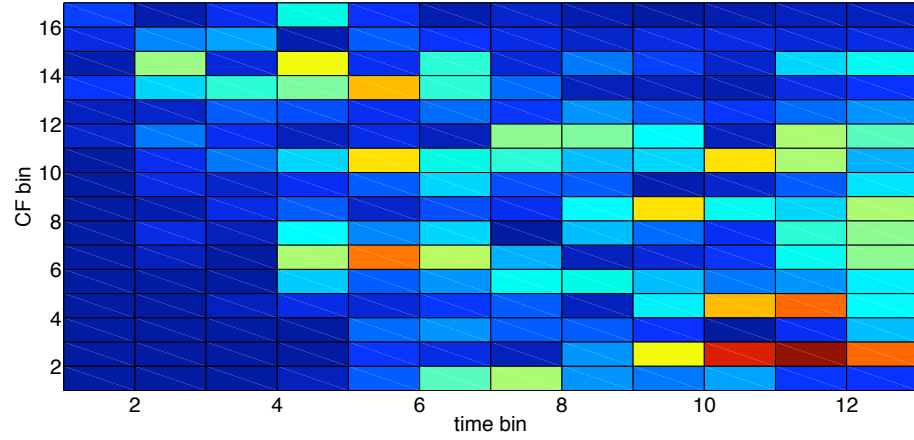
#### 4.3.4 Simplified model of perception of initial stops

To pinpoint the essential aspects of the auditory time-frequency representation underlying the voiced/unvoiced distinction, we computed the differences between bilabial, alveolar and velar spectra at 20 ms and 50 ms VOT. These are shown in Figure 4.14. This reveals that the major differences occur for the four cells with time indices 9 to 12 and CF index of 2. This is intriguing as it is within the vowel, whereas our initial expectation was that it would be somewhere around the transition. Furthermore, the pattern of values across these four cells seems to be specific to the place of articulation. These features are worthy of further investigation with a reduced set of network inputs. Accordingly, we state the following two hypotheses:

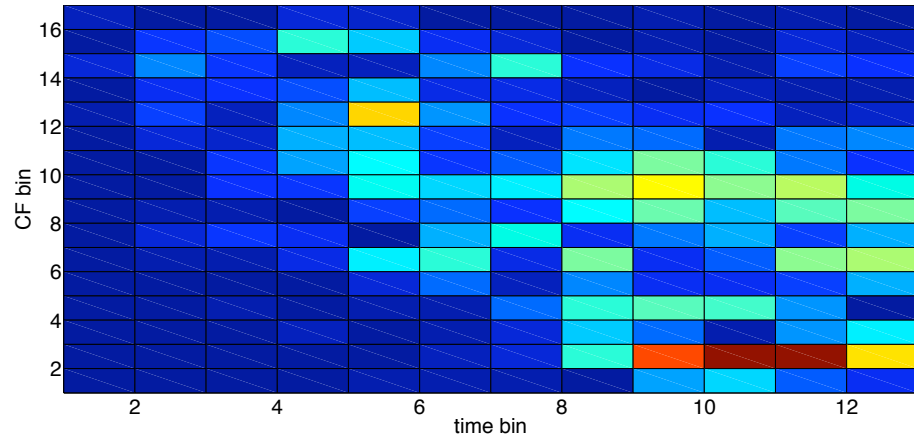
*Hypothesis 4.3.3.* The boundary shift effect will occur with both restricted and complete inputs.

*Hypothesis 4.3.4.* The categorisation of initial stops will occur with both restricted and complete inputs.

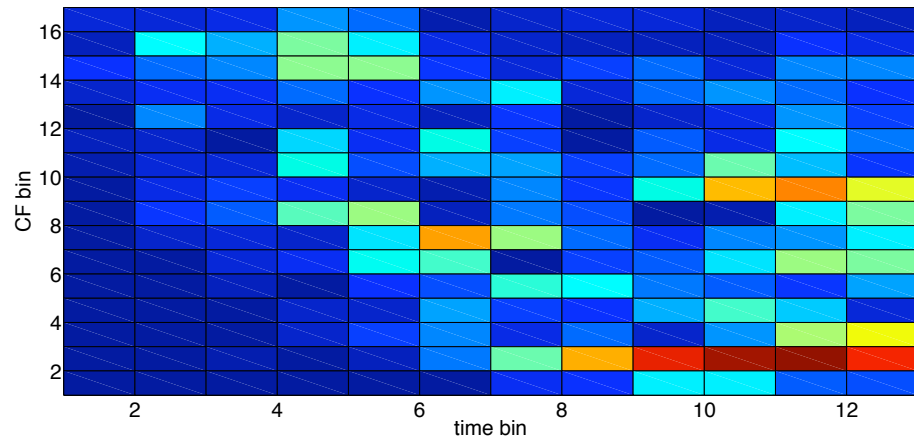
To test the hypothesis that this time-frequency region of the auditory spectrograms is solely and entirely responsible for boundary placement, and hence for the movement of boundary with place of articulation, we built a new model consisting of the auditory preprocessor but with a highly-simplified, 'skeleton' MLP classifier. This skeleton had just 4 inputs, taken from the time-frequency region of interest, and with 5 hidden units as



(a) Bilabial

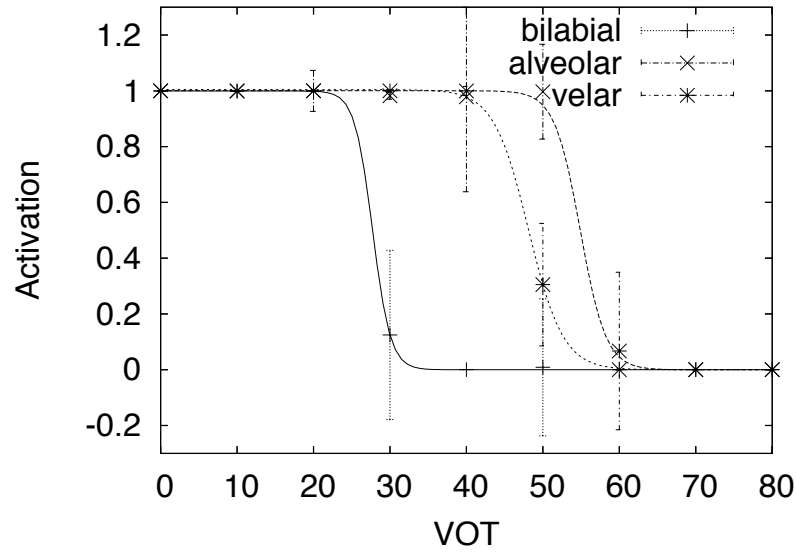


(b) Alveolar

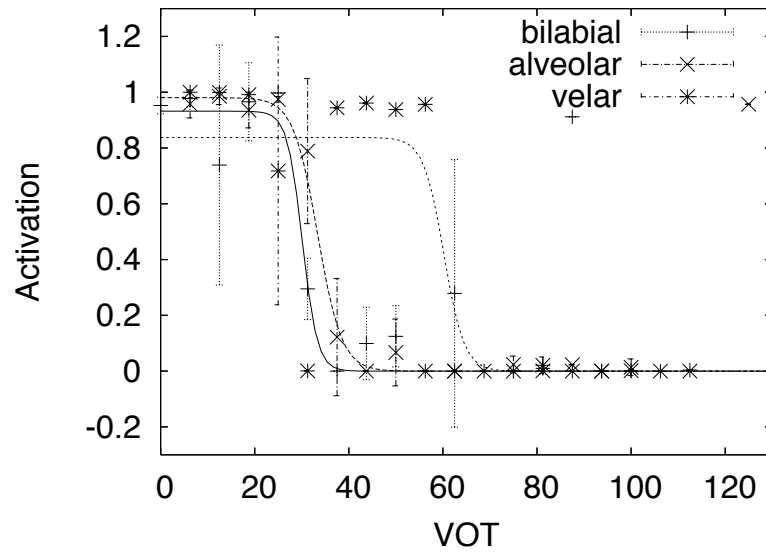


(c) Velar

FIGURE 4.14: Differences of unvoiced/voiced auditory spectrograms for Lisker and Abramson stimuli at 20 ms and 50 ms VOT: (a) bilabial, (b) velar (c) alveolar.



(a) Lisker and Abramson synthetic stimuli, 4 inputs to MLP



(b) Nossair and Zahorian real-speech stimuli, 12 inputs to MLP

FIGURE 4.15: Labelling curves produced by new model with skeleton MLPs: (a) Lisker and Abramson synthetic stimuli with 4 inputs to MLP, (b) Nossair and Zahorian real-speech stimuli with 12 inputs to MLP.

before. Figure 4.15(a) shows the labelling curves for the Lisker and Abramson synthetic stimuli produced by this new model. As can be seen, the essential characteristics of the human and animal psychophysical data are evident in respect of boundary shift. The absolute values of the boundaries are not precise, but this is to be expected given the paucity of the synthetic data (generalisation testing uses just 21 tokens). In any event, the results are comparable in quality to those for the full MLP classifier (Fig. 4.4).



We take this to be a significant and interesting result, effectively de-mystifying Nearey’s “long-standing perceptual puzzle”, at least for these synthetic stimuli. There is, however, a concern in the field that the Lisker and Abramson stimuli are somehow idiosyncratic and of merely historical interest (cf. the demonstration by Kluender (1991) of different effects on different VOT series). So what about real speech?

A reduced or skeleton MLP for the Nossair and Zahorian data set was arrived at in the following fashion. Starting with the same four time-frequency cells as above (time indices 9 to 12 and CF index of 2), we first ascertained that this was inadequate to replicate the boundary shift effect for real speech. With some experimentation, cells contiguous to these were then added as MLP inputs until an acceptable result was obtained. Figure 4.15(b) shows the labelling curves for the Nossair and Zahorian stimuli with just 12 inputs corresponding to time indices 9 to 12 and CF indices 2 to 4. These are not dissimilar to those for the full network (Fig. 4.7). Clearly, the requirement to use 12 inputs in this case (cf. 4 for the synthetic stimuli) is a reflection of the much higher number and variability (both inter- and intra-speaker) of the real speech.

We formally compare the complete neural model (with means and standard deviations shown in Figure 4.7) with the model operating over reduced inputs (Figure 4.15(b)) we find that the different results have a significance of  $\alpha = 0.214$ , for a simple two-tailed  $t$ -test, suggesting that any difference can be attributed to chance. This allows to support the alternative hypothesis 4.2.1: a computational model combining low frequency enhancement, a property of the peripheral auditory system, with a simple neural abstraction will be sufficient to reproduce the observed boundary shift effect in the perception of initial stops.

Taken together, the results for both synthetic and real stimuli indicate that the shift of voicing boundary with place of articulation can be explained by auditory processes occurring at a particular time-frequency locality. There is little evidence in our work that the Lisker and Abramson stimuli are in any way artifactual, or misleadingly unrepresentative of real speech.

## 4.4 Discussion and summary

According to a recent paper by Kingston, Diehl, Kirk, and Castleman (2008), “On examining how phonological feature distinctions are realised phonetically, one is struck by the sheer number of distinct articulatory and acoustic correlates of any minimal contrast” (p.28) such as voicing. It follows that there will be no shortage

of competing explanations of these phenomena, which Kingston, Diehl, Kirk, and Castleman characterize as “associative” (the properties covary reliably), “gestural” (the different properties arise from a single articulatory gesture) and “auditory” (the acoustic properties have the same or similar auditory effect). In this work, by our approach based on use of a computational auditory-neural model, we are most directly positioned in the auditory camp. In particular, the approach has important potential (depending on the fidelity of the modelling) to capture the way that different acoustic properties might produce the same or similar auditory effects. Further, the use of real speech as input to the model ensures that acoustic properties *necessarily* covary realistically, thus overlapping somewhat with the associative position.

In the past, “one important contrastive property” that those in the auditory camp have used to explain the [voice] distinction has been the *low-frequency property* (Kingston and Diehl 1994, p.441). This was previously described by Stevens and Blumstein (1981, p.29) as “presence of low-frequency spectral energy or periodicity over a time-interval of 20–30 ms in the vicinity of the acoustic discontinuity that precedes or follows the consonantal constriction interval” for voiced consonants. In the case of voiceless consonants, this concentration of energy is absent. From this perspective, it is interesting that correct functioning of our model (in terms of replicating the boundary shift effect) is strongly reliant on frequency warping—from hertz to a psychoacoustic scale—that emphasises the low-frequency region. However, on the basis of more recent experimental data, Kingston, Diehl, Kirk, and Castleman (2008) have argued for an account of voicing in which the perceptually relevant property (albeit for intervocalic rather than initial stops) is “the continuation of low-frequency energy across the vowel-consonant border and not merely the amount of low-frequency energy present near the stop” (p.28). This is pleasingly convergent with our modelling results where we are able to replicate the voicing boundary shift effect using *just* low frequency (close to  $f_0$ ) information in the region of the consonant-vowel (in our case) border.

The claim of this chapter is that we have produced the *basis* for a sound explanation of the “unexplained perceptual interaction between ... voicing and place” (Kuhl 1988, p.33). Further work is clearly needed—perhaps using non-speech analogues as input—to tease out the precise mechanisms at play, but this work represents a very promising start. Unlike real listeners in psychophysical studies, computer software is straightforwardly modifiable, giving us the potential to understand its categorisation behaviour in some detail. In the remaining paragraphs, we aim to detail some of the shortcomings of the software modelling approach as well as contrasting it with some competitor explanations of the voicing contrast.

First, we should make the main assumptions of our model explicit. In our approach, segmentation of the input signals is a given, in as much as we construct by hand a zero reference point for the tokens at the initial burst. This sort of assumption is seen in other models (e.g., Nearey 1997); in any event, it should not be too problematic to detect the burst automatically. There is also a closed-set assumption, i.e., the model only ever has to deal with syllable-initial stops of English. However, given the current state of knowledge of speech perception in general, and of causal links between acoustics and audition in particular, some such restriction seems inescapable. Of course, the same accusation could be leveled at almost any theory in the field; they aim at explanation in a severely restricted domain.

Another important consideration is that our model, in addition to the auditory preprocessor, has a strong learning component, which needs to be justified and understood. Kluender and Lotto (1994, p.508) warn: “With the suggestion that a complete model include both auditory and learning processes ... falsifiability becomes at risk ... part of the explanation of speech perception falls out of general auditory processes, and the remaining variance can be ‘mopped up’ by learning processes”. A similar argument against the use of learning (specifically, connectionist) systems to model human performance has been made by Massaro (1988). This view seems to be predicated on the idea that the learning process itself is part of the (explanatory) model. However, this is not the case here. Rather, learning is simply an expedient for finding appropriate parameters to map auditory representations to phonetic percepts; a mapping that must happen somehow. Lacking the knowledge to model the actual processes involved, which may not necessarily be learned in any event, we fall back on a general learning mechanism to ‘plug the hole’. We are not suggesting that the brain does back-propagation!

Let us now consider competitor accounts of the perception of initial stops, and specifically the boundary shift effect. First and foremost, having modelled the effect with real speech data, we believe we can dispense with the idea that boundary movement with place of articulation is artifactual, relying solely on the idiosyncratic nature of the synthetic Lisker and Abramson stimuli. We believe that an explanation of auditory and perceptual phenomena should be in terms of auditory and perceptual processes. As Lieberman (1993, p. 163) writes: “Subjects [*often*] behave as if they are calculating complex probabilities. In fact, the actual mechanism may be surprisingly simple”. Finally, in reviewing competitor accounts, it has been suggested that the shift of boundary with place of articulation is explained by the aeroacoustics of speech production; an auditory account is unnecessary. Certainly, there is an increase of VOT in natural production as place of articulation moves back in the vocal tract, as seen in

the data of Lisker and Abramson (1964, 1970) and Table 4.1. In natural productions, VOT likely depends upon the inertia of the articulators and, hence, the speed with which closure can be released (Diehl and Kluender 1987; Kluender 1991). For bilabials only the lips need move whereas for alveolars the tongue tip is involved and for velars the more massive tongue body must move. Yet one should be careful of simply asserting that ‘perception follows production’. Some differences in production are phonetically distinctive and some are not. The question of why some aeroacoustic phenomena are auditorily-salient and others are not is ignored by this putative explanation, but is a central concern of our auditory/neural model.

It remains the case that computational modelling is a little used tool in phonetics research. In this chapter, we have shown that it has much to offer in terms of exploring the implications for speech perception of current knowledge of audition. Although less well developed here, there is also great potential for generating hypotheses for subsequent experimental test. To quote Kirby (2002, p. 185): “computational modelling will be the core of any future research framework for an explanatory linguistics”. This thesis is now in a position to present a unified model capable of contributing to just such a framework.

## Chapter 5

# Modelling the cultural emergence of speech

Our claim in this chapter is that computational modelling can contribute to a clarification of the questions and issues raised so far. Our thesis has developed a range of tools and methodologies to study speech at the phonetic level. Accordingly, in this chapter, we have constructed a computational model (or ‘artificial agent’), through a unification of these tools, that is able to mimic some aspects of the structure of real human speech. Crucially, this model construction requires us to flesh out the detailed implications of phonetic theory, so introducing an element of rigour absent in abstract thought experiments. The agent is equipped with a biologically-plausible auditory system and vocal tract, resulting in loosely constrained production and perception. Analysis of the model/agent after training reveals that at no time does it establish symbolic phonetic tokens via its cognitive abilities. Rather, complex auditory cues (as opposed to some set of arbitrary, abstract labels) are sufficient to enable the agent to reproduce the perceived speech sounds. We infer from this reproduction that the agent is capable of the direct perception of speech through pattern recognition. After studying the acquisition of a vowel systems and the attempted acquisition of a complete phonetic system we present a population model of speakers and listeners, producing and perceiving ‘real speech’, generated by the vocal tract of Chapter 3 and perceived by the auditory system of Chapter 4. We find that within the complex environment of ‘real speech’ populations of agents are able to converge upon a mutually agreed set of auditory features through a simple process of speaking and listening. Additionally, this process of cultural emergence takes place with agents that are only loosely constrained in their articulatory and auditory abilities. This is in contrast to other phonetic theories that often require stronger constraints on production and perception and the formation of

idealised, abstract cognitive symbols.

## 5.1 Overview

In our opinion, detailed phonetic theories are amenable to a useful degree of test through the construction of an artificial, computational agent system. First and foremost, sufficient is known of articulatory and auditory processes to model the complete cycle of production and perception and these aspects have been successfully modelled and tested in Chapters 2, 3 and 4. To be sure, knowledge of central brain processes is sadly lacking, but there is still some scope for modelling these in relatively simple connectionist terms, only provided we have available suitable and sufficient training data in the form of a database of speech sounds. Further, a computational model is highly flexible in that it is trivially easy to change software. This means that it is in principle possible to implement a range of constraints on articulation and audition such as to cover the gamut from symbolic to dimensional (Port 1990; Port and Leary 2005) theories of speech perception. Viewed in this light, the more physically and physiologically realistic our simulations of the vocal tract and the peripheral auditory system, the ‘weaker’ the constraints. This is based on the assumption that the ‘strong’ constraints required to render simple and transparent “relation[s] between physical and symbolic elements” must come from more central brain processes. We remind the reader that many works in this area (de Boer 2001; Oudeyer 2005b) trivialise the link between articulation, signal and audition, for good reasons of computational convenience, and are hence implicitly (and possibly unwittingly) strongly constrained theories. Only by moving away from these damaging abstractions can we begin to investigate other phonetic theories.

To enable weak audition, the auditory system itself must be capable of speech perception, rendering the signal suitable for our cognitive model, and it must be biologically plausible. This is desirable not because biological plausibility is a good thing in its own right but because our model needs to capture certain features of production and perception. As stated by Dror and Gallogly (1999) “In the context of the distinct types of contributions made by certain computational analyses, the biological plausibility of those analyses is altogether irrelevant”(p.173). We have already shown in Chapter 4 the vital importance of the peripheral auditory system to the perception of speech. Accordingly, we will use Lyon’s cochlear model, detailed in Slaney (1998) and Chapter 4. This auditory model will render the speech signal as 13 frequency bins at each time step, taken as intervals of 10 ms.

The fact, detailed in Chapter 4, that our model seems able to distinguish different

initial stops with information that can be found in the following vowel suggests that we need to eliminate discrete phonetic tokens from our models and instead consider the perception of multiple, high-dimensional, acoustic cues. We are in agreement with Port and Leary (2005) who state “that a fundamental mistake of the generative paradigm is its assumption that phonetic segments are formal symbol tokens” and would argue that this complex view of speech perception warrants further investigation through the adaptation of the classic computational models first detailed in Chapter 2. Accordingly, in this chapter we will develop two models. In the first, an individual agent will be trained on the reduced auditory inputs from the Lisker and Abramson data, selected for its low variance, as detailed in Chapter 4. In the second, a population of agents will communicate with each other, as detailed in Chapter 2, but will only perceive the same four time-frequency bins.

Both the auditory and articulatory system need to be controlled by a suitable cognitive architecture, one that is capable of learning to control the vocal tract and make sense of the real speech signal. To achieve this, we will develop the SOM presented in Chapter 2 into the slightly more complex system detailed in Appendix B.2. This approach is justified in light of the success of our and others’ previous implementations (Oudeyer 2005c). Thanks to our development of vocal and auditory models, the task facing our neural architectures is only slightly more complex than in previous, highly abstract, simulations of emergent phonology. The dimensionality of the auditory space will be increased to 4 to represent the 4 time-frequency bins identified in Chapter 4 as being crucial to the identification of initial stops and the dimensionality of the articulatory space will be increased to 9 to present complete articulatory targets for the 9 parameters controlling the articulatory model.

The remainder of this chapter will be structured as follows. After a brief discussion unifying the tools developed in previous chapters we will, in Section 5.3, detail a model to test loosely constrained production and high-dimensional perception in an individual, perceiving the end-point stimuli of the Lisker and Abramson data. After discussing these results in Section 5.4 this individual is then placed into a population of speakers in Section 5.5, testing their ability to converge on a shared system of communication. The results of this work will then be discussed in Section 5.6.

## 5.2 True empiricism in language models

To capture ‘real speech’ and phonetic emergence we will take the system detailed in Chapter 2 and add the vocal tract model of Chapter 3 and the peripheral auditory system



of Chapter 4. Having established the validity of previous work, a unification of linguistic theory (Fowler 1996; Nearey 1997), vocal tract (Cook 1993), peripheral auditory system (Slaney 1998), and population model (Oudeyer 2001) will be attempted. Details can be found in Appendix B.2. Through this integration, we can enable the agents in a population model to produce ‘real speech’ utterances. These not only introduce a greater level of biological plausibility but also form the lowest level of reasonable model abstraction given current linguistic theory, confronting the challenges of speech perception and emergence.

The population interaction and neural model will initially be based upon our work in Chapter 2. The dimensionality of each self-organising map will be adjusted to account for the input from the auditory system and to enable the manipulation of the nine different dimensions of the vocal tract. The vocal tract model is an exact replication of Cook’s (1993) work. Accordingly, we refer the reader to his work for details of the implementation.

Having defined the vocal tract and neural models, the auditory system will be based upon ‘Lyons Passive Long Wave Cochlear Model’ (Slaney 1998). This model has been used successfully in previous work, e.g., Aleksandrovsky, Whitson, Andes, Lynch, and Granger (1996) and Chapter 4, we feel that it strikes the right balance between abstraction and biological accuracy. It captures certain properties of the auditory system, for example the role of outer hair cells on loudness recruitment, but through necessity abstracts away from “factors such as displacement of the stereocilia due to the influence of Brownian motion and stochastic resonance” (Araújo, Magalhaes, Souza, Yehia, and Loureiro 2005, p. 6).

We should not delude ourselves that the models used here represent a perfect reproduction of the human vocal tract and auditory system. For example, Faundez-Zanuy and McLaughlin (2002) illustrate a number of problems when defining an accurate model of the airflow within the vocal tract. This leads to a call for greater accuracy in synthesis. However, when dealing with an agent system, such as the one detailed in this chapter, a balance has to be struck between accuracy and computational efficiency.

### **5.3 Modelling the emergent perception of real speech**

Before we develop our population model, an individual agent is given the task of learning the six end point stimuli from the Lisker and Abramson data using the reduced input discovered in Chapter 4. As shown in Figure 5.3, the agent possesses an artificial



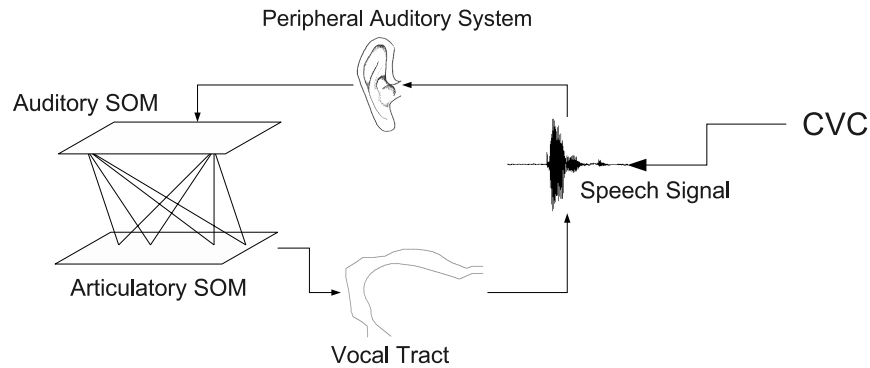


FIGURE 5.1: Illustration of the artificial agents studied in this chapter. Signals are generated from the Lisker and Abramson speech database and from the utterances of the agent. These are processed by the agent's peripheral auditory system which activates the auditory and articulatory spaces, which are self-organising maps (SOMs), via weighted connections. Hebbian learning enables the agent to acquire a mapping from audition to articulation, while the SOMs adjust to represent the perceived signal and control the vocal tract. Once the validity of the individual agent has been established we replace the Lisker and Abramson database with the utterances of other agents in a population model.

auditory system (Chapter 4) plus a model of the vocal tract (Chapter 3) as well as two self-organising maps (SOMs), each of 500 nodes, connected to each other through weighted links (Kohonen 1990). One SOM models the agent's articulatory space, to control the parameters of the vocal tract; and the other models the auditory space, which adjusts to match the output from either its own utterances or the Lisker and Abramson database. Thus, the weighted connections enable, through Hebbian learning, the acquisition of a mapping from audition to articulation.

The model operates as follows. We randomly select an utterance from the appropriate speech sound database, which is then processed by the peripheral auditory system to produce an output that activates the auditory SOM. This in turn activates the articulatory SOM whose output drives the vocal tract model to produce speech output. The agent's task is to match this output to what was initially heard. The difference between the auditory representation of the initial sound selected from the database and that of the agent's uttered response forms the error signal for a Hebbian updating of the weighted links from auditory to articulatory SOMs. We would argue in humans that this error signal becomes internalised after a period of babbling and acquisition and forms an understanding of the intentionality present within the continuous loop of production and perception. As a result, the agent comes to learn which utterance reproduces the perceived sounds within the Lisker and Abramson database or vowels within the corresponding vowel space. In this system, production and perception are

tightly coupled, reflecting our view that this is essential to the proper understanding of speech communication.

Given each individual agent, now equipped with the ability to produce and perceive speech, we will create a population of agents to speak and listen to each other over the course of a simulation. This follows on from work by Oudeyer (2001), in which each agent simply speaks and listens, adjusting its SOM accordingly as it develops a mapping from its own perceived auditory signal to the control of its articulators. By forming a population of these agents and allowing them to speak and listen to each other, Figure 2.4, a shared communication system can emerge.

Having developed this model we define an experimental framework in which to test it. Each agent's SOM is governed by a Gaussian width and learning rate. The learning rate is set to 0.21 and the Gaussian width is 0.4, a reasonable set of parameters discovered through experimentation. First in Section 5.4.1 we will test the abilities of an individual to acquire the Lisker and Abramson stimuli from a reduced set of inputs before developing a population model in Section 5.5. Having formed a population of 10 agents, the simulation runs for 1250 iterations, giving the agents enough time to converge to a stable phonetic system. These utterances are formed from 2 to 4 articulatory targets selected from the chosen agent's articulatory space. This presents the challenges of co-articulation and segmentation to the listening agent. We also record the Euclidean distance of articulatory SOM configurations and the level of clustering within each agent's SOM. This allows us to measure the eventual phonetic similarity between agents and the refining of the lexicon from a set of essentially random articulatory gestures to a discrete, culturally plausible set of phonetic tokens.

## 5.4 Results and analysis

In this section, we first present the results of extensive simulations as described above, showing that our artificial agents are capable of learning important aspects of speech communication. We then analyse the results further to show that there are some shortcomings in terms of sensitivities to parameter settings. These are reduced and robustness improved in a theoretically well-motivated fashion by introducing goals on the part of the agent to achieve effective communication by trading auditory distinctiveness of its utterances versus the articulatory effort required to produce them. We will use these exploratory results to test the following hypotheses.

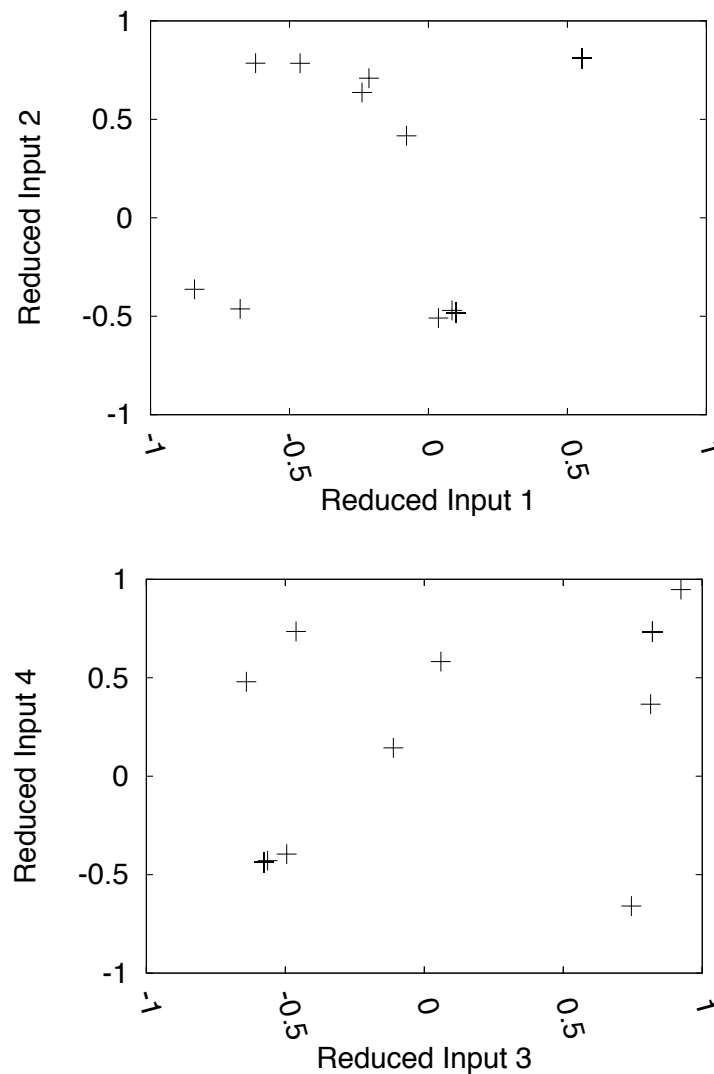


FIGURE 5.2: An example of auditory convergence for an agent trained on the four time-frequency bins derived from the Lisker and Abramson data set. After training there has been a significant reduction in the Euclidean distance between the 500 randomly initialised points.

### 5.4.1 Reproduction of CV utterances from limited auditory input

We will first test the following hypothesis:

*Hypothesis 5.4.1.* The agent developed in this chapter will be unable to reproduce the initial stops from the limited auditory input. If this proves to be true then it will be clear that the limited (non-phonetic) input is not sufficient for perception.

In Figure 5.2 we present an example plot of an agent’s SOM converging to the four reduced inputs, equipped with the peripheral auditory system detailed in Chapter 4, that has converged on a number of auditory features presented by the four time-frequency bins derived from the Lisker and Abramson data set. It is worth noting that although

the auditory space has not converged on each phonetic token the individual agent is still able to partially reproduce each utterance with its own vocal tract.

In Hypothesis 5.4.1 we stated that the agent developed in this chapter will be unable to reproduce the initial stops from the limited auditory input. To determine the level of successful reproduction we measured the Euclidean distance between the formant values of the Lisker and Abramson utterances and the agents' articulations. The average Euclidean distance (summed over  $F_1$ - $F_4$  with standard deviations in brackets) was reduced from 132.8(80.49)  $\text{Hz}^2$  to 84.3(71.03)  $\text{Hz}^2$  after training with a significance of  $\alpha = 7.8 \times 10^{-6}$ , for a standard two-tailed  $t$ -test, showing that this reduction was significant although not perfect.

### 5.4.2 Randomisation of auditory inputs

The convergence of the articulatory space, despite the lack of phonetic idealisations, warrants further investigation.

*Hypothesis 5.4.2.* The articulatory space will still converge in the absence of any meaningful auditory input. If proven this would establish a flaw in our model as we would expect the articulatory space to converge only in response to the auditory input.

Accordingly, we developed a control experiment where the individual agent was exposed to random auditory input. As shown in Figure 5.3 the auditory space fails to converge. It seems that a reduced structured auditory input is necessary for the structured convergence of the articulatory space and the reproduction of the perceived utterances.

Finally we investigated whether the auditory space's convergence on acoustic features was necessary for the successful convergence of the articulatory space, Hypothesis 5.4.2. We found that the articulatory space did not converge in the presence of random auditory input and the difference in final configurations (random auditory input versus Lisker and Abramson auditory input) had a significance of  $\alpha = 5.3 \times 10^{-9}$ .

### 5.4.3 Removal of the peripheral auditory system

*Hypothesis 5.4.3.* The removal of the peripheral auditory system component of the model will have no effect on the accuracy of the agents' utterances.

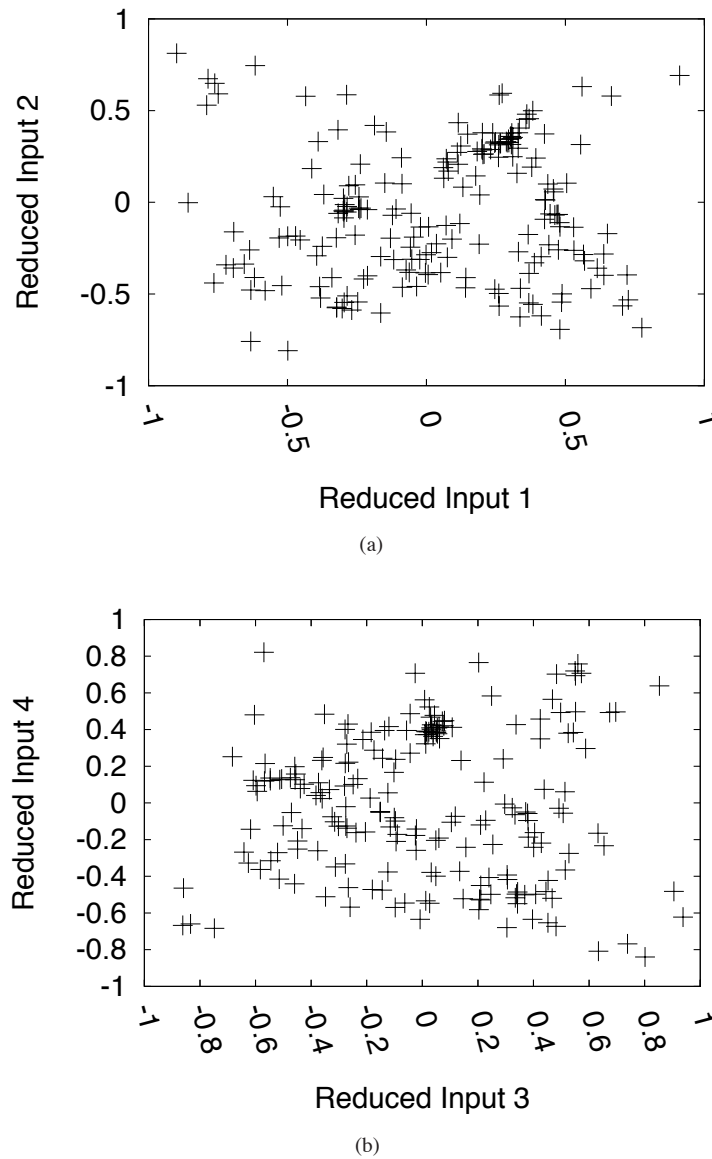


FIGURE 5.3: When presented with random auditory input the auditory space fails to converge – the Euclidean distance between points is not reduced significantly.

In keeping with the methodology detailed in Chapter 4 we replace the peripheral auditory system with a high resolution FFT. As shown in Figure 5.4, the convergence on the auditory feature set is not entirely complete. However, more detailed analysis shows that this difference is not significant. This suggests that the main effect of the peripheral auditory system is in enabling the boundary shift, which is not captured in this model.

In greater detail, we replaced the peripheral auditory system with an FFT analysis to judge the effect on the accuracy of the agents' utterances, as stated in Hypothesis 5.4.3. We found that the Euclidean distance of the auditory features between the simulation

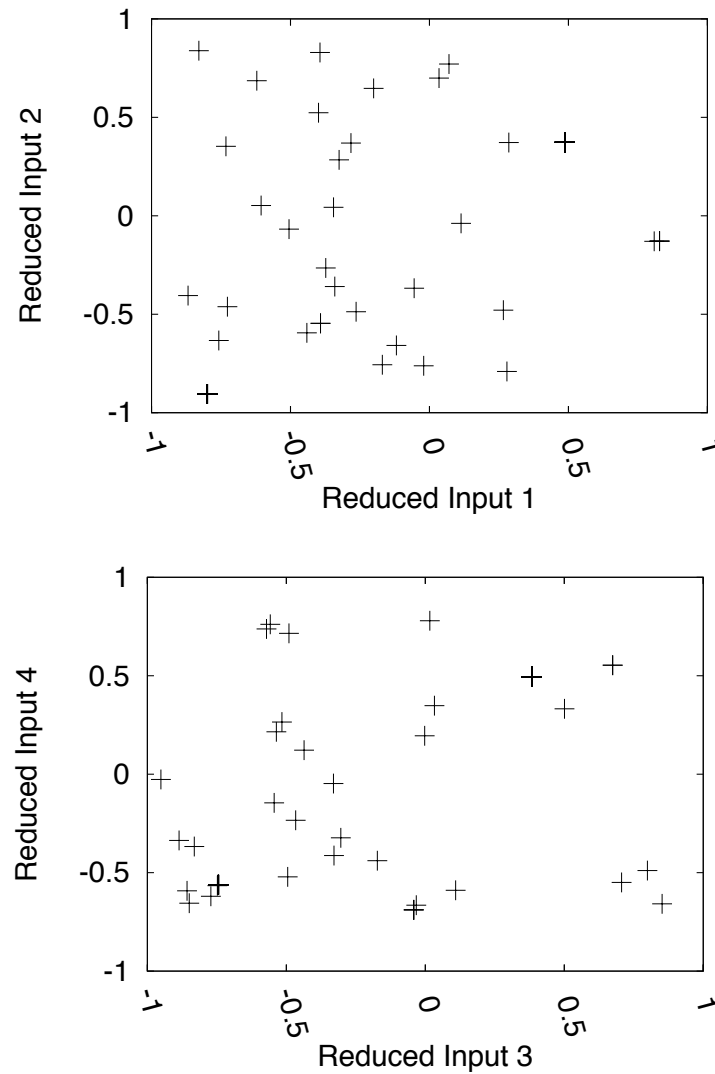


FIGURE 5.4: An example agent equipped with the high resolution FFT detailed in Chapter 4.

and the Lisker and Abramson data reduced from 1.08 to 0.80 with the introduction of the peripheral auditory system. This result had a significance of  $\alpha = 0.49$ , suggesting that this difference is not significant.

#### 5.4.4 Sensitivity to SOM parameter settings

In Chapter 2, the emergence of realistic vowel systems in artificial agents has been found to depend upon parameter settings, notably the width  $\sigma$  of the Gaussian function used in weight adjustment. Intuitively, the lower the sensitivity of the results to parameter variation, the more robust we can take the simulation to be. Accordingly, we have examined this sensitivity as depicted in Figure 5.5. This shows that learning fails, in

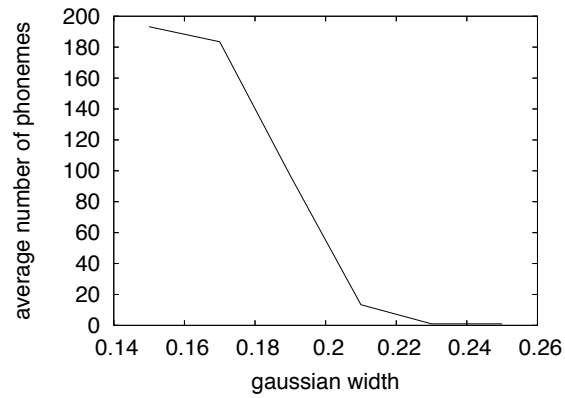


FIGURE 5.5: The effect of a varying  $\sigma$  value on the size of the final articulatory system. As can be seen realistic phonetic systems are only obtained under a narrow range of parameter values.

the sense that there is either no convergence or convergence to a single node, outside of a narrow set of values around  $\sigma = 0.2$  (the value used in all simulations up to this point). This vulnerability undermines the plausibility of the agent system and needs to be addressed.

#### 5.4.5 Articulatory effort and auditory distinctiveness

Chapter 2 has shown that by introducing further realism, by giving the artificial agents a drive to minimise articulatory effort through signal grounding, greater robustness can be gained. In an attempt to overcome the vulnerability inherent in our existing agent system (cf. Figure 5.5), we have added this feature to our simulations. We believe this addition to be theoretically well-motivated. As stated by Lindblom (1990) in H&H theory speakers seek to maximise by hyper-articulation the auditory distinctiveness of their produced speech signal while minimising by hypo-articulation the articulatory effort of production.

We introduce articulatory effort by adding the muscle model from Chapter 3 to the current vocal tract. With this added feature, the agent's utterance is produced as follows. The 10 most highly activated nodes of the articulatory space are first identified, and the effort required to produce them is computed. The agent then selects the minimum-effort utterance as its output.

Introducing auditory distinctiveness within our agent highlights a challenge for existing articulatory theories (Liberman and Mattingly 1985; Fowler 1986): If we perceive the gesture, why do we produce an acoustically distinct signal? But here, no discrete

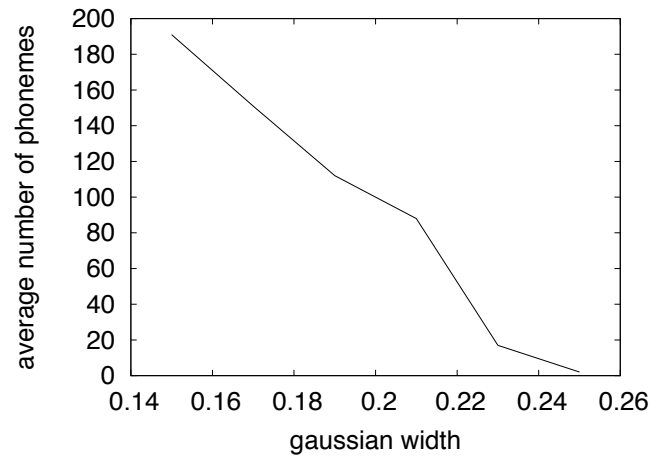


FIGURE 5.6: Sensitivity to parameter variation of a typical agent modified to trade articulatory effort against auditory distinctiveness according to Lindblom’s H&H theory. This extension enables us to sustain emergence of plausible phonetic systems for a Gaussian width  $\sigma$  over a wider range than hitherto.

phonetic tokens form within the auditory space – by focussing upon the reduced inputs identified in Chapter 4 this simply wouldn’t be possible. If no phonetic tokens form we cannot (as we did in Chapter 2) directly optimise the auditory space by pressuring the auditory tokens towards dispersion. To overcome this problem, we might ask if phonetic tokens need to sound distinct in themselves or just distinct enough to convey the intended gesture. We adopt the latter standpoint. Auditory distinctiveness is then introduced by considering the signal’s ability to convey distinct gestures, as follows. At each time step, we take the auditory node maximally activated by the database utterance, and generate 10 alternatives by random sampling around this node. Through the mapping from audition to articulation, these alternatives will cause activation of nodes ( $\leq 10$ ) within the articulatory space. Of these, we select that node that maximises the distance from the remainder as the most distinctive. Finally, the set of auditory nodes over each time step of that utterance that provides the most distinct set of gestures is selected. In this way, the tension inherent in H&H theory is achieved without the formation of abstract phonetic tokens.

With these changes in place, we repeated the simulations detailed in Chapter 2 to give the results shown in Figure 5.6. As can be seen, the new agents designed according to H&H theory are more robust to parameter variation.



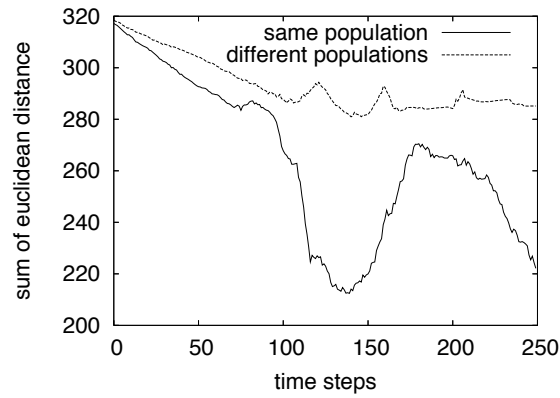


FIGURE 5.7: By recording the Euclidean distance between the articulatory parameters of two randomly selected agents a measure of linguistic similarity is obtained. We can see how the articulatory parameters of each agent converge and diverge as they negotiate and settle upon a final shared set of articulatory parameters. The comparison between an agent within the population and an agent removed from it remains stable and distant.

## 5.5 Language without strong constraints

We now place individual agents into a population and instead of listening to utterances selected from the Lisker and Abramson database, they listen to each other. By forming this population we hope to reproduce the convergence seen in Chapter 2 in this more complex model, so testing the hypothesis:

*Hypothesis 5.5.1.* In a population of agents the simulation will converge to a shared set of auditory features and a set of articulatory tokens. By establishing a shared communication system among a population of agents we will demonstrate that auditory production and perception is possible without having to postulate the existence of abstract phonetic tokens.

We can see from Figure 5.7 that another agent within a population would be able to ‘understand’ a given utterance as both are equipped with similar articulatory prototypes. The Euclidean distance between their articulatory prototypes has been significantly reduced moving from 30.5 at the beginning of the simulation to 15.55 at the end. The fact that these similarities arise from the perception of other agents’ speech signals suggests that each agent is able to decompose each utterance into an appropriate articulatory gesture. The convergence of each agent’s articulatory lexicon into a limited set of prototypes reinforces this view.

We can see from Figure 5.8 how the number of articulatory prototypes has reduced from an initially random 500 possible utterances to a distinct set of approximately 18.

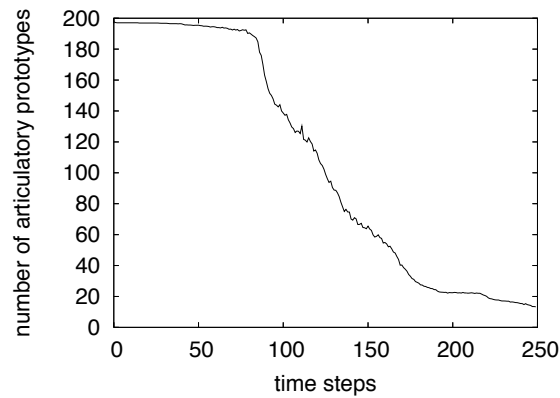


FIGURE 5.8: The number of articulatory prototypes in a given agent's SOM over time

By taking the average of the entire population it is clear that this behaviour is reproduced by all of the agents. The process of speaking and listening to itself causes a partitioning of the articulatory space into a variety of distinct utterances. As is the case for the individual agent, the auditory space remains dispersed, adjusting to match the auditory features provided by the peripheral auditory system but never converging to a distinct set of auditory phonemes. The potential for auditory tokens to emerge is a possibility within the model. Following on from the results of Section 2.2, we could expect to see a limited set of phonemes, each one defined by unique, invariant auditory features. However, this does not happen in this model. The mapping from auditory space to articulatory space drives a convergence within the articulatory space while the auditory system remains dispersed, exploiting and perceiving the complex cues of the speech signal.

To consider the plausibility of the resulting systems, we compared a number of convergences within the population of agents to the set of human phonetic tokens, Figure 5.9. As we can see from these results both the human and artificial systems have a similar shape, a clear cluster of 'popular' phonetic systems is followed by a long tail of increasingly complex systems. It is clear that the average number of distinct phonemes is smaller for the artificial models than the biological systems. We would speculate that there are forces at work, driving greater phonetic diversity, that are not captured by our current model.

Figure 5.9 was produced under ideal conditions. Like previous work (Oudeyer 2001), this model is highly sensitive to parameter variation. As we can see from Figure 5.5, a variation of the Gaussian width governing the SOMs of the agents has a dramatic effect on the resulting phonetic systems. If the Gaussian width is too small, then no convergence takes place. Conversely, a wide Gaussian setting results in only a single articulatory gesture remaining within the population. This susceptibility to parameter

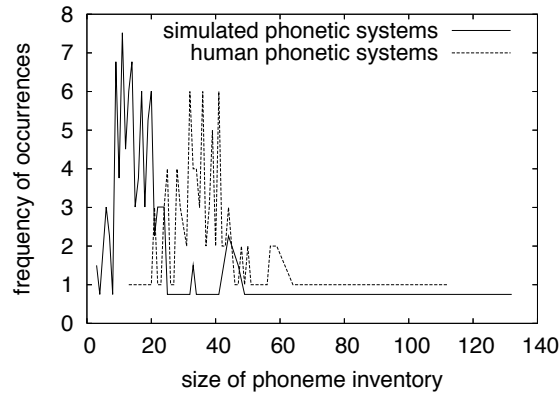


FIGURE 5.9: A comparison of the number of simulated articulatory prototypes and human phonetic systems. The human data were obtained from Tambovtsev and Martindale (2007) and the simulated data from 150 runs of the model under ideal conditions ( $\sigma = 0.21$ ).

variation remains as a weakness of the model and will be addressed in future work.

We now turn our attention to phonetic convergence within the population. By measuring the Euclidean distance between two randomly selected articulatory spaces we can see how the population agrees on a final shared system. The initial divergence after convergence forms one point of interest. We view this as a form of uncertainty before convergence; as each agent listens to others in the population they diverge away from each other. Eventually however, all of the agents agree on a shared system resulting in a final convergence. This convergence can be seen in comparison to the Euclidean distance between agents from different populations. As these two randomly selected agents have never heard each other, their final phonetic systems remain distinct.

We next considered whether a population of agents could converge on a shared set of reduced auditory features and articulatory tokens (Hypothesis 5.5.1). We found that the average Euclidean distance between auditory features, normalised to the range [0–1] was reduced from 0.43(0.023) (for different populations) to 0.32(0.0165) with a significance of  $\alpha = 1.3 \times 10^{-6}$  for a standard two tailed  $t$ -test (for two agents from the same population). While the articulatory prototypes were reduced from 1.33(0.0681) to 0.75(0) with a significance of  $\alpha = 5.1 \times 10^{-11}$ .

## 5.6 Summary

In our analysis, it is important to emphasise a number of points. Firstly, agents do not directly perceive articulatory gestures, they only hear the resulting auditory output. This

is possible because the manipulation of the artificial vocal tracts produces ‘speech’. On hearing these utterances, no phonetic classification within the auditory system takes place, rather the auditory system detects multiple auditory cues derived from the entire utterance and recovers the intended articulatory gestures. This is possible because a mapping between auditory cue and articulatory gesture has been derived through a process of Hebbian learning, activated when the agent listens to its own articulations. This mapping is highly complex, as each auditory cue can contribute to a number of probable gestures. The final articulatory gesture is recovered by the weighted interaction of the auditory cues. The gestures themselves are not discrete points, rather they form basins of attraction, the size of which is defined by the width of the Gaussian placed over the recorded centre point. A ‘recovered’ articulatory gesture is idealised as the centre point of the corresponding basin of attraction, allowing for variation through co-articulation or incomplete Hebbian learning to be accounted for.

In the auditory space, we have considered the role of the peripheral auditory system and we have shown that articulatory gestures can form the objects of speech perception, discarding auditory objects of perception. This became possible once the nature of the peripheral auditory system was considered. If the agents conversed in complete abstractions and not simulations of real speech we would not have been able to investigate these aspects.

Although the simulations presented in this chapter are admittedly simple and preliminary, and we wish to avoid making inflated claims about their significance for understanding speech perception, we do believe that they offer interesting insight into the controversy surrounding the competing theories. Most obviously, we have shown that the computational implementation of a loosely-constrained agent system is able to learn the useful ability to reproduce distinctive sounds, such as could form the basis for aural communication. Further, we have argued that there are no obvious candidates for abstract phonetic-like tokens within the system. Indeed, it can be argued that with an input region reduced to four time-frequency bins it would be impossible for phonetic tokens to form. What then are the objects of perception within our simulation, if not symbolic? The obvious alternative to a symbolic, representational theory of speech perception is the direct perception of complex acoustic cues, with symbols replaced by implicit knowledge of the action-perception link between articulation and audition.

Thus, instead of viewing the speech signal as a static medium for direct perception, it is important to emphasise that the vocal tract is controlled by a speaker with knowledge and intention. The speaker has implicit knowledge of the perceptual apparatus of the listener, since he/she is similarly equipped, and is therefore able to manipulate

the speech signal to exploit its properties—resulting in a loosely constrained signal. He/she tacitly exploits the situation not to convey vocal tract gestures but to convey the ‘affordances’ present within the environment. This notion of *affordances* is central to J. J. Gibson’s theory of direct perception in vision (Gibson 1979), which has a strong link to Fowler’s direct realism theory of speech perception. According to Gibson, “The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill” (p. 127). For the concrete objects of visual perception, this can be thought of as the purposes to which the object can be put, given that the observer is an intentional agent with beliefs, desires, etc. However, the term seems capable of extension to the less concrete domain of speech communication, where the intentional nature of the participants is no less important. Again quoting Gibson, “The richest and most elaborate affordances ... are provided by other animals and, for us, other people” (p. 135). Thus, when constructing a phonetic theory, we should consider both the passive processing of perception and the speaker’s active process of creation. The speaker is trying to create an affordance, and they will use any means available (signal, context, gesture) to achieve this goal.

Speech *is* special, but not in the way that motor theorists believed it to be. It is not special because of any active reconstruction of the gesture (e.g., by analysis-by-synthesis); it is not processed separately and differently to other complex sounds. Rather it is special simply because it is produced by a speaker. The point is trivial but important. Without it we must consider speech to be a passive object in the environment, and the listener simply has to accept the signal ‘as is’ without any consideration of the intentionality of the speaker. Accordingly, the signal must be strongly constrained to facilitate direct perception, leaving no room for pattern recognition. This leaves current direct realist theories open to attacks based on the observed lack of invariance—as when the signal does not reflect the gesture—because the, apparently necessary, strong constraints are broken.

When we return to the Gibsonian notion of the direct perception of affordances, the role of speech perception becomes clear. “As Ludwig Wittgenstein knew, you cannot specify the necessary and sufficient features of the class of things to which a name is given. They have only a family resemblance” (Gibson 1979, p. 134). When applied to speech, it is clear that instead of classification through invariance we are forming “family resemblances” through loosely constrained speech perception, allowing the overarching affordance to be perceived and “to perceive an affordance is not to classify an object” (Gibson 1979, p. 134). In speech communication, we do not classify phonetic objects and we do not classify articulatory gestures. But this does not mean we cannot learn how to use things (speech) and perceive their uses (listen). Ultimately, speech is

an attempt to exploit and manipulate the complex web of affordances present in any human social interaction.

## Chapter 6

# Speech perception as non-symbolic pattern recognition

Despite intensive ongoing research, our understanding of human speech perception (like perception in general) remains largely incomplete, with many competing theories and controversies surrounding it. When communicating by speech, we have the strong intuition and impression that we are utilising a *discrete* compositional code, in which concepts are transmitted from speaker to listener by stringing together basic speech units that can be interpreted at differing hierarchical levels: phonetics and phonology, syntax/grammar and semantics. This is the so-called code model of speech communication (Blackburn 2007). However, the medium of transmission is sound pressure waves in air, and these are inherently *continuous*. Hence, a central problem for the scientific study of speech communication is to determine what are the basic elements of the discrete code, how they are signalled in the continuous pressure wave, and how they are extracted by the perceptual system. Viewed in this light, the problem is one of determining the objects of perception (e.g., Diehl and Kluender 1989); precisely what is it that we attend to when we listen to ('decode') speech? But we should not forget that speech communication is reciprocal; (virtually) every speaker is also a hearer of speech. Although this sort of reciprocity is typically found in any animal signalling system, where the source of the signal is an intentional agent (Dennett 1987), it is not the case for perception in general. When we look at a table, for instance, it has no intention to alter our beliefs, desires, or other aspects of mental content; it is just a table.

We are now in a position to explore inherent tensions in current phonetic theory: specifically, the conflict between *continuous* production and *discrete* perception, and that between passive direct perception and the notion of top-down hypothesis testing that arises from the reciprocal awareness of intentionality.

## 6.1 Overview

Current phonetic theory is divided by two points of contention: the link from production-to-signal-to-audition, and the object of perception/cognition. Traditionally, theorists have attempted to arbitrate between these contentions by conducting experimental phonetics research, often in a search for ‘invariants’ that can be linked to phonetic tokens or labels (Perkell and Klatt 1986). The underlying assumption has been that specifying invariants in either articulation or audition automatically indicate the object of perception/cognition. Yet very slowly, realisation has grown that computational modelling offers an alternative approach, as exemplified in the work in this area of Liljencrantz and Lindblom (1972), Bladon and Lindblom (1981), Lindblom, MacNeilage, and Studdert-Kennedy (1984), Protopapas (1999), Damper and Harnad (2000), de Boer (2000a, 2001, 2003), Oudeyer (2005a, 2005b, 2005c) and others. Typically however, this work has proceeded in an unembodied, abstract fashion avoiding critical phonetic questions. Even when aspects of language are embodied in robotic systems (Steels 1999; Roy 2005), this has been at the syntactic level — giving little consideration to the physical production and perception of the underlying phonetic tokens.

This situation persists because the study of phonetics faces two fundamental challenges (Lindblom 1986b): *invariance* and *segmentation*. Currently no unique set of defining features for a given unit of language has been found. Although a listener will perceive a distinct phoneme, we cannot define a set of acoustic features that will hold true for all occurrences of that phoneme. Secondly, segmentation defines the fact that an utterance cannot be unambiguously segmented into temporal non-overlapping chunks; therefore, co-articulation presents a further barrier to acquisition. These challenges arise, either through insufficient data analysis — we haven’t yet found the acoustic properties that consistently define a phoneme — or through a flaw in the theories underpinning our understanding of speech perception.

In phonetic research, there has long been a connection made between the perceived constraints on signal and articulation and the neural correlates of speech perception. As a result, those holding to an articulatory theory of speech perception look for specifying invariants within articulatory gestures, while those following an auditory theory study the constraints within the acoustic signal and/or its auditory representation. So-called double-weak theory (Nearey 1997) suggests that this approach is not so straightforward as both the articulation and auditory signal are loosely constrained and distinctly lacking in specifying invariants. In this chapter, we ask what this means for the search for the cognitive objects of speech perception and propose a new theory: double-weak direct



realism. This theory combines the loosely constrained signal with the cognitive notion of direct realism, i.e., we directly perceive the world through our senses unmediated by symbolic representation. After investigating the basis for this theory in Section 6.2, we consider existing evidence in Section 6.3 and discussed why current phonetic models are ill-equipped to test this approach in Section 6.4 further justifying our alternative model presented in Chapter 5. This chapter will now attempt to form an account of the data presented in this work.

## 6.2 Theories of speech perception and direct realism

We will now use Nearey's framework to map out the current state of of phonetic research. In doing so we will expose current assumptions and define a place for our alternative account.

### 6.2.1 Nearey's framework: strong and weak relations

With this background in mind, we can now consider the various competing theories of speech perception. Based on Nearey's (1997) framework, we can classify these into strong-articulatory, strong-auditory, double-strong and double-weak. Figure 6.1 contrasts three of these classes of theory schematically. According to Nearey, "the term *strong* is taken to imply a simple, robust, and transparent relation between physical and symbolic elements" (p.3241).

Nearey's (1997) classification system is useful as it clearly delineates the current controversy and defines an alternative, namely the double-weak theory of production and perception. Nearey assesses current speech perception theories on the 'strength' of their connections between the segments of speech: the signal, articulation and perception. The theories that argue for a tightly-constrained connection between all three aspects are considered to be 'double-strong' in nature; there is a clear mapping from articulation to speech token to audition. By comparison Neary cites evidence from both the contra-auditory (Liberman, Cooper, Shankweiler, and Studdert-Kennedy 1967) and the contra-gesturalist (Diehl and Kluender 1989) camps which show that there is a weak connection between the three aspects of speech. This 'double-weak' theory contends that articulation is only 'stylised' by real-time physical constraints while the resulting utterance needs to be reconstructed from multiple, variant acoustic cues. We will now consider each theory in greater detail.

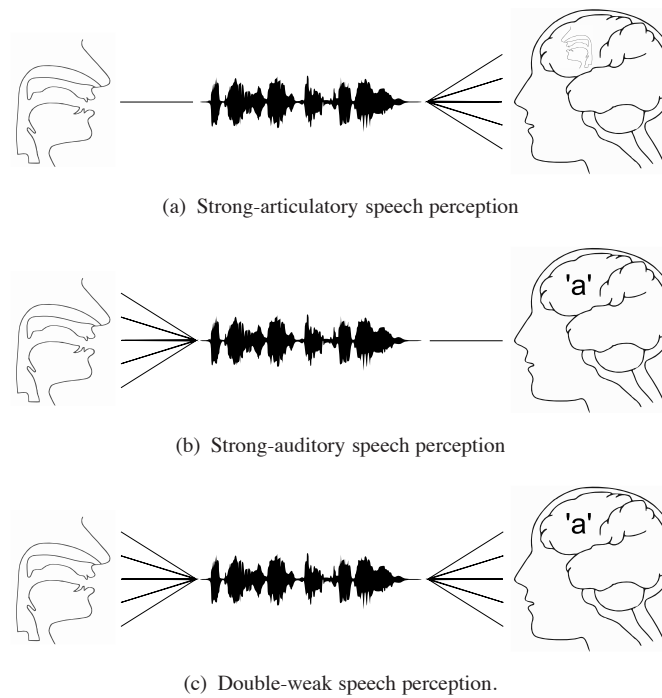


FIGURE 6.1: Conflicting phonetic theories use evidence of strong constraints on articulation or audition to argue for different symbolic systems of perception. For example, the existence of invariant acoustic specifiers would be taken as evidence for the formation of abstract phonetic tokens during speech perception. See text for further explanation of (a), (b) and (c).

Strong-articulatory theory (Fig. 6.1(a)) was historically the early front-runner in the form of motor theory, forever associated with Liberman and his colleagues at Haskins Laboratories (Liberman, Cooper, Shankweiler, and Studdert-Kennedy 1967; Studdert-Kennedy, Liberman, Harris, and Cooper 1970; Liberman and Mattingly 1985; Liberman and Mattingly 1989). Motor theory holds that the objects of perception are symbolic, gestural tokens obtained (e.g., through analysis-by-synthesis) on the part of the listener recreating the speaker's intended vocal tract gesture using an innate, virtual vocal tract synthesiser. (Schematically, this is depicted in Figure 6.1(a) as a vocal tract symbol within the head of the listener.) Motor theory obviously emphasises the fact that (almost) every human listener is also a speaker. In the figure, the assumed strong constraints on articulation are shown as a one-to-one (straight line) mapping from gesture to acoustic signal, whereas the assumed weak constraints on audition are depicted as a one-to-many (multi-line) mapping. In other words, variability in the acoustic signal (arising from the complex acoustic consequences of the gesture) is unproblematic since the invariant gesture is recoverable. Although gestural theories have been rather discredited by researchers in the field of phonetics (Diehl and Kluender 1987; Sussman 1989; Ohala 1996), if not viewed with total scorn, they still have a clear

influence within cognitive science. According to Galantucci, Fowler, and Turvey (2006, p.361), “The motor theory of speech perception . . . is among the most cited theories in cognitive psychology . . . it is perhaps the only theory of speech perception recognized outside the field of speech”.

Strong-auditory theories, by contrast, assume weak constraints on articulation leading to acoustic variability, but this is of little consequence as the variability is in irrelevant aspects rather than the specifying features that enable the listener to recover a token corresponding to that set of features. Hence, the object of perception is an abstract (perhaps ‘phonetic’) token as depicted in Fig. 6.1(b). Strong-auditory theories include Steven’s (2002) well-known quantal theory. According to this, articulations don’t have to be precise in that perturbations of gestures will still yield an acoustic steady state (the quantal region) corresponding to ‘distinctive’ features. Thus, gestural imprecision will not affect perception of the intended token.

Double-strong theories (not depicted), as their name implies, hold that strong relations hold between physics and symbol for both articulation and audition. An example of such a theory is that of Stevens and Blumstein (1978, 1981), whereby distinct vocal tract gestures give rise to robust and transparent acoustic consequences (e.g., the burst spectrum for an initial stop consonant) that directly signals place of articulation.

Double-weak theories, Fig. 6.1(c), posit weak constraints on both articulation and audition. Typical of the double-weak position are connectionist theories like TRACE (McClelland and Elman 1986) and pattern recognition theories, such as the normal a priori probability (NAPP) model of Nearey and Hogan (1986), that assume that the task of extracting invariants from among the irrelevant statistical variation in the acoustic signal can be delegated to a trained pattern classifier. To a large extent, they are a natural conjoining of the strong-articulatory and strong-auditory positions, more so than double-strong theories because, according to Nearey, the evidence for “simple, robust, and transparent relation[s] between physical and symbolic elements” is largely absent. By comparison a double-weak standpoint holds that “only two conditions are necessary for speech to operate as an effective communication system. First, a symbol sequence must be encoded into gestures. Second, the acoustic output of those gestures must provide the listener with auditory cues sufficient to decode the intended symbol sequence” (Nearey 1997, p.3243). In conjoining the two, however, it is necessary to arbitrate on the conflicting issue of the object of perception: strong-articulatory theory takes this to be vocal tract gesture, whereas strong-auditory theory takes it to be an abstract, phonetic-like token. Nearey seems to assume without question that the latter is the case. We speculate that this unstated assumption arises because of the clear

similarity between pattern recognition theories of speech perception and the technology of automatic speech recognition (ASR). In ASR, the goal is very typically seen as the extraction of phonetic equivalence classes from the acoustic signal, and hence it is natural for Nearey to think in exactly these terms. However, this is not essential. It is possible to conceive of a double-weak theory in which the objects of perception are gestures. This alternative seems to fit better our modelling data, presented in Chapter 5. Hence, we now move to discussing a gestural theory, Fowler's (1996) direct realism, that is less susceptible to many of the charges made against motor theory by phoneticians and speech scientists. The latter include the claim that "speech is special" in some meaningful way that has implications for the science of perception (as opposed to being a mere statement of the obvious), and that there exists an innate, virtual analysis-by-synthesis system, required to accommodate the observation that those unable to speak can still decode speech.

### 6.2.2 Direct realism

The direct realism theory of speech perception is due to Fowler (1996). Like motor theory, it is both gestural and a product of Haskins Laboratory. Although related to motor theory, however, it differs in a number of important ways. Significantly, speech perception is not held to be 'special' ... "and there is no more reason to propose a role for the speech motor system in speech perception than to propose an analogous role for the viewer's locomotor system in visual perception of walking" (Fowler 1996, p.1731). Rather, "listeners perceive gestures not by means of a specialised decoder, as in the motor theory, but because information in the acoustic signal specifies the gestures that form it" (<http://www.haskins.yale.edu/CaseStatement/Haskinscase.pdf>). There is, of course, some sense in which this must be trivially true.

Figure 6.2(a) depicts this schematically. The theory is double-strong in that the acoustic signal is seen as a direct, transparent link between gesture and perception. In this schematic, the arrow in the listener's head is intended to indicate implicit knowledge of that link. Crucially, at no point is a symbolic token involved.

Direct realism faces various criticisms from within speech science, arising through its association with motor theory, in that they are often treated as one and the same (e.g., Sussman 1989, Ohala 1996). Other criticisms are more specific. What is the force enabling auditory distinctiveness if we only perceive the gesture? Surely H&H theory would be undermined by the fact that we would be driven to maintain *articulatory*

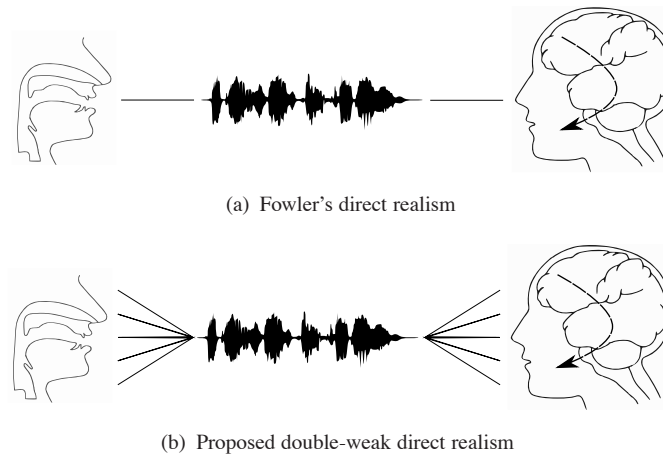


FIGURE 6.2: A comparison of Fowler's direct realism and double-weak direct realism. The phonetic evidence suggests a double-weak approach, while our own work proposes a direct realist cognitive theory. The arrow in the listener's head indicates implicit knowledge of the link between production and perception. See text for further details of (a) and (b).

distinctiveness? Fowler (1996) argues that the acoustic signal still conveys information about the gesture, which accordingly must be sufficiently distinct. But it does not follow that a distinct signal is evidence for a symbolic auditory representation. Another obvious objection, levelled at both direct realism and motor theory, is that those who can't speak can still perceive speech. Motor theorists believe that positing an "innate vocal-tract synthesizer" (Liberman and Mattingly 1985) can overcome this objection, whereas Fowler reemphasises that the direct perception of speech derives from a general theory of perception. She writes that this "inability to reproduce heard gestures does not imply that they did not perceive gestures (any more than the typical person's inability to perform a triple axel implies that he or she cannot see them)" (p. 1738). Direct realism does not have to imply a motor theory of speech perception. It only needs to agree with motor theory in the trivial sense — we obviously 'perceive' the vocal tract as it is the source of the speech signal. It does not need to agree with the claim of motor theory that actual tokens corresponding to vocal tract gestures are involved, e.g., to drive the internal synthesiser used in analysis-by-synthesis. (This is the reason for the different depictions inside the listener head in Figs. 6.1(a) and 6.2(a).) So where direct realism can give theoretical insight is in determining the object of speech perception.

Such disagreements arise because Nearey's classification only considers the means of production, the signal and perception of speech, whereas the major source of disagreement among conflicting theories is the form of the cognitive tokens. Auditory theories hold that these smallest units are resolved as idealised phonetic symbols, whereas motor theory holds that the ultimate forms of perception are gestural

abstractions. Considered in these terms we can see that direct realism and motor theory (lumped together in Nearey's framework) are clearly different, as direct realism considers the perception of speech to be direct "unmediated by processes of hypothesis testing or inference making and unmediated by mental representations" (Fowler 1996, p. 1731) — articulatory or acoustic. Freed from the need to lump all gesturalist theories into the strong-articulatory camp, we can see that direct realism is in fact a double-strong gesturalist theory (as opposed to motor theory's strong-articulatory gesturalist approach). As clearly stated by Fowler: "phonological gestures are the public actions of the vocal tract that cause structure in acoustic speech signals. By hypothesis, they will be found to cause specifiers or invariants in the acoustic signal" (p. 1731).

In this thesis, we propose a view in which speech *is* directly perceived; what is perceived (in the trivial sense) is the vocal tract. Although this appears to agree with Fowler, our view differs in important respects. We question Fowler's naïve realism assertion that invariant "specifying acoustic properties is what allows perception of the phonological properties to be direct" (p. 1731). By "naïve realism", we mean the perceptual version of folk psychology that holds that the perceived world is as common sense would have it appear. We feel that this plays into the hands of a number of arguments against the philosophy of direct realism. Rather we, like Nearey (1997), are "genuinely impressed by the quality of the research by both auditorists and the gesturalists that is critical of the other position" (p. 3242). Given this, we take a double-weak standpoint to the production and auditory perception of the speech signal. However, we do not believe that this necessarily precludes direct realism. As Figure 6.2(b) shows, in this new framework we can conceive of loosely-constrained articulation and perception coupled with the direct perception of speech, leading to a new double-weak direct realism. Perhaps there could be theoretical advantage to decoupling the constraints on speech and the cognitive objects of perception.

Fowler contends that for direct realism to hold, perception must be "unmediated by process of hypothesis testing or inference making and unmediated by mental 'representations' in the literal sense of mental stand-ins for real-world things" (Fowler 1996). We agree that the formation of these symbolic mental representations is strictly unnecessary and in opposition to the wider theory of direct realism. However, we also feel that by eliminating hypothesis testing we are in danger of forming a naïve realist perspective and of discounting the work of various researchers (Nearey 1997) who have failed to find invariant specifiers at the articulatory and acoustic levels.

Whereas naïve realism holds that a perceived object always appears 'as is', unmediated by causal intermediaries, we argue that direct realists must "distinguish between causal

indirectness and cognitive indirectness” (Morvan 2004, p.222). Causal indirectness is perfectly compatible with direct realism and allows for the kind of pattern recognition and hypothesis testing present in double-weak theories like NAPP (Nearey and Hogan 1986). Cognitive indirectness, however, holds out against direct realism by requiring symbolic cognitive intermediaries between the object and the perceiver.

The nature of these symbolic cognitive intermediaries (gestural or phonetic) has been a strong point of contention within phonetic science. However, it is clear that symbolic intermediaries are not mandatory. The signal is not necessarily subservient to the gesture and the gesture is not necessarily subservient to the signal. Both have equal status because of the reciprocal nature of speech communication. The signal must accurately convey the gesture as we understand speech through a tight coupling of production and perception. But the speaker, as an intentional agent, adjusts the gesture so as to produce a recoverable auditory signal, exploiting implicit knowledge of the listener’s auditory system. Instead of the classical examples of directly perceiving a passive environment (as in the visual perception of a table mentioned earlier), we are faced with an intentional actor manipulating the distal objects of direct perception. The implications of this new double-weak direct realism approach will now be explored. We will first consider existing phonetic and neuroscientific evidence before judging the abilities of existing models to test this new theory.

## **6.3 Phonetic and neuroscientific evidence**

Some key studies in experimental phonetics and neuroscience bear on the issues discussed thus far. We will now examine each of these in turn.

### **6.3.1 Phonetic evidence**

Careful reading of the literature reveals that certain key experiments are widely cited both for and against the competing theoretical views detailed in Neary’s framework. In keeping with tradition, we will consider a number of these and assess their impact upon double-weak direct realism. We will mention shadowing response times, trading relations within phonetic perception, and the McGurk effect. Our goal is to highlight the important separation of cognitive aspects in terms of direct realism and articulatory/auditory aspects in terms of double-weak theory.



Regarding shadowing response times, it has been observed (Porter and Lubker 1980) that when completing response tasks, reaction times were significantly faster than when completing choice tasks — which require a simple decision to be made. It was then determined that when repeating a perceived utterance (a choice task), reaction times matched those of a simple response task. When interpreting these results, we are largely in agreement with Fowler: The exploitation of our own gestural knowledge allows for a tight coupling of production and perception, enabling a rapid reaction response, with perception flowing directly and seamlessly to production. There is no time-consuming cognitive break where a signal is resolved into a symbolic phonetic-like token followed by the choice of the correct gesture.

Trading relations describe a phenomenon whereby “settings for one acoustic property could trade for settings of another acoustic property in the phonetic judgments of listeners” (Nitttrouer 2006, p. 1800). This is in accordance with double-weak theory: Via pattern recognition, a phoneme can be recovered from a variety of acoustic cues. These trading relations allow the direct perception of the gesture to continue, because the gesture is the source of the sound.

The well-known McGurk effect (McGurk and MacDonald 1976) shows — by setting acoustic and visual information in opposition — that speech perception is not solely acoustic. It has frequently been cited as support for motor theory, in that the visual (e.g., lip movement) information is essentially articulatory and could be seen as encoding gesture. But equally, this can be seen as the articulatory dimension of a double-weak theory. It is, of course, necessary to suppose that double-weak theory extends to multiple channels of input. That is, the listener will use any available source of information, auditory and visual, to perceive the gesture of the speaker directly (and the speaker implicitly knows this).

This set of perceptual effects provides some evidence for the view of a loosely-constrained signal coupled with direct realism, but the debate will no doubt continue. However, one thing is clear. Any valid alternative should be in reasonable agreement with current experimental evidence.

### **6.3.2 Mirror neurons and perception as action**

Mirror neurons were originally identified about 20 years ago in primates when certain neurons fired both during the production of an action and the perception of the same action by another actor (Rizzolatti and Craighero 2004). To many commentators, this was one of the most important discoveries in neuroscience in the last quarter of the last



century, and it has frequently been cited as strong supporting evidence for motor theory. Rizzolatti and Arbib (1998) point out that these mirror neurons are present in areas that can be associated with the production and perception of speech in humans, charting the development of Broca's area from the primate's F5 region. This could be evidence for the (trivial) perception of gestures in others. Such mirror neurons fire both when speech is perceived and when speech is produced. This lends support to the notion that the production-perception loop cannot be broken and resolved into previously postulated symbolic tokens.

This can be coupled with a plausible account of the evolution of a biological capacity for language, as stated by Rizzolatti and Arbib:

“...natural selection yielded a set of generic structures for matching action observation and execution. These structures, coupled with appropriate learning mechanisms, proved great enough to support cultural evolution of human languages in all their richness. We hold that human language (as well as some dyadic forms of primate communication) evolved from a basic mechanism that was not originally related to communication: the capacity to recognize actions” (p. 193).

The recognition of these actions does not necessarily mean that we form symbolic gestural information through an innate vocal tract synthesiser as in motor theory, rather perception could just as well proceed through direct realism.

As stated previously, although gestural theories cut little ice with the vast majority of speech scientists, they still have had a clear influence within cognitive science — mostly in the form of motor theory. Mirror neurons offer some reason as to why this might be the case. Their existence clearly demands a theory that has a role for the gesture. Double-weak direct realism allows us to move away from motor theory as the default explanation yet provide some form of reconciliation between speech and cognitive scientists. In unifying the cycle of production and perception, we are in agreement with Ohala (1996) that “listeners who need eventually to speak out loud utterances they hear others make would be aided by figuring out what other speakers do with their vocal tracts” (p. 1718). To this end, sounds are optimised to be discriminated and exploited by the auditory system and this system itself is capable of remarkable feats of discrimination for many types of sound. But, ultimately, when perceiving speech we recognise an intentional agent performing an act that we can mimic, directly perceiving the utterance and maintaining “speech perception in relation to the universal character of perception” (Fowler 1996, p. 1732).

## 6.4 Models and limiting abstractions

We propose a view of speech communication in which the speaker has the intention to manipulate the Gibsonian-like affordances perceived by the listener. Far from a notion of naïve realism, this view is dependent upon a recognition of the intentionality of the speaker by the listener. This leads to a natural interpretation in terms of direct perception of speech via complex pattern recognition, so conjoining Gibsonian direct realism (popular among visual psychologists) with the current majority view among phoneticians that speech perception is a process of loosely-constrained auditory pattern recognition. Instead of forming cognitive representations of the external world (either gestural or phonetic), our senses cause the direct perception of affordances through the medium of the acoustic signal.

As we have seen in Chapter 2, shared phonetic systems can rapidly emerge, in abstract computational models that attempt to capture a part of the wider development of common human languages. Having proposed double-weak direct realism we will argue that, because of their reliance on a number of unrealistic abstractions, current models are insufficient to test this new approach.

Work by a range of researchers (e.g., Oudeyer 2001; Cangelosi and Domenico 2002; Smith, Brighton, and Kirby 2003; Tallerman 2005) has been vital in demonstrating the utility of modelling phonetic evolution. However, when we consider their design decisions in light of current phonetic theory, questions can be raised. We must be careful, as although abstractions are necessary in any model, incorrect abstractions could leave valuable scientific questions unanswered and undermine the validity of any observed results.

Therefore, we have attempted to develop a theory to account for the results presented in Chapter 5, which used the tools developed in Chapters 3 and 4 to produce a model of phonetic emergence. This process of implementation demands detail; vague, implausible mechanisms are swiftly exposed, either through biological implausibility or computational impossibility.

To illustrate the interplay of theory, model and reality, Lindblom (2000) asks us to consider the quantal assumption in linguistics and illustrates the possible dangers of only considering the phonemes of the International Phonetic Alphabet, as “without variation we cannot go from the operationally defined to behaviorally real” (Lindblom 2000, p.303). With even the most basic abstractions coming under scrutiny, the modelling approach needs to proceed carefully, justifying its abstractions and considering the utterance in its entirety. We face “a growing realization that accepting

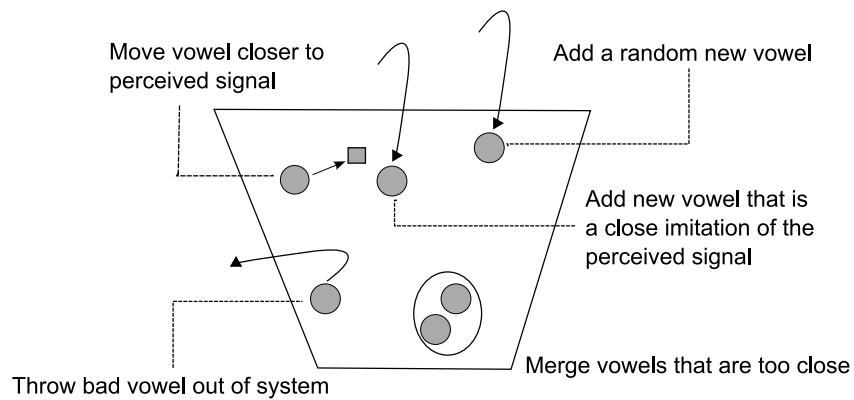


FIGURE 6.3: The de Boer (2000a) model of phonetic emergence, explicitly capturing a variety of forces behind the change and evolution of phonetic tokens.

the priority of form creates an impasse that unnecessarily deprives linguistics of explanatory power” (p.305).

It seems that a theoretical division within the field of phonetics remains due to a lack of conclusive real world data. Given these difficulties, the modelling approach has a clear role to play. When we consider our design decisions, a number of questions are raised. What are the basic units of speech? How are these basic units represented within the mind? Are there strong relations between the symbol and auditory signal or does the listener perform pattern recognition on a loosely constrained output? Phonetic theory shows us that these are significant and current questions. However, current phonetic models (e.g., Figure 6.3) abstract away from these challenges by defining phonemes as discrete points within a coordinate space. When considering current models of speech emergence, it is clear that they follow the ‘double-strong’ theory of speech (Stevens 1998; de Boer 2000a; Oudeyer 2001; Smith, Brighton, and Kirby 2003). The mappings between production and perception are clear and direct, often involving the direct perception of articulatory gestures. This direct perception abstraction does have some basis in motor theory (Liberman and Mattingly 1985; Allott 1998), which contends that some innate specialised module in the human mind is able to reconstruct the distal articulatory symbol from the acoustic signal. Current models, however, have not speculated as to the functioning of this innate module. This is for two reasons. First, most models are constructed to reinforce a linguistic empiricist standpoint; they seek to minimise the role of innate linguistic explanations. Second, the biologically implausible abstraction of directly perceiving abstract symbols allows the functionality of the module to be side-stepped; it becomes an implicit feature of the simulation. For example, with respect to Oudeyer’s (2001) coordinate space model of speech sound emergence, the maintenance of direct perception and the absence of an explicit, symbolic representation causes the model to display the perceptual features of motor

theory (Best 1995). Abstract coordinates pass directly from generation to perception.

How then are we to move from the current double-strong theories of speech perception to the proposed double-weak approach? We would answer that we have attempted this transition as each agent in our simulation comes to understand a speech signal, complete with problems of pattern recognition and phonetic ambiguity, not simply perceiving a direct abstraction, hopefully increasing our understanding of this alternative approach.

# Chapter 7

## Conclusions and future work

In this thesis, we have exploited the potential of computational modelling to demonstrate the possible emergence of a shared communication system despite the production of complex articulatory gestures and the perception of complex auditory patterns. By carefully considering current phonetic theories, we have highlighted the damaging abstractions, (e.g., the reduction of speech to a set of points leading to an inherently double-strong perspective), present within current models of phonetic emergence, Section 6.4.

The development of signal grounding from our initial work revealed the importance of phonetic grounding and clearly motivated the need to capture the subtleties of speech within our model. At the outset, the individual components of this approach established two important points. First, by capturing a direct measure of articulatory effort, it was shown that the two forces present within H&H theory are not always in direct opposition to each other. Rather, the speaker exploits the subtle interplay of articulatory effort and auditory distinctiveness. Given our direct measure, this interplay can now be investigated. Second, the integration of the peripheral auditory system into the model revealed the vital importance of the transformations present in the auditory system when perceiving speech. The careful removal of current, damaging abstractions, e.g., Section 6.4, has led to a number of computational models with a role to play in future research investigating non-symbolic theories of speech perception.

## 7.1 Limitations and future work

At the phonetic level, a number of limitations remain in our model in that the capabilities of the artificial vocal tract are never fully exploited. As a result we have been unable to investigate a number of theories behind syllable formation, for example, MacNeilage's (1998) frame/content theory. To conduct this investigation in future work, our neural model will have to be expanded to gain control of other aspects of speech, like the glottal pulse, and not only the articulatory parameters.

An additional advantage of directly modelling articulatory effort is the fact that we can now consider the emergence of consonants and, ultimately, the formation of syllable structures. This is possible as with improvements to current articulatory models we are no longer constrained by the definition of phonemes in terms of formant values, resulting in the exclusion of everything that can not be formulated as a vowel. However, this is only possible if we are able to obtain a measure of consonant distinctiveness—a challenge for future work.

Beyond the phonetic level, the expansion of this work to other levels of language description will not be trivial. To maintain grounding we would have to consider more than the physical modelling of speech and develop robotic embodiment for the agents as their utterances come to refer to objects in an environment. To move from phonetics to semantics, we should consider the complex interactions of agent, environment and signal. Researchers are beginning to address this challenge and there is a new understanding that “cognition depends upon—is grounded in—the physical characteristics inherited abilities, practical activity, and environment of thinking agents” (Anderson 2003). Current work (Anderson 2003; Steels 2006, 2007; Roy 2007) is attempting to build “a cross-modal bridge between language, perception, and action” (Roy 2007) moving us from animal forms of ‘Cambrian Intelligence’ (Brooks 1999) to increasingly human forms of ‘Neolithic Intelligence’ (Anderson 2003). We believe that in future work, the results of this thesis can be developed to contribute to this current challenge.

## 7.2 Summary of work

Developing symbol grounding towards the specific problem of phonetic perception and emergence, we developed the concept of physical symbol grounding. Demonstrating how the removal of existing modelling abstractions, linking the model to the physical

reality of what it is trying to model, can lead to greater theoretical insights. Specifically, the improved accuracy and robustness of our results led to a justification for the modelling of an increasingly accurate conception of production and perception of speech. It was felt that by confronting the challenges of production and perception, this modelling work could contribute to existing phonetic theories.

To do this, two investigative tools were added and justified: an articulatory vocal tract simulation and a peripheral auditory system. For our vocal tract model, we used Cook's (1993) implementation by encasing it in a muscle model and deriving a direct measure of articulatory effort. This enabled us to present a set of vowel systems based upon Lindblom's H&H theory. The biological plausibility of these systems lent support to this theory and justified the use of this vocal tract in our final multi-agent system.

Next, in developing the peripheral auditory system we produced the *basis* for a sound explanation of the "unexplained perceptual interaction between ... voicing and place" (Kuhl 1988, p.33). Further work is clearly needed—perhaps using non-speech analogs as input—to tease out the precise mechanisms at play, but this chapter represents a very promising start. This study crucially highlights the role of the peripheral auditory system in any model that seeks to capture both the production and perception of real speech. After developing each isolated component we were able to progress to a unified model, testing the new theory of double-weak direct realism.

After developing our vocal tract and peripheral auditory system, we tested both the perception of real speech and the emergence of a phonetic system among a population of speakers. Through simply speaking and listening to each other, a shared set of auditory features and articulatory tokens emerged among our agents. Within this model, each agent's vocal space converged to a shared set of gestures but the auditory space did not.

To develop this contrast, between audition and articulation, we consider phonetic, neurological and phonetic evidence in light of these results. By isolating a single agent and exposing it to speech, we were able to analyse its ability to perceive complex speech signals.

In summary, we have progressively built up a model of speech production and perception that does not assume the cognitive presence of "uniform letter-like units (whether phones or phonemes)," but instead demonstrates the emergent possibilities of "rich auditory patterns of speech plus any coupled visual, somatosensory and motor patterns" (Port, forthcoming). At no point have we been forced to define symbolic mappings between production and perception. We do not even have to perform phoneme recognition in the auditory space. Recovering the 'cues' of speech is sufficient

for intended articulatory gestures to be recovered.

### 7.3 Wider implications and ASR

In this thesis we have argued that spoken language processing cannot be viewed as the passive perception of an abstract signal. Rather, speech should be viewed as an intentional act, which is comprehended as such by both speaker and listener. To capture the wider implications of this comprehension, we propose that speech should be placed in the context of a wide range of ‘interaction affordances’. Expanding upon the direct realist position (Gibson 1979), namely the direct perception of environmental affordances, these interactions are defined as the set of affordances produced by an intentional agent within the environment. As a result of this intentionality this set is open to manipulation and as such require the perceiver to model the production in order to allow direct perception to proceed. This has consequences for all aspects of spoken language processing and we will demonstrate this by briefly detailing a new approach to automatic speech recognition (Moore 2007a) and considering wider aspects of conversation.

In previous sections, we have highlighted the conflict between phonetic theory and perceptual evidence. By considering the listener’s understanding of the speaker and the speaker’s understanding of the listener we present a solution to this problem. The current debate between modern phonetics and motor theory seems to rage over the symbolic end points of phonetic perception. We propose that perception proceeds through an unbroken loop between speaker and listener; perception is direct and dependent upon the perceiver’s understanding of production. Furthermore, this act of perception is not passive (as in, say, the direct perception of a table). Rather the speaker possesses intentionality and exploits their knowledge of the listener and the wider context to fulfil those intentions, manipulating the set of interaction affordances and adjusting their signal accordingly. The difference between this and the direct perception of a table is fundamental. A table attempts none of these things, it is just a table.

To further illustrate this point, we can conceive of a three-level hierarchy within the natural world:

1. **Interactions within the environment:** devoid of intentionality, interactions can be placed solely within the realms of physics.



2. **Interactions between agents and the environment:** proposed by Gibson (1979), and extended to the realm of speech perception by Fowler (1996), an agent directly perceives the affordances of the surrounding environment, i.e., what the environment affords that particular individual.
3. **Interactions between agents:** each is aware of the other's intentionality and so seeks to adjust the set of affordances that they themselves present to the other. The speaker, aware of the listener's capabilities, adjusts their own articulatory gestures to compensate and convey the desired intention. The listener aware of the speaker as an intentional agent perceives the intended interaction affordance through hypothesis testing over a weakly constrained signal.

Separately motivated and developed, the PRESENCE ASR system represents the embodiment of this new approach. Moore (2007a) parodies the difference between PRESENCE and traditional approaches to ASR as two different approaches to heating a room. In the first, a thermostat is installed and the system adjusts the heating according to the deviation from the desired temperature. In the second, a wide range of factors pose a fundamental challenge for our engineer: doors are opened, people enter the room, ambient temperature changes over time. To account for this, sensors are fitted to doors and windows and a wide variety of statistical heating models are proposed to account for noise and variance. Why has the speech community taken the second approach? We propose that it is because we have not yet invented the 'thermometer'.

Within the framework of interaction affordances, this 'thermometer' can be thought of as an error signal derived from the intentionality of the listener. The participants of a conversation are seeking to fulfil some purpose, taken from the set of currently available interaction affordances, and it is the mismatch between the desired and the current perceptual state that allows the listener to refine, actively and continuously, the process of perception. These refinements are possible because the listener models the speaker's intentionality and conversely the speaker models the listener; far from traditional symbolic conceptions of perception, this continuous recursive process establishes an unbroken loop between perception and production.

Finally, we briefly consider the wider implications of spoken language processing as the exploitation of interaction affordances. For example, emotion can be seen as the manipulation of the range of interaction affordances available to the agent. Within conversation, my emotional state affects your emotional state and your emotional state changes the set of interaction affordances that are available to me. In these terms, emotion becomes a strategy for goal fulfilment when coupled with an understanding of the space of possible actions afforded by another.

In this thesis, we have adopted the view that spoken language understanding cannot be divorced from the wider aspects of human behaviour. We contend that these aspects can be conceived as the continuous, recursive manipulation of interaction affordances. This has implications for all aspects of spoken language, from the role of emotion as the manipulation of available interaction affordances to the theories underlying fundamental phonetic perception. By separating ourselves from passive, symbolic notions of perception and instead conceiving of speech as a continuous, shared, process we reveal a wide range of promising research implications.

## 7.4 Conclusion

This work can be developed to consider the perception of real speech by obtaining an error signal from a separate speaker model. At present, the agent's own vocal tract stands in for the speaker, and this is a reasonable abstraction as all of the agents are identical, although in future work a richer understanding of feedback and intentionality will have to be developed.

Despite these drawbacks, the development of double-weak direct realism has implications for both the engineering and scientific aspects of AI. At the engineering level, some researchers (Moore 2007a, 2007b) are already considering the advantages of 'closed-loop' automatic speech recognition systems. If we build systems that have a model of the speaker and their intentionality then "The advantage of a negative feedback closed-loop control system is that it is capable of maintaining a controlled variable at a prescribed value in the face of an infinite number of possible disturbances" (Moore 2007b, p. 1178). Equipped with a model of the speaker and the environment, then ASR systems can begin to perceive speech without suffering from the high level of variance present in natural speech in noisy environments.

Scientifically, both this new ASR approach and current neurological evidence hold sway as they account for such a wide variety of findings. There is now a danger of a growing chasm between phonetics and the rest of the scientific world; double-weak direct realism seeks to bridge this gap. By providing a theory consistent with both bodies of evidence, the complex auditory pattern recognition remains in play and those working in speech and neuroscience can now consider an alternative to motor theory.

Starting with a fundamental consideration of AI, we have tested a new approach to phonetics with a detailed computational model. We have challenged the connection between specifying phonetic invariants and symbolic cognitive representations and

investigated the subtle implications of H&H theory. In future work, the viability of double-weak direct realism will be tested both by the success of a new generation of automatic speech recognition systems and the implications of non-symbolic models of cognition. We have provided a valid alternative to existing phonetic theories, driven by detailed computational modelling. It is our hope that this body of work will form a significant contribution to both speech science and artificial intelligence.

# Appendix A

## Removing ‘mind-reading’ from the iterated learning model

The use of an iterated learning method, in which a culture’s language is produced by successive generations of agents, allows language to be considered in terms of cultural evolution. Iterated learning simplifies a compositional language as the compression of an object space. This compression is motivated by a poverty of stimulus as not all objects in the space will be encountered by an individual in its lifetime. However, in this methodology every agent has a complete understanding of the surrounding object space, which weakens the comparisons made to natural language evolution. By defining each agents’ internal meaning space as a self organising map the meaning space can remain personal and potentially unique. This strengthens the parallels to real language as the agents’ omniscience and mind reading abilities are removed. Additionally, this improvement causes the compression of the language to be motivated through a poverty of memory as well as a poverty of stimulus. Analysis of this implementation shows that a maintenance of a more compositional (structured) language remains. The effect of an implicit generalisation parameter is also analysed and when each agent is able to generalise over a larger number of objects a more stable compositional language emerges.

### A.1 Overview

The notion that language is a system of compression driven to adjust itself so that it can be learnt by the next generation is a relatively new idea in the field of linguistics. However, a variety of simulations (Kirby and Hurford 1997; Kirby 2001; Kirby 2002;

Kirby 2002; Smith, Brighton, and Kirby 2003) have illustrated its potential and provide an alternative to established ‘innate’ accounts of language (Gillis and Durieux 1995; Hauser, Chomsky, and Fitch 2002). Currently, existing versions of this ‘iterated learning’ (IL) model suffer from a number of criticisms, this paper will address some of these and establish how current conclusions can remain valid after a number of modifications.

In a general IL model, an agent selects an object from its environment and produces a meaning-signal pair which is then perceived by a listener. This meaning-signal pair is formed through a weighted connection between a meaning node and a signal node, it is then used to adjust the weighted connections between the meaning space and signal space of a listening agent. Through this means, a language evolves across a number of generations. If each agent is only given the associated signal for a small subset of possible objects, it is forced to generalise across the remaining object space, causing the formation of a stable compositional language.

## A.2 Criticisms of the iterated learning approach

In IL, the agent’s meaning space represents the mind of a person using a language. However, in a range of areas this analogy breaks down. As each agent is created with a perfect knowledge of the surrounding object space, this perfect knowledge is never found in reality. We need to consider the nature of the object space and the agents’ ability to generalise across it. Also a learning agent directly observes each meaning-signal pair; this introduces an element of ‘mind-reading’ as the learner now knows exactly what the adult agent was thinking when he produced a signal. Obviously, this weakens the IL implementation’s strength as a simulation of language evolution. Kirby supports this clear criticism; “the ready availability of signals with meanings neatly attached to them reduces the credibility of any results derived from these models.”(Kirby 2002)

We will now develop a new IL model to address these criticisms. We consider the iterated learning approach to be a language,  $\lambda$ , which is able to describe every object found in the object space,  $N$ , through compression. This compression is possible by forming a compositional language, which describes common features of objects in the space.

As illustrated in Figure A.1(a), a compositional meaning node is able to define partially a number of objects. In the original implementations, that number is automatically

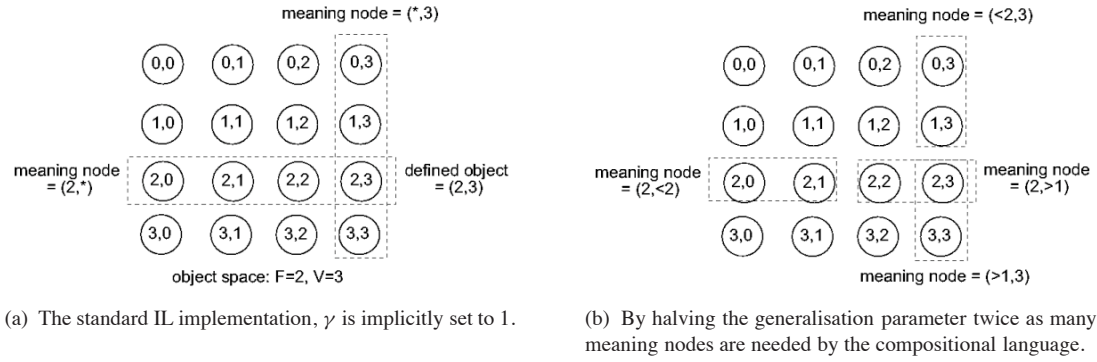


FIGURE A.1: An adjustment of the previously implicit generalisation parameter showing the effect on the level of compression that each compositional meaning node can achieve.

determined by the number of values,  $V$ , in the object space, e.g., in Figure A.1(a) each compositional meaning node is able to define partially four objects. This number represents the strength of the generalisation parameter  $\gamma$  and must be considered significant to the structure of the final compositional language.

In Figure A.1(b), we can see how the compression value of each compositional node can be halved. In order to understand the role of the environment in the emergence of language we need to consider what happens when the generalisation parameter,  $\gamma$  is not equal to  $V$ . That is to say that generalisation becomes independent of the size of the meaning space. To do this we will define the agent’s meaning space as a self organising map and  $\gamma$  as a radius around a selected object, removing the two stated criticisms of IL and allowing us to develop a variable generalisation parameter.

### A.3 Applying a self organising map to iterated learning

In the agents’ environment an example object can be defined as:

$$x_k = \{1, 2\} \quad (\text{A.1})$$

forming a simple coordinate space. This can be defined in the meaning space as:

$$m_j = \{1, 2\} \quad (\text{A.2})$$

or defined as:

$$m'_j = \{1, *\}$$

and

$$m'_{j+1} = \{*, 2\} \tag{A.3}$$

In the above example,  $m_j$  forms a holistic signal as this individual meaning node is only capable of defining one object. While the combination of  $m'_j$  and  $m'_{j+1}$  will form a compositional signal as features from the object space are defined by the two meaning nodes. These are then combined to define an individual object. These feature definitions can then be used in other combinations to describe other objects. We will maintain this aspect of traditional IL by redefining generalisation as a variable radius around a perceived object.

The weightings on the connections between the meaning and signal space determines the mapping from meaning to signal and signal to meaning. The object space,  $N$ , that each agent talks about is represented by a simple coordinate system and a subset of these coordinates are drawn from the object space according to a uniform probability distribution. Each object in turn is mapped directly to the appropriate meaning node in the agent’s meaning space. The signals,  $l_i$ , are generated by mapping from this meaning space to the signal space, and are represented as characters from an alphabet,  $\Sigma$ . A signal  $l_i$  can then be defined as:

$$l_i = \{(s_1, s_2, s_l, s_l) \mid s_l \in \Sigma \wedge 1 \leq l \leq l_{max}\} \tag{A.4}$$

It is clear from Eq. A.4 that we need a sufficient number of signal nodes in the signal layer to express any of the nodes in the meaning space.

Formally, the object space consists of:

$$N = \{x_1, \dots, x_k, \dots, x_n\} \tag{A.5}$$

where each element is given as:

$$x_k = \{(f_1, f_2, \dots, f_F) \mid 1 \leq f_i \leq V\} \tag{A.6}$$

and  $F$  represents the number of dimensions in the object space and  $V$  represents the range of values.

When required to produce an utterance, an agent will select an object,  $x_k$ , and each node in the meaning space,  $m_j$ , competes to have the shortest euclidean distance from this point. Formally, if we define the closest node as  $m(x_k)$  then:

$$m(x_k) = \arg \min_j \|x - m_j\|, j = 1, 2, \dots, l \quad (\text{A.7})$$

The winning node moves closer to the selected point, better defining the object space as a whole. In addition, a number of neighbouring nodes move closer to the object, allowing the network as a whole to represent the experienced object space. The extent to which these nodes move is determined by a Gaussian function,  $h_{j,k}$ , centred around the selected object (Haykin 1999).

$$h_{j,k} = \exp\left(-\frac{d_{j,k}^2}{2\sigma^2}\right) \quad (\text{A.8})$$

In order to form a compositional signal, we need to build valid decomposition sets from the meaning space. This is governed by the generalisation parameter,  $\gamma$ , which we define as a radius centred around the specified object,  $x_k$ . We can then define a set,  $K(x)$ , containing all of those meaning nodes which fall inside this radius. Formally then:

$$\forall x \mid \|x - m_j\| \leq \gamma \mid m_j \in K(x) \quad (\text{A.9})$$

Then considering all possible decompositions in turn the system will pick the corresponding signal with the highest combination of weight values. According to:

$$g(\langle l_i \rangle) = \sum_{i=1}^{N_S} \sum_{j=1}^{K(x)} \omega(K(x)_j) \cdot W_{K(x)_j N_{Si}} \quad (\text{A.10})$$

where “ $\omega(K(x)_j)$  is a weighting function which gives the non-wildcard proportion of  $x$ ” (Kirby 2002), favouring compositional meaning nodes.

All meaning and signal nodes that correspond to a possible decomposition of the object are activated,  $a_{si}$  and  $a_{mj}$ . If two active nodes are connected the weight on that connection is increased, whereas if there is a connection between an active node and an inactive node the weight is decreased. Weights between two inactive nodes remain unchanged. As shown by Oliphant (1999) this kind of learning is required for agents to



acquire a language system. The learning rule displayed by this Hebbian network can be formalised as follows:

$$\Delta w_{ij} = \begin{cases} +1 & \text{iff } a_{si} = a_{mj} = 1 \\ -1 & \text{iff } a_{si} \neq a_{mj} \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.11})$$

where  $\Delta w_{ij}$  is the weight change at the intersection between  $s_i$  and  $m_j$ ,  $s_i \in N_S$  and  $m_j \in N_M$ .

While listening to each utterance, the weight values of the agent are adjusted allowing it to form an understanding of its current language. This hypothesis then allows it to generalise to objects it has not encountered before, resulting in a meaningful expression. Therefore, a poverty of stimulus causes the language to generalise across an object space. Additionally by having a limited number of nodes form the meaning space, the agent does not have an infinite memory resource to draw upon, forcing compression through limited memory as well as limited stimulus.

Using this model, we will vary  $\gamma$  in order to assess how this affects the stability,  $S$ , of the final compositional language, where

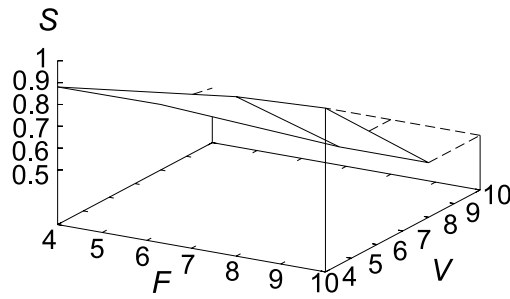
$$S = \frac{S_c}{S_c + S_h} \quad (\text{A.12})$$

The higher the value of  $S$  the more compositional the language. The value  $S_c$  represents the proportion of compositional languages and  $S_h$  defines the proportion of holistic languages (Kirby 2002).

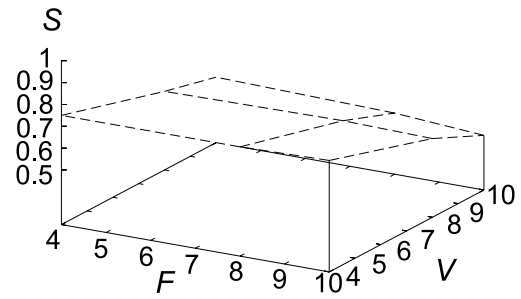
As the meaning space is now able to learn the structure of the object space, each agent’s meaning space can be undefined at birth and the agent will need to learn the structure of the object space as each object is encountered. Consequently, the meaning space gradually comprehends the object space but also remains potentially unique to each agent as a different subset of objects is encountered.

## A.4 Results

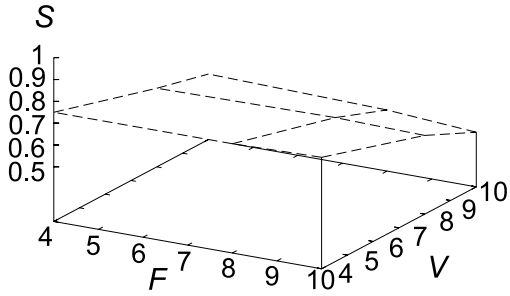
First we run the new self-organising map and iterated learning (SOM-IL) implementation under the same conditions as the previous implementation, see Figures A.2(a),



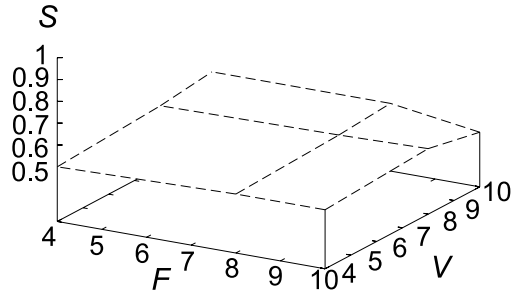
(a) Stability of the resulting languages when each agent is exposed to 10% of the object space.



(b) Stability of the resulting languages when each agent is exposed to 20% of the object space.

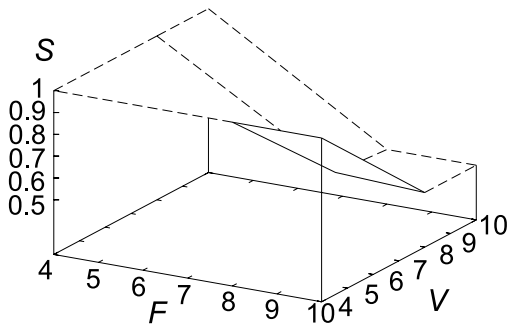


(c) Stability of the resulting languages when each agent is exposed to 50% of the object space.

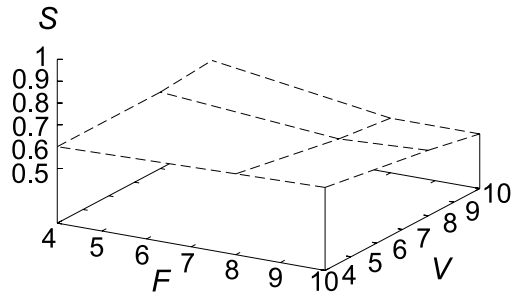


(d) Stability of the resulting languages when each agent is exposed to 90% of the object space.

FIGURE A.2: Under the new IL model results similar to the ones obtained in previous implementations are maintained.



(a) Stability of the resulting languages when the previously implicit generalisation parameter,  $\gamma$ , is doubled and each agent is exposed to 10% of the object space.



(b) Stability of the resulting languages when the previously implicit generalisation parameter,  $\gamma$ , is halved and each agent is exposed to 10% of the object space.

FIGURE A.3: Previously unacknowledged generalisation parameter has a clear effect on the language, with a greater generalisation ability leading to a higher level of stability.

A.2(b), A.2(c) and A.2(d). Then we consider the effect of varying the generalisation value,  $\gamma$ , Figures A.3(a) and A.3(b).

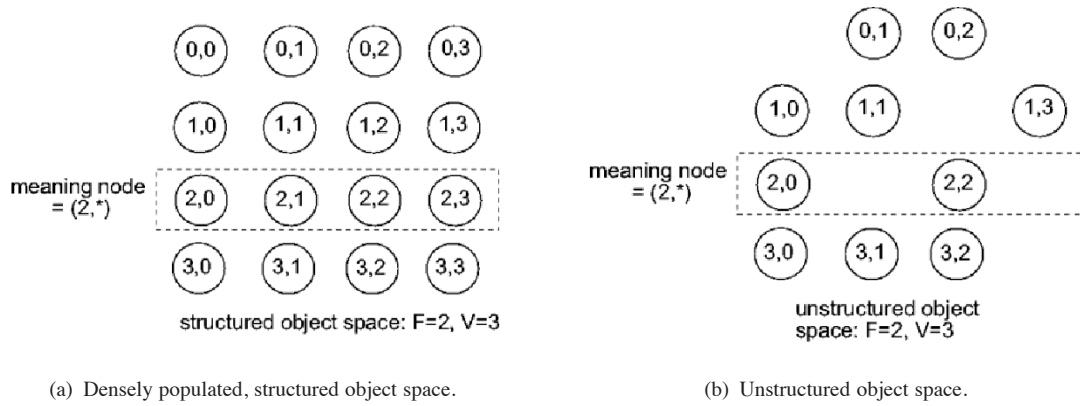


FIGURE A.4: In a structured object space each meaning node can generalise over a greater number of objects, this will affect the stability of the language.

## A.5 Analysis

The results show that compositional languages emerge under a similar set of circumstances to Smith, Brighton, and Kirby’s (2003) previous IL implementation. Therefore, the requirements for a tight bottleneck and a structured meaning space remain in this implementation. However, varying the generalisation parameter does have an effect. The higher the generalisation value, the greater the stability,  $S$ , of the compositional language and accordingly the lower the generalisation value, the lower the stability.

The amenability of the meaning space to generalisation must be considered when analysing these results. We need to explain why unstructured environments display compositional languages with a low stability in an unstructured environment. Figure A.4 demonstrates that the potential generalisation of each meaning node is not as effective as fewer objects are located in each generalisation area, the compositional meaning node can only generalise across two objects in the unstructured object space.

## A.6 Summary

Although, it can be argued that the SOM-IL meaning spaces are simply another example of meaningless symbol shuffling, the meaning spaces within this system possesses certain properties analogous to real cognitive spaces. Specifically, real meaning spaces are limited in the amount of memory resource they have available. The human mind does not have the capacity to define and understand everything in its environment. The SOM itself does not have a high enough capacity to define completely the agent’s

environment — forming a further motivation to generalise.

Additionally, each meaning space is unique to that individual; we all understand different things and we all have different understandings of the same things. By having the SOM define the object space throughout the course of the agent’s lifetime each final SOM configuration could be potentially unique. Additionally, the learning displayed by the SOM demonstrates the second property of a real meaning space, namely, change over time with each new encountered object. It is clear that this is another property of the human mind. This ever-changing meaning space also remains hidden from the general population and the individual who possess the meaning space. In previous IL implementations, meaning-signal pairs were presented for learning and by making these meanings available to the listening agent, a form of ‘mind-reading’ was taking place. This weakened the comparisons to real language evolution.

Significantly, the poverty of stimulus encountered both in reality and in this simulation has been shown to be essential in the evolution of a structured language. By improving the realism of the meaning space simulation this requirement has been maintained and further reinforces the validity of this approach.

However, this field is in its infancy. There is a wide range of work still to be done and many questions remain unanswered. The initial promise shown by this approach should lead to further fascinating pieces of research. The complex interplay of factors that leads to the emergence of language still poses many unanswered questions.

# Appendix B

## Detailed models

In this Appendix, we detailed the models used through this thesis. We first present Oudeyer’s agent model before detailing our own. We then present an articulatory muscle model developed from Hill’s equations.

### B.1 Oudeyer’s agent model

In Oudeyer’s work, each agent has an artificial vocal tract, an artificial ear (cochlear model), and an artificial ‘brain’, or neural model. These will now be detailed in turn.

#### B.1.1 Vocal tract model

Following de Boer (2001), Oudeyer uses a vocal tract simulation controlled by three parameters, namely lip rounding  $r$ , tongue height  $h$  and tongue position  $p$ . Each parameter is constrained to reflect the anatomical range of the corresponding articulator movement. We can derive formant values as follows:

$$\begin{aligned}
F1 = & ((-392 + 392r)h^2 + (596 - 668r)h + (-146 + 166r))p^2 + ((348 - 348r)h^2 \\
& + (-494 + 606r)h + (141 - 175r))p + ((340 - 72r)h^2 + (-796 + 108r)h \\
& + (708 - 38r))
\end{aligned}$$

$$\begin{aligned}
F2 = & ((-1200 + 1208r)h^2 + (1320 - 1328r)h + (118 - 158r))p^2 \\
& + ((1864 - 1488r)h^2 + (-2644 + 1510r)h + (-561 + 221r))p \\
& + ((-670 + 490r)h^2 + (1355 - 697r)h + (1517 - 117r))
\end{aligned}$$

$$\begin{aligned}
F3 = & ((604 - 604r)h^2 + (1038 - 1178r)h + (246 + 566r))p^2 + ((-1150 + 1262r)h^2 \\
& + (-1443 + 1313r)h + (-317 - 483r))p + ((1130 - 836r)h^2 \\
& + (-315 + 44r)h + (2427 - 127r))
\end{aligned}$$

$$\begin{aligned}
F4 = & ((-1120 + 16r)h^2 + (1696 - 180r)h + (500 + 522r))p^2 + ((-140 + 240r)h^2 \\
& + (-578 + 214r)h + (-692 - 419r))p + ((1480 - 602r)h^2 \\
& + (-1220 + 289r)h + (3678 - 178r))
\end{aligned}$$

Although it would be possible to produce sounds (i.e., synthetic vowels) exhibiting these formant values, which were then ‘heard’ by the ‘speaker’ and other agents, this is not done in Oudeyer’s simulations or in Chapter 2. Rather, a short-cut is taken in which auditory parameters are calculated from the formant values.

### B.1.2 Cochlear model

A cochlear (ear) model, designed by Boë, Schwartz, and Vallée, is employed to process the formant values, placing the result in a 2-D auditory space. The model perceives the first formant directly and derives an ‘effective’ second formant,  $F2'$  (Carlson, Granström, and Fant 1970), as follows:

$$F2' = \begin{cases} F2 & \text{if } F3 - F2 > c \\ \frac{(2-w_1F2+w_1F3)}{2} & \text{if } F3 - F2 \leq c \text{ and } F4 - F2 \geq c \\ \frac{w_2F2+(2-w_2)F3}{2} - 1 & \text{if } F4 - F2 \leq c \text{ and } F3 - F2 \leq F4 - F3 \\ \frac{(2+w_2)F3-w_2F4}{2} - 1 & \text{if } F4 - F2 \leq c \text{ and } F3 - F2 \geq F4 - F3 \end{cases}$$

where  $c$  is as a constant of value 3.5 Barks (Chistovich and Lublinskaya 1979), and  $w_1$  and  $w_2$  are defined as:

$$w_1 = \frac{c - (F3 - F2)}{c}$$

$$w_2 = \frac{(F4 - F3) - (F3 - F2)}{F4 - F2}$$

The above equations assume frequency is represented on the Bark scale. Conversion to this scale from hertz frequency is done using the following conversion formula (Traunmüller 1990):

$$f_{\text{Bark}} = \frac{26.81}{1 + 1960/f_{\text{Hz}}} - 0.53$$

### B.1.3 Neural model

The neural model is based on two self-organising maps (Kohonen 1990). The self-organising map (SOM) defining the articulatory space captures the configurations of the vocal tract in terms of parameters  $r$ ,  $h$  and  $p$ . The auditory space codes for the range of acoustic cues in terms of the first formant  $F1$  and second ‘effective’ formant  $F2'$ . Each agent’s neural model is then established by forming weighted connections between the nodes of the auditory and articulatory spaces.

When activated, the  $j$ th node in the articulatory space produces a vector  $\mathbf{v}_j = (r_j, h_j, p_j)$  forming a point in  $[0, 1]^3$  space coding articulatory configuration. A sequence of these vectors,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  where  $n$  is a random number between 2 and 4, is then fed to the vocal tract model. This produces an articulatory trajectory (‘utterance’) of from 2 to 4 configurations. All remaining neurons are then modified according to:

$$\mathbf{v}_k(t+1) = \mathbf{v}_k(t) + G_k(\mathbf{v}_j)(\mathbf{v}_j - \mathbf{v}_k(t)) \quad \left\{ \begin{array}{l} k = 1..N, k \neq j, \\ \text{where } N \text{ is the number} \\ \text{of neurons in each map} \end{array} \right.$$

Each articulatory neuron is updated by a Gaussian activation function:

$$G_k(\mathbf{v}_j) = \exp\left(\frac{d_{j,k}^2}{2\sigma^2}\right) \quad (\text{B.1})$$

where  $d_{j,k}^2 = \|\mathbf{v}_j - \mathbf{v}_k\|^2$

This update mechanism causes the nodes to converge on points in the articulatory space. The location of these points of convergence is determined by the agent's choice of articulation and the utterances that it is exposed to. The articulatory space can then be modified by the auditory space through the weighted connections between the two. The connections between the perceptual neuron  $i$  and the articulatory neuron  $j$  are characterised by the weight  $w_{i,j}$  (initially random).

The auditory space is able to achieve a similar convergence, since on perceiving an utterance a vector containing acoustic cues  $\mathbf{s}$  (derived from the 'speech signal') is placed in the perceptual space and the neurons updated by:

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) + G_i(\mathbf{s})(\mathbf{s} - \mathbf{v}_i(t))$$

The articulatory space is then further updated through the weighted connections by characterising  $d_{j,k}^2$  as:

$$d_{j,k}^2 = \sum_i^N w_{i,j} G_i(\mathbf{s})$$

Taking the functional dependence of  $G(\cdot)$  on  $\mathbf{s}$  as implicit, for simplicity, the weights are updated by a Hebbian learning rule:

$$\Delta w_{i,j} = \alpha (G_i - \langle G_i \rangle) (G_j - \langle G_j \rangle)$$

where  $\alpha$  is set to some small random number and  $\langle G_j \rangle$  represents the average gaussian activation over the previous time steps.



## B.2 Our agent model

Similar to Oudeyer’s work, each agent has an artificial vocal tract, an artificial ear (cochlear model), and an artificial ‘brain’, or neural model. The cochlear model and vocal tract are direct implementations of work by Cook (1993) and Slaney (1998). Details and the differences to Oudeyer’s work are detailed and illustrated below.

The self-organising map (SOM) defining the articulatory space captures the articulatory configurations of the vocal tract in terms of eight parameters ( $r_1 \dots r_8$ ). The auditory space codes for the range of acoustic cues determined by the output of the cochlear model at each time step. Each agent’s neural model is then established by forming weighted connections between the nodes of the auditory and articulatory spaces.

When activated, the  $j$ th node in the articulatory space produces a vector  $\mathbf{v}_j = (r_{1j} \dots r_{8j})$  forming a point in  $[0, 1]^8$  space coding an articulatory configuration. A sequence of these vectors,  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  where  $n$  is a random number between 2 and 4, is then fed to the vocal tract model. This produces an articulatory trajectory (‘utterance’) of from 2 to 4 configurations. All remaining neurons are then modified according to:

$$\mathbf{v}_k(t+1) = \mathbf{v}_k(t) + G_k(\mathbf{v}_j)(\mathbf{v}_j - \mathbf{v}_k(t)) \quad \left\{ \begin{array}{l} k = 1..N, k \neq j, \\ \text{where } N \text{ is the number} \\ \text{of neurons in each map} \end{array} \right.$$

Each articulatory neuron is updated by a gaussian activation function:

$$G_k(\mathbf{v}_j) = \exp\left(\frac{d_{j,k}^2}{2\sigma^2}\right) \quad (\text{B.2})$$

where  $d_{j,k}^2 = |\mathbf{v}_j - \mathbf{v}_k|^2$

This update mechanism causes the nodes to converge on points in the articulatory space. The location of these points of convergence is determined by the agent’s choice of articulation and the utterances that it is exposed to. The articulatory space can then be modified by the auditory space through the weighted connections between the two. The connections between the perceptual neuron  $i$  and the articulatory neuron  $j$  are characterised by the weight  $w_{i,j}$  (initially random).

The auditory space is able to achieve a similar convergence, since on perceiving an utterance a vector containing acoustic cues  $\mathbf{s}$  (derived from the ‘speech signal’) is placed in the perceptual space and the neurons updated by:

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) + G_i(\mathbf{s}) (\mathbf{s} - \mathbf{v}_i(t))$$

The articulatory space is then further updated through the weighted connections by characterising  $d_{j,k}^2$  as:

$$d_{j,k}^2 = \sum_i^N w_{i,j} G_i(\mathbf{s})$$

Taking the functional dependence of  $G(\cdot)$  on  $\mathbf{s}$  as implicit, for simplicity, the weights are updated by a Hebbian learning rule:

$$\Delta w_{i,j} = \alpha (G_i - \langle G_i \rangle) (G_j - \langle G_j \rangle)$$

where  $\alpha$  is set to some small random number and  $\langle G_j \rangle$  represents the average gaussian activation over the previous time steps. This increase in complexity allows us to perceive complex auditory cues and produce articulatory targets for the artificial vocal tract.

### B.3 Muscle model

To develop our notion of articulatory effort we can capture the muscular effort,  $E$ , of articulation by coding the muscles of the vocal tract according to Hill’s muscle model (Umberger, Karin, and Philip 2003):

$$\dot{E} = (h_{AM} + h_{SL} + \dot{w}) \quad (\text{B.3})$$

where  $h_{AM}$  represents the activation and maintenance heat rate,  $h_{SL}$  the shortening/lengthening heat rate,  $\dot{w}$  represents the mechanical work rate. The maintenance and activation heat rate can then be represented as follows:

$$h_{AM} = 1.28(\%_{FT}) + 25 \quad (\text{B.4})$$

Symbol	Description
$\%_{FT}$	% Fast Twitch
$\%_{ST}$	% Slow Twitch
$P_{CSA}$	Cross Section
$\sigma$	Tension

TABLE B.1: The basic parameters for the muscle model.

As shown by Table B.1, the proportion of fast twitch muscles in the tissue under consideration is determined by parameter  $\%_{FT}$ . The work rate of the muscle is then determined by the force,  $f$ , and velocity of the muscle,  $v$ , divided by its mass,  $m$ :

$$\dot{w} = -\frac{f \cdot v}{m} \quad (\text{B.5})$$

The term  $\dot{h}_{SL}$  varies, dependent upon whether the muscle is shortening or lengthening. When shortening,  $\dot{h}_{SL}$  can be defined as:

$$\begin{aligned} \dot{h}_{SL} &= -\alpha_{S(ST)} \times 12 \left( \frac{1 - \%_{FT}}{100} \right) - \alpha_{S(FT)} \times 12 \left( \frac{\%_{FT}}{100} \right) \\ \text{where } \alpha_{S(ST)} &= \frac{100}{12 - \%_{ST}} \\ \text{and } \alpha_{S(FT)} &= \frac{153}{12 - \%_{FT}} \end{aligned}$$

When lengthening,  $\dot{h}_{SL}$  is defined as follows:

$$\begin{aligned} \dot{h}_{SL} &= \alpha_L \times 12 \\ \alpha_L &= 4 \times \alpha_{S(ST)} \end{aligned}$$

To determine the force exerted by a muscle we consider the muscle's maximum force and modify the existing Hill's constants,  $a$  and  $b$ . These constants are determined from experimental evidence (Hill 1938):

$$f = \frac{F_{\max} \times b - 0.2a}{0.2 + b} \quad (\text{B.6})$$

The maximum force of a muscle is then determined by its tension,  $\sigma$ , and cross sectional area,  $P_{CSA}$ :

$$F_{\max} = \sigma \times P_{CSA} \quad (\text{B.7})$$

$$V_{\max} = \frac{F_{\max} \times b}{a} \quad (\text{B.8})$$

$$b = 12a \quad (\text{B.9})$$

$$a = 0.1 + 0.4 \times \%FT \quad (\text{B.10})$$

where  $V_{\max}$  represents the maximum velocity of the muscle.

# References

- Aleksandrovsky, B., J. Whitson, G. Andes, G. Lynch, and R. Granger (1996). Novel speech processing mechanism derived from auditory neocortical circuit analysis. In *Proceedings of Fourth International Conference on Spoken Language Processing, ICSLP'96*, Volume 1, Philadelphia, PA, pp. 558–561.
- Allott, R. M. (1998). Motor theory of language origin: The diversity of languages. In J. Wind, A. Jonker, R. Allott, and L. Rolfe (Eds.), *Studies in Language Origins*, Volume 3, pp. 125–160. Amsterdam: John Benjamins.
- Anderson, M. (2003). Embodied cognition: A field guide. *Artificial Intelligence* 149(1), 91–130.
- Araújo, L. C., T. N. Magalhaes, D. P. M. Souza, H. C. Yehia, and M. A. Loureiro (2005). A brief history of auditory models. In *10 Simpósio Brasileiro de Computação Musical*, Volume 1, Belo Horizonte, Brazil.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences* 22(4), 577–609.
- Belpaeme, T., S. Cowley, and K. F. MacDorman (Eds.) (2007). Symbol Grounding: Special issue of *Interaction Studies* 8(1). John Benjamins.
- Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience*, pp. 171–204. Baltimore, MD: York Press.
- Blackburn, P. L. (2007). *The Code Model of Communication: A Powerful Metaphor in Linguistic Metatheory*. Dallas, TX: SIL International.
- Bladon, R. A. W. and B. Lindblom (1981). Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America* 69(5), 1414–1422.
- Boë, L., J.-L. Schwartz, and N. Vallée (1995). The prediction of vowel systems: perceptual contrast and stability. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Recognition*, pp. 147–167. Chichester, UK: John Wiley & Sons.

- Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems* 6(1), 3–15.
- Brooks, R. A. (1991). The role of learning in autonomous robots. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, Santa Cruz, CA, pp. 5–10.
- Brooks, R. A. (1999). *Cambrian Intelligence*. Cambridge, MA: Bradford Books/MIT Press.
- Cangelosi, A. and P. Domenica (Eds.) (2002). *Simulating the Evolution of Language* (1st ed.). London: Springer-Verlag.
- Cangelosi, A., A. Greco, and S. Harnad (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science* 12(2), 143–162.
- Carlson, R., G. Fant, and B. Granström (1975). Two-formant models, pitch and vowel perception. In G. Fant and M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech*, pp. 55–82. London: Academic Press.
- Carlson, R., B. Granström, and G. Fant (1970). Some studies concerning perception of isolated vowels. In *Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR)*, Volume 2–3, pp. 19–35.
- Chistovich, L. A. and V. V. Lublinskaya (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustic study of the perception of vowel-like stimuli. *Hearing Research* 1(3), 185–195.
- Cook, P. R. (1993). SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal* 17(1), 30–44.
- Crystal, D. (1980). *A First Dictionary of Linguistics and Phonetics*. London: André Deutsch.
- Damper, R. I. (1997). A biocybernetic simulation of speech perception by humans and animals. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Volume 2, Orlando, FL, pp. 1638–1643.
- Damper, R. I. (1998). Auditory representations of speech sounds in a neural model: The role of peripheral processing. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 98)*, Anchorage, AL, pp. 2196–2201.
- Damper, R. I. (2000). Emergence and levels of abstraction. *International Journal of Systems Science* 31(7), 811–818.

- Damper, R. I., S. R. Gunn, and M. O. Gore (2000). Extracting phonetic knowledge from learning systems: Perceptrons, support vector machines and linear discriminants. *Applied Intelligence* 12(1/2), 43–62.
- Damper, R. I. and S. Harnad (2000). Neural network models of categorical perception. *Perception and Psychophysics* 62(4), 843–867.
- Damper, R. I., M. J. Pont, and K. Elenius (1990). Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus. Technical Report STL-QPSR 4/90, Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Stockholm. Also published in W. A. Ainsworth (Ed.), *Advances in Speech, Hearing and Language Processing, Vol. 3 (Part B)*, pp. 497–546. Greenwich, CT: JAI Press, 1996.
- Darling, A. M., M. A. Huckvale, S. Rosen, and A. Faulkner (1992). Phonetic classification of the plosive voicing contrast using computational modelling. *Proceedings of the Institute of Acoustics* 14(6), 289–295.
- Davidsson, P. (1993). Toward a general solution to the symbol grounding problem: combining machine learning and computer vision. In *Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?*, Raleigh, NC, pp. 157–161.
- de Boer, B. (1999). *Self-Organisation in Vowel Systems*. Ph. D. thesis, Vrije Universiteit Brussel, AI-Lab.
- de Boer, B. (2000a). Emergence of vowel systems through self-organisation. *AI Communications* 13(1), 27 – 39.
- de Boer, B. (2000b). Self-organization in vowel systems. *Journal of Phonetics* 27.
- de Boer, B. (2001). *The Origins of Vowel Systems*. Oxford: Oxford University Press.
- de Boer, B. (2003). Conditions for stable vowel systems in a population. In *European Conference on Artificial Life*, Dortmund, Germany, pp. 415–424.
- de Boer, B. (2005). Evolution of speech and its acquisition. *Adaptive Behavior* 13(4), 281–292.
- de Saussure, F. (1983). *Course in General Linguistics*, Volume 1, Chapter 1, pp. 65 – 70. London: Duckworth.
- Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
- Diehl, R. L. and K. R. Kluender (1987). On the categorization of speech sounds. In S. Harnad (Ed.), *Categorical Perception: the Groundwork of Cognition*, pp. 226–253. Cambridge, UK: Cambridge University Press.

- Diehl, R. L. and K. R. Kluender (1989). On the objects of speech perception. *Ecological Psychology* 1(2), 121 – 144.
- Dietrich, E. (1990). Computationalism. *Social Epistemology* 4(2), 135–154.
- Dooling, R. J., K. Okanoya, and S. D. Brown (1989). Speech perception by budgerigars (*melopsittacus undulatus*): The voiced/voiceless distinction. *Perception and Psychophysics* 46, 65–71.
- Dooling, R. J., S. D. Soli, R. M. Kline, T. J. Park, C. Hue, and T. Bunnell (1987). Perception of synthetic speech sounds by the budgerigar (*melopsittacus undulatus*). *Bulletin of the Psychonomic Society* 25, 139–142.
- Dror, I. E. and D. P. Gallogly (1999). Computational analyses in cognitive neuroscience: In defense of biological implausibility. *Psychonomic Bulletin & Review* 6(2), 173–182.
- Faundez-Zanuy, M. and S. McLaughlin (2002). Nonlinear speech processing: Overview and applications. *International Journal of Control and Intelligent Systems* 30(1), 1–10.
- Fodor, J. (1975). *The Language of Thought*. New York, NY: Crowell.
- 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.
- Fowler, C. A. (1996). Listener do hear sounds, not tongues. *Journal of the Acoustical Society of America* 99(3), 1730–1741.
- Galantucci, B., C. A. Fowler, and M. T. Turvey (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review* 14(3), 361–377.
- Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Laurence Erlbaum Associates. (Pagination refers to 1986 second edition).
- Gillis, S. and G. Durieux (1995). A computational model of P and P: Dresher and Kaye (1990) revisited. In M. Vernips and F. Wijnen (Eds.), *Approaches to Parameter Setting*, Volume 5 of *Amsterdam Series in Child Language Development*, Amsterdam, Netherlands, pp. 135–173.
- Guenther, F. H., F. T. Husain, M. A. Cohen, and B. G. Shin-Cunningham (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America* 106(2), 2900–2912.
- Guenther, F. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* 102, 594–621.
- Guenther, F. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, and Hearing Research* 47(1), 46–57.



- Guenther, F. H. and M. N. Gjaja (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America* 100(2), 1111–1121.
- Hagan, M. T. and M. Menhaj (1994). Training feed-forward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks* 5(6), 989–993.
- Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.
- Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think* 2(1), 12–78.
- Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: Bradford Books/MIT Press.
- Hauser, M. D., N. Chomsky, and W. T. Fitch (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(22), 1569–1579.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation* (2nd ed.). New Jersey: Prentice Hall.
- Hill, A. (1938). The heat of shortening and the dynamic constants of muscle. *Proceedings of the Royal Society of London. Series B* 126(843), 136–195.
- Huckvale, M. and I. Howard (2005). Teaching a vocal tract simulation to imitate stop consonants. In *Interspeech*, Lisbon, Portugal.
- International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the international phonetic alphabet*. Cambridge, UK: Cambridge University Press.
- Jia, S. and D. Z. Z. He (2005). Motility-associated hair-bundle motion in mammalian outer hair cells. *Nature Neuroscience* 8(8), 1028–1034.
- Kingston, J. and R. L. Diehl (1994). Phonetic knowledge. *Language* 70(3), 419–454.
- Kingston, J., R. L. Diehl, C. L. Kirk, and W. A. Castleman (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics* 36(1), 28–54.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure – an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5(2), 102–110.
- Kirby, S. (2002). Natural language from artificial life. *Artificial Life* 8(2), 185–215.
- Kirby, S. and J. Hurford (1997). Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands and I. Harvey (Eds.), *Fourth European Conference on Artificial Life*, Brighton, England, pp. 493–503.

- Kirby, S. and J. Hurford (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*, Chapter 6, pp. 121–148. London, UK: Springer-Verlag.
- Kluender, K. R. (1991). Effects of first formant onset properties on voicing judgements resulting from processes not specific to humans. *Journal of the Acoustical Society of America* 90(1), 83–96.
- Kluender, K. R., R. L. Diehl, and P. R. Killeen (1987). Japanese quail can learn phonetic categories. *Science* 237, 1195–1197.
- Kluender, K. R. and A. J. Lotto (1994). Effects of first formant onset frequency on [–voice] judgements result from auditory processes not specific to humans. *Journal of the Acoustical Society of America* 95(2), 1044–1052.
- Kluender, K. R. and A. J. Lotto (1999). Virtues and perils of an empiricist approach to speech perception. *Journal of the Acoustical Society of America* 105(1), 503–511.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE* 78(9), 1464–1480.
- Kuhl, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical Perception: the Groundwork of Cognition*, pp. 355–386. Cambridge, UK: Cambridge University Press.
- Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution* 3(1-2), 19–43.
- Kuhl, P. K. and J. D. Miller (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America* 63(3), 905–917.
- Kuhl, P. K. and D. M. Padden (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics* 32(3), 542–550.
- Kuhl, P. K. and D. M. Padden (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America* 73(3), 1003–1010.
- Ladefoged, P. and I. Maddieson (1996). *The Sounds of the World's Languages*. Oxford: Blackwell Publishers.

- Lakoff, G. (1993). Grounded concepts without symbols. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Society*, Boulder, CO, pp. 161–164.
- Liberman, A. M. (1996). *Speech: A Special Code*. Cambridge, MA: MIT Press.
- Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review* 74, 431–461.
- Liberman, A. M., P. C. Delattre, and F. S. Cooper (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech* 29(11), 153–167.
- Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54, 358–368.
- Liberman, A. M. and I. G. Mattingly (1985). The motor theory of speech perception revised. *Cognition* 21(1), 1 – 36.
- Liberman, A. M. and I. G. Mattingly (1989). A specialization for speech perception. *Science* 243(4890), 489–494.
- Lichtwark, G. A. and A. A. Wilson (2005). A modified Hill muscle model that predicts muscle power output and efficiency during sinusoidal length changes. *Journal of Experimental Biology* 208(15), 2831–2843.
- Lieberman, D. A. (1993). *Learning: Behavior and Cognition* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- Liljencrantz, J. and B. Lindblom (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language* 48(1), 839–862.
- Lindblom, B. (1986a). On the origin and purpose of discreteness and invariance in sound patterns. In J. S. Perkell and D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes*, Chapter 23, pp. 493–511. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lindblom, B. (1986b). Phonetic universals in vowel systems. In J. Ohala and J. Jaeger (Eds.), *Experimental Phonology*, Chapter 2, pp. 13–44. New York: Academic.
- Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle and A. Marchal (Eds.), *Speech production and speech modeling*, pp. 403–439. Kluwer.
- Lindblom, B. (1998). Systemic constraints and adaptive change in the formation of sound structure. In J. Hurford and C. Knight (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*, pp. 242–264. UK: Edinburgh University Press.

- Lindblom, B. (2000). Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica* 57(2–4), 297–314.
- Lindblom, B., P. MacNeilage, and M. Studdert-Kennedy (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, and Ö. Dahl (Eds.), *Explanations for Language Universals*, pp. 181–203. New York, NY: Mouton.
- Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America* 57(6), 1547–1551.
- Lisker, L. (1986). ‘Voicing’ in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1), 1547–1551.
- Lisker, L. and A. Abramson (1964). A cross-language study of voicing in initial stops. *Word* 20(3), 384–422.
- Lisker, L. and A. Abramson (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of 6th International Congress of Phonetic Sciences, Prague, 1967*, pp. 563–567. Academia, Prague.
- Lucero, J. C., T. R. Maciel, D. A. Johns, and K. G. Munhall (2005). Empirical modeling of human face kinematics during speech using motion clustering. *Journal of the Acoustical Society of America* 118(1), 405 – 410.
- Lucero, J. C. and K. G. Munhall (1999). A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America* 106(5), 2834–2842.
- MacNeilage, P. F. (1998). The frame/content theory of the evolution of speech production. *Brain and Behavioral Sciences* 21(4), 499–546.
- MacWhinney, B. (1998). Models of the emergence of language. *Annual Reviews of Psychology* 49, 199–227.
- Maddieson, I. (1984). *Patterns of Sounds*. Cambridge, UK: Cambridge University Press.
- Maeda, S. (1982a). A digital simulation method of the vocal-tract system. *Speech Communication* 1(3 – 4), 199–229.
- Maeda, S. (1982b). The role of the sinus cavities in the production of nasal vowels. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 2, Paris, France, pp. 911–914.
- Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society of Industrial and Applied Mathematics* 11(2), 431–441.

- Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language* 27(2), 213–234.
- Mays, W. (1951). The hypothesis of cybernetics. *British Journal for the Philosophy of Science* 2(7), 249–250.
- McClelland, J. L. and J. L. Elman (1986). The TRACE model of speech perception. *Cognitive Psychology* 18(1), 1–86.
- McGurk, H. and J. MacDonald (1976). Hearing lips and seeing voices. *Nature* 264(5588), 746–748.
- Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America* 53(4), 1070–1082.
- Minsky, M. (1974). A framework for representing knowledge. Technical report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.
- Minsky, M. (1990). Analogical vs. logical or symbolic vs. connectionist or neat vs. scruffy. In P. H. Winston (Ed.), *Artificial Intelligence at MIT: Expanding Frontiers*, Volume 1, pp. 219–243. Cambridge, MA: MIT Press.
- Moore, B. C. J. and A. J. Oxenham (1998). Psychoacoustic consequences of compression in the peripheral auditory system. *Psychological Review* 105(1), 108–124.
- Moore, R. K. (2007a). PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Transactions On Computers* 56(9), 1176.
- Moore, R. K. (2007b). Spoken language processing: Piecing together the puzzle. *Speech Communication* 49(5), 418–435.
- Morvan, P. L. (2004). Arguments against direct realism and how to counter them. *The American Philosophical Quarterly* 41(3), 221–234.
- Nagano, A., B. R. Umberger, M. W. Marzke, and K. G. M. Gerritsen (2005). Neuromusculoskeletal computer modeling and simulation of upright, straight-legged, bipedal locomotion of *australopithecus afarensis* (al 88-1). *American Journal of Physical Anthropology* 126(1), 2–13.
- Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101(6), 3241–3254.
- Nearey, T. M. and J. T. Hogan (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves. In J. J. Ohala and J. J. Jaeger (Eds.), *Experimental Phonology*, pp. 141–161. Orlando, FL: Academic Press.

- Newell, A. (1973). Computer models of thought and language. In R. C. Shank and K. M. Colby (Eds.), *Artificial Intelligence and the Concept of Mind*, pp. 1–60. San Francisco, CA: Freeman.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science* 4(2), 135–183.
- Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
- Newell, A. and H. Simon (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM* 19(3), 113–126.
- Nittrouer, S. (2006). Children hear the forest. *Journal of the Acoustical Society of America* 120(4), 1799–1802.
- Nossair, Z. B. and S. A. Zahorian (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America* 89(6), 2978–2991.
- Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America* 99(3), 1718–1725.
- Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior* 7(3–4), 371–384.
- Oudeyer, P.-Y. (2001). Origins and learnability of syllable systems: a cultural evolutionary model. In P. Collet, C. Fonlupt, J. K. Hao, E. Lutton, and M. Schoenauer (Eds.), *Artificial Evolution, Fifth International Conference*, pp. 143–155. Le Creusot, France.
- Oudeyer, P.-Y. (2002). Origins and learnability of syllable systems: A cultural evolutionary model. In *Artificial Evolution*, pp. 17–34.
- Oudeyer, P.-Y. (2005a). How phonological structures can be culturally selected for learnability. *Adaptive Behavior* 13(4), 269–280.
- Oudeyer, P.-Y. (2005b). The self-organization of combinatoriality and phonotactics in vocalization systems. *Connection Science* 17(3–4), 325–341.
- Oudeyer, P.-Y. (2005c). The self-organization of speech sounds. *Journal of Theoretical Biology* 233(3), 435–449.
- Perkell, J. S. and D. H. Klatt (1986). *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pfeifer, R. and C. Scheirer (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.



- Port, R. F. (1990). Representation and recognition of temporal patterns. *Connection Science* 2, 151–176.
- Port, R. F. (2009). Rich memory and distributed phonology. *Language Sciences*. In press.
- Port, R. F. and A. P. Leary (2005). Against formal phonology. *Language* 81(4), 927–965.
- Porter, R. and J. Lubker (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research* 23(1), 593–602.
- Protopapas, A. (1999). Connectionist models of speech perception. *Psychological Bulletin* 125(4), 410–436.
- Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: Bradford Books/MIT Press.
- Rizzolatti, G. and M. A. Arbib (1998). Language within our grasp. *Trends in Neurosciences* 21(5), 188–194.
- Rizzolatti, G. and L. Craighero (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27, 169–192.
- Roy, D. (2005). Grounding words in perception and action: Computational insights. *Trends in Cognitive Science* 9(8), 389–396.
- Roy, D. (2007). A computational model of three facets of meaning. In M. D. Vega, G. Glennberg, and G. Graesser (Eds.), *Symbols, Embodiment and Meaning*. Oxford: Oxford University Press.
- Sanguineti, V., R. Laboissiere, and D. J. Ostry (1998). A dynamic biomechanical model for neural control of speech production. *Journal of the Acoustical Society of America* 103(3), 1615 – 1627.
- Sanguineti, V., R. Laboissiere, and Y. Payan (1997). A control model of human tongue movements in speech. *Biological Cybernetics* 77(1), 11–22.
- Schwartz, J.-L., L.-J. Boë, N. Vallee, and C. Abry (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics* 25(3), 279–282.
- Schwartz, J.-L. and P. Escudier (1989). A strong evidence for the existence of a large scale integrated spectral representation in vowel perception. *Speech Communication* 8(3), 235–259.
- Sellers, W. and P. Manning (2007). Estimating dinosaur maximum running speeds using evolutionary robotics. *Proceedings of the Royal Society B–Biological Sciences* 274(1626), 2711–2716.

- Slaney, M. (1998). Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, Palo Alto, CA.
- Smith, K., H. Brighton, and S. Kirby (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems* 6(4), 537–558.
- Soli, S. D. (1983). The role of spectral cues in the discrimination of voice-onset time differences. *Journal of the Acoustical Society of America* 73(6), 2150 – 2165.
- Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication* 1(1), 1–34.
- Steels, L. (1998). Synthesizing the origins of language and meaning using coevolution, self-organization and level formation. In J. Hurford and C. Knight (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*. UK: Edinburgh University Press.
- Steels, L. (1999). *The Talking Heads Experiment. Volume 1. Words and Meanings*. Laboratorium, Antwerpen.
- Steels, L. (2003). The evolution of communication systems by adaptive agents. In E. Alonso, D. Kudenko, and D. Kazakov (Eds.), *Adaptive Agents and Multi-Agent Systems*, Volume 2636 of *Lecture Notes in Artificial Intelligence*, pp. 125–140. Springer-Verlag Berlin.
- Steels, L. (2006). Semiotic dynamics for embodied agents. *IEEE Intelligent Systems* 21(3), 32–38.
- Steels, L. (2007). The symbol grounding problem has been solved. so what’s next? In M. D. Vega, G. Glennberg, and G. Graesser (Eds.), *Symbols, Embodiment and Meaning*. Oxford: Oxford University Press.
- Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds. In T. F. Myers, J. Laver, and J. Anderson (Eds.), *The Cognitive Representation of Speech*, pp. 61–75. North-Holland.
- Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.
- Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111(4), 1872–1891.
- Stevens, K. N. and S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America* 64(5), 1358–1368.



- Stevens, K. N. and S. E. Blumstein (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the Study of Speech*, pp. 1–38. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper (1970). Motor theory of speech perception: A reply to Lane’s critical review. *Psychological Review* 77(3), 234–239.
- Summerfield, A. Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America* 72, 51–61.
- Sun, R. (2000). Symbol grounding: A new look at an old idea. *Philosophical Psychology* 13(2), 149–172.
- Sussman, H. (1989). Neural coding of relation invariance in speech: Human language analogs to the barn owl. *Psychology Review* 96(4), 631–642.
- Sussman, H. M., H. A. McCaffrey, and S. A. Matthews (1991). An investigation of locus equations as the source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America* 90(3), 1309–1325.
- Tallerman, M. (2005). *Language Origins: Perspectives on Evolution*. Oxford: Oxford University Press.
- Tambovtsev, Y. and C. Martindale (2007). Phoneme frequencies follow a yule distribution. *SKASE Journal of Theoretical Linguistics* 4(2), 1–12.
- Trautmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America* 88(1), 97–100.
- Umberger, B. R., G. M. G. Karin, and E. M. Philip (2003). A model of human muscle energy expenditure. *Computational Methods of Biomechanical and Biomedical Engineering* 6(2), 99–111.
- Vogt, P. (2002). The physical symbol grounding problem. *Cognitive Systems Research* 3(3), 429–457.
- Vogt, P. and B. de Boer (2009). Language evolution: Computer models for empirical data. *Adaptive Behavior*. (to appear).