The Acquisition of Vowel Normalization during Early Infancy: Theory and Computational Framework

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Andrew R. Plummer, B.S.,M.S.,M.A.

Department of Linguistics

The Ohio State University

2014

Dissertation Committee:

Prof. Mary E. Beckman, Adviser

Prof. Eric Fosler-Lussier

Prof. William Schuler

ABSTRACT

Vowel normalization is a computation that is meant to account for the differences in the absolute direct (physical or psychophysical) representations of qualitatively equivalent vowel productions that arise due to differences in speaker properties such as body size types, age, gender, and other socially interpreted categories that are based on natural variation in vocal tract size and shape. In this dissertation, we address the metaphysical and epistemological aspects of vowel normalization pertaining to spoken language acquisition during early infancy. We begin by reviewing approaches to conceptualizing and modeling the phonetic components of early spoken language acquisition, forming a catalog of phenomena that serves as the basis for our discourse. We then establish the existence of a vowel normalization computation carried out by infants early in their spoken language acquisition, and put forward a conceptual and technical framework for its investigation which focuses attention on the generative nature of the computation. We then situate the acquisition of vowel normalization within a broader developmental framework encompassing a suite of vocal learning phenomena, including language-specific caretaker vocal exchanges, perceptual warping, and multisensory matching and narrowing. We demonstrate the applicability of the technical formulation through the creation of a virtual environment for vocal learning which provides the means to model the acquisition of vowel normalization, along with other aspects of vocal learning. We conclude with a discussion of the broader implications of the conceptual and technical formulation.

ACKNOWLEDGMENTS

Beginning with the principal contributors, I would like to thank Mary Beckman for providing the motivation and many of the subtle mechanisms for unlearning the host of epistemological prejudices that I brought with me into linguistics from mathematics, particularly with respect to the nature of progress in a developing science such as ours. I would like to thank Eric Fosler-Lussier for shaping the hard structural component of this odd endeavor, without which I would have no means to even attempt to demonstrate the merit of what I have put forward in this dissertation. And I would like to thank Carl Pollard, who right from the beginning of my time in the program served as a clear example of how to tenaciously seek alternatives to linguistic dogma. As openness and entrepreneurship are far more important than intelligence, computing prowess, analytic ability, or any of the cruder "tagline" and "showmanship" characteristics prized in academia (and especially linguistics) these days, it is difficult to overstate the value of these contributions. Indeed, this collaborative creation of an open scientific mind is very likely the only significant achievement associated with this dissertation, and surely the only one that will mean anything to anyone as time goes by.

The contributors, of course, extend well beyond the principal group. I would like to thank William Schuler for bringing his expertise in computational modeling to bear on the later stages of the dissertation project, and for his helpful comments and genuine interest in the topic. I would also like to thank Misha Belkin for bringing the manifold learning

approach to my attention, and for his advice on how best to make use of it in modeling the acquisition of vowel normalization. I would like to thank Ilana Heintz for being a welcoming and encouraging student mentor within the HSD research group. The work in this dissertation is in many ways an extension of the research that she spearheaded. I would also like to thank Pat Reidy and Jeff Holliday, not only for their camaraderie over the years, but also their efforts to implement the auditory models used in this dissertation, and investigate their role in acquisition.

Beyond those at Ohio State, I would like to thank the group of researchers who carried out the perceptual categorization experiments that made the work in this dissertation possible, including Ben Munson, Catherine McBride-Chang, Chanelle Mays, Asimina Syrika, Kiyoko Yoneyama, and Hyunju Chung. I am also grateful to Ben for his expert readiness in vouching for me as a hireable researcher. The stimuli for the aforementioned experiments were provided by Lucie Ménard, whom I would like to thank, along with Louis-Jean Boë, for providing me with the matlab version of the Vlab articulatory model.

This dissertation would not have been possible without the OSU linguistics faculty, staff, and students. I would like to thank Brian Joseph and Chris Brew for their encouragement and willingness to engage with me on many issues ranging over the long history and general nature of linguistics and its practice. I am also very grateful to the entire faculty for providing me with guidance and instruction during my time in the department. I would like to thank Jim Harmon for all of his hard work maintaining the linguistics computing infrastructure. I am very grateful to Julie McGory, Claudia Morettini, Julia Papke, and the undergraduate interns for their supply of clues that led me through various bureaucratic labyrinths. I am also very grateful to Hope Dawson for helping me through teaching in the final year of dissertation writing. I would like to thank my Oxley mates Dennis Mehay, Dominic Espinosa, Raja Rajkumar, Steve Boxwell, Kirk Baker, Jianguo Li, D. J. Hovermale, Murat Yasavul, Rory Turnbull, and Liz McCullough for providing the human resources necessary to make coming into the office a good experience. I am very grateful to the entire student body for creating a pleasant working environment.

I am also grateful to the students for putting together countless cordial social engagements, providing much needed distraction from the difficulties of the program. Specifically, I would like to thank Dominic Espinosa and his family, and Abby Walker and her extended group of friends, for creating opportunities to get away from the department and linguistics issues for a time. I would also like to thank the LBS for many nights of carousing around Columbus.

Finally, I would like to thank those closest to me. I am very grateful to my parents, Jude and Patricia, my sisters Tanya, Elizabeth, and Kate, my brothers-in-law Glenn and David, my niece Emily and my nephew Matthew, for providing the basis of strength and examples of perseverance I needed to keep moving forward. I am also grateful to my entire family for their love and support on my long and complicated journey out into the world. And I would like to thank Marivic Lesho, along with Hatchet and Seaweed and the faculty of Snugglemore College, for much needed love, support, and encouragement, especially during the rather difficult last few years of graduate school.

This research was supported in part by NSF grants BSC-0729306, BSC-0729277, and BSC-0729140 to Mary Beckman, Eric Fosler-Lussier, Benjamin Munson, and Jan Edwards for “DHB/Collaborative Research: Using Machine Learning to Model the Interplay of Production Dynamics and Perception Dynamics in Phonological Acquisition”, and by an Ohio State University Center for Cognitive Science seed grant to Mikhail Belkin, Mary

Beckman, and Eric Fosler-Lussier for research on “Using manifolds to model phonological learning in infancy and early childhood”.

VITA

April 13, 1982	Born in Miami, FL, USA
December, 2004	B.S., Mathematics Florida International University, Miami, FL, USA
December, 2006	M.S., Mathematics Georgia State University, Atlanta, GA, USA
May, 2007	Graduate Certificate, Linguistics Florida International University, Miami, FL, USA
March, 2012	M.A., Linguistics The Ohio State University, Columbus, OH, USA

PUBLICATIONS

Journal Articles and Thesis

Johannes H. Hattingh, Ernst J. Joubert, Elizabeth Jonck, and Andrew R. Plummer. “Total restrained domination in unicyclic graphs.” *Utilitas Mathematica* **82** (2010) 81-95.

Johannes H. Hattingh and Andrew R. Plummer. “A note on restrained domination in trees.” *Ars Combinatoria* **94** (2010) 477-483.

Johannes H. Hattingh, Ernst J. Joubert, Marc Loizeaux, Andrew R. Plummer, and Lucas van der Merwe. “Restrained domination in unicyclic graphs.” *Discussiones Mathematicae Graph Theory* **29** (1) (2009) 71-86.

Johannes H. Hattingh, Elizabeth Jonck, Ernst J. Joubert, and Andrew R. Plummer. “Nordhaus-Gaddum results for restrained domination and total restrained domination in graphs.” *Discrete Mathematics* **308** (2008) 1080-1087.

Johannes H. Hattingh and Andrew R. Plummer. “Restrained bondage in graphs.” *Discrete Mathematics* **308** (2008) 5446-5453.

Hua-Ming Xing, Johannes H. Hattingh, and Andrew R. Plummer. “On the domination number of Hamiltonian graphs with minimum degree six.” *Applied Mathematics Letters* **21** (2008) 1037 - 1040.

Johannes H. Hattingh, Elizabeth Jonck, Ernst J. Joubert, and Andrew R. Plummer. “Total restrained domination in trees.” *Discrete Mathematics* **307** (2007) 1643 - 1650.

Andrew R. Plummer. “S4 enriched multimodal categorial grammars are context-free.” *Theoretical Computer Science* **388** (2007) 173 - 180.

Andrew R. Plummer. “Characterizations in Domination Theory.” Master’s Thesis, Georgia State University, 2006.

Conference Papers and Presentations

Andrew R. Plummer, Lucie Ménard, Benjamin Munson, and Mary E. Beckman. “Comparing vowel category response surfaces over age-varying maximal vowel spaces within and across language communities.” *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, August 2013.

Andrew R. Plummer, Benjamin Munson, Lucie Ménard, and Mary E. Beckman. “Examining the relationship between the interpretation of age and gender across languages.” *Presented at the Annual Meeting of the Acoustical Society of America*, June 2013.

Andrew R. Plummer. “Bolzano-Lewis possible worlds semantics: An improvement over its successors.” *Presented at the Annual Meeting of the North American Association of the History of the Language Sciences (NAAHoLs)*, January 2013.

Andrew R. Plummer. “Aligning manifolds to model the earliest phonological abstractions in infant-caretaker vocal imitation.” *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2012.

Andrew R. Plummer and Carl J. Pollard. “Agnostic possible worlds semantics.” *Proc. of the Conference on the Logical Aspects of Computational Linguistics (LACL)*, July 2012.

Andrew R. Plummer. “Manifold alignment, vocal imitation, and the perceptual magnet effect.” *Presented at the International Child Phonology Conference (ICPC)*, June 2012.

Andrew R. Plummer. “Galen’s critique of Rationalism and Empiricism and its relevance for modern linguistics.” *Presented at the Annual Meeting of the North American Association of the History of the Language Sciences (NAAHoLs)*, January 2012.

Andrew R. Plummer, Mary E. Beckman, Mikhail Belkin, Eric Fosler-Lussier, and Benjamin Munson. “Learning speaker normalization using semisupervised manifold alignment.” *Proc. of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2010.

FIELDS OF STUDY

Major Field: Linguistics

Studies in:

Language Acquisition	Prof. Mary E. Beckman
	Prof. Eric Fosler-Lussier
	Prof. William Schuler
Computational Modeling	Prof. Eric Fosler-Lussier
	Prof. Prof. William Schuler

TABLE OF CONTENTS

Appendices:

LIST OF FIGURES

Figure **Page**

PREFACE

Since the general topic of vowel normalization is rather old, and its sinuous offshoots twist their way quite far into adjacent areas of inquiry, it is useful to begin with a brief description of what this dissertation purports to model. It is quite common in current computational modeling of spoken language acquisition to make two simplifying assumptions – the first is the identification of *acquisition* with *learning*, and the second is the identification of learning with *statistical learning* – resulting in the identification of acquisition with statistical learning. Furthermore, the raw input to statistical learning is often assumed to be either phenomenological representations, e.g., intuitive "features" that the researcher has conscious access to, or their scientific realist counterparts, e.g., the acoustic correlates of such features. Models that adopt these assumptions are typically lauded as "simple" and "plausible" by the working linguist, even though they seem to be anything but.

It is our position that none of these assumptions will lead the researcher very far in understanding how infants develop spoken language. Accordingly, we take acquisition to be distinct from learning, where the scope of the former concerns the formation of structures that facilitate the organization of experience, i.e., the scope of the latter. Moreover, we take learning to be far broader than statistical learning, and the input to learning to be different from the phenomenological base assumed by linguists. Having stated this position, in accordance with its title, this dissertation is meant to provide a theory and computational framework for the *acquisition of vowel normalization*. It must be stressed that we are not proffering a theory or model of speech perception or production, and we are not proffering a theory or model of the learning of aspects of spoken language (e.g., vowel category learning). These subjects are related to our present study, and we make use of the body of literature concerning each of them, when relevant, yet, we are not proffering a theory or model of any one of them, per se, but rather of their relationship to and their roles in the acquisition of vowel normalization.

With this in mind, we proceed with a brief description of the manner in which we mean to conduct our investigation of the acquisition of vowel normalization. Our general approach is two-tiered, using aspects of a research program for the biological study of behavior elucidated by the Ethologist Niko Tinbergen, and a more specific research strategy from basic cognitive psychology outlined by Artificial Intelligence researcher Patrick Winston in a recent presentation given at an MIT symposium on Brains, Minds, and Machines. Tinbergen's (1953) influential program has provided a useful infrastructure for organizing investigation of animal social and behavioral phenomena. Based on Julian Huxley's "three major problems of Biology," to which Tinbergen (1963) added a fourth, the approach involves an integrated view of social and behavioral phenomena from four vantages:

1. The Mechanistic Perspective – attempting to characterize the physiological and psychological mechanisms that implement the phenomena.
2. The Ontogenetic Perspective – attempting to characterize the developmental and environmental aspects influencing the phenomena.
3. The Phylogenetic Perspective – attempting to reconstruct the evolutionary history relevant to the phenomena.
4. The Functional Perspective – attempting to characterize the effects of the phenomena on survival and reproduction.

In this dissertation, we will conduct our investigation primarily from the mechanistic perspective, working toward modeling the psychophysical and cognitive mechanisms involved in the acquisition of vowel normalization. Given the nature of the phenomena, we make use of the ontogenetic and functional perspectives as well. Occasionally, computational models (e.g., Oudeyer, 2005; Zuidema and de Boer, 2009; de Boer and Zuidema, 2010) attempt to reason from the phylogenetic perspective, even though, at present, the phylogeny of speech and language is a highly controversial topic plagued by lack of evidence. These models typically adopt a simple "natural selection" approach to the evolution of natural language sound systems, where a richer conceptualization is likely needed (see Gould, 1997a,b). Accordingly, we will not pursue our investigations from this perspective.

The different perspectives are not mutually exclusive, and overlap substantially. Key ontogenetic aspects of the phenomena include the gradual growth, change in shape, and increase in self-control of the infant's articulatory system between three and eight months of age. This development necessitates modification of the infant's cognitive representation of the articulatory system, as well as the infant's own vowel vocalizations, since an increase in articulatory size and control alter the nature of the repertoire of vocalizations the infant can produce. Key functional aspects of the phenomena include the nature of the interaction between infants and adult caretakers. Both ontogenetic and functional aspects are closely tied to psychophysical and cognitive mechanisms, and we model them from the mechanistic perspective.

Our approach from the mechanistic and ontogenetic perspectives naturally suggests a research strategy that makes extensive use of computational modeling. However, computational modeling of cognitive phenomena is a very broad area of research whose history has been affected significantly by shifts in funding priorities from advances in basic science to immediate, short-term deliverables. Research within academic institutions mostly followed the shift, and much of the focus of academic computational modeling, particularly that of speech and language phenomena, has focused on “get-rich-quick” approaches to research. In response to this shift, there has been a recent push for a “return to basics” in the computational cognitive modeling community. One aspect of this is a simple research strategy, reviewed and outlined by Winston (2011), that emphasizes the fundamentals of cognitive psychology:

1. Characterize the phenomena and the problems to be addressed.
2. Computationally formulate the characterized problems.
3. Propose computational solutions to the computational problems, and implement them.
4. Attempt to crystallize principles from repeated application of the above steps.

Following the research strategy, we will begin by providing a characterization of the phenomena we are interested in modeling, and the problems to be addressed. The emphasis of the characterizations will be clarity of concepts, and facilitation of computational formulation to better characterize the phenomena and the problems we aim to address. It is important to note that the characterization of the phenomena is itself a work in progress, with computational formulation being refined as more is learned. We then propose computational solutions to the problems, providing detailed implementations of the solutions. Finally, we attempt to extract principles concerning the phenomena based on our computational models, and discuss the implications of these potential principles for the study of vowel category acquisition, and acquisition generally.

Aspects of the biological and psychophysical research we are conducting essentially require adopting a comparative biology approach, and by employing Tinbergen's conceptual organization for animal social and behavioral phenomena in our investigation of human behavior, we are extending this comparative approach to include a social component. Moreover, in attempting to further characterize the phenomena relevant to the acquisition of vowel normalization, we will draw concepts and examples from animal cognition studies, extending the scope of the approach to include comparative psychology. The multitude of methodological sources reflects the need for "round-table" interaction between the many research communities whose foci intersect in the current work. Our explicit methodological consideration is meant to facilitate our own investigations, but more importantly to solicit greater involvement of these research communities in the research program within which our own work lies.

Assuming that spoken language acquisition, and linguistics in general, fall within the scope of biological, psychophysical, and social investigation, and thus (at least partially) within Tinbergen's conceptual program, it is useful to bear in mind his admonition about the state of Ethology in the early 1960s:

> "It just is a fact that we are still very far from being a unified science, from having a clear conception of the aims of the study, of the methods employed and of the relevance of the methods to the aims. Yet for the future development of Ethology it seems to me important to continue our attempts to clarify our thinking, particularly about the nature of the questions we are trying to answer" (Tinbergen, 1963, p. 410).

With this in mind, this dissertation is meant to clarify thought about the nature of acquisition, and the questions this research community is trying to answer.

In this connection we briefly discuss the nature of modeling contributions, making use of a few key examples from biology relevant to the purpose at hand, in keeping with our use of Tinbergen's approach. The goal is simply to illustrate the point that modeling is a complex endeavor whose benefits are to be measured in a manner befitting the complexity. Too often within linguistics models are judged using metrics too simple to reflect all the benefits that models have to offer, and too often the merits of models are inflated by metrics too simple to reveal substantial shortcomings. Below are two key properties of models that, in our view, must be taken into consideration in their evaluation:

- Predictive Power – the identification of the epistemological consequences of a model that may then be used to formulate new, testable statements concerning the phenomena of interest.
- Interpretive Power – the organization of a conceptualization or metaphysics undergirding a model that illuminates aspects of the phenomena of interest otherwise left in the dark.

Is is now very common within current computational modeling of linguistic phenomena to tout the predictive power of models, particularly with respect to some quantitative evaluation, effectively achieving the status of community dogma. While quantitative evaluation and predictive power are indeed valuable components of modeling, it is important to keep in mind that even though the quantitatively-based predictive power of a model may be very strong, it is not thereby guaranteed to yield any real insight into the phenomena of interest. Moreover, the predictive power may not be located entirely within the information provided by quantitative analysis.

The following example, based on computational modeling of cardiac phenomena carried out by Denis Noble, demonstrates aspects of the predictive power of a model that are often underappreciated. During the migration of physicists into biology during the 1940s, biological processes involving excitable cells were modeled using nonlinear differential equations, e.g., Hodgkin and Huxley's (1952) model of nerve cell action potential. Noble (2002) recalls that the success of the approach with nerve cells "created an unrealistic expectation for the rapid application of the same principles elsewhere" (p. 1155), including the mechanical contraction of cardiac cells. It was known that depolarization of cardiac cells is sustained for a longer duration than that of nerve cells, and early cardiac cell models "sought insight" into this "most obvious difference between electrical activity in heart and nerve" (ibid). However, Noble's (1962) early cardiac model based on Hodgkin-Huxley equations struggled with the longer depolarization period, the "main defect" being that "it included only one voltage-gated inward current" (p. 1156), which was sodium-based. The reason for this, Noble notes, is that "[c]alcium currents had not then been discovered" (ibid), and so

> "the only way in which the model could be made to work was to greatly extend the voltage range of the sodium 'window' current by reducing the voltage dependence of the sodium activation process...In effect, the sodium current was made to serve the function of both the sodium and the calcium channels" (ibid).

The "clear prediction here," Noble recalls, is that "either sodium channels in the heart are quantitatively different from those in the nerve, or other inward current-carrying channels must exist," with the current understanding that "[b]oth predictions are correct" (ibid).

Together with the measurement procedure in Deck and Trautwein (1964), the simulations "rapidly lead to the discovery of the cardiac calcium current." In this case, in addition to the quantitative prediction, the computational modeling predicted a richer metaphysics, which was then corroborated through experimentation. Reflecting on the modeling development, Noble (2002) comments that "[s]imulation is a necessary tool of analysis in attempting to understand biological complexity" (p. 1155), while stressing that modeling complex biological phenomena requires “many years of interaction between experiment and theory” (ibid). Too often computational modeling within linguistics, and especially language acquisition, ignores the kinds of conceptual/metaphysical predictions suggested by models in favor of attending to predictions catering to quantitative evaluative improvement.

Aside from reminding us of the importance of metaphysical considerations, and of communication between the computational modeler and the experimenter, the example also serves to remind us that progress has to have a point of departure. Noble’s heartbeat model began with the “unrealistic expectation” that cardiac cells behave just like nerve cells, even as they “sought insight” into the differences in behavior. Research advances often begin with exactly this kind of initial assumption. Similarly, modeling of the articulations of infants begins with the unrealistic assumption that the infant articulatory system is a (nonuniformly) scaled-down version of the articulatory system of an adult. It is our position that progress can be made in this fashion, even as we seek insight into the differences between infant and adult articulatory systems, and how an infant overcomes the differences in spoken language acquisition, even with the knowledge that the model of infant articulation is unrealistic.

Given the amount of promotion of predictive power, it is easy to forget to consider the interpretive power that models offer. The above example makes sense only since it had long been established that the heart beats, and that there existed mechanisms that bring about the beating action and ways of quantifying this action and its components. It easy to forget that there was a time when this was not the case, and that the process by which this conceptualization crystallized was slow and complicated. We illustrate with the following bit of history. The 17th century anatomist William Harvey (1628) is most widely credited with establishing that blood and the heart constitute a biological system wherein blood circulates through the body with the heart acting as a pump. As the story goes, Harvey's *De Motu Cordis* (1628, translated by Leake, 1928) effectively dispelled the "occult properties" surrounding the role of blood as one of the "four humors" in the human body going back to Aristotle. Yet, at the time of writing, Harvey was very committed to interpreting his experimental results within the Aristotelian metaphysics. It was only after the publication of *De Motu Cordis* that the heart-as-a-pump interpretation began to take shape. Moreover, this interpretation, prefigured in *De Motu Cordis* even if not formulated explicitly, built on the work of a number of Harvey's predecessors (see the Prefaces in Leake, 1928, for more on the history). Over time, Harvey's model, building on those of his predecessors, provided an interpretation that greatly illuminated the phenomena that he was investigating, surpassing the understanding provided by the "four humors" interpretation.

In this case, the model is essentially an analogy to a mechanical entity. It is unclear how to evaluate such a model in concrete terms, let alone quantitative ones. Moreover, it would be another two centuries after the model's inception before the interpretation it yielded became clear enough to produce progress within clinical practice. The merit of the model lies in its interpretive power, the clarity it provided concerning the phenomena Harvey was interested in. The example reminds us that it is important to keep this kind of modeling aspect in mind.

CHAPTER 1: INTRODUCTION

This dissertation explores the phonetic component of language acquisition and proposes a framework for modeling critical aspects of acquisition that are set in place well before an infant's first words. The main focus of the dissertation is vowel normalization – a cognitive computation that is meant to account for the differences in the representations of absolute position within quantitative parametric (physical or psychophysical) spaces of qualitatively equivalent vowel productions that arise due to differences in speaker properties such as body size types, age, gender, and other socially interpreted categories that are based on natural variation in vocal tract size and shape. Dating back to Martin Joos's proposal in his monograph on acoustic phonetics (Joos, 1948, see p. 63), nearly all previous accounts of vowel normalization have rested on the presupposition that accounting for qualitative equivalence despite the absolute differences involves computation of some transform or translation within some primary sensory domain so as to produce an invariant representation directly within that domain. In this dissertation, we present a framework for the investigation of vowel normalization that does not rest on this presupposition, and argue, rather, that what is going on in the phonetic component of language acquisition, including vowel category acquisition, is the creation of a "self," collections of "others," and the computation of a "likeness" function that facilitates mapping the representations of others onto the representations of the self. More generally, we take this mapping to be a potential basis for other "likeness" functions that can map the representations of some coherent group of others onto representations of some different coherent group of others. We also put forward a computational modeling methodology for investigating this claim.

Specifically, we model the physiological and cognitive structures separated out below for specialized investigation. These *delimited phenomena* are further divided compartmentally into primary and tributary phenomena as specified below.

PRIMARY PHENOMENA: These phenomena concern the acquisition of a system of cognitive structures that facilitate the perception and production of vowels, by an infant between the ages of three and six months, from a community of fluent speakers of a given language. The major components of the primary phenomena are listed below.

1. The acquisition of a cognitive structure that provides an infant with the means for representing auditory (sensory) information, which is assumed to be vowel-like, coming from both the infant and the speech community, in a way that facilitates the infant's organization of the auditory information into structures for use in further cognitive computation. In the minimal case, the infant must be able to represent and organize their own productions and those of one adult speaker.
2. The acquisition of a cognitive structure that provides an infant with the means for representing articulatory (motor) information derived from the infant's own vowel productions in a way that facilitates the infant's organization of the articulatory information into structures for use in further cognitive computation.
3. The acquisition of a cognitive structure that provides an infant with the means to relate auditory structures and articulatory structures in a manner that facilitates further computation involving the articulatory and auditory cognitive structures themselves.
4. The acquisition of a "normalization" computation for computing equivalences between representations of qualitatively similar vowels that may differ absolutely in

representation due to speaker variation. The normalization computation must be flexible enough to respond accordingly to highly differentiated vowel vocalizations that arise due to cross-speech community differences in vowel productions.

5. The acquisition of a cognitive structure which an infant uses to interpret interactions with the members of a speech community in order to facilitate the acquisition procedures described above.

TRIBUTARY PHENOMENA: The primary phenomena necessitate consideration of (at least) the following: the infant's auditory system and the infant's articulatory system. More specifically, we need to take into account the following:

1. The influence of vowel-intrinsic dynamic properties of vocalizations and their effects on the acquisition of the infant's auditory cognitive structure after interpretation by the auditory system; including the indication of the boundaries of the representational space, and indication of especially salient areas of the representational space that yield significant distinguishability, potentially facilitating normalization.
2. The ontogeny of the articulatory system during the infant's first six months, its potential configurations, and moreover, the strategies used by the infant in production and the manner in which they are represented in the infant's articulatory cognitive structure are relevant for interaction with the auditory cognitive structure.

In the remainder of this chapter, we provide a brief characterization of the primary phenomena, along with previous models, in order to frame the problems to be addressed. In the remainder of this dissertation, we propose a framework for investigating the primary phenomena that accommodates the more refined metaphysics we believe is needed (Chapter 2). We then apply the framework in development of a methodology for modeling vocal learning during early infancy that serves as the basis of the acquisition of vowel normalization (Chapter 3). We then extend the methodology to incorporate intermodal aspects of spoken language acquisition during early infancy (Chapter 4). We conclude with a discussion of the specific issues addressed in this dissertation, and future directions for the work (Chapter 5).

1.1 Basis for Cognitive Phonetics

1.1.1 Representation and Categorization

Rigorous study of the acoustic properties of vowels began to take shape in the early to mid 19th century, particularly in the classic works of Helmholtz and Lord Rayleigh, with further organization and conceptual progress taking place after the turn of the century, e.g., in the works of Russell (1928) and Chiba and Kajiyama (1941). By the middle of the 20th century, the acoustic spectrum and the resonance information it contained had become the standard representation for vowels in the study of their perception by humans. The landmark study of American English vowels carried out by Peterson and Barney (1952), for example, makes extensive use of the first and second formant frequencies in the visualization of vowel data and in reasoning about relationships between vowel categories. Their experimental results on human classification of vowel tokens still serve as a basis for comparison for modern computational models of vowel categorization.

Helmholtz and Lord Rayleigh had also been interested in the perception of vowels, which they knew to involve the physiology of the auditory system as well as a more subjective psychological component, leading Helmholtz to posit psychophysical representations of vowels distinct from their acoustic representation. Steady progress was made on the physiology since the mid 19th century, culminating with mathematical models of the interpretive effects of the auditory system on vowel signals. The models, in their simplest forms, are transformations (in the mathematical sense) from an acoustic representation (typically a vector in a Hertz frequency domain) to a psychophysical representation (based on a psychophysical scale, e.g. the "mel-scale" developed by Stevens et al., 1937, or the "Bark scale" developed by Zwicker, 1961). Revisions of these transformations were proposed as more information on the auditory system was learned. The "ERB transform" and its resultant scale, put forward by Moore and Glasberg (1996) and described in greater detail in Section 3.3.1, is the current basis of a number of psychophysical models of the effects of the auditory system on sound entering the ear, yielding a number of high-dimensional "auditory representations."

Regardless of which kind of representation is employed, basic computational modeling of vowel categorization is typically carried out using a categorization algorithm that computes similarities over some collection of vowel tokens based on their representational distances from each other. In recent decades, "statistical learning" algorithms have come to be viewed as models of vowel category acquisition, with the "simpler" algorithms being taken to model the acquisition procedure carried out by infants. Vowel category acquisition in particular has been taken to involve "plastic" representations which respond and conform to a distribution of perceived vowel tokens. Guenther and Gjaja (1996), for example, make use of self-organizing maps (Kohonen, 1982) which consist of a set of neural units represented by weight vectors. During learning, the weight vectors adapt to a given distribution of vowel tokens, and afterwards are used to determine the model's "perception" of a given vowel token. The ability of the model to simulate aspects of vowel category acquisition, e.g., the "perceptual magnet effect" reported in Kuhl (1991), precipitated the development of similar models (e.g., Lake et al., 2009; Feldman et al., 2009) attempting to broaden understanding of category acquisition.

Guenther and Gjaja's (1996) model is itself situated within a larger architecture (Guenther, 1995; Guenther et al., 1998, 2006; Tourville and Guenther, 2011) which attempts to model the acquisition of a cross-modal mapping relating auditory targets with articulatory trajectories. Guenther's (1995) broader acquisition model, which incorporates articulatory representations derived from Maeda's (1990) articulatory model, is composed of a set of adaptive neural structures over an "auditory perceptual space" whose regions represent possible speech sound targets, an "articulatory configuration" space defined by the seven parameters of the articulatory model, and a series of learned transformations between them, modified by articulatory and auditory feedback.

Guenther's (1995) model has been extended to incorporate vocal tract development in children between 12 and 60 months of age (Callan et al., 2000). The vocal tract representations were adapted from Maeda's (1990) articulatory model using mid-sagittal MRI scans of males between 3 and 45 months of age. Callan et al. (2000) maintain focus on the acquisition of a relationship between auditory representations and articulatory representations, while assuming that vowel categories are already clearly delineated prior to the adaptation of the auditory-articulatory mapping during acquisition. As articulation is thought to shape vowel category learning (Kamen and Watson, 1991; Perkell, 1996; Honda, 1996), an alternative approach is likely needed. The same applies to Lake et al.'s (2009) and Feldman et al.'s (2009) models, which focus exclusively on auditory representations, as well as Rasilo et al.'s (2013) and Hörnstein's (2013) models, which takes articulatory representations as the basis of category learning. In this connection, Guenther's (1995) model has also been extended to include mirror neuron processes involved in speech recognition, and a richer feedback system in which perceptual and motor representations are “brought into register” or “aligned” through activation of these neurons (Guenther and Vladusich, 2012). In this dissertation, the alignment of representations across modalities is assumed to be a the basis for the creation of representations that are critical for the acquisition of spoken language.

1.1.2 Intermodal Representation

Westermann and Miranda (2002, 2004) and Oudeyer (2002) model the initial stages of the acquisition of vowel categories as the integrated self-organization of an articulatory representation and an acoustic representation. The relationship between articulation and acoustics is represented as a pair of Hebbian transformations between their respective representations. Although the models incorporate a form of cross-modal learning, they face (at least) two conceptual difficulties. The articulatory models employed in both seem useful mostly as a first-step in the modeling of sensorimotor coupling when contrasted with the articulatory models in Callan et al. (2000) and Guenther et al. (2006). More generally, all of the aforementioned models lack a key aspect of cross-modal perception likely to be the crux of cross-modal learning. In order to draw out the issues, it is useful to recall a bit of history on intermodal perception, generally, and its potential role in spoken language acquisition.

In the first half of the 20th century, the focus of developmental research turned to integration of the numerous modalities involved in language acquisition, most recognizably in the work of Jean Piaget, who was concerned with the “coördination between hearing and sight” and “between hearing and phonation” (Piaget, 1936, p. 87). Early conceptualizations of the integration assumed that the modalities were initially separate and came to be linked over the course of developement. Since the 1960s, there has been significant refinement, modification, and expansion in research techniques and theoretical organization resulting in a broad variety of approaches to cross-modal perception and learning. Specifically, integration of the relevant modalities is assumed to exist at birth, while the nature of the integration and its developmental properties have become the center of investigation (see Lewkowicz and Lickliter, 1994, for further details). With respect to spoken language, the majority of cross-modal research focused on auditory-visual relations, though steps have been taken toward furthering understanding of auditory-articulatory relations.

To illustrate, early efforts into investigating auditory-articulatory relations attempted to establish their existence and reason about their nature using results obtained on relations between the auditory and visual modalities. Kuhl and Meltzoff (1982), for example, presented 18 to 20 week-old infants with audio-visual vowel stimuli using a loudspeaker situated between two video screens, each of which featured a face that would articulate a vowel syncronously with the loudspeaker playing a vowel sound recorded by an adult female. One of the faces articulated a vowel that matched the vowel recording, whle the other articulated a vowel that did not. Results showed that infants looked more at the face which articulated the vowel that matched the recording, suggesting that 5 month old infants are already in posession of auditory-visual integration of speech sounds. Kuhl and Meltzoff (1982) also reported observing that infants "produced sounds that resembled the adult female's vowels" and "seemed to be imitating the female talker, 'taking turns' by alternating their vocalizations with hers" (p. 1140). Together, the results and observation are taken to "reflect a knowledge of the relationship between audition and articulation," and adduce the "the infant's intermodal representation of speech" (ibid).

In light of results of this nature, Meltzoff and Kuhl (1994) argue more generally for the existence of intermodal representations, defined as "a higher order phonetic representation of speech that acts as a mediator between nonidentical information in the two [or more] modalities" (p. 360) relevant for speech representation. The term "intermodal" is used as to distinguish this concept from that denoted by the term "cross-modal," which is the weaker concept of a relation formed over representations from two or more modalities. If intermodal representations that are more complex than cross-modal representations indeed exist, it follows that this kind of representation is missing from models that focus solely on cross-modality (e.g., Guenther, 1995; Callan et al., 2000; Westermann and Miranda, 2002, 2004; Oudeyer, 2002; Howard and Messum, 2011; Rasilo et al., 2013; Hörnstein, 2013). Guenther et al. (2006) present a reasonable step forward in modeling intermodal representation, yet, their model ignores important intermodal phenomena, such as "multisensory narrowing" that occurs during early spoken language acquisition (see Lewkowicz and Ghazanfar, 2009, and Section 4.1) as well as the broader socio-biological phenomena discussed in Section 1.2. In this dissertation, we take intermodal representations over auditory and articulatory representations to be a key aspect of the acquisition of vowel normalization, as they are created by the infant in order to model social agents interacting with the infant.

1.1.3 Relations over Representations

Consideration of intermodal representation draws attention to relations over representations. Each of the cross-modal models discussed in the previous section are in some sense pairings of auditory and articulatory representations, hence they take for granted the notion that representations can be organized in this fashion. The relations explicitly considered span multiple modalities yet, relations within a single modality are also commonly assumed. For example, within the "task dynamics" approach to speech production (Saltzman, 1986; Saltzman and Munhall, 1989; Saltzman, 1995) the articulatory modality is modeled as a set of relations between representations across multiple "reference frames," each similar in nature to Guenther's (1995) articulatory configuration space, though capturing different aspects of speech production (see Chapter 2 for greater detail on reference frames). Moreover, within most models, cross-modal or otherwise, different kinds of relations are assumed between representations within a single reference frame, again capturing different aspects of the phenomena.

Nearey (1989), for example, discusses several relations over representations, including "vowel-intrinsic spectral change," which is taken to be an important cue for vowel identification (Nearey and Assmann, 1986). This concept is typically modeled in terms of a set of tuples over an "acoustic reference frame," such as the F1-F2 plane, where each tuple corresponds to a single vowel production. To illustrate, (Nearey and Assmann, 1986) use ordered pairs derived by extracting 30ms sound clips centered at the 24% and 64% time points of naturally produced Canadian English vowels. Similar approaches use either the differences between pairs values in the tuples (e.g., Neel, 2008; Fox and Jacewicz, 2009), or curves that represents the overall spectral movement for a vowel estimated from the tuple values (e.g., Zahorian and Jagharghi, 1993; McDougall, 2006). We demonstrate the incorporation of relations over representations of this nature within our model in Chapter 4.

Relations based on spatial and metric concepts are also typically assumed. For example, Guenther's (1995) model assumes that speech sounds correspond to regions within an auditory perceptual space, where the term "region" is defined using the "position" of representations within the space, and "distance" between representations across the space. It is these concepts that permit the grouping of representations into regions. The position of a representation within a space is typically determined in terms of the orthogonal arrangement of axes corresponding to "features" or "parameters" of speech sounds (e.g., formants), and the distance between representations is determined based on their positions within the space. Within Guenther's (1995) model, inter alia, speech sounds are identified with groupings of representations based on these concepts alone, which is potentially problematic for modeling phenomena which necessitate other means of relating representations, a key example being the formation of "equivalence relations" over representations of speech sounds produced by different speakers, as discussed in the next section. Guenther's (1995) model, like many others (e.g., Callan et al., 2000; Westermann and Miranda, 2002, 2004; Oudeyer, 2002; Guenther et al., 2006) assumes that a relation over representations of this kind exists prior to vowel category acquisition.

In this dissertation, we take relations over representations called "manifolds" to play the principle role in category acquisition. The assumption is based on the study of manifolds (Seung and Lee, 2000; Belkin and Niyogi, 2003; Ma and Fu, 2012) in organizing representations of physical phenomena taken to have a manifold structure. For example, "vowel manifolds" (Jansen and Niyogi, 2006, 2007) formed over vowel representations are argued to facilitate the learning of speech sounds, while "cognitive manifolds" are argued to be used by infants to relate representations within and across modalities (see Plummer, 2012a, and Chapter 4). The latter is achieved through "manifold alignment" (Wang, 2010), which maps representations on two (or more) manifolds to a mediating "latent space" (Ham et al., 2005; Ma and Fu, 2012), where categorization may take place (Plummer et al., 2010), or further cognitive computations. Manifolds are discussed in detail in Chapter 2.

1.1.4 Vowel Normalization

Infants lacking adequate auditory access to the ambient speech of others are known to acquire abnormal spoken language. For instance, hearing-impaired infants, at four months, possess a smaller repertoire of consonant-like vocalization types (Stoel-Gammon and Otomo, 1986) and, at seven months, produce vocalizations with fewer canonical syllables (Oller and Eilers, 1988) than their normal-hearing counterparts. In addition to the ambient speech of others, infants may require auditory access to their own vocalizations. Normal-hearing children who in infancy had tracheostomies for extended periods of time exhibit slower speech development relative to other developmental cognitive components (Bleile et al., 1993).

These considerations highlight an important aspect of vowel category acquisition that needs to be addressed in modeling. The speech productions of adult speakers of a language are quite different from those of an infant (Davis and MacNeilage, 1995; Kuhl and Meltzoff, 1996) or older child (Lee et al., 1999) learning that language. These acoustic differences result in differences in the psychophysical interpretations, and thus auditory representations, of "categorically similar" vowels, at least at the periphery of the auditory system. Moreover, the differences between infant and adult productions constitutes only a small part of the variation in the ambient speech resulting from the multitude of speakers contributing to it. Nevertheless, infants are able to compute the categorical similarities across different speakers (Kuhl, 1979, 1983), and relate these similarities to the equivalences over representations of their own productions, even across reference frames (i.e., frames whose representations incorporate bone conduction, or other organism-internal influences). These computations are likely crucial to spoken language acquisition.

Attempts have been made to account for (or more accurately, eliminate) representational differences due to speaker variation within the distributional approach to vowel system acquisition. Callan et al. (2000) and Guenther et al. (2006) address the issue of speaker variation by representing vowels using log ratios over formant values: $\log(F2)/\log(F1)$, $\log(F3)/\log(F2)$, $\log(F2)/\log(F1)$, and $\log(F1)/SR$, where SR is a "sensory reference," calculated using the geometric average of all values of F0 for a given talker (Miller, 1989), thereby reducing differences in the representation of categorically similar vowel tokens. Ishihara et al. (2009) use a probabilistic model of normalization based on a standard statistical learning algorithm. The normalization computation amounts to a simple translation in the standard F1,F2-space (in mels). Adank et al. (2004) reviews a large set of similar "standard" formant-based normalization algorithms, evaluating their performance at eliminating variation due to anatomical/physiological variation, while preserving phonetic and sociolinguistic variation.

Hindle (1978) describes normalizations of this nature as "technical answers" to the question "How can the measured formant values for different speakers be transformed so that different speakers' versions of the same phoneme coincide?" (pp. 162-3). By the same token, Hindle brings attention to the "psychological aspect of the normalization problem," asking "What is a speaker doing when he equates two vowels spoken by different speakers and having different formant values?" (p. 162). Following this second line of thought, Sussman (1986) addresses the question by suggesting that normalization is "hard-wired" into an infant's neural circuitry, providing the cognitive means for handling speaker variation. Ames and Grossberg (2008) implement a "biologically inspired" normalization computation, carried out by "orthogonal strips" along the auditory cortex, to account for speaker variation as a preliminary step in category acquisition.

Whether "technical" or "psychological," these approaches assume that vowel normalization is prespecified and fixed, and unresponsive to component-wise and functionally differentiated aspects of the input they operate over. Moreover, they presuppose the goal of the computation is some invariant representation. This assumption and presupposition, however, seem to be in conflict with what is known about the phenomenon from observation and experimentation (e.g., the context-sensitivity in Ladefoged and Broadbent, 1957, the cross-language effects in Johnson, 2005, and the long-term ontogenetic effects in Kohn and Farrington, 2012). A vowel normalization computation is a far richer phenomenon than previously recognized (see Johnson, 1990a,b, 2005).

In this dissertation we take a sharp break from Hindle's dichotomous framing of normalization, instead taking such a computation to be characterized in terms of an infant's cognitive development and complex interaction with the environment. Specifically, we view the creation of an equivalence relation on vowel productions as derivative of a more fundamental computation involving "alignment" of the manifolds used to organize the articulatory, auditory, and more abstract representations an infant develops during early vocal learning. In the next section, we lay out a socio-biological framework for characterization of the acquisition of a vowel normalization computation.

1.2 Socio-biological Learning

1.2.1 Vocal Imitation in Phonetic Category Acquisition

In addition to motivating the role of normalization in vowel category acquisition, the early vocalizing carried out by infants suggests the importance of the social interaction between caretaker and infant during the early critical period in the development of the cognitive structures involved in vowel category acquisition. Although the "imitation games" played by an infant and caretaker seem to substantially affect the infant's acquisition of speech and language, their nature and structure is by no measure well-understood, and their relation to an infant's cognitive development is even more obscure. We begin with a short catalog of the concepts involved in characterizing vocal imitation in the development of speech production and perception in infants, the main entries being differentiated infant and adult vocalizations, constraints on relations between them, and an infant's cognitive representation of the relations.

A natural point of departure is *Descent of Man*, where Darwin (1871) observed that a kind of vocal imitation takes place between young and their parents in some species, noting in particular that "[b]irds imitate the songs of their parents, and sometimes of other birds" (p. 62). Darwin generally noted that "[t]he parents of many animals" make use of "the principle of imitation in their young" in order "to educate them" (ibid) by providing interaction likely necessary at key stages of development. Assuming the general principle applies to vocal imitation, Darwin's observations may be interpreted as an inceptive characterization of a nontrivial signaling relation between conspecifics impacting vocal learning.

In addition to sensorimotor learning, Jean Piaget's seminal work in developmental psychology brought attention to the aspects of infant intention involved in vocal imitation. Piaget (1945) described a vocal imitation "learning process" in stages determined by the level of intention present in an infant's imitative actions. The first stage, for example, occurring during the first few weeks of life, involves no intention on the part of the infant, and is characterized as a "level of pure reflexes" wherein an infant vocalizes simply in reflexive response to hearing others. The second stage, occurring primarily during two and three months of age, builds on the reflexive nature of the first stage through "vocal contagion" wherein "the voices of others stimulate the infant's voice, whether it be crying or some other sound," and in addition, "the other voices must either reproduce certain familiar sounds already uttered by the child, or certain intonations known to him" (p. 10). Piaget's stages suggest the importance of differentiated infant vocalizations in acquisition. The second condition on vocal contagion also highlights the importance of differentiated caretaker vocalizations.

Taking a comparative biological approach, Masataka's (2003) investigation of the imitative vocal interactions between infants and their caretakers draws heavily from studies of imitative behavior in nonhuman primates. Importantly, the development of the infant's vocal tract is an explicit factor in the learning of imitation, and the variation in the acoustic properties of signals produced by different individuals, caretakers and infants included, is explicitly recognized as an important aspect of vocal imitation. Moreover, differences in the forms of infant vocalizations over the course of their development are taken to be an integral part of imitation, even driving it at times. For example, during the third month of life an infant's vocal tract begins a radical transformation wherein the larynx begins to lower, allowing the infant to produce more adult-like vocalizations. Turn-taking vocal exchanges between infants and caretakers are known to elicit a higher proportion of infant-produced "syllabic sounds," which exhibit "greater oral resonance, pitch variation, and possible consonant-vowel contours" to "vocalic sounds," which have "greater nasal resonance" and are perceived as being "more uniform in pitch" Bloom et al. (1987, p. 215). The differentiated vocalizations also receive differentiated responses, as mothers respond contingently to the different infant vocalizations, with "play-like" vowel vocalizations in response to vowel-like infant vocalizations, and with imitation in response to "consonant-vowel clusters" (Gros-Louis et al., 2006). The social interaction based on vocal exchanges composed of differentiated infant and adult vocalizations is likely crucial to phonological learning (Goldstein and Schwade, 2008). These results establish conditions on the signaling link between conspecifics in vocal learning hinted at by Darwin 130 years earlier.

Additionally, Masataka (2003) takes steps toward formulating a characterization of the cognitive components corresponding to the vocal imitative interaction. Masataka (2003, p. 125) takes "perception" to be the "discrimination of an object or an event through one or more sensory modalities, separating them [sic] from the background or from other objects or events," and "perceptual characterization" to be "a process by which an individual may treat non-identical objects or events as equivalent" (citing Edelman, 1987, p. 26). The "equivalence emerges in the mapping between two disjunctive processes," (p. 125) such as the production and perception of vocalizations (or the perception of self-productions and perception of the productions of others). Furthermore, "the mapping – the categories – self-organize through their reciprocal interaction with one another" (citing Thelen and Smith, 1994, p. 143). Masataka (2003) concludes that, in each infant, "through interaction with caregivers...some global mapping [i]s selectively strengthened," constituting "the development of imitation in infants" (p. 125).

The "emergence" of categories via turn-taking vocal imitation is typically modeled using statistical learning algorithms. de Boer's (2000) model focuses on the emergence of vowel categories at the community-level via cooperative imitation-based interactions between agents in the community. During the interactions, an initiator agent produces a vowel token, an imitator agent then tries to produce a maximally similar token, which is then judged a successful imitation or not by the initiator. Repeated interactions result in the emergence of community-level vowel categorization systems. Oudeyer (2001, 2002) uses a similar model primarily concerned with the general emergence of aspects of speech and language at the community level. Other models (e.g., Bailly, 1997; Ishihara et al., 2009; Guenther et al., 2006; Heintz et al., 2009; Ananthakrishnan and Salvi, 2011; Howard and

Messum, 2011; Rasilo et al., 2013; Hörnstein, 2013) take a similar approach, though focusing more on the internalization of the imitative exchanges within a single learner. However, all of these models adopt a view of vocal imitation based on statistical learning that may not accord with experimental results that reveal the phenomenon to possess subtle structure, e.g., contingent turn-taking between infant and caretaker, with highly differentiated socio-functional input likely crucial to phonetic category acquisition. More broadly, the aforementioned models lack explicit modeling of an infant's creation of representations of the social agents in their environments, which may be a precondition on vocal learning.

1.2.2 Social Aspects: Vocal Imitation to Vocal Learning

Preliminary characterization of the complex phenomenon vocal imitation in phonetic category acquisition requires characterization of the different kinds of vocalizations produced by infants and their adult caretakers, and how these vocalizations are related in vocal exchanges, along with characterization of an infant's cognitive organization of the relation. As rich as vocal imitation may be, it constitutes only a small part of the more general phenomenon of vocal learning, about which very little is known. The broader social nature of the vocal exchanges between infants and caretakers and the broader aspects of its cognitive organization within the mind of the infant have only come into view in the last few decades, and approaches to modeling them are only just beginning to take shape. With this in mind, we recount a few broad socio-cognitive aspects, taking the following as our starting point: assuming that infants and adults are partaking in vocal imitative exchanges, why are they doing so at all?

Beginning with adults, Bloom and Lo (1990) showed that they prefer the more "speech-like" syllabic vocalizations of infants that are produced in turn-taking vocal exchanges. Importantly, these "speech-like" vocalizations are potential loci for "phenotypic matching"

– members of a species seeking their own behavioral qualities in conspecifics, or more broadly, the behavioral qualities of kin. According to Masataka (2003), “[p]henotypic matching is known to occur extensively in nonhuman primates” which are “thought to achieve matching through various communicative behaviors,” e.g., postnatally learned dialects in New World primates (p. 89). Fitch (2004) recounts that human “language seems more complex than necessary for communication, in the sense that our ability to recognize regional or class dialects far exceeds the needs of semantic communication” (p. 290), suggesting that “the existence of extra-propositional dialectal variation increases the ability of kin to recognize each other and to share information” (p. 290). Adults may be looking to transfer their speech and other communicative behavioral qualities to infants, and doing so during the specialized interaction using the preferred vocalizations.

Concerning the infants, old insight can be found in the “new approaches to the mind” that formed within the field of social psychology in the late 19th century, which focused investigation on imitation “as the means by which animals and children become part of a group” (Smith, 1997, pp. 489-90), e.g., a speech community in the case of children (as briefly noted in Whitney, 1875, Chapter 1). G. H. Mead (1909) recognized that imitation itself is a complex phenomenon, noting that “the existence of a development of self and knowledge of others is a precondition of imitation,” and that “this precondition results from a reciprocal interaction of stimuli and action, that is, a social process prior to the differentiation of self” (as characterized by Smith, 1997, p. 490). If true, vocal imitation may have to meet the same precondition. And if so, it may be that infant vocalizations during the first six months of life constitute far more than vocal practice for imitation. Rather, they, together with the vocalizations of adult caretakers, provide the representational basis for the development of a “vocal self” as differentiated from “vocal others.” Development of the former likely occurs through the acquisition of an auditory-articulatory intermodal mapping, and the latter through interpretation of a caretaker's differentiated feedback. This conceptualization is by no means new, as it accords with 17th century physician Luigi Settala's observation that "hearing must be regarded as the first thing that opens the way to speech," and following this initial experiential opening, "the intellect is rendered apt both for reception of the species represented by the voice and speech, and for expression through speech" (quoted in Wollock, 1997, p. 258), that is, the infant's formation of auditory representations of "other" conspecifics, and development of the means to communicate with them.

Along these lines, recent experimental results suggest the importance of individual speaker identification, hence the representation of others, to the formation of more abstract phonetic representations (e.g., Perrachione et al., 2011). Representation of "self" has recently been more generally adduced in modeling of the cognitive development of intentional agents (e.g., the 'like me' framework in Meltzoff, 2007). Importantly, the essential distinction between "self" and "others" entails derivation of a mapping that "allows the infant to see the behavior of others as commensurate with their own" (Meltzoff, 2007, p. 26). The nature of such a mapping depends on the infant's interpretation of what behavior in others is or is not commensurate with their own. For example, following an infant's vowel production, a caretaker may respond with a vowel productions of their own together with a positive social signal to indicate to the infant that their production was a "good" example of a vowel within the caretaker's vowel system. The infant's interpretation of the exchange then guides the formation of the likeness mapping. Within this view, vowel category acquisition emerges only as a byproduct of the acquisition of the normalization computation. This conceptualization provides an ontogenetic explanation for Johnson's (1990a) insight that “information that hearers use to evaluate vowel quality includes not only acoustically available information (such as vowel spectrum and F_0), but also computed information about the person [or group of people] doing the talking” (p. 252). More generally, it seems to be in line with the “shared manifold” hypothesis put forward by Gallese (2001) on the basis of the discovery of mirror neurons and their potential role in spoken language acquisition, inter alia.

Even more generally, infant vocalizations viewed in light of the results discussed above are characterized by their “social function,” that is, their role “as a signal for establishing and maintaining social contact with others” (Hsu et al., 2013, pp. 3-4), though they are likely multi-functional, serving also “as a symptom reflecting an infant’s motivational and affective state,” and “as a symbol representing an awareness of and information about the context in which vocalizations are produced” (p. 3). Moreover, the symptom-symbol-signal functional differentiation is likely decoupled from the differentiated infant vocalization forms that are observed in infant-adult vocal exchanges, with the coupling learned during the first year of life whence “the specific patterns of coupling are often idiosyncratic and unique to individual infant-caregiver dyads” (p. 5). Indeed, Hsu et al.’s (2013) investigation of the effects of social games between infants and caretakers on infant vocalizations support the decoupling. These results provide a glimpse at the high degree of abstraction involved in an infant’s developmental organization of early auditory comprehension, and how much further there is to go in simply characterizing it.

1.3 Aspects of the Theory of Vowel Normalization Acquisition

In light of the discussion in previous sections, we adopt the following approach to the acquisition of vowel normalization.

1. The acquisition of a cognitive structure that provides an infant with the means for representing auditory (sensory) information, which is assumed to be vowel-like, coming from both the infant and the speech community is modeled in terms of manifold formation over auditory representations. Manifold formation is defined and discussed in Chapter 2, while auditory representations are characterized in Section 3.3.1. Models involving "multifold" auditory representations are demonstrated in Chapter 4.
2. The acquisition of a cognitive structure that provides an infant with the means for representing articulatory (motor) information derived from the infant's own vowel productions is modeled in terms of manifold formation over articulatory representations. Articulatory representations are characterized in Section 4.2.
3. The acquisition of a cognitive structure that provides an infant with the means to relate auditory structures and articulatory structures is modeled as manifold alignment, which is defined and discussed in Chapter 2. The modeling itself is demonstrated in Chapter 4.
4. The acquisition of a "normalization" computation yielding equivalences between representations of qualitatively similar vowels that may differ absolutely in representation due to speaker variation is also modeled as manifold alignment. Normalization restricted to auditory representations is demonstrated in Chapter 3, while intermodal normalization is demonstrated in Chapter 4.
5. The acquisition of a cognitive structure which an infant uses to interpret interactions with the members of a speech community is modeled as a weighted pairing operation over representations, which is defined in Chapter 2. Cross-linguistic investigation the influence of the pairing operation on the acquisition of vowel normalization, hence vowel categorization, is carried out in Chapter 3 through the creation of a "virtual

environment for vocal learning." The environment consists of models of caretaker agents representing five different language communities (American English, Cantonese, Greek, Japanese, and Korean) derived from vowel category perception experiments, and models of infant agents that "vocally interact" with their caretakers. The "social" and "auditory" aspects of the vocal learning environment are described in Chapter 3, while the "articulatory" and "intermodal" aspects are presented in Chapter 4.

CHAPTER 2: CONCEPTUAL AND TECHNICAL BASIS

In this chapter, we narrow the scope of the previous chapter to focus on key aspects in modeling the acquisition of vowel normalization. We take as our point of departure the ideas surrounding one of the "major breakthroughs" in linguistics recounted by Hockett (1965), the "quantization hypothesis," which constitute the seldom-cited-but-often-invoked historical predecessors to the kind of work carried out in this dissertation (Section 2.1). With the conceptual background established, we then describe a computational model for investigation of the approach, assuming essentially no background and building from the ground up (Section 2.2). We then formulate the computational model, providing a mathematical specification of the concepts involved in the implementation (Section 2.3).

2.1 Aspects of the Theory

In his 1965 Presidential Address to the Linguistic Society of America, Charles Hockett selected from the history of linguistics four achievements which he described as "major breakthroughs" central to the field. One of the achievements, which he termed the "quantization hypothesis," was put in motion by the "discovery of phonetics," referring to the technical advances of the late 19th century in the recording of speech productions and the study of their corresponding acoustics. The "discovery," Hockett writes, led to the understanding that "[t]he 'space' of all possible speech sound, either in articulation or acoustically, is a multidimensional continuum" (p. 192) in which "[t]here is no end to the fineness of difference that can be observed" between speech sounds. As this understanding took shape, Hockett explains, it necessitated more sophisticated thought concerning the “neat discrete” nature of speech sounds.

Hockett attributes early progress on this matter to Otto Jespersen, citing a key passage in Jespersen’s classic *Language: Its Nature, Development, and Origin.* Jespersen’s (1922) passage, attempting to address articulatory variation and sound change as well the nature of speech sounds, appeals to “a sort of conception of an average pronounciation...which we aim at” in speech production, determined with respect to “the only measure at our disposal” which is “that we are or are not understood” (p. 166). Hockett’s comments on this passage suggest that Jespersen may not have been aware of its significance to the development of the quantization hypothesis, and that Jespersen even rejected the hypothesis once it came to prominence. Nevertheless, Hockett makes clear reference to the passage in his crystallization of the quantization hypothesis later in his address:

> “In any speech community, only certain DIFFERENCES of speech sounds are functional. This breaks the continuous multidimensional space of all possible speech sound into a finite number of regions; in at least some environments it matters, to hearer and thus to speaker, whether a given sound falls into one region or the next, but does not matter where it falls within a single region. Successive articulations aimed into the same region show a scatter, clustered around a LOCAL FREQUENCY MAXIMUM. The local frequency maxima are, or create, the ‘sort of conception of an average pronunciation’ of which Jespersen spoke. The frequency maxima, then, are the neat discrete functioning units of the phonological system of the language. (pp. 194-5)

Hockett next sketches a formulation of basic notions from phonetics and phonology with respect to the quantization hypothesis, using an “acoustic space over all possible speech sound with the smallest number, say n, of pairwise orthogonal parameters (axes)” (p. 195). Within the sketch, a speech signal corresponds to a “time-dependent vector” that “traces a trajectory” through the acoustic parameter space. Moreover,

> “the time-dependent vector that is the speech signal for a given speaker – both for what he says and for what he hears from others – spends more time in some

> regions of acoustic space than others. This yields a DENSITY DISTRIBUTION defined for all points in the space...Within the acoustic space there will be a finite number of points at which the density distribution is a local maximum." (p. 201)

Hockett takes these local maxima (or the local frequency maxima adapted from Jespersen) for a particular language to be the "acoustic allophones" of that language, and the projections of its acoustic allophones onto the coordinate axes of the acoustic space are taken to be its "distinctive features."

Following a review of Noam Chomsky's contributions to linguistics, the fourth "major breakthrough," Hockett "bring[s] the quantization hypothesis to fruition," (p. 200) discussing a specific version of his own formulation sketch, based on the work of Chomsky and Morris Halle, along with a critique of it. Focus on the Chomsky and Halle take on the quantization hypothesis is narrowed to their assumption that "[t]he parameters of the speech signal (the coordinate axes [of the acoustic space]) are universals, the same for all languages" (p. 201). Hockett then takes aim, insisting that "Chomsky and Halle are wrong when they assume that a single fine-grained quantization of acoustic space can yield a finite set of points of reference valid for all languages" (p. 201-2), on the basis that this simplification is too simple. Hockett adduces this claim with a few observations problematic for Chomsky and Halle:

> "The number of contrasting distinctive features along any one axis can and does vary from language to language, and even when the number is the same the exact locations are typically different; the quantizations differ from language to language in such a way as to render the units of different languages incommensurate...Furthermore, in course of time in any one community, as the density distribution varies for all the speakers of the community in pretty much the same way, the local maxima slowly wander about." (p. 202)

The broader conclusion Hockett draws from the criticism is that the assumption is too narrow in scope to deal with variability and sound change, calling for a more general view of the components of the acoustic parameter space.

Hockett's discussion of the quantization hypothesis touches on a number of issues that will feature within this dissertation, e.g., the concept of a continuous multidimensional space of speech sound, the treatment of speech signals as trajectories within such spaces, formulating parameter spaces for talking about these points, defining functions over the vectors in these parameter spaces in order to formulate phonetic and phonological concepts, and interpreting these spaces, points, vectors, and functions with respect to what is known about the spoken language phenomena we are interested in, all while attempting to avoid making our theories and models too simple. It is worth while to sort out which aspects of the conceptualization and which components of the formulation sketch above we will build on, and which need to be modified or simply rethought.

Hockett's statement of the quantization hypothesis mentions a "continuous multidimensional space of all possible speech sound," and its assignment of an n-dimensional acoustic parameter space, along with an assignment of a "time-dependent vector" that "traces a trajectory" through the acoustic parameter space to each speech sound. It is useful to review a specific instantiation of such an assignment, and since we limit our attention to vowel signals in this dissertation, the instantiation centers on the delimited continuous multidimensional space of all possible vowel sound. The acoustic parameter space used in Peterson and Barney's (1952) analysis of vowel categorization has coordinate axes that correspond to parameters derived from formant patterns of vowel signals. An acoustic parameter space of this type is often called a *formant space*. Each vowel sound within the space of all vowel sound is assigned to a single point (i.e., a trivial time-dependent vector) within the formant space. Before discussing this particular instantiation further, we take the opportunity to generalize it a bit, and expand the conceptual stock.

Although Hockett refers to “the ‘space’ of all possible speech sound, either in articulation or acoustically,” the parameter space he describes as an example is strictly acoustic. We need a general way of talking about speech sound that may not necessarily be acoustic in nature, and moreover departs from the notion that there is a single “space” of all such sound. In this connection, we follow Guenther (2003) in an approach dating back to Lewin (1936), which makes use of the concept “reference frame.” Guenther’s (2003) neural model of the processing and interpretation of speech sounds uses a collection of reference frames each of which “can be thought of as a coordinate frame that best captures the form of information represented in a particular part of the nervous system” (p. 209). The model also uses reference frames that represent information from other domains including the physical and the psychophysical. We adapt the notion to include cognitive domains as well. We will discuss reference frames further in Section 2.2, providing a mathematical formulation in Section 2.3.

In this dissertation reference frames are assumed to be Euclidean spaces, and a reference frame whose axes serve as interpretations of parameters are called “parameter spaces.” We dispense with the notion of a single space of all possible speech sound, opting in favor of a set of physical, psychophysical, and cognitive reference frames, each of which best captures the form of vowel phenomena information relevant to our investigation of the acquisition of vowel normalization.

We postpone discussion of cognitive reference frames until later in this section. The physical reference frames used in this dissertation are essentially those used in Peterson and Barney (1952) only with a larger number of parameters. The reason for the increase in parametric complexity is provided by Peterson and Barney (1952) themselves in an observation often overlooked within the computational modeling community. Within an acoustic space with two parameters, taken to be F1 and F2, the authors note that “[ɝ] produces extensive overlap in the [ʊ] region in a graph involving only the first two formants,” but “may be isolated from other vowels readily by means of the third formant” (p. 183). The problem is more general than this one instance, and thus we will typically use parameters corresponding to the first three formant frequencies of vowel sounds.

The two psychophysical reference frames in this dissertation are meant to capture information representing the manner in which the human articulatory and auditory systems produce and perceive vowel sounds, as a first step in modeling the organization of this information in the mind of a developing infant. For example, we may take an “articulatory reference frame” to be composed of vectors representing an infant’s psychophysical organization of the motor action of the articulatory system. We derive such vectors from the assignment of coordinates to mid-sagittal vocal-tract images as described in Section 4.2, though this should be viewed only as a first step toward a more sophisticated dynamical systems approach (Saltzman, 1986; Saltzman and Munhall, 1989; Saltzman, 1995). Similarly, we may take an “auditory reference frame” to be composed of vectors representing an infant’s psychophysical organization of energy level information from the basilar membrane’s response to vowel signals. Further details about this kind of vector representation of the response of the auditory system to vowel signals are provided in Section 3.3.1.

The use of multiple reference frames for talking about vowel signals necessitates description of a manner with which to relate them. Within Guenther’s (2003) approach to reference frames, “[i]nteractions between brain regions can be thought of as transformations of information between the corresponding reference frames” (p. 209). As before, we extend the notion of “transformation” beyond the physical and psychophysical conceptualization so that it includes the interactions involving cognitive reference frames. We extend Hockett’s formulation sketch by taking transformations to be mathematical functions between reference frames.

Before proceeding, we return to Hockett’s concept of a space of all possible speech sound, to illustrate the care needed when linking such a concept to a reference frame. Peterson and Barney’s (1952) formant space illustrates the crucial point that not all parameter settings within such a space have a sensical interpretation with respect to the space of all possible vowel sound. By definition, the first formant frequency of a vowel must be less than or equal to its second formant frequency, and so on for higher formant frequencies. Moreover, the articulatory system imposes hard limits on the formant patterns that humans can produce. Thus, within any formant space only a subset of its constituent parameter settings make sense with respect to representing vowel signals. To further illustrate, consider an articulatory parameter space over real-valued parameters corresponding to parts of the tongue, the lips, and the larynx. Due to the morphology of the vocal tract, only a subset of the parameter settings constituting the articulatory parameter space make sense with respect to physically possible arrangements of these articulators.

In the case of Peterson and Barney’s (1952) formant space the subset corresponding to the space of all vowel sound possesses an integrated character, often taken to be that of a shape, as evinced by its name “vowel triangle,” in use for at least 150 years (see Russell, 1928, Chapter 13). We take this intuitive characterization as a starting point for the addition of the notion of “shape” to Hockett’s initial conceptual stock. Specifically, we assume that the space of all possible vowel sound is conceivable as a shape, such as a triangle or tetrahedron, and that this shape is reflected in the subsets of reference frames used to describe it. The geometric conceptualization within a formant space has a mechanical correlate within articulatory parameter spaces. A subset of such a space corresponding to the possible arrangements of the articulators is often called a “task space,” and is naturally interpreted as a kind of shape embedded within the parameter space. To illustrate the interpretation, think of a ball attached to a rod fixed to a swivel such that the rod can swivel freely both horizontally and vertically. The reference frame for the ball at the end of the rod in this case is simply three-dimensional Euclidean space, while the task space is the two-dimensional surface of the sphere centered at the swivel, with radius equal to the length of the rod.

The “shapes” described above can be given a roughly unified treatment by assuming that they are objects called “manifolds.” A manifold can be thought of generally as a shape that can be divided up into small regions, each of which can be mapped to a region in a Euclidean space, and hence assigned coordinates, where the dimensionality of the Euclidean space is constant across regions of the manifold. For example, an atlas of the surface of Earth contains maps of regions of Earth situated within a plane. In this *intrinsic* view, “a manifold exists in and of itself, and needn’t lie in any higher dimensional space” (p. 39 Weeks, 2002). Another approach to thinking about manifolds holds that they are embedded within some ambient space. This *extrinsic* view focuses attention on the ambient space in which a manifold is embedded, in addition to the manifold itself. To contrast the approaches, within the intrinsic view, a plane may be conceived of in and of itself, without attending to some ambient higher-dimensional space within which it may be embedded. Within the extrinsic view, the plane is conceived of as embedded within a higher dimensional ambient space.

In this dissertation, we adopt the extrinsic view of manifolds, along with the assumption that they are embedded within reference frames. Applying the view to the discussion of vowel sound, we assume that vowel signals lie on “vowel manifolds” (Jansen and Niyogi, 2006, 2007) embedded within some ambient reference frame. To illustrate, we take the “vowel triangle” subset of Peterson and Barney’s (1952) formant space to be a manifold whose ambient space is simply the formant space. Similarly, we may conceive of “articulatory manifolds” embedded within articulatory reference frames (Plummer, 2012b), “auditory manifolds” embedded within auditory reference frames, and so on. That is, we may conceive of each of the several “perceptual manifolds” (Niyogi, 2004) embedded within the appropriate psychophysical reference frame. Again, we extend the conceptualization beyond the physical and the psychophysical, and posit the existence as well of ***cognitive manifolds***, which an infant may use to abstract over their organizations of psychophysical information. We describe a computational approach to manifolds in Section 2.2, followed by a formulation of the approach in Section 2.3.

The next issue we raise forms the basis of much of the remainder of this dissertation, and despite its importance to the titular topic of Hockett’s address, namely sound change, he seems to have completely omitted discussion of it. Recall that, according to Hockett,

> “the time-dependent vector that is the speech signal for a given speaker – both for what he says and for what he hears from others – spends more time in some regions of acoustic space than others. This yields a DENSITY DISTRIBUTION defined for all points in the space...Within the acoustic space there will be a finite number of points at which the density distribution is a local maximum.” (p. 201).

Peterson and Barney’s (1952) vowel data were produced by children, women, and men, having substantially different vocal tract lengths. Accordingly, the authors observe, “the children's formants are highest in frequency, the women's intermediate, and the men's formants are lowest in frequency" (p. 183). That is, the children's vowel data occupies a region of formant space different from that occupied by the women's and men's vowel data. Consider a limiting case, focusing on one infant and one adult and the vowel signals each may produce. A sample of such vowel signals is depicted within a formant space in Figure 2.1 (left). Suppose the language they speak has a three vowel system, composed of only the corner vowels /i/, /u/, and /a/. Moreover, suppose that the green, orange, and yellow points in Figure 2.1 (middle) represent regions within the formant space where the adult's speech signal "spends more time," corresponding to these corner vowels. Similarly, suppose that the green, orange, and yellow points in Figure 2.1 (right) represent regions within the formant space where the infant's speech signal "spends more time," also corresponding to these corner vowels. By Hockett's formulation the regions represented by the green, orange, and yellow points in each of the infant's and adult's vowel sample sets each have a local maximum, and these local maxima essentially represent the vowel inventory of the language of the adult and the infant. In this simplified example involving only two speakers, given the locations of their vowel signals, we have five local maxima, and hence a five vowel inventory. Yet, the language has only three vowels, a contradiction. This problem only gets worse as we introduce the vowels of more speakers into the formant space.

The general problem that formant frequencies, and hence the local maxima, of different speakers do not match up nicely in formant space was recognized at least as early as the latter part of the 19th century. R. J. Lloyd (1890) proposed a solution to the problem based on the relationship between the formant frequencies of a vowel signal. Specifically,

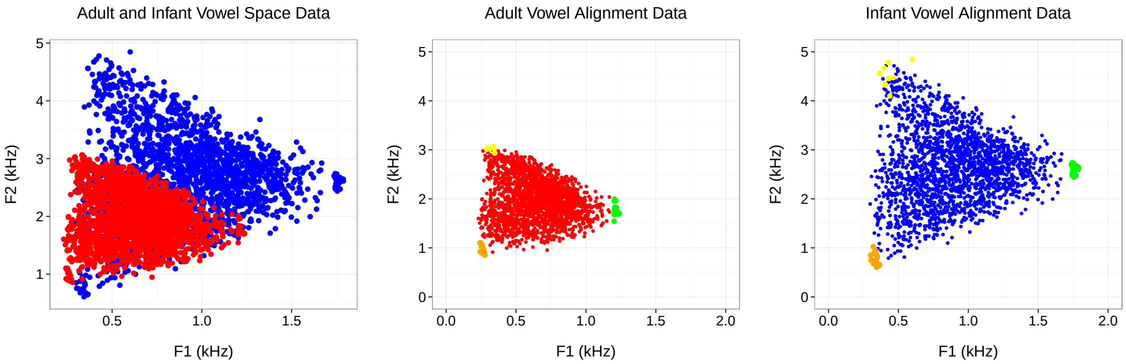


Figure 2.1: Formant space representations of infant (blue) and adult caretaker (red) vowel signals.

Lloyd holds that "[v]owel-quality is not conferred by the absolute pitch of one or more concomitant resonances, but by the relative pitch of two or more" (p. 177). That is, although formant patterns of vowel signals of the same quality exhibit great variation across individuals, one aspect of them that remains invariant is some relation between their formant frequencies, and it is with respect to this invariance that local maxima be determined. The problem was also recognized and treated in the landmark work by Chiba and Kajiyama (1941), who proffered a similar, qualified invariance approach wherein a "*vowel is characterized by its relative formants, provided the centres of the formants are situated within certain frequency regions fixed for a given vowel*" (pp. 193-4).

While the solutions above operate over an acoustic reference frame, Joos (1948), who provided an illustration of the problem using a formant space containing vowel triangles corresponding to three different speakers (p. 60), suggested an alternative approach. His proposed solution differs from both Lloyd's and Chiba and Kajiyama's in that the acoustic differences in formant patterns produced by different speakers is overcome through some sufficient amount of experience with a sufficiently broad set of the vowels of each speaker, which then are used to recover information about their respective vocal tracts. That is, the invariant aspect of a vowel is represented within some articulatory reference frame, and this representation determines the vowel's quality. A similar solution was put forward nearly a decade later in the analysis-by-synthesis approach to speech recognition (Stevens, 1960; Halle and Stevens, 1962) based on Jakobson et al.'s (1952) distinctive features. Joos' approach remains distinct, at least in conceptualization of speech recognition as carried out by humans, as he had assumed that "all listeners to the language have been socially trained" to overcome the problem, a critical difference leading to vastly different theoretical consequences (p. 59).

During the 1950s, the problem received greater and more sophisticated scrutiny, and as a result, a richer metaphysics based in large part on Ladefoged and Broadbent's (1957) characterization of different kinds of information present in vowel signals. In addition to the anatomical and physiological information about speaker vocal tracts, covered in each of Lloyd (1890), Chiba and Kajiyama (1941), Joos (1948), and Peterson and Barney (1952), vowel signals also contain information related to a speaker's group memberships, characterized by regional and economic background, as well as information related to different modes of speech that vary in level of formality. The investigation into the second kind of information and how it varies in speech signals generally evolved into the discipline of sociophonetics (see the seminal work in Labov, 1963, 1966, 1972). By the time Hockett gave his 1965 address, the absolute differences in representations of speech signals due to differences in speaker properties such as body size types, age, gender, and other socially interpreted categories that are based on natural variation in vocal tract size and shape coalesced under the concept "speaker variation," and the systematic reduction of the various kinds of speaker variation in vowel signals through the use of transformations over parameter spaces, acoustic or other, had crystallized into the concept “vowel normalization.”

Numerous approaches to vowel normalization were put forward in the following decades, leading to the development of taxonomies. The earliest distinction between approaches resided in which reference frames were involved in the normalization procedure. Chiba and Kajiyama’s (1941) relative formant method involves only an acoustic reference frame and a simple transformation over it, and accordingly may be termed an “acoustic normalization.” Joos’s (1948) approach, however, uses both acoustic and articulatory reference frames, thus falling outside of the acoustic normalization class. Within the class of acoustic normalization procedures, Ainsworth (1975) introduced a distinction between those that require the use of the information from more than one vowel from a speaker, termed “extrinsic methods,” and those that rely only on the information in each individual vowel, termed “intrinsic methods” (see Clopper, 2009, for a modern account). The distinction is not to be conflated with the intrinsic-extrinsic distinction for manifolds described above, especially since we set it aside.

Ladefoged and Broadbent’s (1957) characterization of variation types also led to a kind of taxonomy of normalization procedures based on what kind of variation information they preserve or eliminate, and how well they do so. Hindle (1978) compared and evaluated three approaches to vowel normalization, what he termed “technical answers” to the variation problem, focusing on how well each clustered qualitatively similar vowel signals and separated qualitatively dissimilar vowel signals produced by different speakers, and by how well each approach preserved sociolinguistic variation in this clustering and separating. Nearly four decades later, Adank et al. (2004) carried out a similar evaluation of technical answers, though by this time the number had increased to 12, speaking to the difficulty and complexity of speaker variation.

In preparatory discussion of the technical answers to variation, Hindle broached the "psychological aspect of the normalization problem," asking "[w]hat is a speaker doing when he equates two vowels spoken by different speakers and having different formant values?" (p. 162). Although Hindle set aside this aspect of vowel normalization, it became central to the psychophysical and cognitive study of vowel categorization, prompting investigation of normalization procedures operating over psychophysical and cognitive reference frames. Within the psychophysical approach, the peripheral auditory system is assumed to perform normalization within, for example, a bark-transformed acoustic space (e.g., Ménard et al., 2002) or an auditory reference frame (e.g., Smith et al., 2005), yielding invariant representations for categorization further down the processing line. Within the cognitive approach, it is often assumed that domain-general cognitive computations yield a kind of normalization. For example, exemplar approaches (e.g., Johnson, 1997) attempt to reduce normalization to the storage of variation, while myriad emerging Bayesian approaches reformulate normalization in terms of some optimal computation over a rational listener's beliefs about speaker intentions.

In this dissertation, we take *vowel normalization* to be a cognitive computation "in which interspeaker vowel variability is reduced in order that perceptual vowel identification may then be performed by reference to relative vowel quality rather than absolute [psychophysical] parameters of vowels" (Johnson, 1990a, p. 230). Given that infants appear to be reconciling the absolute differences between their representations of adult vowels and their own by six months of age (Kuhl, 1979, 1983; Kuhl and Meltzoff, 1996), we focus our attention on the acquisition of this computation by an infant during the earliest stages of spoken language acquisition, well before the infant has acquired any words of the ambient language that might guide the determination of what signals are qualitatively similar or different.

Within the last 30 years, researchers have turned to the study of the role of vowel normalization in spoken language acquisition, beginning with early attempts to situate the normalization transformations within the brain. Sussman (1986), for example, attaches a neurological meaning to a variant of the relative formant normalization put forward by Lloyd and Chiba and Kajiyama, positing that a “combination-sensitive two formant field” in the brain computes the ratios F1/F2, F1/F3, and F2/F3 of a given vowel, to recover its invariant information. Sussman (1986) moreover assumes that normalization is “hard-wired” into an infant’s neural circuitry, providing the cognitive means for handling speaker variation. A similar approach to normalization in acquisition is also taken by Callan et al. (2000), Guenther and Vladusich (2012), and Heintz et al. (2009). More elaborate models have been proposed, such as Ishihara et al.’s (2009) probabilistic transformation within formant space, and Ananthakrishnan and Salvi’s (2011) prespecified links between nodes of self-organizing maps. Ames and Grossberg’s (2008) vowel category acquisition model uses a “biologically inspired” normalization computation, carried out by “orthogonal strips” along the auditory cortex, to account for speaker variation as a preliminary step in category acquisition.

The normalizations mentioned above, regardless of their taxonomic status, are prespecified and fixed, and generally unresponsive to distributional aspects of the vowel signals within the reference frames they operate over. They make ontological commitments to an external or universalist, direct transformation interpretation of normalization. Within this “direct transformation” approach, the infant learns a single, pre-specified and fixed transformation. This commitment, however, seems to be in conflict with what is known about the phenomenon from observation and experimentation, e.g., the context-sensitivity in Ladefoged and Broadbent (1957); Johnson (1990a), the cross-language effects in Johnson (2005), and the long-term ontogenetic effects in Kohn and Farrington (2012).

In this dissertation, therefore, we model the acquisition of vowel normalization instead using both psychophysical and cognitive reference frames, and transformations that operate over them. That is, we believe that Hockett's criticism of the Chomsky-Halle approach to distinctive features extends to the aforementioned approaches to vowel normalization. Just as Hockett (1965) dismissed their assumption that "a single fine-grained quantization of acoustic space can yield a finite set of points of reference valid for all languages" (p. 201-2), we dismiss the assumption that a single transformation over such a space can yield a finite set of points of reference for classification of the vowels of all languages. We dismiss the assumption that the transformations involved in normalization are fixed or prespecified, adopting instead an approach that builds on Joos's (1948) assumption that listeners are "socially trained" to carry out vowel normalization, and take the transformations involved in the acquisition of vowel normalization to vary from language to language as a result of the social training infants receive during the earliest stages of life. Our approach assumes an internal, potentially idiosyncratic interpretation of the transformations, which are constructed from an infant's "alignment" of representations of their own vowel productions and those of the other speakers in their environment. Moreover, we believe that Hockett's basis for the criticism extends to the vowel normalization computation as a whole. Just as Hockett recognized that

> "[t]he number of contrasting distinctive features along any one axis [of a parameter space] can and does vary from language to language, and even when

> the number is the same the exact locations are typically different; the quantizations differ from language to language in such a way as to render the units of different languages incommensurate," (p. 202)

in this dissertation, we take the position that the very reference frames within which vowels are represented for categorization vary from language to language, and are constructed during the earliest stages of spoken language acquisition. The position is by no measure new, as Joos (1948) points out:

> "It is easy enough to say that the listener can do something in his brain that is directly equivalent to shifting and distorting one of the acoustic triangles of Fig. [2.1] until it coincides with another of them. But if we think that that process, when carried out upon ACOUSTIC material, is simple and straightforward, we are being misled by a specious appearance. True,...the use of two formants as an abbreviated description of vowel color cannot properly be objected to as artificial, inorganic. But our formant charts go a great way beyond this: THEY PLOT TWO FREQUENCIES CROSSWISE OF EACH OTHER, and for this there can be no justification in the nature of sound as sound, for frequency is one dimension, not two. [The orthogonal arrangement of these dimensions] is a semantic device for facilitating discussion, not a hypothesis about the nature of speech sound. If this...makes it seem easy to slide and twist one acoustic triangle of Fig. [2.1] until it fits another of them, that proves nothing whatever. There may actually be a simple and economical sort of brain activity that does just this, but Fig. [2.1] is not the way to prove it." (pp. 62-3)

In this connection, early results on vowel categorization in infants with respect to such variation remain instructive. Experimental investigation of the perceived similarity of "discriminably different but phonetically identical" (Kuhl, 1983, p. 263) vowels simulated to reflect the spectral differences between the corresponding natural vowel productions of men, women, and children "provided strong support for the notion that 6-month old infants recognize equivalence classes that conform to vowel categories" (p. 281, see also Kuhl, 1979). Yet, while the acoustic characteristics of the vowels used in the experiments are well-understood (/a/ and /i/ in Kuhl, 1979; /a/ and /ɔ/ in Kuhl, 1983), Kuhl (1983) notes that "it is nevertheless difficult to specify the sorting rule [infants] use in acoustic terms” and moreover, “[w]e cannot, in fact, identify the exact nature of the information used by adult listeners in the categorization of vowels produced by different talkers” (p. 281). That is, we cannot assume that the neat physical properties that researchers attribute to acoustic phenomena correlate with any properties that a listener (infant or adult) might make use of in “sorting” auditory internalizations along a particular set of psychophysical or cognitive dimensions. Rather, it seems that phonetic category acquisition involves the selection of such dimensions to use in imposing order on auditory internalizations derived from multiple talkers in some manner facilitating categorization. If so, we cannot assume that the input to some phonetic category mapping is neatly normalized along one or more cue dimensions.

In the remainder of this chapter, we put forward a framework for the investigation of the acquisition of vowel normalization based on the idea that infants map their psychophysical representations of the vowels of individual speakers to mediating cognitive reference frames, guided by vocal imitative interaction with their caretakers, as a first step in phonological acquisition. We proffer a modeling methodology which involves the alignment of the cognitive manifolds, that the infant builds using the psychophysical representations of the vowels of individual speakers (Section 2.2). We then conclude with a formal presentation of the main algorithm involved in implementation of the methodology (Section 2.3).

Our approach to the study of the acquisition of vowel normalization centers on the “minimal requisites” involved in acquisition. We assume that there are two agents involved: an *infant learner*, or simply an *infant*, and an *adult caretaker*, or simply an *adult*. Our further assumptions are of three kinds, those pertaining to the infant, those pertaining to the adult, and those pertaining to the nature of their vocal interaction during the early stages of language acquisition.

Regarding the infant, we assume that the infant is endowed with an auditory reference frame, and the ability to represent vowel signals within this frame. The representations may be derived from vowel signals perceived by the infant, or of a purely mental origin, such as a projection or estimation of a potential vowel signal not necessarily perceived. We also assume that the infant is endowed with an articulatory reference frame, and the ability to represent vowel signals within this reference frame. The representations may be derived from vowel signals produced by the infant, or of a purely mental origin, such as a projection or estimation of a potential vowel signal not necessarily produced by the infant, but say by the adult.

We assume that the infant is able to construct manifolds over representations within these psychophysical reference frames, and furthermore that the infant is able to construct transformations over multiple reference frames using these manifolds. This assumption is, for the most part, in line with the assumption of transformations between reference frames in Guenther (1995), Callan et al. (2000), Guenther and Vladusich (2012), Westermann and Miranda (2002, 2004), and Oudeyer (2002). With this machinery, the infant may also link manifolds within the articulatory and auditory reference frames, effectively bridging representations within the distinct modalities, in line with notions of cross-modal perception (e.g, Davenport, 1976) and corresponding experimental results (e.g., Kuhl and Meltzoff, 1982). Transformations derived from manifolds embedded within distinct psychophysical reference frames are termed "intermodal transformations." The term is used in generality in this chapter, with specific examples formulated in Chapter 4. We assume that intermodal transformations are constructed via *babbling* taken to be vowel signals produced and perceived by the infant, yielding pairs consisting of a representation of the vowel signal in the articulatory reference frame and a representation in the auditory reference frame. This machinery exercises the infant's *internal model* (Wolpert et al., 1995) of vowel production which may be used to refine intermodal transformations through various feedback mechanisms.

We assume that the infant is able to differentiate between the vowel signals produced by the infant, and the vowel signals produced by the adult. Experimental results in speech perception (e.g., Winters et al., 2008) and pathology (e.g., Perrachione et al., 2011) suggest the importance of individual speaker identification, hence the representation of "other" conspecifics such as an adult caretaker, to the formation of more abstract phonetic representations. More general modeling of the cognitive development of intentional agents (e.g., the 'like me' framework Meltzoff, 2007) adduces the need for a representation of "self," as distinct from others. Applying the general model to the domain of spoken language acquisition (as in Howard and Messum, 2011) suggests the importance of the distinction between self and others in vowel normalization. The essential distinction between self and others entails carrying out a normalization computation that "allows the infant to see the behavior of others as commensurate with their own" (Meltzoff, 2007, p. 26). Commensuration computations of this nature have been hypthosized to exist at a general cognitive level, e.g., within the "shared manifold" approach to representation and relations between models of the self and others (Gallese, 2001), based on the potential role or mirror neurons in higher-order cognition. We assume that the infant is able to construct distinct manifolds over representations within a single reference frame that respect this differentiation between self and others. This assumption may be obviated through the use of an auditory reference frame whose representations also encode organism-internal information, e.g., that from bone conduction in addition to signal from the eardrum, that is available only for the representation of self. However, at present, we keep to the assumption of a single general auditory reference frame. We further assume that the infant is able to construct transformations using these manifolds.

We assume that the infant is able to construct reference frames distinct from the endowed psychophysical reference frames, termed “cognitive reference frames.” We similarly assume that an infant may construct manifolds over representations within cognitive reference frames, and furthermore construct transformations using these manifolds. Within our approach, cognitive reference frames serve as the domains within which vowel categorization takes place. These terms are again used in generality in this chapter, with specific examples formulated in Chapters 3 and 4.

Regarding the adult, we assume that the adult is in possession of a “distribution” (roughly in Hockett’s sense) over an acoustic parameter space, in this case a formant space, that indicates the locations of the “best examples” of the vowel categories that constitute the allophones of the adult’s language. We further assume that the adult is a fully functioning, healthy speaker of the language, capable of interpreting and producing all of the vowels of the language. Specifically, given a vowel signal produced by the infant, we assume that the adult is capable of interpreting the vowel signal with respect to the adult’s own vowel categories, assigning a category to the infant’s vowel signal, and moreover, responding with productions of vowel signals from within that category.

Regarding the nature of the vocal interaction between the adult and infant, we assume that the infant is attempting to internalize a modified version of the distribution of the adult. Experimental results suggest that the vocal interactions between the infant and adult involve at the very least: (i) structured turn-taking between the infant and adult (Masataka, 2003), and (ii) adult responses differentiated according to the nature of infant vocalizations (Gros-Louis et al., 2006; Goldstein and Schwade, 2008). These richly-structured individuated

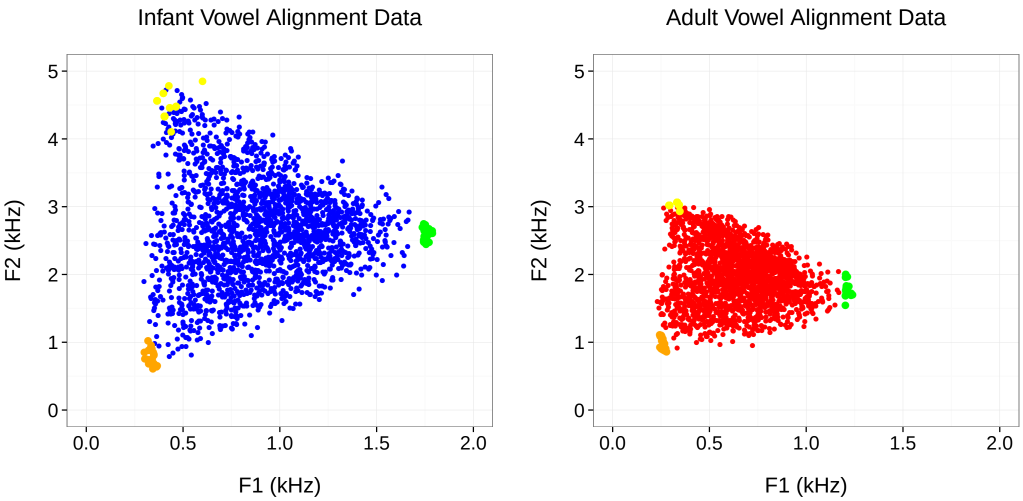


Figure 2.2: The infant vowel data V_I (left) and adult vowel data V_A (right), including "good" examples of the infant and adult caretaker's vowels used for alignment.

instances of vocal interaction provide "evidence for [the child] to deduce a correspondence between his output and the speech sound equivalent within [the mother's] L1 that she produces" (Howard and Messum, 2011, p. 87).

Finally, we assume that during the earliest stage of spoken language acquisition, within a set of reference frames, infants construct manifolds over their representations of their own vowel productions, and those of their adult caretakers. Vowel normalization is the "alignment" of the manifolds constructed by the infant within the reference frames. The manifold alignment is guided by vocal imitative exchanges between the infant and adult caretaker.

We close this section with a brief, simplified example of the modeling approach. Suppose an infant has access to the vowel data V_I, derived from the infant, and V_A, derived from an adult caretaker, within an acoustic reference frame (Figure 2.2). Moreover, suppose the yellow points are, according to the adult caretaker, "good" examples of /i/, the

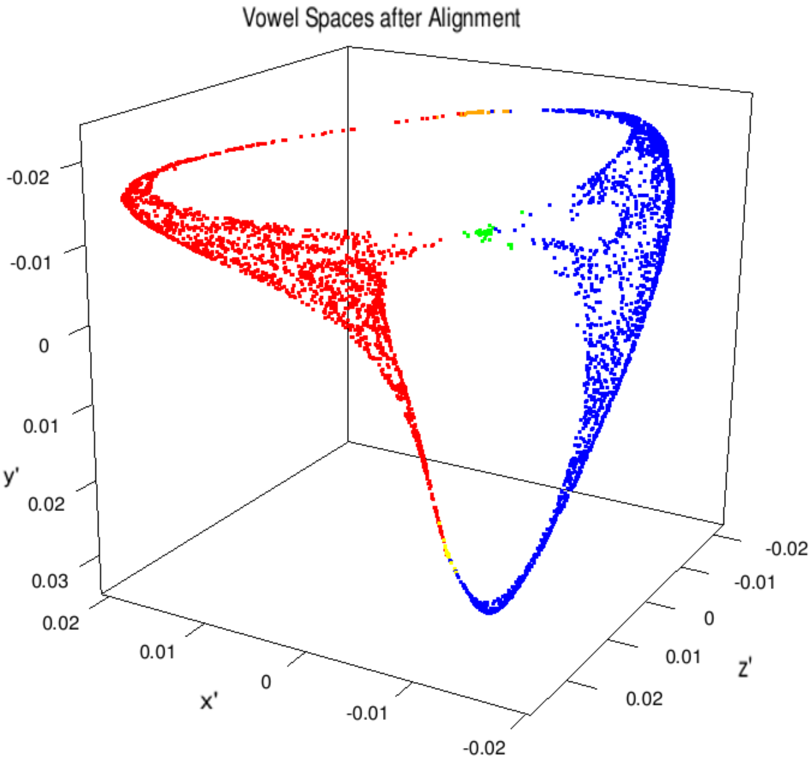


Figure 2.3: Ordered triples in a reference frame in which reflecting the alignment of the adult and infant vowel spaces using the "good" examples of the infant and adult caretaker vowels.

orange points are "good" examples of /u/, and the green points are "good" examples of /a/, approximating a very simple "goodness distribution" over V_A. Assume the infant has produced a vowel signal. The adult interprets the infant's vowel with respect to the goodness distribution over V_A. Suppose the adult interprets the infant's vowel as a good example of a vowel category within the adult's language. The adult may respond with a vowel production that they take to be a good example of that vowel category, along with a corresponding positive gesture, such as a smile or a touch of the infant's stomach. We assume that the infant is able to interpret the level of positivity in the adult's social response, and couple the representation of the adult's vowel signal response with the representation of the vowel signal the infant produced to elicit the response. That is, the vocal interaction in this case results in the infant possessing a pairing of its own vowel signal with the adult's response signal, along with some interpretation of the level of positivity in the adult response. In this fashion, the adult is in some sense imparting their notion of category goodness to the infant by responding in a positive manner to what they perceive to be good examples of the infant's /i/, /u/, and /a/, with good examples of their own /i/, /u/, and /a/. The infants pairing of "good" productions with "good" adult responses guides the alignment of the manifolds the infant constructs over V_I and V_A, respectively, yielding the aligned representations in a constructed cognitive reference frame depicted in Figure 2.3.

2.2 Computational Modeling

In this section, we exposit the main algorithm we use in modeling vowel normalization with some simple examples, introducing the technical concepts and vocabulary along the way. We begin with a quasi-mathematical description, working toward a more formally satisfying formulation in the following section. This section is an elucidation of an algorithm presented in Ham et al. (2005), and generalized in Chapter 5 of Ma and Fu (2012).

Suppose we have two shapes $\mathcal{P}$ and $\mathcal{Q}$ embedded in a space, say the hexagons in Figure 2.4 (left), and we want to devise a method to bring them into alignment in a particular way that is advantageous to us for performance of some task, such as matching up the corners from each shape. In order to do this, we need to specify a language that we can use to describe the shapes themselves, the plane that they are embedded in, and the procedures that constitute our desired method.

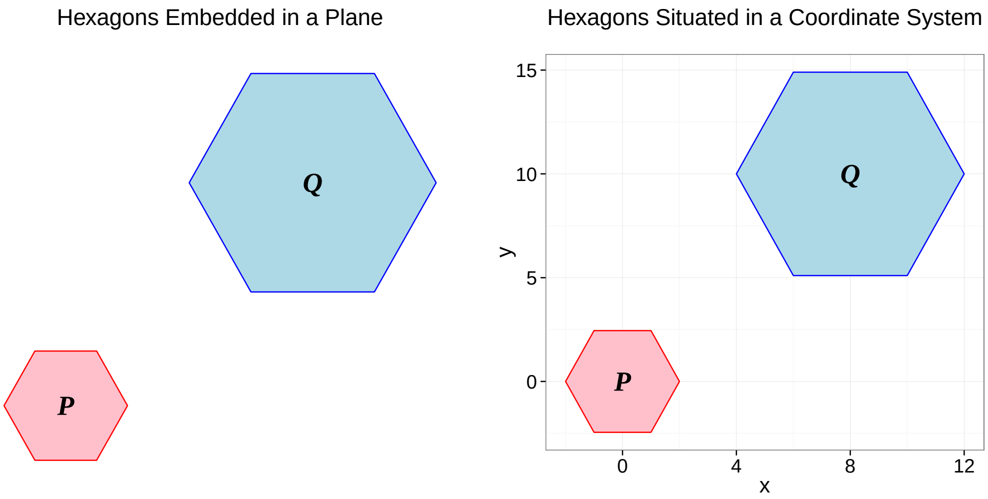


Figure 2.4: (Left) Hexagons $\mathcal{P}$ and $\mathcal{Q}$ embedded in a plane. (Right) Hexagons $\mathcal{P}$ and $\mathcal{Q}$ situated within a coordinate system.

The usual language used to describe the shapes $\mathcal{P}$ and $\mathcal{Q}$, and the plane they are embedding in, is that of a "coordinate system." Each point in the plane is assigned an "ordered pair" of real numbers x and y, typically written (x, y), and the shapes are identified with certain collections of these ordered pairs. For example, the coordinate system in Figure 2.4 (right) assigns the ordered pair $(0, 0)$, the origin, to the center point of $\mathcal{P}$. More broadly, the collection of ordered pairs in the red region constitutes $\mathcal{P}$, while those in the blue region constitute $\mathcal{Q}$. It is important to keep in mind that the shapes themselves have meaning outside of the choice of a coordinate system, and the one we have chosen is one of many, and is somewhat arbitrary.

Since we are interested in matching up the corners of $\mathcal{P}$ and $\mathcal{Q}$, let P and Q denote the collections ordered pairs corresponding to the corner points from $\mathcal{P}$ and $\mathcal{Q}$, respectively. The ordered pairs in P are labeled p_1 through p_6, and those in Q are labeled q_1 through q_6, as depicted in Figure 2.5 (left). We can represent the collection P as a data table (left,

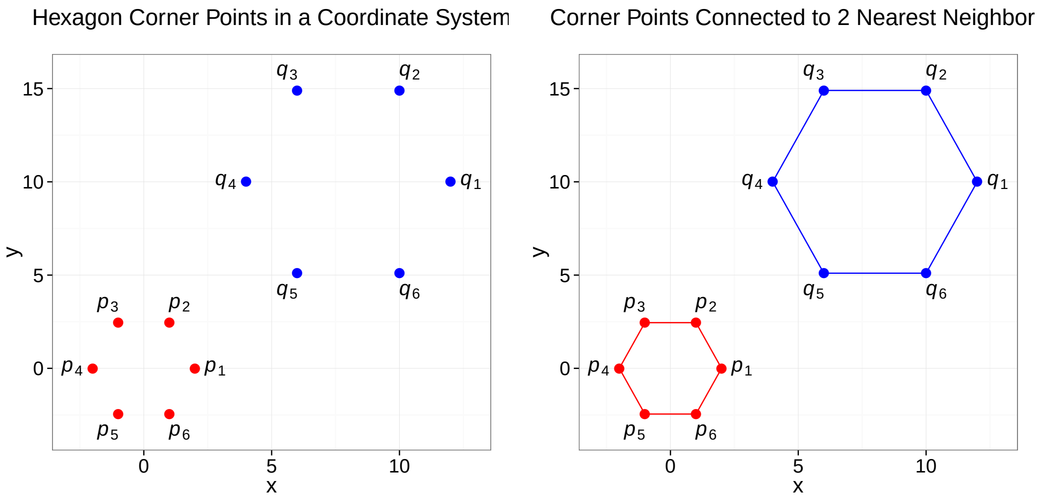


Figure 2.5: (Left) Coordinates are assigned to the corner points of the hexagons $\mathcal{P}$ and $\mathcal{Q}$. (Right) Ordered pairs corresponding to the corners of $\mathcal{P}$ and $\mathcal{Q}$ are depicted as connected to their two nearest neighbors by line segments.

below) whose rows are the ordered pairs p_1 through p_6, and we do the same with Q and q_1 through q_6 (right, below):

	X	Y
p_1	2	0
p_2	1	2.45
p_3	-1	2.45
p_4	-2	0
p_5	-1	-2.45
p_6	1	-2.45

	X	Y
q_1	12	10
q_2	10	14.9
q_3	6	14.9
q_4	4	10
q_5	6	5.1
q_6	10	5.1

Now that we have some numerical data representing $\mathcal{P}$ and $\mathcal{Q}$, we need a method for aligning them that makes use of it. There are a number of different ways to do so, but we will keep to the one we will use in the remainder of this dissertation. We begin as follows. For each corner point, we want to have some information about the other points that are near it in some sense. One way to obtain such information is to take each ordered pair p in P and compute its nearest neighbors in P (not including the trivial nearest neighbor p itself) according to the standard Euclidean distance computation, and similarly, do the same for each ordered pair q in Q with respect to Q. In Figure 2.5, each ordered pair in P is depicted as connected to its two nontrivial nearest neighbors in P, as is each ordered pair in Q. The choice of computing the two nearest neighbors of each point is one of expositional convenience. The choice of how many to compute may be determined empirically or on principle.

The next step is to extract a particular kind of “relation” from this two-nearest-neighbors computation, which indicates that p_1 is related to p_2 and p_6, but not to p_3, p_4, p_5 or p_1 itself, and similarly for the ordered pairs in Q, but does not necessarily depend on the coordinate system. The kind of relation we use is called a “graph,” which we can succinctly summarize with an “adjacency matrix,” constructed as follows. Given that there are six ordered pairs in P, we construct a matrix A_P (left, below) that has six rows and six columns, and populate A_P so that the entry at the ith row and jth column is 1 if p_i is a (nontrivial) nearest neighbor of p_j, and 0 otherwise. Doing the same for Q yields the adjacency matrix A_Q (right, below):

$$A_P = \begin{array}{c} \\ P1 \\ P2 \\ P3 \\ P4 \\ P5 \\ P6 \end{array} \!\!\begin{array}{c} \begin{array}{cccccc} P1 & P2 & P3 & P4 & P5 & P6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \qquad A_Q = \begin{array}{c} \\ Q1 \\ Q2 \\ Q3 \\ Q4 \\ Q5 \\ Q6 \end{array} \!\!\begin{array}{c} \begin{array}{cccccc} Q1 & Q2 & Q3 & Q4 & Q5 & Q6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

The adjacency matrices A_P and A_Q are discrete summarizations of the shapes $\mathcal{P}$ and $\mathcal{Q}$, and they constitute the basis of our alignment method. Each row (and column) in an adjacency matrix corresponds to a “vertex” in the graph it represents, and a nonzero entry at a particular row and column indicates that an “edge” exists between the vertices that correspond to that row and column. Graphs corresponding to the adjacency matrices A_P and A_Q are depicted below in Figure 2.6 (left). The graph corresponding to A_P is composed

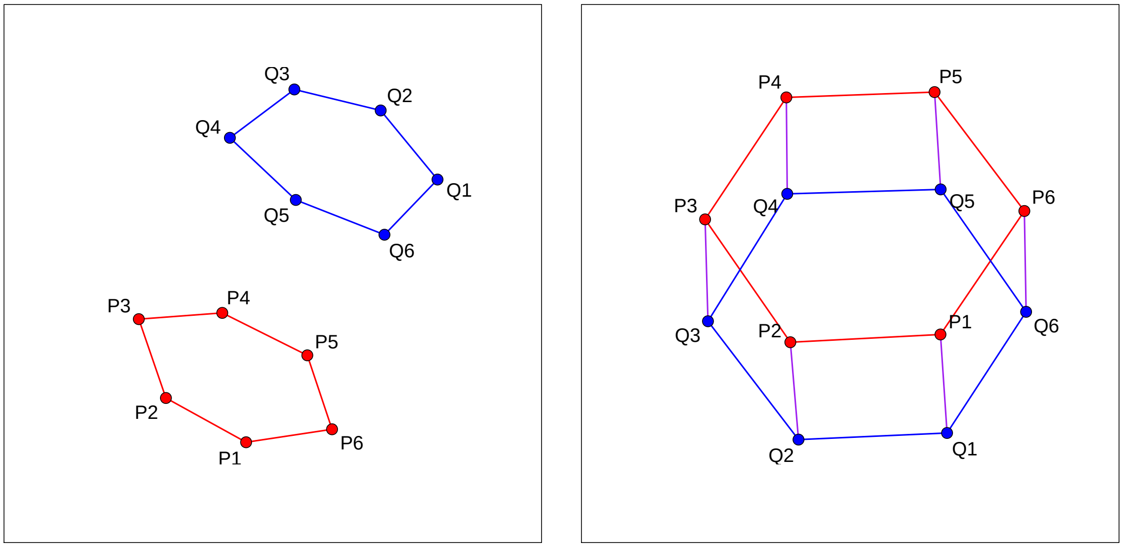


Figure 2.6: (Left) Graphs corresponding to the adjacency matrices A_P and A_Q. (Right) Graph corresponding to the combined adjacency matrix C_{PQ}.

of the six vertices $P1, P2, \ldots, P6$, along with the depicted edges between them, while A_Q is composed of the six vertices $Q1, Q2, \ldots, Q6$, and the depicted edges. Herein, we identify a graph with its adjacency matrix.

The next step is to specify which vertices in A_P we want to match up with which vertices in A_Q. To do this, we construct two "alignment matrices" A_{PQ} and A_{QP} in the following way. Since A_P contains six vertices, we endow A_{PQ} with six rows, and since A_Q also has six vertices, we endow A_{PQ} with six columns. We arrange A_{QP} in transposed fashion, with six rows corresponding to the vertices in A_Q, and six columns corresponding to the six vertices in A_P. We initialize both A_{PQ} and A_{QP} to have all zero entries, and then place a 1 at the ith row and jth column of A_{PQ}, if we want Pi to align with Qj, along with a 1 at the jth row and ith column of A_{QP}. Assuming that we want Pi to align with Qi for

$1 \leq i \leq 6$, we have:

$$A_{PQ} = \begin{array}{c} \\ P1 \\ P2 \\ P3 \\ P4 \\ P5 \\ P6 \end{array} \begin{array}{c} \begin{array}{cccccc} Q1 & Q2 & Q3 & Q4 & Q5 & Q6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \qquad A_{QP} = \begin{array}{c} \\ Q1 \\ Q2 \\ Q3 \\ Q4 \\ Q5 \\ Q6 \end{array} \begin{array}{c} \begin{array}{cccccc} P1 & P2 & P3 & P4 & P5 & P6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

We then combine the alignment matrices A_{PQ} and A_{QP} with the adjacency matrices A_P and A_Q, to form a combined adjacency matrix C_{PQ} as follows:

$$C_{PQ} = \begin{pmatrix} A_P & A_{PQ} \\ A_{QP} & A_Q \end{pmatrix}.$$

The matrix C_{PQ} is a useful summary of the local geometric information about the shapes $\mathcal{P}$ and $\mathcal{Q}$, as well as the information that is used to match up their corners. A graph corresponding to C_{PQ} is depicted in Figure 2.6 (right).

We next take the combined adjacency matrix C_{PQ}, and compute its "graph Laplacian" (see Chung, 1997). The graph Laplacian of an adjacency matrix is useful for approximating functions that depend on the information that the adjacency matrix summarizes. Given the adjacency matrix C_{PQ}, we compute its graph Laplacian as follows: i) create a diagonal matrix D that has 12 rows and 12 columns, one row and column for each row in C_{PQ}, ii) for the i row in C_{PQ}, sum the entries and place the sum as the ith diagonal entry in D, and then iii) subtract C_{PQ} from D. The resulting matrix, denoted L_{PQ} is the graph Laplacian for C_{PQ}:

$$L_{PQ} = \begin{pmatrix} 3 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & -1 & 3 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 3 & -1 & 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & -1 & 3 \end{pmatrix}.$$

The alignment computation involves finding "eigenvectors" of the graph Laplacian L_{PQ}. In this context, an eigenvector **v** of the matrix L_{PQ} is an $n \times 1$ matrix, with at least one nonzero entry, such that there exists a real number λ, called an "eigenvalue" satisfying the matrix multiplication equation $L_{PQ}\mathbf{v} = \lambda\mathbf{v}$. That is, the effect of multiplying **v** on the left by L_{PQ} is the same as multiplying each entry of **v** by λ. Since L_{PQ} has 12 rows and columns, it has 12 eigenvectors, and 12 (not necessarily distinct) eigenvalues. Moreover, since L_{PQ} is a symmetric matrix with real numbers as entries, all of its eigenvalues are nonnegative real numbers. Thus we can order the eigenvalues of a matrix using the familiar "less-than-or-equal-to" ordering.

Now, the eigenvalues of L_{PQ}, in ascending order, are 0, 1, 1, 2, 3, 3, 3, 3, 4, 5, 5, and 6, and their corresponding eigenvectors are arranged into a matrix E_{PQ}, given below (with some suppression):

$$E_{PQ} = \left(\begin{array}{cccccccc}
\scriptstyle 0 & \scriptstyle 1 & \scriptstyle 1 & \scriptstyle 2 & \scriptstyle 3 & \scriptstyle \cdots & \scriptstyle 5 & \scriptstyle 6 \\
-0.29 & -0.41 & 0.00 & -0.29 & 0.00 & \cdots & 0.41 & -0.29 \\
-0.29 & -0.20 & -0.35 & -0.29 & -0.20 & \cdots & -0.20 & 0.29 \\
-0.29 & 0.20 & -0.35 & -0.29 & -0.47 & \cdots & -0.20 & -0.29 \\
-0.29 & 0.41 & 0.00 & -0.29 & 0.16 & \cdots & 0.41 & 0.29 \\
-0.29 & 0.20 & 0.35 & -0.29 & 0.47 & \cdots & -0.20 & -0.29 \\
-0.29 & -0.20 & 0.35 & -0.29 & 0.05 & \cdots & -0.20 & 0.29 \\
-0.29 & -0.41 & 0.00 & 0.29 & 0.16 & \cdots & -0.41 & 0.29 \\
-0.29 & -0.20 & -0.35 & 0.29 & 0.47 & \cdots & 0.20 & -0.29 \\
-0.29 & 0.20 & -0.35 & 0.29 & 0.05 & \cdots & 0.20 & 0.29 \\
-0.29 & 0.41 & 0.00 & 0.29 & 0.00 & \cdots & -0.41 & -0.29 \\
-0.29 & 0.20 & 0.35 & 0.29 & -0.20 & \cdots & 0.20 & 0.29 \\
-0.29 & -0.20 & 0.35 & 0.29 & -0.47 & \cdots & 0.20 & -0.29
\end{array}\right).$$

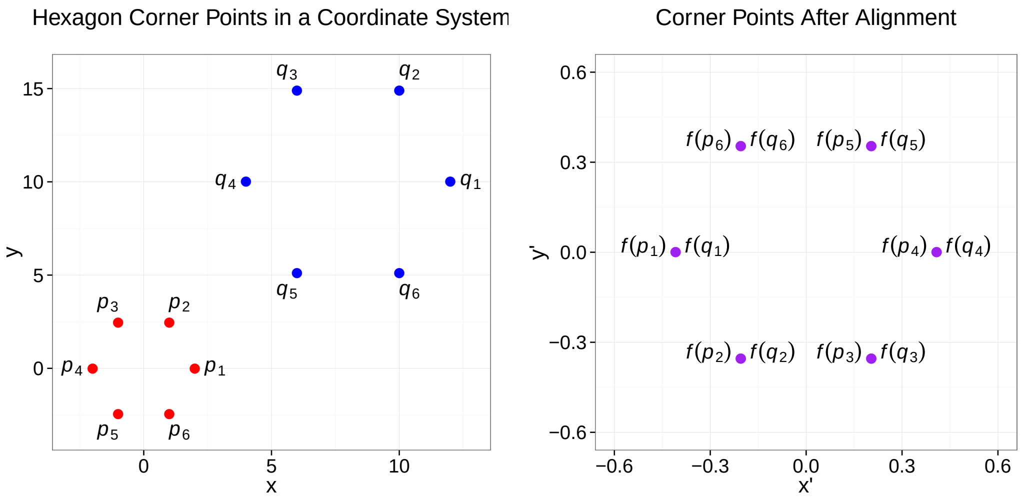
The first column of E_{PQ} consists of a trivial eigenvector corresponding to a zero eigenvalue. The second and third columns, however, contain nontrivial eigenvectors. The first six entries of the second and third columns constitute ordered pairs for the corners of the shape $\mathcal{P}$ in a new "reference frame" or way of viewing the shapes. Similarly, the last six entries in the second and third columns constitute ordered pairs for the corners of the shape $\mathcal{Q}$ in the new reference frame. We can associate the ordered pairs in P and Q with the first and last six entries, respectively, in the second and third columns of E_{PQ}:

		X′	Y′
p_1	$\mapsto$	-0.41	0.00
p_2	$\mapsto$	-0.20	-0.35
p_3	$\mapsto$	0.20	-0.35
p_4	$\mapsto$	0.41	0.00
p_5	$\mapsto$	0.20	0.35
p_6	$\mapsto$	-0.20	0.35

		X′	Y′
q_1	$\mapsto$	-0.41	0.00
q_2	$\mapsto$	-0.20	-0.35
q_3	$\mapsto$	0.20	-0.35
q_4	$\mapsto$	0.41	0.00
q_5	$\mapsto$	0.20	0.35
q_6	$\mapsto$	-0.20	0.35

The original coordinate system for the ordered pairs corresponding to the corner points of $\mathcal{P}$ and $\mathcal{Q}$ (left) and the ordered pairs yielded by the alignment association (right), denoted by f, are visualized in Figure 2.7. The association f is called a *transformation*.

Before moving on, we summarize the alignment procedure we just carried out:


Hexagon Corner Points in a Coordinate System

Corner Points After Alignment

Figure 2.7: (Left) The ordered pairs for the corners of $\mathcal{P}$ and $\mathcal{Q}$ in the original coordinate system. (Right) The ordered pairs yielded by the alignment association f for the corner points of $\mathcal{P}$ and $\mathcal{Q}$.

1. Given two hexagons $\mathcal{P}$ and $\mathcal{Q}$ embedded in a plane, we situated them within a coordinate system.
2. We selected some salient parts of the hexagons, i.e., their corners, and collected together the ordered pairs corresponding to the corners of $\mathcal{P}$ into a set P, and those for $\mathcal{Q}$ into a set Q.
3. We obtained some local geometric information about the corner points of $\mathcal{P}$ and $\mathcal{Q}$ by computing the (nontrivial) nearest neighbors of their corresponding ordered pairs in P and Q respectively.
4. We created graphs A_P and A_Q, represented as adjacency matrices, from the nearest neighbor computations, which discretely summarize the local geometric information about the corners of $\mathcal{P}$ and $\mathcal{Q}$.

5. We then specified how to align $\mathcal{P}$ and $\mathcal{Q}$ by specifying which vertices to connect between A_P and A_Q, using alignment matrices A_{PQ} and A_{QP}, and creating a combined adjacency matrix C_{PQ}.
6. We then computed the graph Laplacian of C_{PQ}, computed its eigenvectors, and used the entries of the eigenvectors corresponding to the second and third smallest eigenvalues to assign coordinates in a new reference frame to the corners of $\mathcal{P}$ and $\mathcal{Q}$, which coincided as desired.

We next present a more sophisticated example of the above procedure that involves i) two different kinds of shapes, ii) much more data corresponding to different areas of the shapes, iii) aligning more than the corners of shapes, iv) constructing alignment matrices that specify how to align only a portion of the data, and v) weighting the alignment matrices to reflect the importance of the alignment. This next example is the first step in a progressive demonstration of the flexibility and generality of this procedure that will take shape over the course of the dissertation.

Suppose we have a rectangle $\mathcal{R}$ and a square $\mathcal{S}$ embedded in a plane, as shown in Figure 2.8 (left), and we want to bring them into alignment in some fashion similar to what we did above. Specifically, suppose we want to match up their corners, and, additionally, suppose we also want to match up some portion of their middle regions.

We begin by situating $\mathcal{R}$ and a square $\mathcal{S}$ is a coordinate system, as shown in Figure 2.8 (right). Within this coordinate system, we can easily and completely describe $\mathcal{R}$ and $\mathcal{S}$ using sets of ordered pairs. However, these sets are infinite, and we need finite sets for our computational procedure. Rather than focusing exclusively on a small set of ordered pairs corresponding to salient points, such as the corners of $\mathcal{R}$ and $\mathcal{S}$, we can select a large (yet

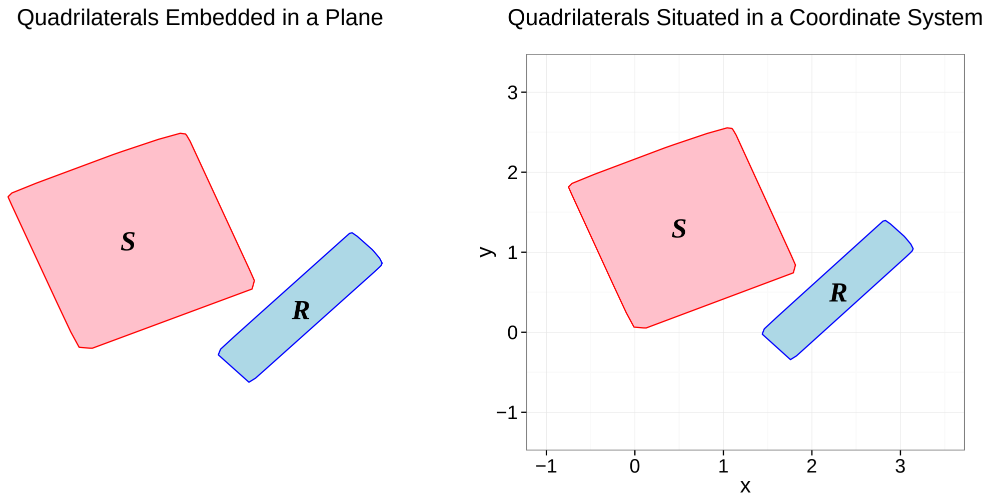


Figure 2.8: A rectangle $\mathcal{R}$ and square $\mathcal{S}$ embedded in a plane (left), and situated within a coordinate system (right).

finite) set of ordered pairs corresponding to many points within them to represent them in our alignment.

Let R and S be the sets of ordered pairs representing points on $\mathcal{R}$ and $\mathcal{S}$, respectively, each of which contains 3000 ordered pairs (as depicted in Figure 2.9, left). We proceed as before, obtaining local geometric information about $\mathcal{R}$ and $\mathcal{S}$ by the (nontrivial) nearest neighbors of each ordered pair in R with respect to the ordered pairs in R, and similarly for each ordered pair in S with respect to the pairs in S, in each case computing 20 nearest neighbors. Based on this computation, we obtain graphs A_R and A_S, that discretely summarize the local geometric information about $\mathcal{R}$ and $\mathcal{S}$, respectively.

Now, we are still interested in matching up the corners of $\mathcal{R}$ with the corners of $\mathcal{S}$, as well as their middle regions. Since this involves more data than we had in the hexagon example above, we use a slightly more technical description. The “cartesian product” of the two sets R and S is the set of all ordered pairs of the form (r, s), where r is in R and s

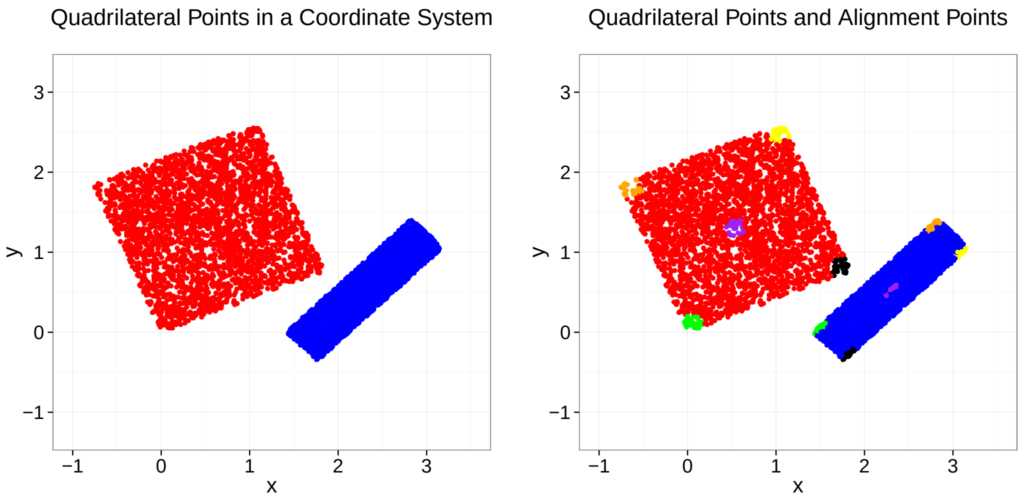


Figure 2.9: (Left) Coordinates are assigned to points of the rectangle $\mathcal{R}$ and square $\mathcal{S}$. (Right) Ordered pairs corresponding to the corners and middle of $\mathcal{R}$ and $\mathcal{S}$ are depicted by the different colors. Ordered pairs of the same color compose the alignment pairs in $\chi_{R\times S}$.

is in S, and is denoted $R \times S$. A "subset" of ordered pairs from $R \times S$, is a set such that all of its members are in $R \times S$. We select a subset of $R \times S$, denoted $\chi_{R\times S}$, that specifies which ordered pairs from R match up with which in S. Figure 2.9 (right) depicts the set $\chi_{R\times S}$. For example, each orange ordered pair in R is matched up with an orange ordered pair in S, and the alignment pair they form is an element in $\chi_{R\times S}$. The same holds for each color depicted. In this particular example, there are 30 orange alignment pairs, as well as 29 green, 31 yellow, 30 black, and 36 purple alignment pairs.

We use $\chi_{R\times S}$ to create our alignment matrices A_{RS} and A_{SR} in the following way: for every ordered pair (r, s) in $\chi_{R\times S}$, place a 1 in the row and column entry in A_{RS} corresponding to r and s, respectively, and place a 1 in the row and column entry in A_{SR} corresponding to s and r, respectively. Now, before combining A_{RS} and A_{SR} with A_R and A_S, we can emphasize (or de-emphasize) the importance of the alignment by weighting the alignment

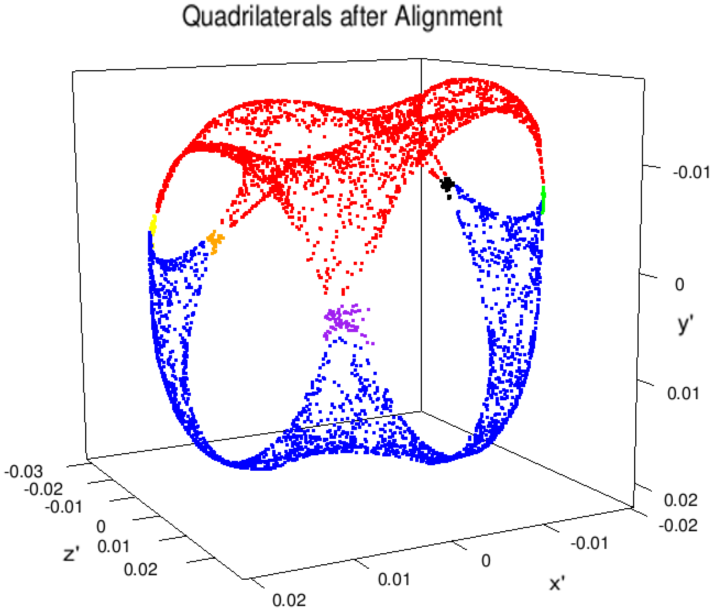


Figure 2.10: Ordered triples in a reference frame in reflecting the alignment of the quadrilaterals $\mathcal{R}$ and $\mathcal{S}$.

matrices. This amounts to multiplying A_{RS} and A_{SR} by a nonnegative real number, say μ. Setting μ to zero results in total de-emphasis of the alignment, while setting μ to a number between zero and one diminishes the importance, and setting μ to any number greater than one inflates the importance. The choice of μ can be determined empirically or on principle. In this case, we set $\mu = 20$ for expositional convenience, and construct the combined adjacency matrix:

$$C_{RS} = \begin{pmatrix} A_R & \mu A_{RS} \\ \mu A_{SR} & A_S \end{pmatrix}.$$

We then compute the graph Laplacian L_{RS} of C_{RS} and its eigenvectors, arranging them from left to right in columns of a matrix E_{RS}, in accordance with the ordering on the eigenvalues of L_{RS} from least to greatest. Discarding the first column of E_{RS} which corresponds to a zero eigenvalue, the rows of the second, third, and fourth columns of E_{RS} constitute

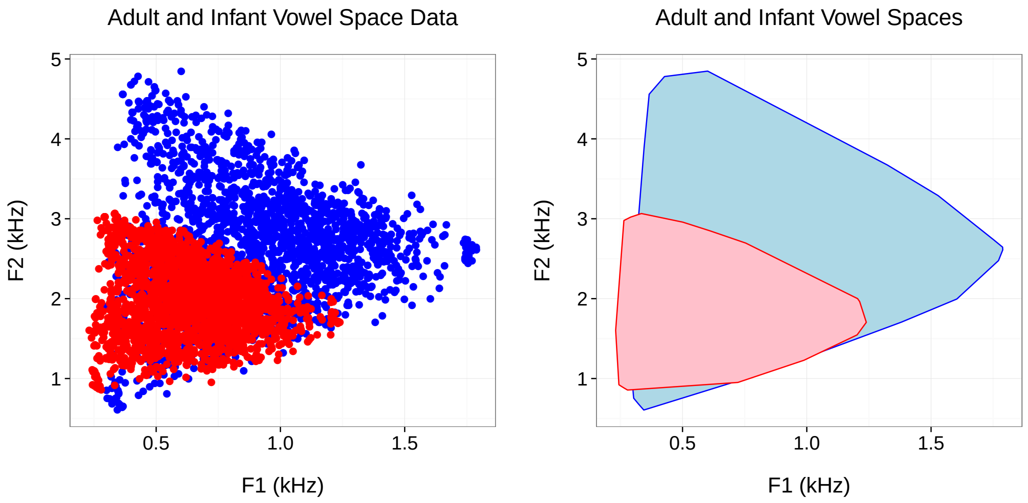


Figure 2.11: (Left) Ordered pairs assumed to be sampled from the adult (red) and infant (blue) vowel spaces. (Right) Convex hulls approximating the shapes of the vowel spaces of an adult (red) and an infant (blue), embedded in a plane.

ordered triples in a new reference frame for $\mathcal{R}$ and $\mathcal{S}$, with the first 3000 rows representing $\mathcal{R}$ and the last 3000 representing $\mathcal{S}$. The ordered triples are depicted in Figure 2.10. Notice that the corners and middle portions match up as desired.

We close this section with a brief example applying the alignment procedure to vowel normalization within an acoustic reference frame. There are two key differences between this application and the ones above. The first is that we do not know the shapes that are involved and assumptions must be made about them. The second is that the alignment matrices are not constructed using ordered pairs provided by an oracle, but rather a different source involving interaction between two agents.

Suppose we are given the two sets of vowel data depicted in Figure 2.11 (left), both of which are produced by the VLAM articulatory synthesizer (see Section 4.2), one of which corresponds to an infant articulatory system (blue) and the other to an adult articulatory

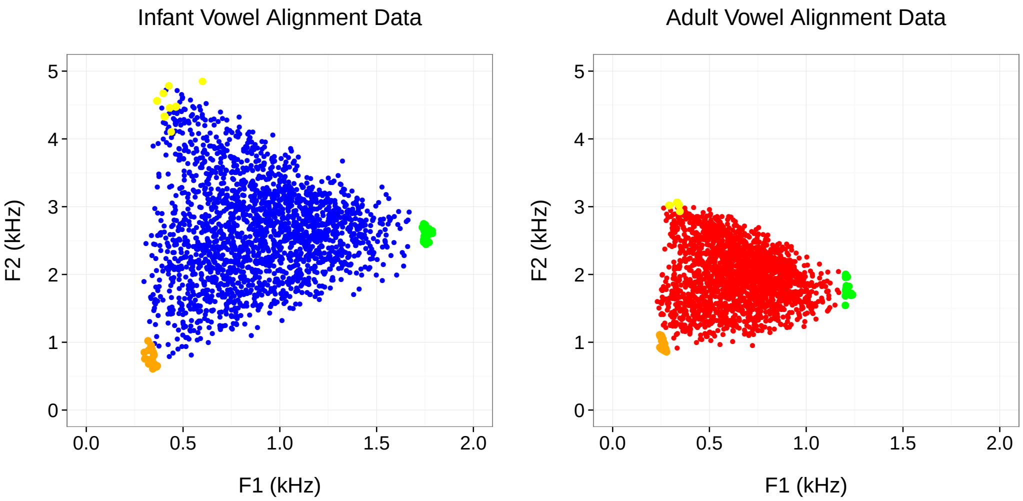


Figure 2.12: The infant vowel data V_I (left) and adult vowel data V_A (right), including “good” examples of the infant and adult caretaker’s vowels used for alignment.

system (red). The coordinate system is chosen so that the abscissa of an ordered pair represents a value for the first formant frequency of a vowel signal, and the ordinate the second formant frequency. Importantly, we are assuming that the vowel data correspond to points on shapes, typically called “vowel spaces,” and furthermore that these shapes are embedded within a reference frame, which we are taking to be a plane. Corresponding to our conceptualization of the coordinate system, we are assuming that the reference frame is acoustic.

Let V_A and V_I denote the set of adult and infant vowel data. In this case, each set has 2060 ordered pairs. As usual, we construct adjacency matrices A_{V_A} and A_{V_I}, using a 20 nearest neighbors computation. At this point, we need to construct alignment matrices that indicate how to match up the vowel spaces represented by V_A and V_I. Importantly, we are assuming that the alignment is provided through interaction between the two agents whose vowel spaces are involved in the computation. In this simplified example, we are

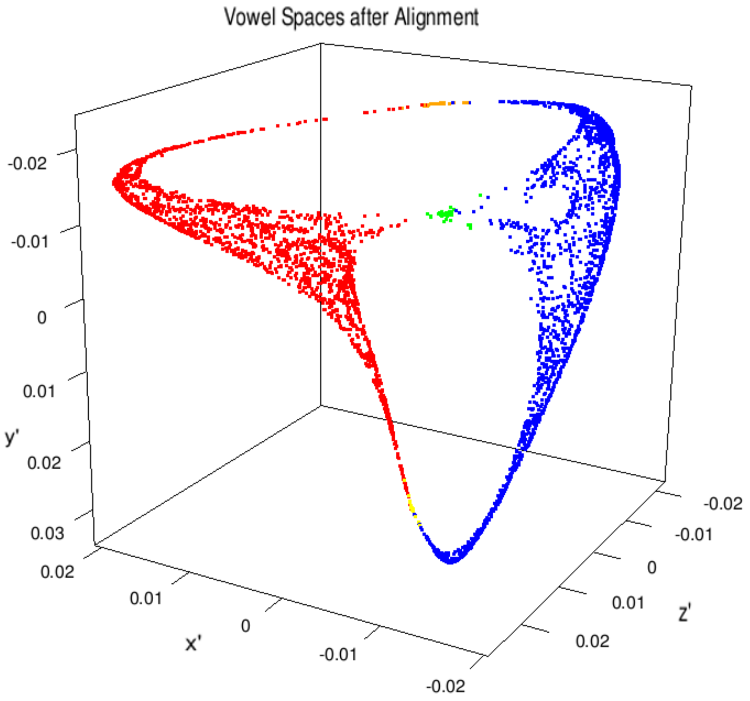


Figure 2.13: Ordered triples in a reference frame reflecting the alignment of the adult and infant vowel spaces.

assuming that the adult vowel space has three main salient regions, the first is a region roughly corresponding to the vowel /i/, the second corresponding to the vowel /u/, and the third corresponding to the vowel /a/. Moreover, we are assuming that within these regions there are subregions that the adult agent associates with "good" examples of that vowel. For example, in Figure 2.12 (right) the 20 yellow ordered pairs are "good" examples of /i/, the 20 orange ordered pairs are "good" examples of /u/, and the 20 green pairs are "good" examples of /a/.

We are further assuming that the adult agent assigns an interpretation of goodness to the infant vowel signals, and that the adult's interpretation of the infant vowel space has three main salient regions, again corresponding roughly to the vowel /i/, /u/, and /a/.

Moreover, we are assuming that within these regions there are subregions that the adult agent associates with “good” examples of that vowel. Accordingly, in Figure 2.12 (left) the 20 yellow ordered pairs are “good” examples of the infant’s /i/, the 20 orange ordered pairs are “good” examples of /u/, and the 20 green pairs are “good” examples of /a/. Finally, we are assuming that the adult is in some sense imparting their goodness regions to the infant by responding in a positive manner to what they perceive to be good examples of the infant’s /i/, /u/, and /a/, with good examples of their own /i/, /u/, and /a/. The good infant vowels, together with the good adult responses constitute the set of alignment pairs. Assuming that the infant produces vowels corresponding to the yellow, orange, and green ordered pairs in Figure 2.12 (right), and the adult considers them good, and responds with the yellow, orange, and green ordered pairs in Figure 2.12 (left), these pairs are used to construct alignment matrices $A_{V_AV_I}$ and $A_{V_IV_A}$.

We are identifying the level of positivity in the adult’s response with the weighting of the alignment matrices. Assuming that the adult responds with positivity level μ_{pos}, we have the combined adjacency matrix

$$C_{V_AV_I} = \begin{pmatrix} A_{V_A} & \mu_{pos}A_{V_AV_I} \\ \mu_{pos}A_{V_IV_A} & A_{V_I} \end{pmatrix}.$$

Taking the first three nontrivial eigenvectors of the graph Laplacian of $C_{V_AV_I}$ to be the ordered triples of the vowel data in a new reference frame, we have the alignment desired by the adult agent, as show in Figure 2.13.

2.3 Mathematical Formulation

In this section, we provide the mathematical formulation of the procedure described in the previous section. A list of basic terms and concepts are provided in Appendix C. For the analytic justification of the algorithms the reader is directed to Rosenberg (1997), Belkin and Niyogi (2003), Ham et al. (2005), and Ma and Fu (2012).

A *reference frame* is simply a set X, called a *carrier set*, together with a set of relations involving the elements of X, called *structures*. In this dissertation, we focus on reference frames that are real coordinate spaces endowed with the usual inner product and its corresponding norm and metric. In the remainder of this dissertation, we call a reference frame taken to be $\mathbb{R}^n$ a *reference frame of dimensionality n*. Thus a reference frame R of dimensionality n contains the carrier set of ordered n-tuples of real numbers, and the structures inherited from the vector space, i.e., vector addition and scalar multiplication, and the dot product $\cdot$ defined as

$$(\mathbf{x}, \mathbf{y}) \mapsto \sum_{i=1}^{n} x_i y_i,$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are in $\mathbb{R}^n$. The dot product is used to define the *norm of* $\boldsymbol{x}$, denoted $||\mathbf{x}||$ to be the non-negative square root of $\mathbf{x} \cdot \mathbf{x}$, which is then used to define the distance between vectors $\mathbf{x}$ and $\mathbf{y}$ as

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Let R be a reference frame of dimensionality n, and let X be a finite subset of the carrier set $\mathbb{R}^n$ of cardinality n_X. A *vertex set representing X* is a set of vertices, denoted $V(X)$, that is in bijective correspondence with the set X. A *neighborhood relation derived from X* is an irreflexive binary relation on $V(X)$, denoted $N(X)$. An *adjacency relation derived from X*, denoted $E(X)$, is the symmetric closure of a neighborhood relation derived from X. An ordered pair composed of a vertex set representing X and an adjacency relation derived from X is called a *graph derived from* X. Let $G(X) = (V(X), E(X))$ be a graph derived from X. We represent $G(X)$ in the following way. Let I_X be a bijection that maps each vertex in $V(X)$ to a natural number in the set $\{1, 2, \ldots, n_X\}$, and define the *indexed adjacency relation* $I_X(E(X))$ as the relation composed of all and only those elements of the form $(I_X(v), I_X(u))$ where $(v, u) \in E(X)$. We then define the $n_X \times n_X$ *adjacency matrix for* $G(X)$ as

$$A_X(i, j) =_{def} \begin{cases} 1 & \text{if } (i, j) \in I_X(E(V)); \\ 0 & \text{otherwise.} \end{cases}$$

The definitions given above generalize the manner in which we obtained graph representations of data in the previous section. Recall that $P \subseteq \mathbb{R}^2$ is a set of ordered pairs in a reference frame of dimensionality 2. The set $\{P1, P2, P3, P4, P5, P6\}$ is a vertex set representing P, which we now denote $V(P)$. The 2-nearest-neighbors computation over P was used to produce a binary relation over $V(P)$, which is irreflexive and hence a neighborhood relation derived from P, now denoted $N(P)$. As it happens, $N(P)$ is its own symmetric closure, and so $N(P)$ is an adjacency relation. This is due to the fact that the adjacency matrix A_P constructed from the 2-nearest neighbors computation was itself symmetric. The graph $G(P) = (V(P), E(P))$ derived from P is depicted in red in Figure 2.6 (left).

Let $E(X)$ be an adjacency relation derived from X. A *weighting function over E(X)* is a function from $E(X)$ to the set of nonnegative real numbers $\mathbb{R}_+$ such that $w(v, u) = w(u, v)$ for each $(v, u) \in E(X)$. An ordered triple composed of a vertex set representing X, an adjacency relation derived from X, and a weighting function over $E(X)$ is called a *weighted graph derived from* X. Let $G(X) = (V(X), E(X), w)$ be a weighted graph derived from X, and let $I_X(E(X))$ be the indexed adjacency relation over $E(X)$. We define the $n_X \times n_X$ *weighted adjacency matrix for* $G(X)$ as

$$W_X(i, j) =_{def} \begin{cases} w(i, j) & \text{if } (i, j) \in I_X(E(V)); \\ 0 & \text{otherwise.} \end{cases}$$

Every graph $G(X) = (V(X), E(X))$ derived from X is trivially a weighted graph derived from X using the weighting function that maps each element of $E(X)$ to 1. Thus every adjacency matrix for $G(X)$ is trivially a weighted adjacency matrix for $G(X)$. Indeed, the adjacency matrix A_P is a weighted adjacency matrix for the graph $G(P)$.

Let $G(X)$ be a weighted graph derived from X. Let $D_{G(X)}$ be the $n_X \times n_X$ diagonal matrix whose ith diagonal entry is the row sum $\sum_{j=1}^{n} W_X(i,j)$. The *graph Laplacian of* $G(X)$ (see Chung, 1997), denoted $L_{G(X)}$, or simply L_X, is the matrix

$$L_X =_{def} D_{G(X)} - W_X.$$

The graph Laplacian is a discrete approximation of the Laplace-Beltrami operator on a Riemannian manifold (see Rosenberg, 1997). Let $spec(L_X)$ denote the (multi)set of eigenvalues of L_X, called the *spectrum* of L_X. Since L_X is a positive semidefinite matrix, the n_X (possibly nondistinct) eigenvalues in its spectrum are all nonnegative. We can order the eigenvalues in $spec(L_X)$ in the following way:

$$0 \leq \lambda_1, \leq \lambda_2 \leq \cdots \leq \lambda_{n_X}.$$

Let $\mathbf{h}_i$ be the eigenvector corresponding to the eigenvalue λ_i, and let λ_ℓ be the smallest nonzero eigenvalue in $spec(L_{XY})$. An *m-dimensional eigenmap derived from G(X)* is an $n_X \ \times \ m$ matrix $E_X =_{def} (\mathbf{h}_\ell \ \mathbf{h}_{\ell+1} \ \ldots \ \mathbf{h}_{\ell+m})$.

Recall that R is a reference frame of dimensionality n, and that X is a finite subset of the carrier set $\mathbb{R}^n$ of cardinality n_X. Let $G(X) = (V(X), E(X), w_{E(X)})$ be a weighted graph derived from X with weighted adjacency matrix W_X. Now, let R' be a reference frame of dimensionality k (which may or may not be identical to R), and let Y be a finite subset of the carrier set $\mathbb{R}^k$ of cardinality n_Y. Moreover, let $G(Y) = (V(Y), E(Y), w_{E(Y)})$ be a weighted graph derived from Y with weighted adjacency matrix W_Y. An *alignment* *for X and Y* is a bijection between a set $X' \subseteq X$ and a set $Y' \subseteq Y$. Let v_X be the bijection mapping X to $V(X)$, and v_Y the bijection mapping Y to $V(Y)$. Given an alignment a for X and Y, an *alignment relation derived from X and Y*, denoted $A(X,Y)$, is the symmetric closure of the relation whose elements are all and only those ordered pairs $(v_X(\mathbf{x}), v_Y(\mathbf{y}))$ where $(\mathbf{x}, \mathbf{y}) \in a$. It is easy to see that $A(X,Y)$ is irreflexive, and is thus an adjacency relation derived from the disjoint union of Y and Y.

Let $w_{A(X,Y)}$ be a weighting function over $A(X,Y)$. A *combined weighted graph derived from X and Y* is the ordered triple $G(X,Y) = (V(X,Y), E(X,Y), w_{E(X,Y)})$ where $V(X,Y) = V(X) \cup V(Y)$, $E(X,Y) = E(X) \cup E(Y) \cup A(X,Y)$ and $w_{E(X,Y)} = w_{E(X)} \cup w_{E(Y)} \cup w_{A(X,Y)}$. We can view $G(X,Y)$ as a weighted graph derived from $X \cup Y$, and thus we can construct an $(n_X + n_Y) \times (n_X + n_Y)$ adjacency matrix A_{XY} and weighted adjacency matrix W_{XY} for $G(X,Y)$ in the manner already described. Without loss of generality, we assume that the indexing of the adjacency relation $E(X,Y)$ assigns the vertices of $V(X)$ to the first n_X rows and columns, and the vertices of $V(Y)$ to the last n_Y rows and columns, of A_{XY} and W_{XY}.

The definitions given above generalize the manner in which we obtained the combined adjacency matrix representations of data in the previous section. By matching up Pi with Qi ($1 \leq i \leq 6$), and constructing the alignment matrix A_{PQ} and its transpose A_{QP}, we constructed an alignment relation $A(X,Y)$. By combining the adjacency matrices A_P and A_Q with the alignment matrices to form the combined adjacency matrix C_PQ, we constructed a weighted graph $G(P,Q) = (V(P,Q), E(P,Q), w_{A(P,Q)})$ derived from P and Q, where $w_{A(P,Q)}$ is the weighting function that maps all arguments to 1. The graph $G(P,Q)$ is depicted in Figure 2.6 (right).

Let $G(X,Y) = (V(X,Y), E(X,Y), w_{A(X,Y)})$ be a weighted graph derived from X and Y with weighted adjacency matrix W_{XY}. Let L_{XY} be the graph Laplacian of $G(X,Y)$, and let $\mathbf{h}_i$ be the eigenvector corresponding to the eigenvalue λ_i, and let λ_ℓ be the smallest nonzero eigenvalue in $spec(L_{XY})$. An *m-dimensional eigenmap derived from G(X,Y)* is an $(n_X \times n_Y) \times m$ matrix $E_{XY} =_{def} (\mathbf{h}_\ell\ \mathbf{h}_{\ell+1} \ldots \mathbf{h}_{\ell+m})$. An *m-dimensional eigenmap of X with respect to* E_{XY} is the matrix composed of the first n_X rows of E_{XY}. Similarly, an *m-dimensional eigenmap of Y with respect to* E_{XY} is the matrix composed of the last n_Y rows of E_{XY}. These definitions are direct abstractions of the alignment computations over the data in the previous section (see Belkin and Niyogi, 2003; Ham et al., 2005).

Recall that R is a reference frame of dimensionality n, and that X is a finite subset of the carrier set $\mathbb{R}^n$ of cardinality n_X. Let R_m be a reference frame of dimensionality m. A *transformation from R to* R_m *with respect to X* is simply a function from X to R_m. An m-dimensional eigenmap derived from $G(X)$ is thus a transformation from R to R_m with respect to X. Recall that R' is a reference frame of dimensionality k (which may or may not be identical to R), and Y is a finite subset of the carrier set $\mathbb{R}^k$ of cardinality n_Y. A *transformation from R and* R' *to* R_m *with respect to X and Y* is a function from the disjoint union of X and Y to R_m. An m-dimensional eigenmap derived from $G(X,Y)$ is thus a transformation from R and R' to R_m with respect to X and Y. Moreover, an m-dimensional eigenmap of X with respect to E_{XY} is a transformation from R to R_m with respect to X, and an m-dimensional eigenmap of Y with respect to E_{XY} is a transformation from R' to R_m with respect to X. These defintions are again direct abstractions over the concepts discussed in the previous sections. Examples are constructed in Chapters 3 and 4.

CHAPTER 3: VOCAL LEARNING

In this chapter, we lay out a conceptual foundation for an approach to vocal learning and its relation to the acquisition of vowel normalization. The basis for our exposition is an article reviewing parallels between birdsong and human speech, Doupe and Kuhl (1999) bring together "work in developmental biology, ethology, linguistics, cognitive psychology, and computer science, as well as work in neuroscience" for "critical assessment of the hypothesis" that "the acquisition of song in birds provide[s] insights regarding learning of speech in humans" (p. 568). We begin by recounting aspects of the review that aid in defining and delimiting vocal learning and inform our modeling of the acquisition of vowel normalization. In particular, we focus on characterizing the "experience" relevant to vocal learning and acquisition – a nontrivial concept often oversimplified. In this chapter, we restrict our view to "auditory vowel normalization" – an intramodal progenitor of a more general normalization computation described in Chapter 4.

We then take a sharp break from the standard conceptualization of vowel normalization in favor of one in which the computation is taken to be a "generative" operation over objects called manifolds. To adduce the generative conceptualization of vowel normalization, we present a framework for investigating its acquisition, which involves (i) a methodology for modeling caretaker interpretations of infant vocalizations (Section 3.2), and (ii) a methodology for modeling the internalization of a relationship between an infant's interpretation of their own vocalizations and differentiated caretaker responses to those vocalizations (Section 3.3). The methodology is expressed through the creation of a vocal learning environment consisting of caretaker and infant agents constructed from cross-linguistic perceptual categorization results. The potential of the approach is demonstrated via its computational structural output (Section 3.4).

3.1 Comparative Aspects of Vocal Learning

The ontogenetic and phylogenetic depths of social signaling among conspecifics and its impact on their means of signal interpretation should not be underestimated. It is now known that even the brainless and nuclei-envious prokaryotes are in possession of complex intraspecies communication systems enabling cell-to-cell signaling that ranges over entire communities (see Miller and Bassler, 2001; Waters and Bassler, 2005), with astonishing consequences. During the phenomenon termed "quorum-sensing," bacteria

> produce and release chemical signal molecules termed autoinducers whose external concentration increases as a function of increasing cell-population density. Bacteria detect the accumulation of a minimal threshold stimulatory concentration of these autoinducers and alter gene expression, and therefore behavior, in response. (Waters and Bassler, 2005, p. 320)

That is, the degree of social signaling among communities of these organisms significantly affects their biological state, even to the point of altering gene expression.

In this connection, although seemingly a truism, it nevertheless needs to be stated that all organisms are faced with the task of sorting out the vast amount of variation they sense in their external environments, especially that pertaining to the signals of other organisms, both non- and con-specific. To illustrate the point, quorum-sensing bacteria "routinely exist in fluctuating environments containing complex mixtures of chemicals, some of which are signals and some of which presumably do not convey meaningful information" (Waters and Bassler, 2005, p. 338). Each such species is equipped with "quorum-sensing signal detection and relay apparatuses" that "are complex and often consist of multiple circuits organized in a variety of configurations" (p. 338). Moreover, specialized quorum-sensing architectures "may be critical for filtering out noise from molecules in the environment that are related to the true signals and/or noise from signal mimics produced by other bacteria in the vicinity" (p. 327). That is, even bacteria have means to sort out the variability in their signaling in order for the "true signal" to be detected.

It is useful to consider in greater detail the quorum-sensing system of a particular species, *Vibrio harveyi*, depicted in Figure 3.1. Below, we repeat the summary of the system provided by Waters and Bassler (2005, pp. 325-7), with slight re-organization and structural emphasis:

External Signals: The *V. harveyi* quorum-sensing system detects an external, conspecific chemical signal composed of three "autoinducers": an AHL (acyl-homoserine lactone) signal termed HAI-1, a furanosyl borate diester known as AI-2, and an unidentified molecule termed CAI-1.

Internalization: The system also uses three cognate "receptors" functioning in parallel to channel information into a shared regulatory pathway: the autoinducer HAI-1 binds to a membrane-bound sensor histidine kinase (LuxN), AI-2 is bound in the periplasm by the protein LuxP; the LuxP-AI-2 complex interacts with another membrane-bound sensor histidine kinase, LuxQ, and CAI-1 interacts with a membrane-bound sensor histidine kinase, CqsS.

Internal Computation: At low cell density, in the absence of appreciable amounts of autoinducers, the three sensors – LuxN, LuxQ, and CqsA – act as kinases, autophosphorylate, and subsequently transfer the phosphate to the cytoplasmic protein LuxU. LuxU passes the phosphate to the DNA-binding response regulator protein LuxO. Phospho-LuxO, in conjunction with a transcription factor termed σ^{54}, activates transcription of the genes encoding five regulatory small RNAs (sRNAs) termed Qrr 1-5 (for Quorum Regulatory RNA). The Qrr sRNAs interact with an RNA chaperone termed Hfq, which is a member of the Sm family of eukaryotic RNA chaperones involved in mRNA splicing. The sRNAs, together with Hfq, bind to and destabilize the mRNA encoding the transcriptional activator termed LuxR.

Behavior: LuxR is required to activate transcription of the luciferase operon *luxCDABE*. Thus, at low cell density, because the *luxR* mRNA is degraded, the bacteria do not express bioluminescence. At high cell density, when the autoinducers accumulate to the level required for detection, the three sensors switch from being kinases to being phosphatases and drain phosphate from LuxO via LuxU. Unphosphorylated LuxO cannot induce expression of the sRNAs. This allows translation of *luxR* mRNA, production of LuxR, and expression of bioluminescence.

At the most general level, this quorum-sensing signal detection system has an integrated character internal to the organism, and operates over input initially external to the organism. Thus characterization of the system involves at least characterization of the organism-external input, and the organism-internal structures used in organizing the input. The external input is chemical in nature, and has three components (the three autoinducers), each of which has its own chemico-structural characterization. The organism-internal structures are of two kinds, those at the boundary of the organism, in contact with the environment, and those entirely internal to the organism, not in contact with the external world. An external signal is initially "internalized" by the boundary structures of the internal system, and then "processed" using a complex structure completely internal to the organism. Interestingly, the first kind of internal structure contributes its own signals (phosphate) to the detection process, while the second kind of internal structure makes no use of the external signal or its components, relying only on input from the boundary structures, which differ in nature and structure from the external input.

In some sense, quorum-sensing in bacteria can be viewed as a minimal case in the broader study of normalization, involving "noise reduction" in order to recover a "true signal" that influences organism-internal behavior toward one of two possibilities. The leap from our earliest relatives to more recent ones, e.g., zebra finches and their ability to

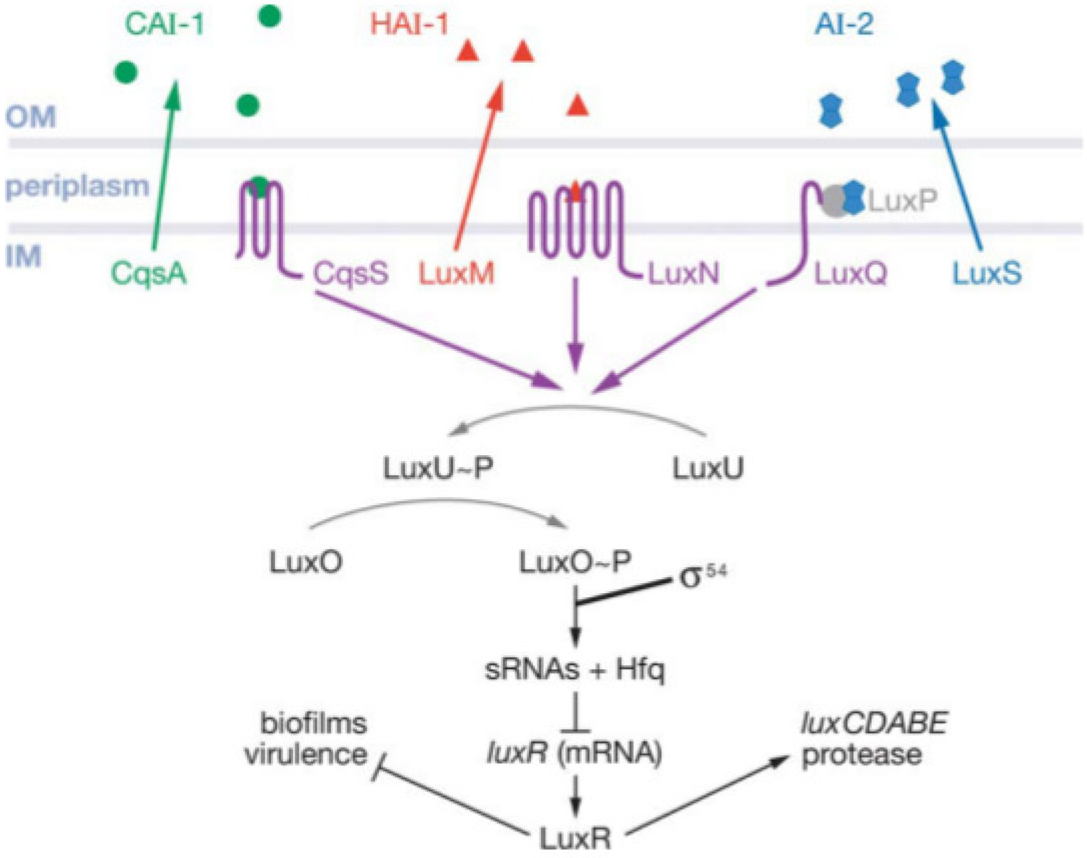


Figure 3.1: From Waters and Bassler (2005), page 326. *Vibrio harveyi* produces and responds to three distinct autoinducers. The sensory information is fed into a shared two-component response regulatory pathway. The arrows indicate the direction of phosphate flow in the low-cell-density state. CAI-1, HAI-1, and AI-2 are respectively represented by green circles, red triangles, and blue double pentagons. OM, outer membrane; IM, inner membrane.

normalize cross-gender vowel productions (Ohms et al., 2010), brings with it substantial complication of the minimal case, with respect to (i) organism-external input: the nature and structure of sociality and the signals involved, and (ii) organism-internal structures: the boundary structures for signal internalization, and (presumably) the presence of means to represent internalized signals, which may be accompanied by more sophisticated means of organizing variation in external signals. In the remainder of this section, we review attempts to characterize (i), and how it might be related to (ii), with special attention to ontogeny.

At the conclusion of the 1990s, the Presidentially proclaimed "Decade of the Brain," a host of reviews of the progress made within the brain and cognitive sciences were produced by leading researchers looking to assess its extent, and organize further investigation. One facet of the organization involved clarifying connections with adjacent disciplines, including the study of communication systems in species capable of producing and perceiving conspecific vocalizations. In a 1999 *Annual Review of Neuroscience* review article, Allison Doupe and Patricia Kuhl took a comparative, multidisciplinary approach, addressing contemporary understanding of the "common themes and mechanisms" found in the vocal communication systems of humans and song birds, focusing on the parallels bearing on the complex phenomenon termed "vocal learning." Indeed, the review may also be seen as an attempt to characterize what must be taken into account when studying vocal learning in a species capable of carrying it out. It is useful to separate out the phenomena and corresponding conceptualizations Doupe and Kuhl address that we incorporate within our modeling of the acquisition of vowel normalization. In this chapter, we focus on those relevant to what we take to be an intramodal vowel normalization computation over internal auditory representations of acoustic events external to the infant. Normalization involving auditory-articulatory intermodal abstraction is treated in Chapter 4.

Doupe and Kuhl (1999) (hereafter D&K) begin their comparative review by recognizing that "[m]any animals produce complex communication sounds," and take as their point of departure the "few of them [that] can and must learn these vocal signals" (p. 573). Within this select few, D&K point out, certain mammals are known to exhibit aspects of vocal learning, e.g., cetaceans' acquisition of a "vocal repertoire" and their demonstration of "vocal mimicry" (citing McCowan and Reiss, 1997), though humans constitute the main representative of the class in this regard, being "consummate vocal learners." Moreover, while certain groups of birds "meet the criteria for vocal learning," including "songbirds, the parrot family, and some hummingbirds," the representative group is taken to be "passerine (perching) songbirds," whose vocalizations are referred to as "birdsong" (p. 573), or simply "song." The vocal learning exhibited by passerines and humans was selected for comparison not only due to the richness of the "basic phenomenology of learning of song or speech" (p. 574) in each respective case, but also due to the "striking similarity" across cases.

Having selected the key organisms for comparison, D&K proceed with a description of the "evidence for vocal learning" concerning their respective intraspecies communication systems, the "most important" being the existence of "group differences in vocal productions that clearly depend on experience" (p. 574). With respect to humans, this evidence consists of the fact that "people learn the language to which they are exposed" and that "even within a specific language, dialects can identify the specific region of the country in which a person was raised" (p. 574). Similar evidence holds for songbird species, and may be even stronger given the results of "cross-fostering experiments, in which birds of one species being raised by another will learn the song, or aspects thereof, of the fostering species" (pp. 574-5). Early work on the subject by Immelmann (1969) shows that male zebra finches reared by a Bengalese finch foster father will learn the song of the foster parent, even in the presence of vocalizing conspecifics. While these results serve as evidence for vocal learning, they also focus attention on the nature of the "experience" that is critical to it. Rather than learning simply due to exposure to the distribution of conspecific song, the finches exhibited "a predisposition to learn the song of birds to which an 'emotional' relationship of some kind exists" (p. 67), suggesting that the experience involved in vocal learning cannot be characterized solely in terms of passive "auditory experience." In this dissertation, we assume this on principle.

The evidence that both humans and songbirds learn using the vocalizations to which they are exposed is supported by cases where lack of exposure results in impoverished acquisition. D&K identify documented cases of humans raised in social isolation (e.g., "Genie" in Fromkin et al. (1974), or the "Wolf Boy" in Lane (1976), along with a host of others) exhibiting abnormal spoken language, along with more compelling evidence from "songbirds collected as nestlings, and raised in isolation from adult song" who themselves "produce very abnormal songs" (p. 575). Thorpe (1958) showed that when these "Kaspar Hauser" Chaffinches "are themselves grouped together in isolated communities" they "build up, by mutual stimulation and imitation, complex but highly abnormal songs" (p. 568). The motivation to create group likeness in vocalization persists even in the absence of normal song repertoire. This aspect of song acquisition is further highlighted by species which seem to require social interaction for successful song acquisition. Zebra Finches do not learn song from passive auditory experience of audio recordings of conspecific vocalizations. Rather, they require "interaction" of some kind with an auditory tutor: in the absence of visual contact with a live tutor, physical/social interaction such as grooming or feeding suffices, while in the absence of a live tutor all together, interaction with the recordings, such as pressing a button to elicit vocalizations, also suffices (see Catchpole and Slater, 1995, for review).

D&K also draw attention to the likely need for humans and songbirds to hear their own vocalizations during vocal learning. Some species of songbirds exhibit "a separation between the period of hearing adult song and the onset of vocalizations," which provides means to investigate song quality in birds that "have had adequate tutor experience" (p. 576) but are unable to hear themselves, e.g., after being deafened. Konishi's (1965) study of White-Crowned Sparrows, for example, revealed that "the bird must hear its own sound in order to reproduce" the patterns of conspecific song "heard during the critical period" (p. 772) prior to vocalizing. That human infants need to hear their own vocalizations during vocal learning is more difficult to document because the period of initial vocal learning is extended and overlaps with the onset of vocalization in normal development, yet, there are sources of evidence, including normal-hearing children who had tracheostomies for extended periods of time. Infants undergoing the procedure exhibit slower speech development relative to other developmental cognitive components (Bleile et al., 1993). Kamen and Watson's (1991) investigation of vowel production in children with long-term tracheostomy revealed significant group differences in the spectral characteristics of their productions compared to those of matched controls exhibiting normal development. Notably, the acoustic spaces of trach children were considerably reduced relative to their non-trach counterparts, with the reduction occurring mainly in the regions associated with the corner vowels /i/ and /a/. These differences are attributed to the positioning of the tracheal cannula, which inhibits both fronting of the tongue and lowering of the vocal tract. The cannula also results in aphonia, preventing auditory feedback that facilitates exploration of the extremal regions of acoustic space. The lack auditory feedback may also preclude the assignment of referential value to the child's voice through vocal interaction with a caretaker. In this connection, Locke and Pearson (1990) suggest that the lack of normal speech may be explained by a "pragmatic component," given the comparatively normal development of infants born without tongues (MacKain, 1983).

It is worth stressing the importance of the differentiation between auditory experience derived from others and that derived from an infant's own productions. If the need for an infant to hear both (i) the vocalizations of others, and (ii) their own vocalizations is a condition on successful spoken language acquisition, and more specifically vowel category acquisition, then distribution-based models of vowel categorization must take this into account. Any model whose starting point is the "discovery" of statistical patterns in the distribution of experienced vocalizations will derive categories regardless of whether the vocalizations exclude those of the infant, or include only those of the infant. The former case covers nearly all current models of phonetic category acquisition, raising serious concerns about what they are modeling. Moreover, if true, the condition requires models of vowel category acquisition to incorporate the differences between the vowel productions of infants and those of their adult caretakers, rather than averaging them away. These differences must be recognized at some level of category acquisition modeling, if the learner is to differentiate between their own productions and those of others. In this dissertation, we adopt the following conceptual basis:

- Representational differences between vowels produced by an infant and their adult caretakers exist at the acoustic and auditory levels of the infant's perception of speech, and likely their cognitive representations.
- Cognitive organization of the representational differences is a prerequisite for category acquisition, and constitutes the basis of an infant's development of a cohesive auditory representation of the self and of their caretakers, in the sense discussed in Section 1.2.2. Specifically, these cohesive auditory representations are auditory manifolds.

This approach shifts focus from passive organization of undifferentiated auditory experience to the active organization of broader, and highly differentiated auditory experience, which is taken to include the "emotional" and "pragmatic" aspects mentioned above, and discussed further below.

In support of this view, we turn to experimental investigation of infant auditory experience factoring into spoken language acquisition. Greater scrutiny of the phenomena over the last few decades has yielded a fairly rich taxonomy of distinct yet overlapping kinds of vocalizations the infant experiences, either from others or from the self. Key differences in the kinds of vocalizations that constitute auditory experience are drawn out during vocal exchanges between infants and their caretakers, particularly those which exhibit a "turn-taking" structure involving contingent adult responses to infant vocalizations.

Comparative evidence suggesting that structured conspecific vocal exchanges involving differentiated vocalizations exist between adults and infants comes from results on nonhuman primates. Squirrel monkeys forage in large groups and emit calls, termed "chucks," to remain in vocal contact when separated visually by large distances (see Masataka and Biben, 1987). Their chuck vocalizations are of "two different types...*i.e.* chucks occurring independently of one another and those which are dependent upon the preceding calls" (p. 313, ibid). Analysis of the distribution of chuck calls among a group of 10 captive squirrel monkeys revealed the following temporal rule: the dependent response chucks should be given within the next 0.5 sec. of the preceding call, while independent chucks should not be produced during that period (ibid). Moreover, vocalization pairs consisting of an independent chuck followed by a response chuck correlate positively with affiliative interactions between group members (Masataka, 2003). Similar results were found for "coo calls" in Japanese macaques, whose "response" coos were moreover modified to match acoustic features of the prompting coos, potentially "fulfil[ling] a phatic function, which is presumably an underlying function of human conversation" (Masataka, 2003, p. 104).

Concerning humans infants, Masataka (2003) reports that "temporal rules for affiliative exchanges play an important role in maintaining their interaction with caregivers from a very early age," and that "the skill of conversational turn-taking...has been regarded as one of the major milestones in the early interactional development before the onset of true language" (p.44). Furthermore, infants likely have already learned the subtle structural aspects of the turn-taking activity by three to four months of age. During interactions "with very young infants, mothers are likely to socially stimulate them contingently upon their spontaneous behaviours," and Masataka's (1993) results show that "contingency did not affect the infant's rate of vocalization, but influenced its quality and timing" (p. 310). Specifically, it was found that "after vocalizing spontaneously, the infant waits for the mother's response," with the amount of time waited depending on "recent experience with the mother" (p. 310).

Structured turn-taking vocal exchanges characterized by contingent caretaker responses are a key locus for different infant and adult vocalization types. During turn-taking vocal exchanges between mother-infant dyads involving infants ranging between 3 and 4 months of age, Bloom et al. (1987) found a higher ratio of “syllabic” infant vocalizations, characterized by their “greater oral resonance, pitch variation, and possible consonant-vowel contours,” to “vocalic” vocalizations, which exhibit “greater nasal resonance” and uniformity in pitch (p. 215). Masataka (1993) reports like findings, noting that “syllabic utterance increased when the mother responded contingently to the infant” (p. 311). Goldstein et al.’s (2003) similar results for infants between 6 and 11 months of age highlight a parallel between the “social shaping,” or “selective reinforcement of vocal precursors by social companions,” that “biases learning toward certain vocal forms and facilitates the development of crystallized song” (p. 8030) in passerine songbirds, and the vocal development the infants undergo. Their results dovetail with Gros-Louis et al.’s (2006) findings that “mothers respond with play vocalizations significantly more to vowel-like sounds” produced by infants, while responding “to consonant-vowel vocalizations [of infants] with imitations significantly more than they did to vowel-like sounds” (p. 117).

Turn-taking exchanges also provide input infants use in “[l]earning the relation between their vocalizing and social responding from others” (Goldstein et al., 2009, p. 636). Using a “still-face paradigm” consisting of “a brief naturalistic face-to-face interaction between an adult and infant,” immediately after which, “the adult assumes a neutral expression and looks at the infant without speaking or changing expressions (a “still face”),” followed by the adult “engag[ing] in a second naturalistic interaction episode” (p. 637), Goldstein et al. (2009) found that by “5 months, infants have learned that their prelinguistic vocalizations elicit reactions from others,” including “unfamiliar social partners,” who infants expect “to respond to their vocalizations” (p. 643). Importantly, “[v]ocalization has acquired instrumental value” that is crucial to “the beginning of a developmental cascade of socially guided learning” (p. 643). Follow-up studies showed that “[i]nfants who learned the effects of their vocalizations on adults by 5 months appear to have advantages for later language learning" (p. 642). In addition, through "differential maternal responding" to an infant's vocalizations "mothers encourage the use of particular sounds, giving them meaning and framing the interaction" (Gros-Louis et al., 2006, p. 117). In particular, "imitations and expansions" over infant vocalizations "provide infants with information about the 'meaning' of their vocalizations" – behavior which is known to "correlate positively with language development" (ibid, citing Girolametto et al., 1999; Tamis-LeMonda et al., 2001).

The results discussed above suggest the following extension of our conceptual basis. Specifically, we assume that:

- Infants have the ability to cognitively represent a pairing of their vocalizations with the vocal responses of adult caretakers during turn-taking exchanges, and assign to each pairing the social meaning the caretakers impose on their responses.
- Infants use this pairing to derive mappings between auditory manifolds representing a self and their caretakers. This computation is called *vowel normalization*. In the simplest case, an auditory manifold representing a self is "aligned" with an auditory manifold representing a single caretaker.

From this basis, we take the view that vowel category acquisition is an emergent phenomenon which arises from a vowel normalization computation. Toward adducing this claim, the remainder of this section reviews aspects of the mainstream conceptualization of vowel category acquisition, while drawing out potential problems with mainstream modeling. In this connection, we turn to D&K's review of phonetic category acquisition, focusing on the relevant history leading up to the "new view" proffered in Kuhl (2000) and extended in Kuhl (2007). In the next section, we situate these aspects within our own approach.

D&K begin with the “classic experiments” carried out in the 1970s (e.g., Eimas, 1975a,b; Streeter, 1976) based on “tests in which a conditioned head turn is used to signal infant discrimination,” showed that “early in postnatal life, infants respond to the differences between phonetic units used in all of the world’s languages, even those they have never heard,” revealing the “exquisite sensitivity of infants to the acoustic cues that signal a change in the phonetic units of speech, such as the VOT differences that distinguish /b/ from /p/ or the formant differences that separate /b/ from /g/ or /r/ from /l/” (p. 580). Moreover, perception experiments using “a computer-generated series of sounds that continuously vary in small steps, ranging from one syllable (e.g., /ba/) to another (/pa/), along a particular acoustic dimension (in the case of /pa/ and /ba/, the VOT)” revealed that both infant and adult listeners “tend not to respond to the acoustic differences between adjacent stimuli in the series but perceive an abrupt change in the category – the change from /ba/ to /pa/ – at a particular VOT” (p. 580). This “categorical perception” of speech sounds, however, is known to differ between infants and adults, with the latter group only exhibiting the phenomenon for native language sounds (citing Miyawaki et al., 1975), while the former also “demonstrate the phenomenon for languages they have never heard before” (p. 580, citing Streeter, 1976; Lasky et al., 1975) well into the sixth month of life, with rapid shift toward language-specificity shortly thereafter.

In this connection, D&K further recount that these “initial studies demonstrating categorical perception of speech sounds in infants and its narrowing with language exposure led many speech theorists to take a strongly nativist or selective view of speech learning” (p. 587). The view held that infants were “biologically endowed” with some innate specification of all possible phonetic units, and “the subsequent decline in speech discrimination was seen as a process of atrophy of the prespecified phonetic representations in the absence of experience" (p. 587). By the mid 1990s, however, two lines of research began to coalesce, which, according to Kuhl (2007), together constitute evidence against the selectionist view and suggest that infants "are engaged in some other kind of learning process...that is not fundamentally subtractive in nature" (Kuhl, 2000, p. 11852).

One line of research D&K reference involved "more detailed study on the changes in infant phonetic perceptions brought about by experience" which suggests "that perceptual learning is not in fact simple sensory memory of the sound patterns of the language" (p. 587). Rather, perceptual learning of speech sounds seems to involve (at least)

> "a complex mapping in which perception of the underlying acoustic dimensions of speech is warped to create a recognition network that emphasizes the appropriate phonetic differences and minimizes those that are not used in the language." (p. 587)

Warping of consonant perception was revealed using "large grids of systematically varying consonant-vowel syllables spanning the phonetic boundary between American English /r/ and /l/" (p. 588) and "perceptual similarity" ratings of all possible pairs of stimuli from the grid provided by American English-speaking listeners (see Iverson and Kuhl, 1996). Analysis of the ratings "indicated that although the real physical distances in the grid were equal," the "American listeners perceived many of the sounds as if they were closer to the best, most prototypical examples of /r/ and /l/" as well as "a larger than actual separation between the two categories" (p. 588). This phenomenon, termed the "perceptual magnet effect" due to the more prototypical sounds acting "as magnets for surrounding sounds," also occurs for vowel categories (Kuhl, 1991) and is generally characterized by a "shrinking" of within-category perceptual distances, together with a "stretching" of between-category distances. Moreover, the magnet effect is language-specific, so that "in each language group, perception is distorted to enhance perception of that language" (D&K, p. 587). Importantly, infants demonstrate the perceptual magnet effect by 6 months of age (Kuhl, 1991; Kuhl et al., 1992), with the effect for vowel categories likely independent of the effects of categorical perception (see Iverson and Kuhl, 2000). In the next section, we show how this independence can be captured within our approach.

The second line of research involves exploration of the hypothesis that infants are "sensitive to the distributional frequencies of the sounds they hear in ambient language" (Kuhl, 2007, p. 112). Saffran et al.'s (1996) now highly conspicuous experimental results concerning the hypothesis suggest that infants are able to segment pseudowords such as [bidaku] from a continuous signal without the support of word-level prosodic cues based on two minutes of exposure to spliced-together syllables such as [bidakupado...]. Bates and Elman (1996) prominently lauded the results, which they took to be proof "that infants can use simple statistics to discover word boundaries in connected speech, right at the age when systematic evidence of word recognition starts to appear in real life," and moreover that "infants are capable of extracting statistical regularities from only 2 min of spoken input" given "no reward or punishment other than the pleasure of listening to a disembodied human voice" (p. 1849). Similar work, along with similar conclusions, soon followed, e.g., Maye et al.'s (2002) results on infant discrimination of VOT stimuli based on exposure to unimodal/bimodal distributions over a VOT scale, which they interpret as demonstrating that "infants of 6 and 8 months can harness a...powerful learning mechanism in the service of phonetic categorization" (p. B109) along acoustic dimensions, generally. In light of such results, Kuhl (2007) concludes that "phonetic learning can be altered by the distributional patterns in language input" (p. 112), and together with perceptual magnet effect cast substantial doubt on selectionism.

These two lines of evidence have also substantially influenced the course of research in phonetic category acquisition in recent years. The study of the perceptual magnet effect has influenced the investigation of a broader array of "perceptual narrowing" phenomena (see Lewkowicz and Ghazanfar, 2006; Lewkowicz and Hansen-Tift, 2012), while distributional learning has spread across the entire cognitive scene. We take the set of phenomena related to the perceptual magnet effects as interesting, and in need of further experimental investigation and theoretical accounting, and, we do so with due circumspection, given the scant nature of current understanding. We consider "statistical learning mechanisms," however they may be construed (e.g., in terms of Bayesian learning, Exemplar learning, etc.), with a level of caution proportional to the level of enthusiasm the approach has generated.

Enthusiasm surrounding statistical learning mechanisms and their role in human language acquisition is based on the notion that their existence "flies in the face of received wisdom" on the topic (see Bates and Elman, 1996). Despite the enthusiasm (or perhaps due to it), and in light of the mischaracterized implications of the existence of such mechanisms, it is important to recognize that the experimental results discussed above are little more than potential hints concerning the nature and structure of the mind. Importantly, the manner in which such experiments are formulated and reported cannot be taken as a direct formulation of cognitive objects. Iverson and Kuhl (2000) make explicit the distinction between the perceptual magnet effect as a "perceptual phenomenon of sensitivity minima near best exemplars" and the "hypothesized mechanisms" which may be its cause. The vocabulary used to formulate the magnet effect for experimentation (e.g., "prototypical exemplars") may have led to its being "strongly linked" to a mechanism where

> phoneme categories are represented in terms of prototypes (i.e., a single abstract exemplar that represents all the members of a category) and that phonemes are at least partially perceived in terms of their distance from these prototypes, (Iverson and Kuhl, 2000, p. 874)

though the authors explicitly note that “the perceptual magnet effect may, in fact, be due to other mechanisms” (p. 874) such as “categorization processes based on multiple stored exemplars” (p. 875, citing Lacerda, 1995), or “experience-related distortions in auditory perception” (p. 875, citing Guenther and Gjaja, 1996), the latter of which makes no use of prototypes or exemplars at all. By the same token, the “sensitiv[ity infants demonstrate] to the distributional frequencies of the sounds they hear in ambient language” (Kuhl, 2007, p. 112) is a perceptual phenomenon quite separate from the “simple” and “powerful” statistical “learning mechanisms” so eagerly hypothesized to be its cause. The vocabulary used to formulate the experimentation (e.g., “distributions,” “frequency,” etc.) may have again created a strong, yet potentially baseless link to a particular set of posited mechanisms.

To suss out the potential problem, we briefly return to bacteria quorum-sensing, and reformulate the investigation in terms of a decision theoretic vernacular. Recall that the phenomenon involves the recovery of a conspecific signal from the myriad signals in the environment external to the organism that influences organism-internal behavior toward one of two possibilities, say bioluminescence, corresponding to a critical density of conspecifics, or non-bioluminescence, corresponding to a lack of critical density of conspecifics. Suppose the bacteriologists formulated experimentation in the following manner: characterize the autoinducer levels as “data” that the bacteria use in “deciding” whether or not a quorum is present, i.e., whether or not to bioluminesce. This formulation might lead modelers to posit that the bacteria possess a “decision mechanism” that selects between the two hypotheses: i) a quorum is present, or ii) or quorum is not present, which results in the surface behavior: bioluminescence or non-bioluminescence. This decision mechanism might then be modeled using a corpus of autoinducer and external state recordings to train and test some decision algorithm, say, a Bayesian classifier or some Exemplar-based "computational model." Successful modeling then amounts to reporting which parameter settings of the Bayesian classifier or Exemplar model match the experimental data concerning the relationship between autoinducer levels and bioluminescence.

Although this approach may be very simple and implementable, while yielding predictions, it is difficult to see how it would shed any light on any of the structures that are already known to play a role in the organism's quorum-sensing, or how they relate to one another. Moreover, it is even more difficult to see how it could have revealed these structures and relations, if they were not already known to exist. Nevertheless, an analogous approach is taken quite seriously in modeling phonetic category acquisition, and is widely and strongly believed to be making real progress in revealing its properties. On this note, we turn to a real example.

A recent issue of the journal *Developmental Science* promoted the statistical learning approach, with one of the main contributions coming from Richard Aslin, an author on the catalyst paper Saffran et al., 1996, devoted to the "computational principles of language acquisition." Specifically, McMurray et al. (2009) take Maye et al.'s (2002) results as a point of departure for investigating the "sufficiency" of the hypothesis that "statistical learning may be an important mechanism for the acquisition of phonetic categories in an infant's native language" using "a computational model based on a mixture of Gaussians (MOG) architecture" (p. 369). The authors interpreted the results of the "computational work" as "provid[ing] further evidence for the plausibility of unsupervised learning of speech categories via a statistical learning mechanism" combined with "another core mechanism (competition)," and "yield[ing] not only successful data-driven learning that approximates the developmental timecourse, but also novel insights about the...nature of early speech categories" (p. 377). It is important, however, to construct a more circumspect appraisal of work of this nature.

Regarding the modeling, McMurray et al. (2009) state that their MOG model "uses a simple architecture, makes few theoretical assumptions, and adds constraints only when needed to account for the data" (p. 370). On closer inspection, however, it is clear that the architecture only appears to be simple and that there are a number of hidden theoretical assumptions and imposed constraints that may have little or nothing to do with the data. The hidden theory includes, inter alia, the assumption of a Euclidean space to serve as the space of cue values, including the Euclidean distance metric on that space, as well as the assumption of a specific statistical manifold whose geometry constrains the possible values of the MOG parameters, including the Riemannian metric on that space. This exemplifies the growing tendency to conceptualize statistical models as independent of their underlying probability theory, giving them a "simple" appearance supporting their "psychological plausibility." The authors also claim that the model is "capable of learning the correct number of categories" for any given contrast in a language, a presumed improvement over previous work (e.g., de Boer and Kuhl, 2003) in which "the number was specified *a priori*" (p. 371). Rather, the modeling architecture begins with an "array of K Gaussians [categories]...to serve as the initial state" of the model, where "K is relatively high (e.g. 10-20)," and then reducing the number of Gaussians as exposure to data increases. That is, the authors have built the number of categories into the model in a slightly less overt way. Moreover, their improved architecture imposes a "fundamentally subtractive" mechanistic constraint on the phenomena that the distributional conceptualization was taken to be evidence against.

Regarding the phenomena, further assumptions are imposed, as McMurray et al. (2009) take “the computational problem the system is trying to solve” to be “learning the mapping between continuous inputs and categories” (p. 370). It is by no measure clear that any subsystem of phonetic acquisition is principally concerned with a mapping of this nature. Rather, it may be the case that category acquisition is itself a derivative phenomenon, resulting from a more general set of computations by which the self is related to others; the position we take in this dissertation. Moreover, even if acquisition does encompass such a mapping, there is little reason to believe that it possesses the properties suggested by their model, e.g., that it is acquired “iteratively (i.e., learning occurs after *each* input)” (p. 371), and that learning ought to be unsupervised. To address these issues in greater detail, it is useful to separate (i) the mapping itself from (ii) the set of algorithms that computes it, considering each in turn.

With respect to the mapping, McMurray et al. (2009) keep to a common Positivist dogma, characterizing the input in terms of a physicalist description of generic, undifferentiated auditory experience derived from externalizations. This characterization seems rather inappropriate in light of the nature of an infant’s early auditory experience, even when limiting consideration to caretaker vocalizations. Results in Kuhl et al. (2003) on the role of “live tutoring” in phonetic learning of a foreign language during infancy suggest that “learning from complex language input relies on more than raw auditory sensory information” and “is influenced by the presence of a live person” (p. 9100), paralleling results on the acquisition of song by zebra finches. Kuhl (2007) interprets the results more generally, taking them to indicate that “[e]xposure to a new language in a live social interaction situation induces remarkable learning in 9-month-old infants, but *no* learning when the exact same language material is presented to infants by a disembodied source” (p. 116). If true, this “flies in the face” of Saffran et al.’s (1996) claim about powerful statistical learning mechanisms operating over “nothing more than...a disembodied human voice.” Rather, it seems that learning is “socially gated,” leading to the formulation of the hypothesis that “[s]ocial interaction is *essential* for natural speech learning” (Kuhl, 2007, p. 110).

Regarding the computation of the mapping, McMurray et al.’s (2009) assumption that category acquisition involves “unsupervised learning” which “occurs after each input” is simply baseless given the influence of contingent caretaker responses influences vowel category acquisition.

Similar argumentation holds for most of the recent “cognitively plausible,” “ecologically valid,” and “data-driven” statistical learning approaches to category acquisition coming from the statistical learning paradigms (e.g., Guenther et al., 2006; Rasilo et al., 2013; Hörnstein, 2013). The corresponding models typically rely on large amounts of externalization data of some kind to drive “domain-general” statistical algorithms taken to be the organizational means that moreover model the whole of the developing human mind, and in many cases the human scene more broadly. It seems likely that a different modeling approach may be needed for the computation of the mapping, more in line with the notion that “that perceptual learning is not in fact simple sensory memory of the sound patterns of the language,” but rather

> “a complex mapping in which perception of the underlying acoustic dimensions of speech is warped to create a recognition network that emphasizes the appropriate phonetic differences and minimizes those that are not used in the language.” (Doupe and Kuhl, 1999, p. 587)

In the remainder of this chapter, we put forward a framework for modeling the acquisition of vowel normalization, which we take to have the following structural components.

External Signals: We take the external signals involved to be the acoustic signals produced by an infant and a single caretaker, together with social signals that correspond to the acoustic signals produced by the caretaker, as well as the infant. We are explicitly differentiating between infant and caretaker acoustic signals, and within each group, further differentiating according to a social metric.

Internalization: We take the internalization of signals to involve the creation of auditory representations over the acoustic signals derived from the infant and caretaker productions, as well as the creation of representations derived from interpretation of the caretaker's social signals.

Internal Computation: We take cognitive computation to involve: i) the creation of manifolds over auditory representations, derived either from infant and caretaker acoustic signals or from internal self-signaling, ii) the creation of a pairing of auditory representations of infant productions with auditory representations of caretaker responses derived from turn-taking vocal exchanges, along with the assignment to each pair of an interpretation of the social signal imposed on by the caretaker on their response, and iii) the alignment of auditory manifolds using the pairing. In this fashion, the spaces that are ultimately used for vowel categorization are language-specific, and moreover, dyad-specific, rather than fixed across languages.

Behavior: The resulting behavior, e.g., categorical perception of vowel signals, the perceptual magnet effect, etc., is based on emphasis of regions in the aligned manifolds made salient as a result of the internal computations. Importantly, the categorical behavior is a derivative aspect of the internal computation. As we will see in Chapter 4, the main output of the computations is a conceptual structure allowing for general equivalence computations between the infant and conspecifics.

We model the components listed above through the creation of a "virtual environment for vocal learning." The environment consists of models of caretaker agents representing five different language communities (American English, Cantonese, Greek, Japanese, and Korean) derived from vowel category perception experiments, and models of infant agents that "vocally interact" with their caretakers. We develop a model of caretaker social and vocal signaling in response to infant vowel productions (Section 3.2). We then characterize the internalization of these signals and the internal computations over them (Section 3.3), demonstrating how they yield the external behavior (Section 3.4). The "articulatory" and "intermodal" aspects of vocal learning are left to Chapter 4. The main computations carried out within the vocal learning environment are summarized in Appendix A (Figures A.1 through A.3 depict the computations described in the following sections).

3.2 Maximal Vowel Spaces and Caretaker Responses

Models of adult caretakers within the vocal learning environment are based on analysis (Plummer et al., 2013a) of a set of cross-language vowel categorization experiments (Munson et al., 2010). Seven sets of vowel stimuli were generated by an age-varying articulatory synthesizer, one for each of the seven ages: 6 months, 2, 4, 5, 10, 16, and 21 years. The stimuli were categorized by members of 5 different language communities: Cantonese (n=15), English (n=21), Greek (n=21), Japanese (n=21), and Korean (n=20). Each listener assigned each stimulus a vowel category from the listener's native language, along with a "goodness rating" (Miller, 1994, 1997) indicating how good the listener felt that stimulus was as an example of the assigned category. A statistical methodology, based on a smoothing spline approach (Wahba, 1990; Gu, 2002) to additive modeling (Friedman and Stuetzle, 1981; Buja et al., 1989; Hastie and Tibshirani, 1990; Wood, 2006), provides means for estimating a set of "vowel category response surfaces" over the "maximal vowel space" for each age, based on a listener's identification responses and associated goodness ratings for the 38 stimuli for that age. The details of the approach, along with its corresponding operations, are covered in the remainder of this section.

3.2.1 Modeling Basis: Perceptual Experimental Results

The *Variable Linear Articulatory Model* (VLAM, Boë and Maeda, 1998) is a computational model of the articulatory system and its speech production capacities. Midsagittal representations are wrought by configuring "articulatory blocks" (Lindblom and Sundberg, 1971; Maeda, 1990, 1991) corresponding to jaw height, tongue body position, tongue dorsum position, tongue apex position, lip protrusion, lip height, and larynx height. The VLAM is age-varying and capable of representing vocal tract lengths ranging from those of infants to young adults (see Figure 3.2, top), calibrated in accordance with age-related "organic variation" (Beck, 1996; Goldstein, 1980). Further details on the VLAM are given in Section 4.2.

Given an age in years, the set of all articulatory configurations of the VLAM at that age that do not result in occlusion of the oral cavity yield a corresponding *maximal vowel space* (MVS, Boë et al., 1989; Schwartz et al., 2007) for that age. We adapt the term to our framework and take an MVS to be the set of formant patterns corresponding to the collection of articulatory configurations. Each formant pattern within an MVS is identified with a *formant vector* whose components are the first three formant frequencies of that formant pattern. We fix the following age index $\textsc{vlamAges} = \{0.5, 2, 4, 5, 10, 16, 21\}$ for indexing the MVSs discussed below. For each $a \in \textsc{vlamAges}$, let $\textsc{mvs}(a)$ be a dense sampling of the MVS for age a, subject to the following constraints: minimal constriction area of each articulatory configuration was fixed at 0.3 cm^2 for ages 2 and over, and 0.15 cm^2 for the 6 m.-o., in accordance with previous modeling (Ménard et al., 2002), and lip area was constrained from 0.1 cm^2 to 8 cm^2 for all ages. Each $\textsc{mvs}(a)$, depicted in Figure 3.2 (bottom), contains approximately 5,000 formant vectors.

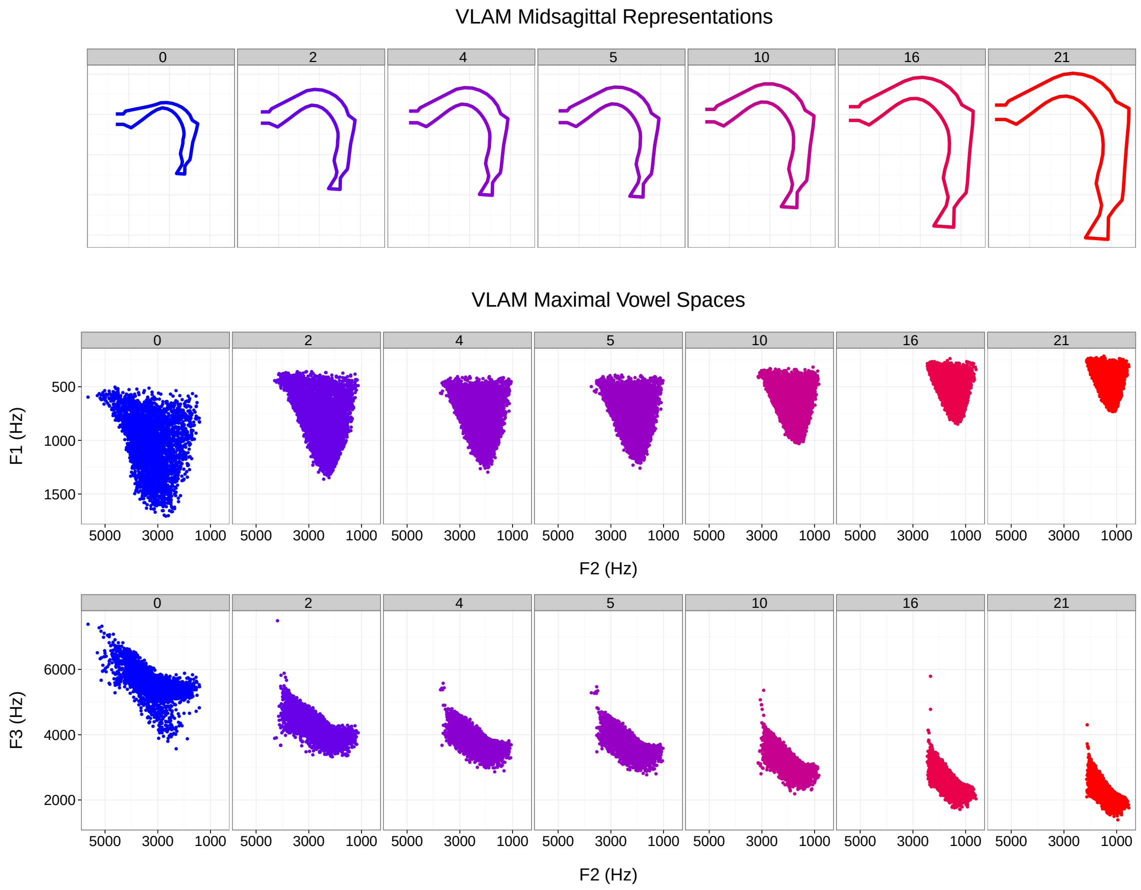


Figure 3.2: (Top) VLAM midsagittal representations of a neutral vocal tract for ages 0.5 (denoted 0), 2, 4, 5, 10, 16, and 21, in years, along with the corresponding densely-sampled MVS within formant space (Bottom).

The stimuli used in the perceptual experiments (Munson et al., 2010) were *vowel prototypes* (Vallée et al., 1995), or simply *prototypes*, selected from the MVSs for each $a \in$ VLAMAGES. The selection process (Ménard and Boë, 2000; Ménard et al., 2002) is meant to yield a set of cross-linguistically relevant prototypes in accordance with the dispersion-focalization theory (Schwartz et al., 1997a,b). For each $a \in$ VLAMAGES, a set of 38 prototypes, denoted by P(a), were selected from MVS(a). Prototypes in P(a) are indexed $\mathbf{p}^i$, $1 \leq i \leq 38$, though the superscript is often dropped. The sets P(0.5) and P(10) of

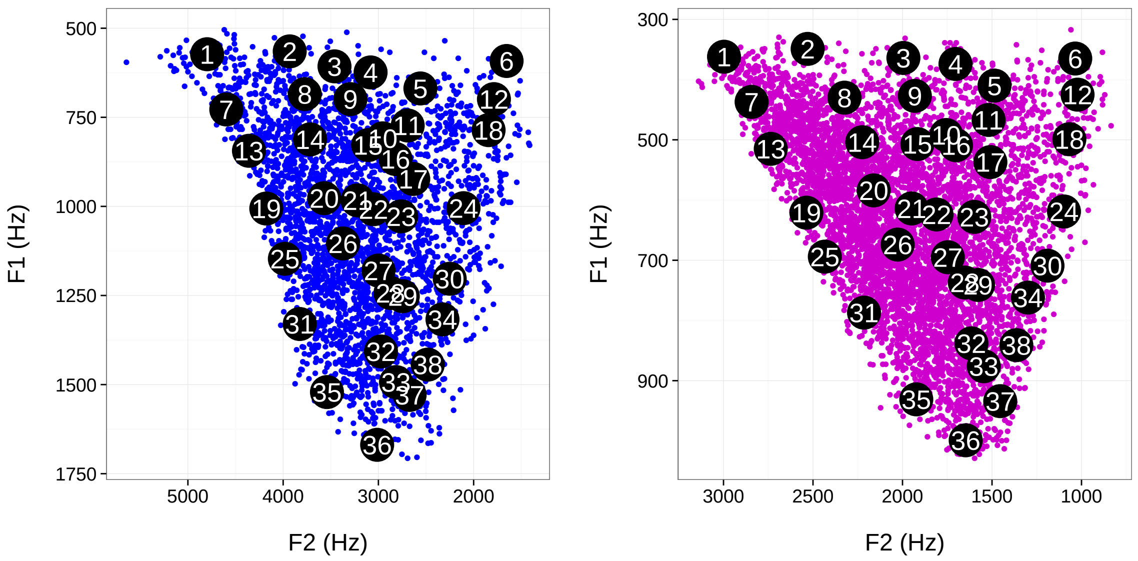


Figure 3.3: (Left) Age 0.5 vowel prototypes P(0.5) within an age 0.5 maximal vowel space MVS(0.5). (Right) Age 10 vowel prototypes P(10) within an age 10 maximal vowel space MVS(10). Note the difference in scales.

prototypes for ages 0.5 and 10, respectively, are depicted within MVS(0.5) and MVS(10) in Figure 3.3.

The perceptual experiments (Munson et al., 2010) elicited responses from native speakers of Cantonese (n=15), American English (n=21), Greek (n=20), Japanese (n=21), and Korean (n=20) for all 38 prototypes in P(a) for all 7 ages. Cantonese- and American English-speaking listeners categorized each prototype by clicking on any of 11 keywords representing the monophthongal vowels in each language. Listeners for the other languages categorized by clicking on a symbol or symbol string that unambiguously represented a (short monophthongal) vowel in isolation, choosing among 7 vowels (Korean-speaking listeners) or among 5 vowels (Greek- and Japanese-speaking listeners).

In addition to assigning a category to each prototype, each listener provided a visual analog scale (VAS, Massaro and Cohen, 1983; Miller, 1994, 1997) value indicating the "goodness" of that prototype as a representative of the assigned category. The VAS values ranged from 90-535, (90 best, 535 worst), though it is convenient to range normalize (we use min-max range normalization, though others are viable), and order reverse the scale, which hereafter ranges from 0-1 (1 best, 0 worst).

The formal method for generalizing over a subject's response to prototypes in $\textsc{p}(a)$ is as follows. Consider, say, subject 12 from the Greek language community (G), which we denote s^G_{12}. Let $C_G = \{\mathsf{i}, \mathsf{e}, \mathsf{a}, \mathsf{o}, \mathsf{u}\}$ be the set of vowel categories for language community G, and let $\textsc{vas}$ denote the interval $[0, 1]$ of possible VAS values for vowel prototype category ratings. Let $a \in \textsc{vlamAges}$. For each $\mathbf{p} \in \textsc{p}(a)$, let c and γ, respectively, be the category and VAS goodness rating assigned to $\mathbf{p}$ by s^G_{12}. We can then define a function

$$R(s^G_{12}, a) : \textsc{p}(a) \to C_G \times \textsc{vas}$$

such that $\mathbf{p} \mapsto (\mathsf{c}, \gamma)$. That is, the codomain of this function is a set of ordered pairs called *responses* whose first component is a category in C_G, and second component a VAS "goodness" value. In the case of a "no-response" from s^G_{12} we may augment C_G and $\textsc{vas}$ with an element NA.

We can extend $R(s^G_{12}, a)$ to reflect implicit judgments about how well each prototype represents categories not assigned to that prototype. For each $\mathsf{c} \in C_G$, let

$$\gamma_{\mathsf{c}} : C_G \times \textsc{vas} \to C_G \times \textsc{vas}$$

such that $(\mathsf{c}', \gamma) \mapsto (\mathsf{c}, \gamma)$ if $\mathsf{c}' = \mathsf{c}$, and $(\mathsf{c}, \alpha(1 - \gamma))$ otherwise, where $0 \leq \alpha \leq 1$. The *response category extension parameter* α allows us to make a more general model than in previous work (Plummer, 2012b) (which corresponds to a choice of $\alpha = 0$), to be more compatible with general Signal Detection Theory approaches (Wickens, 2002) (which might be approximated by choosing $\alpha = 1$). To exemplify, suppose s^G_{12} gives

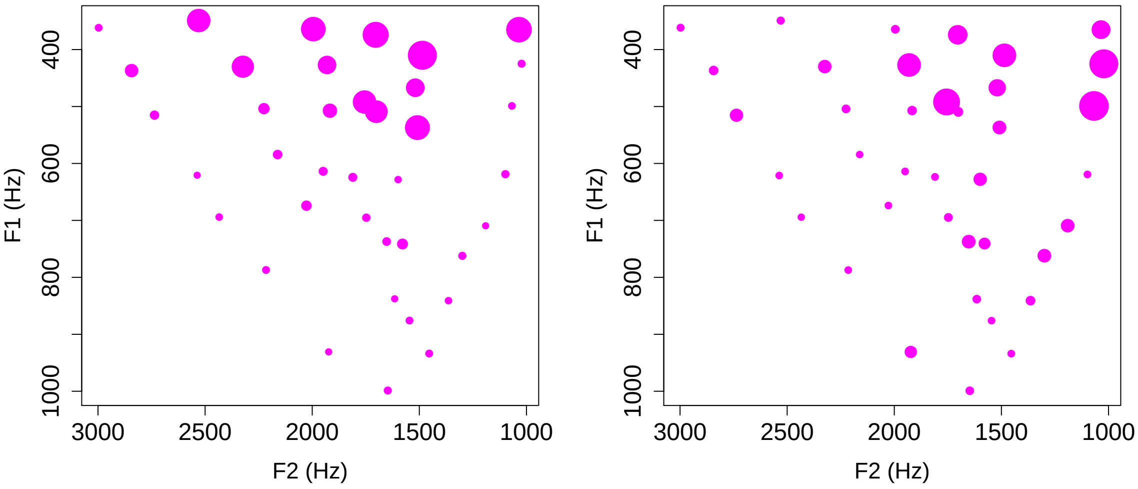


Figure 3.4: Individual category response surfaces for $R'_{\textsf{u}}(s^J_{20}, 10)$ (left) and $R'_{\textsf{u}}(s^G_{12}, 10)$ (right) where the response category extension parameter $\alpha = 0.5$. The size of the points indicates a larger ICRF value.

response $r = (\textsf{i}, 0.9)$ and $\alpha = 0.5$. Then $\gamma_{\textsf{i}}(r) = (\textsf{i}, 0.9)$, while $\gamma_{\textsf{o}}(r) = (\textsf{o}, 0.05)$. We can then compose each $\gamma_{\textsf{c}}$ with $R(s^G_{12}, a)$ to obtain functions

$$R_{\textsf{c}}(s^G_{12}, a) =_{def} \gamma_{\textsf{c}} \circ R(s^G_{12}, a) : \textsc{p}(a) \to C_G \times \textsc{vas}$$

such that, for each $\textsf{c} \in C_G$, each $\mathbf{p} \in \textsc{p}(a)$ is mapped to a response reflecting its goodness as an example of $\textsf{c}$.

Let $\textsc{lang} = \{\mathrm{C}, \mathrm{E}, \mathrm{G}, \mathrm{J}, \mathrm{K}\}$ be an index set over denotations of the five language communities. Given a language community $\ell \in \textsc{lang}$, let C_ℓ denote the set of vowel categories for ℓ. Let n_ℓ denote the number of subjects for ℓ. Given an age $a \in \textsc{vlamAges}$, a subject s^ℓ_τ, where $1 \leq \tau \leq n_\ell$, and a category $\textsf{c}^\ell \in C_\ell$, the function $R_{\textsf{c}}(s^\ell_\tau, a)$ is called an *individual category response function for* s^ℓ_τ *at age a*, or simply an *individual category response function* (ICRF). We construct ICRFs for each subject and category from each language community in LANG, for each age $a \in$ VLAMAGES. Individual category response surfaces for $R'_{\mathsf{u}}(s^J_{20}, 10)$ and $R'_{\mathsf{u}}(s^G_{12}, 10)$ are depicted in Figure 3.4.

3.2.2 Vowel Category Response Surfaces

Our formal method for extending the domain of an ICRF $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)$ from P(a) to all of MVS(a) is as follows. The basic idea is to use a regression technique to construct a "surface" of responses over MVS(a) using $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)$. The response surface value for a formant vector $\mathbf{f} \in$ MVS(a) is meant to approximate s^{ℓ}_{τ}'s VAS goodness rating of $\mathbf{f}$ as an example of vowel category c for age a.

The regression is carried out using smoothing spline-based additive models (Buja et al., 1989; Hastie and Tibshirani, 1990; Wood, 2003, 2006). In the statistical formulation, we begin with a *response variable* Y and observed *response vector* $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, and *design variables* X_j with observed *design vectors* $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})^T$, where $1 \leq j \leq p$. Design vectors are arranged in a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, whose rows are denoted $\mathbf{x}^i$, $1 \leq i \leq n$. We are interested in deriving a *fit* $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$ such that $\mathbf{y} = \hat{\mathbf{y}} + \epsilon$, for residuals $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$, that bears a smooth relationship to the $\mathbf{x}_j$, though not necessarily that of a least-squares line. One of the simplest ways to obtain such a fit is through the use of smooth functions $g_j(X_j)$, and an *additive predictor* $\beta + \sum_{j=1}^{p} g_j(X_j)$.

To illustrate, consider the univariate case of estimating a smooth function g from the n observations (y_i, x_i) such that $y_i = g(x_i) + \epsilon_i$, where ϵ_i is a random error term. We can estimate g using a "thin-plate regression spline" method (Wood, 2003, 2006), which involves finding a function $\hat{g}$ minimizing

$$\sum_{i=1}^{n} [y_i - h(x_i)]^2 + \lambda J_{md}(h).$$

The term on the left determines closeness of fit, while the term on the right controls the smoothness of the fit. The *smoothing parameter* λ, which can be estimated along with g, controls the trade off between these terms: as $\lambda \to \infty$ the fit approaches a straight line, while $\lambda = 0$ yields an unpenalized regression spline estimate. The operator J_{md} has the general form:

$$\int \cdots \int_{\mathbb{R}^d} \sum_{v_1+\cdots+v_d=m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial {x_1}^{v_1} \ldots \partial {x_d}^{v_d}} \right)^2 dx_1 \ldots dx_d.$$

where d is the number of arguments to h, and m is the desired order of partial derivative. The univariate case $d = 1$ considered above generalizes easily to additive predictors involving more than one design variable, but also to smooth functions over more than one variable. At present, we use the computationally tractable “thin-plate regression spline” method for response surface regression available in the mgcv R package (Wood, 2003, 2006), though others are available for implementation, as are other smoothing spline techniques (e.g., the R package bruto).

Given an $\textsc{mvs}(a)$, let F^a_1, F^a_2, and F^a_3 be variables whose observed values $f^a_{k,1}$, $f^a_{k,2}$, and $f^a_{k,3}$ constitute the first, second, and third components, respectively, of the formant vectors in $\textsc{mvs}(a)$. Let $\mathbf{f}^a_1$, $\mathbf{f}^a_2$, $\mathbf{f}^a_3$ be the vectors of observed values for F^a_1, F^a_2, and F^a_3, respectively. Form the data matrix $\mathbf{F}^a = (\mathbf{f}^a_1\ \mathbf{f}^a_2\ \mathbf{f}^a_3)$. Similarly, form a data matrix $\mathbf{P}^a$ from the rows of $\mathbf{F}^a$ that correspond to the prototypes in $\textsc{p}(a)$. Thus each formant vector in $\textsc{mvs}(a)$ is indexed by its row position in $\mathbf{F}^a$, and each prototype in $\textsc{p}(a)$ by its row position in $\mathbf{P}^a$. Let P^a_1, P^a_2, and P^a_3 be variables whose observed values are the first, second, and third columns of $\mathbf{P}^a$.

Given an ICRF $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)$, for each $\mathbf{p}^i \in \textsc{p}(a)$, let γ_i be the second component of the response $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)(\mathbf{p}^i) = (\mathsf{c}, \gamma)$, i.e., $\gamma_i = \gamma$, and let $\mathbf{y}_{\mathsf{c}} = (\gamma_1, \ldots, \gamma_{38})^T$. We can now estimate smooth functions $g_1(P^a_1)$, $g_2(P^a_2)$, and $g_3(P^a_3)$ for an additive predictor $\sum_{j=1}^{3} g_j(P^a_j)$ using $\textsc{p}(a)$ as a data matrix and $\mathbf{y}_{\mathsf{c}}$ as response vector, yielding an additive

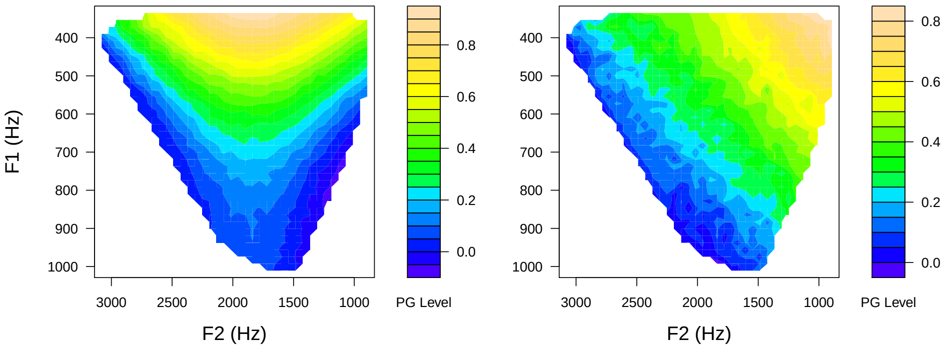


Figure 3.5: Vowel category response surfaces for $R'_{\mathsf{u}}(s^J_{20}, 10)$ (left) and $R'_{\mathsf{u}}(s^G_{12}, 10)$ (right) where the response category extension parameter $\alpha = 0.5$. The "PG Level" is the predicted goodness level, approximating a subject's goodness rating of a formant vector as an example of the category u.

model $\mathbf{y}_c = \sum_{j=1}^{3} g_j(P^a_j) + \epsilon$. More importantly, we can now predict category goodness ratings for each $\mathbf{f} \in \textsc{mvs}(a)$ by applying the additive model to $\mathbf{F}^a$. Given $\mathbf{f}^k \in \textsc{mvs}(a)$, let γ_k be its predicted goodness value under the additive model derived from $R_{\mathsf{c}}(s^\ell_\tau, a)$ and $\mathbf{y}_{\mathsf{c}}$, and let $r^k = (\mathsf{c}, \gamma_k)$. Pairing each $\mathbf{f}^k$ with r^k, we can define

$$R'_{\mathsf{c}}(s^\ell_\tau, a) : \textsc{mvs}(a) \rightarrow C_\ell \times \textsc{vas},$$

which approximates an extension of $R_{\mathsf{c}}(s^\ell_\tau, a)$ from $\textsc{p}(a)$ to all of $\textsc{mvs}(a)$. The response surface value for a formant vector $\mathbf{f} \in \textsc{mvs}(a)$ is meant to approximate s^ℓ_τ's VAS goodness rating of $\mathbf{f}$ as an example of vowel category c for age a.

Given an age $a \in \textsc{vlamages}$, a category $\mathsf{c}^\ell \in C_\ell$, and a subject s^ℓ_τ, where $1 \leq \tau \leq n_\ell$, the function $R'_{\mathsf{c}}(s^\ell_\tau, a)$ is called a *vowel category response surface for s^ℓ_τ at age a*, or simply a *vowel category response surface* (VCRS). Figure 3.5 depicts the VCRSs $R'_{\mathsf{u}}(s^G_{12}, 10)$ and $R'_{\mathsf{u}}(s^J_{20}, 10)$ where the response category extension parameter $\alpha = 0.5$.

We use a similar method for projecting an ICRF $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)$ to an arbitrary maximal vowel space $\textsc{mvs}(a_0)$, where a_0 is no longer assumed to fall within $\textsc{vlamAges}$. The primary reason for this is that we adopt the use of a modern version of the VLAM, called the "Vlab" (see Section 4.2.2), whose MVSs lack perceptual categorization results. The set of possible vocal tract ages that the Vlab is capable of simulating is denoted $\textsc{vlabAges}$, and the definition of an MVS is extended to include those yielded by the Vlab. The basic idea is to combine the regression technique specified above with a range normalization over the prototypes $\textsc{p}(a)$ and $\textsc{mvs}(a_0)$. Specifically, given an ICRF $R_{\mathsf{c}}(s^{\ell}_{\tau}, a)$ and a maximal vowel space $\mathsf{mvs}(a_0)$ for an arbitrary age setting a_0 of the Vlab, we minmax range normalize the set $\mathsf{mvs} \cup \mathbf{p}(a)$, and regress a vowel category response surface onto the range normalized formant vectors from $\textsc{mvs}(a_0)$. Let $\mathbf{r}^k$ be the range normalized formant vector corresponding to the formant vector $\mathbf{f}^k \in \textsc{mvs}$. The *projected* response surface value assigned to $\mathbf{f}^k$ for a vowel category c is the vowel category response surface value assigned to $\mathbf{r}^k$. Carrying this out for each $\mathbf{f}^k \in \textsc{mvs}$, we can define a function

$$P'_{\mathsf{c}}(s^{\ell}_{\tau}, a, a_0) : \textsc{mvs}(a_0) \to C_{\ell} \times \textsc{vas}.$$

Given ages $a \in \textsc{vlamAges}$ and $a_0 \in \textsc{vlabAges}$, a category $\mathsf{c}^{\ell} \in C_{\ell}$, and a subject s^{ℓ}_{τ}, where $1 \leq \tau \leq n_{\ell}$, the function $P'_{\mathsf{c}}(s^{\ell}_{\tau}, a, a_0)$ is called a *projected vowel category response surface for s^{ℓ}_{τ} from age a to a_0*, or simply a *projected vowel category response surface* (VCRS). Figure 3.6 depicts the PVCRSs $P'_{\mathsf{u}}(s^{G}_{12}, 10, 10)$ and $P'_{\mathsf{u}}(s^{J}_{20}, 10, 10)$, with response category extension parameter $\alpha = 0.5$. Each PVCRS is taken to be a VCRS, thus methods defined for VCRSs in the sections that follow apply to PVCRSs as well. Moreover, to simplify presentation, we adopt the notation $Q^{\ell}(\mathsf{c}, a)$ for a VCRS for category c from language ℓ over a maximal vowel space $\textsc{mvs}(a)$, with details on its provenance (e.g., subject, age in $\textsc{vlamAges}$, projection properties, etc.) provided only as needed.

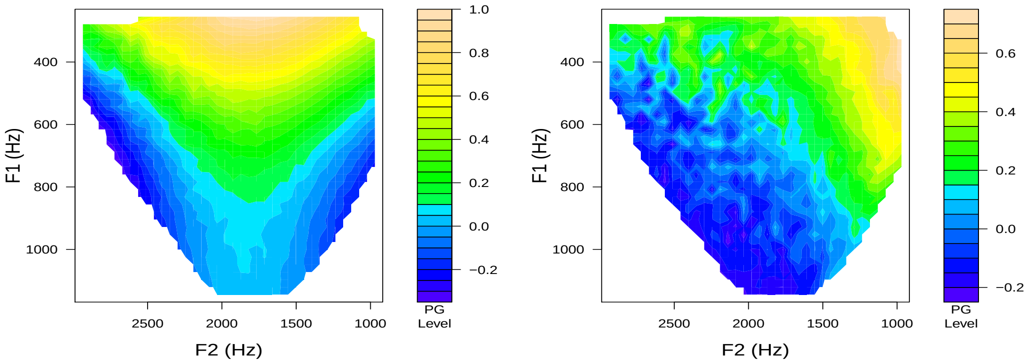


Figure 3.6: Projected vowel category response surfaces with response category extension parameter $\alpha = 0.5$, for $P'_{\mathsf{u}}(s^J_{20}, 10, 10)$ (left) and $P'_{\mathsf{u}}(s^G_{12}, 10, 10)$ (right).

3.2.3 Response Surface Computations

We now define a method for comparing VCRS patterns within and across language communities. Let $\ell_1, \ell_2 \in \textsc{lang}$. Given categories $\mathsf{c}_1 \in C_{\ell_1}$ and $\mathsf{c}_2 \in C_{\ell_2}$, consider the corresponding VCRSs $Q^{\ell_1}(\mathsf{c}_1, a)$ and $Q^{\ell_2}(\mathsf{c}_2, a)$. We begin by defining similarity between $Q^{\ell_1}(\mathsf{c}_1, a)$ and $Q^{\ell_1}(\mathsf{c}_2, a)$ in a simple manner. Let $\mathbf{z}^{\mathsf{c}_1}$ be the vector whose kth component is the predicted VAS value from the response $Q^{\ell_1}(\mathsf{c}_1, a)(\mathbf{f}^k) = (\mathsf{c}_1, \gamma_k)$, i.e., γ_k. Construct $\mathbf{z}^{\mathsf{c}_2}$ in similar fashion. The *distance between* $Q^{\ell_1}(\mathsf{c}_1, a)$ and $Q^{\ell_1}(\mathsf{c}_2, a)$ is

$$||\mathbf{z}^{\mathsf{c}_1} - \mathbf{z}^{\mathsf{c}_2}||_1 =_{\text{def}} \sum_{k=1}^{|\textsc{mvs}(a)|} abs(z^{\mathsf{c}_1}_k - z^{\mathsf{c}_2}_k)$$

where $z^{\mathsf{c}_1}_k$ and $z^{\mathsf{c}_2}_k$ are the k components of $\mathbf{z}^{\mathsf{c}_1}$ and $\mathbf{z}^{\mathsf{c}_2}$, respectively, and $|\textsc{mvs}(a)|$ the cardinality of $\textsc{mvs}(a)$.

Distance between VCRSs can be used to reason about differences in vowel category perception within a given language community (the case where $\ell_1 = \ell_2$ and $\mathsf{c}_1 = \mathsf{c}_2$), as well as differences across communities (the case where $\ell_1 \neq \ell_2$). Plummer et al. (2013a) present an application involving a comparison of VCRS patterns concerning the point vowels [i], [a], and [u] within and across the Greek (G) and Japanese (J) language communities. The statistical methodology was shown to be useful in capturing both categorical and sociophonetic differences between the vowel systems of the different language communities. Importantly, it was sensitive enough to pick up on the vowel category differences between two vowel systems that have roughly the same set of categories, and quantify the corresponding differences in vowel category perception. In light of this, we next define a method which serves as the basis for the modeling "vocal interaction" within our vocal learning environment.

Given $\ell \in \textsc{lang}$, a category $\mathsf{c} \in C_\ell$, and ages $a_0, a_1 \in \textsc{vlabAges}$, consider the corresponding VCRSs $Q^\ell(\mathsf{c}, a_0)$ and $Q(\mathsf{c}, \ell, a_1)$. A *response pairing over* $\textsc{mvs}(a_0)$ *and* $\textsc{mvs}(a_1)$ *for* c, denoted $T_\mathsf{c}(a_0, a_1)$ is a set of ordered pairs $(\mathbf{f}^j, \mathbf{f}^k)$ where $\mathbf{f}^j \in \textsc{mvs}(a_0)$ and $\mathbf{f}^k \in \textsc{mvs}(a_1)$. Given $T_\mathsf{c}(a_0, a_1)$, a *category transfer function over* $T_\mathsf{c}(a_0, a_1)$ is a function

$$C(T_\mathsf{c}(a_0, a_1)) : T_\mathsf{c}(a_0, a_1) \to \{\mathsf{c}\} \times \mathbb{R}$$

which assigns to each $(\mathbf{f}^j, \mathbf{f}^k) \in T_\mathsf{c}(a_0, a_1)$ an ordered pair (c, g) where the *transfer weight* g is a nonnegative function of the response surface values $Q^\ell(\mathsf{c}, a_0)(\mathbf{f}^j)$ and $Q^\ell(\mathsf{c}, a_1)(\mathbf{f}^k)$. Given $\ell \in \textsc{lang}$, and a set $\{T_\mathsf{c}(a_0, a_1)\}_{\mathsf{c} \in C_\ell}$ of response pairings $T_\mathsf{c}(a_0, a_1)$ over $\textsc{mvs}(a_0)$ and $\textsc{mvs}(a_1)$ with category transfer functions $\{C(T_\mathsf{c}(a_0, a_1))\}_{\mathsf{c} \in C_\ell}$, a *response pairing over* $\textsc{mvs}(a_0)$ *and* $\textsc{mvs}(a_1)$ for language ℓ is defined as $T(a_0, a_1) = \bigcup \{T_\mathsf{c}(a_0, a_1)\}_{\mathsf{c} \in C_\ell}$, and the corresponding *category transfer function over* $T(a_0, a_1)$ is defined as $C(T(a_0, a_1)) = \bigcup \{C(T_\mathsf{c}(a_0, a_1))\}_{\mathsf{c} \in C_\ell}$. We use response pairings to model turn-taking vocal exchanges that take place between infants and caretakers during early spoken language acquisition.

3.2.4 Caretaker Structures

Within our vocal learning environment, caretaker agents, or simply caretakers, are modeled as structures (Q, T) where Q is a set of VCRSs, and T is a set of response pairings over the VCRSs in Q.

Definition 3.1 (Caretaker Agent)**.** Given $\ell \in \textsc{Lang}$ and $a_0 \in \textsc{VlabAges}$, a *caretaker of age* a_0 *from language community* ℓ, is a structure $c^{\ell}_{a_0} = (Q, T)$, where Q is a set of VCRSs $Q^{\ell}(\mathtt{c}, a)$ over $\textsc{mvs}(a)$ for each category in C_{ℓ} and $a \in \textsc{VlabAges}$. The age a_0 identifies $\textsc{mvs}(a_0)$ as the caretaker's own MVS. The set T is accordingly composed of response pairings $T(a_0, a)$ over $\textsc{mvs}(a_0)$ and $\textsc{mvs}(a)$ for $a \in \textsc{VlabAges}$, together with category transfer functions $C(T(a_0, a))$.

The basic idea behind our language-specific, and moreover caretaker-specific model of vowel category transfer is as follows: suppose an infant agent "produces" a formant vector $\mathbf{f} \in \textsc{mvs}(a)$ which is "perceived" by a caretaker $c^{\ell}(a_0)$. If the VCRS value $Q^{\ell}(\mathtt{c}, a)(\mathbf{f})$ is high enough for some $\mathtt{c} \in C_{\ell}$, the caretaker may respond with a formant vector $r(\mathbf{f}) \in \textsc{mvs}(a_0)$ where $Q(\mathtt{c}, \ell, a_0)(r(\mathbf{f}))$ is also high. A response pairing over $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$ derived from $Q^{\ell}(\mathtt{c}, 0.5)$ and $Q^{\ell}(\mathtt{c}, 10)$ for subject s^{J}_{20} is depicted in Figure 3.7, which we denote $T^{J}_{20}(10, 0.5)$. The top row depicts $\textsc{mvs}(0.5)$ (left) and $\textsc{mvs}(10)$ (right), and gives a gross indication of the formant vectors that comprise the response pairing. The middle and bottom rows depict the pairing in greater detail. The left most column depicts the pairs for vowel category I, with the pairing indicated by the numbering. The middle column depicts the pairs for vowel category A, and the right most column depicts for vowel category U. The corresponding category transfer function values are indicated by color

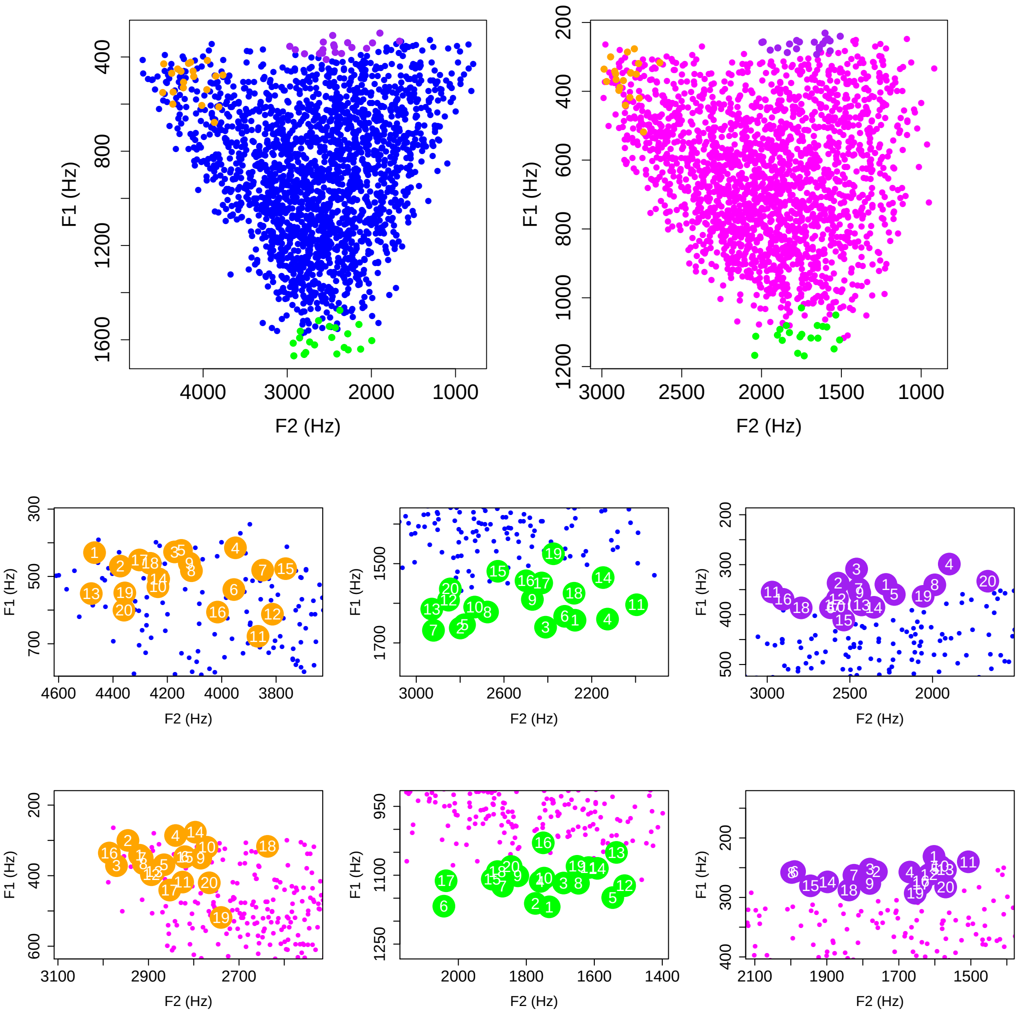


Figure 3.7: A response pairing (middle and bottom rows) over MVS(0.5) (top, left) and MVS(10) (top, right) derived from $Q^J(\mathsf{c}, 0.5)$ and $Q^J(\mathsf{c}, 10)$ for subject s^J_{20} ($\mathsf{c} \in \{$i,a,u$\}$).

(paired with a constant weight of 1). In the next section, we model an infant's interpretation of a response structure, and its influence on the acquisition of vowel normalization.

3.3 Creating and Aligning Auditory Manifolds

In this section, we model infant agents in terms of the structures they use for organizing vocal interaction with caretakers and the internal computations carried out over their representations of acoustic and social signals. Specifically, we model an infant's internalization of the acoustic and social signals provided by adult caretaker responses in terms of caretaker response pairings, while the broader organization of representations is modeled using manifolds. We model an infant's internal representation of acoustic signals from a caretaker's productions as well as their own in terms of a model of the infant's auditory system.

The model of an infant's auditory system is essentially a sequence of transformations acting on acoustic representations of vowel signals. Within this approach, the outer ear is typically modeled as a fixed filter that modifies a sound wave as it moves from a free field to the auditory canal (see Shaw, 1974). The middle ear is typically modeled as a fixed filter that modifies a sound wave as it moves from the auditory canal to the inner ear (Moore et al., 1997). Since the models are intended for adult auditory systems, and since they are constant across acoustic input, we omit them from our infant auditory model. Modeling the action of the inner ear and auditory nerve is somewhat more complicated. In this dissertation we will keep to a simplified model based on characteristic frequencies of regions of the basilar membrane. The biophysical and psychoacoustic bases for the model, along with its technical formulation, are presented in Reidy (2013), thus we keep to the main computational points concerning frequency sensitivity, and frequency scale compression. We assume familiarity with basic signal processing terminology.

3.3.1 Auditory Modeling

It has long been understood that modeling of speech perception must take into account the effects of the auditory system on acoustic signals (see Tonndorf, 1981, for a selection of classic papers). Recently, the effects have been argued to factor into spoken language acquisition (e.g., Holliday et al., 2010; Kallay and Holliday, 2012). In this connection, we use an auditory model that reflects these effects as our model of the internalization of acoustic signals.

We first review a bit of anatomy and physiology of the auditory system. The outer ear is composed of the pinna and the auditory canal, meatus, or ear canal. The outer ear significantly modifies an incoming sound, especially at high frequencies, and is important in our ability to localize sound. The middle ear is composed of the ear drum, malleus, incus, and stapes. The primary function of the middle ear is to transfer sound from the air in the auditory canal to the fluids in the cochlea. As the ear drum vibrates in response to incoming sound, the malleus and incus cause the stapes to make contact with an 'oval window' in the cochlea, efficiently transferring the sound from the outer to inner ear.

Turning to the inner ear, the cochlea is divided along its length by two membranes, Reissner's membrane and the basilar membrane. The start of the cochlea, where the oval window is situated, is called the base, while the other end is called the apex. An incoming sound is transferred through the oval window by the stapes, and travels through the scala vestibuli. As the incoming sound travels through the scala vestibuli, causing Reissner's membrane and the basilar membrane to vibrate. The sound then exits through the scala tympani and finally out the round window. The organ of Corti is composed of the basilar membrane, the tectorial membrane, outer and inner hair cells, and the auditory nerve, which contains auditory neurons. The Organ of Corti transduces pressure waves to action potentials of auditory neurons. As the basilar membrane vibrates in response to incoming sound, the hair cells press against the tectorial membrane and cause the neurons in the auditory nerve to fire.

Taking the inner ear as our point of departure, von Békésy (1947) showed that regions of the basilar membrane vibrate differently in response to a simple harmonic wave. Given a simple harmonic wave with frequency f, each region on the basilar membrane for which there is movement in response to the wave oscillates with frequency f. Yet, some regions vibrate with amplitude greater than others. Given a region along the basilar membrane, the frequency that yields the greatest amplitude at that region is called the *characteristic frequency* of that region. Let r be a region along the basilar membrane with characteristic frequency f_r. Regions very near to r respond to f_r with amplitudes close to that at which r responds to f_r, whereas regions far from r respond with much lower amplitudes.

Letting $\mathcal{B}_{f_r}$ denote a bandpass filter whose center frequency is f_r, we model the basilar membrane as a filter bank:

$$\mathcal{B} = \{\mathcal{B}_{f_r} \mid r \text{ is a region on the basilar membrane}\}.$$

Each filter $\mathcal{B}_{f_r}$ is itself modeled using an *equivalent rectangular bandwith (ERB) filter* – a filter that is perfectly rectangular, whose bandwidth is the width of the rectangle, scaled to have the same height and area of $\mathcal{B}$, whose center frequency is the center point of the bandwidth, assumed to be the center frequency of $\mathcal{B}_{f_r}$. Note that a single ERB filter may correspond to different bandpass filters, leaving us free to select from its set of corresponding filters as more is learned about the basilar membrane.

In order to specify the ERB of a given characteristic frequency f_r, Patterson (1976) derives a "critical band" for the region r along the basilar membrane that is most responsive to frequencies near f_r. The length of the critical band is taken to be the bandwidth of a bandpass filter whose center frequency is f_r, which is derived using the following function:

$$\mathsf{ERB}(f) = 0.107939(2\pi f) + 24.7.$$

Given a set C of n characteristic frequencies corresponding to regions along the basilar membrane, we can specify an n-channel filter bank:

$$\{\mathcal{B}_{\mathsf{ERB}(f_r)} \mid f_r \in C\}.$$

To derive a set of n characteristic frequencies that adequately covers the basilar membrane, we make use of the following function:

$$\mathsf{ERB}_{\mathsf{num}}(f) = \frac{1}{0.107939} \log_{10}\left(\frac{2\pi f}{0.00437} + 1\right).$$

This function carries the compression of the frequency scale, while matching frequencies with physical locations along the basilar membrane. Based on the normal frequency range of human we fix our minimum and maximum $\mathsf{ERB}_{\mathsf{num}}$s at 3 and 39, respectively. We then choose $n - 2$ points between these values such that they are evenly spaced. Call the collection of these points, together with the 3 and 39, $\mathsf{ERB}_{\mathsf{num}}(C)$. For each $c \in \mathsf{ERB}_{\mathsf{num}}(C)$, we compute $f_r = \mathsf{ERB}^{-1}_{\mathsf{num}}(c)$, and call the set of f_r frequencies C. We thereby derive an n-channel *ERB filter bank*

$$\mathcal{E} = \{\mathcal{B}_{\mathsf{ERB}(f_r)} \mid f_r \in C\}.$$

Specifying a filter bank this way allows us to choose from a variety of implementations. The *gamma function* is defined below, as an extension of the factorial function (translated 1 unit horizontally) to the whole of the complex plane.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

A *gammatone* is defined as $\gamma(t) = am(t)\cos((2\pi f)t + \varphi)$, where $m(t)$ is an *amplitude modulator* given as

$$m(t) = c\left(\tfrac{t}{\beta}\right)^{n_\gamma - 1} exp\left(\tfrac{-t}{\beta}\right).$$

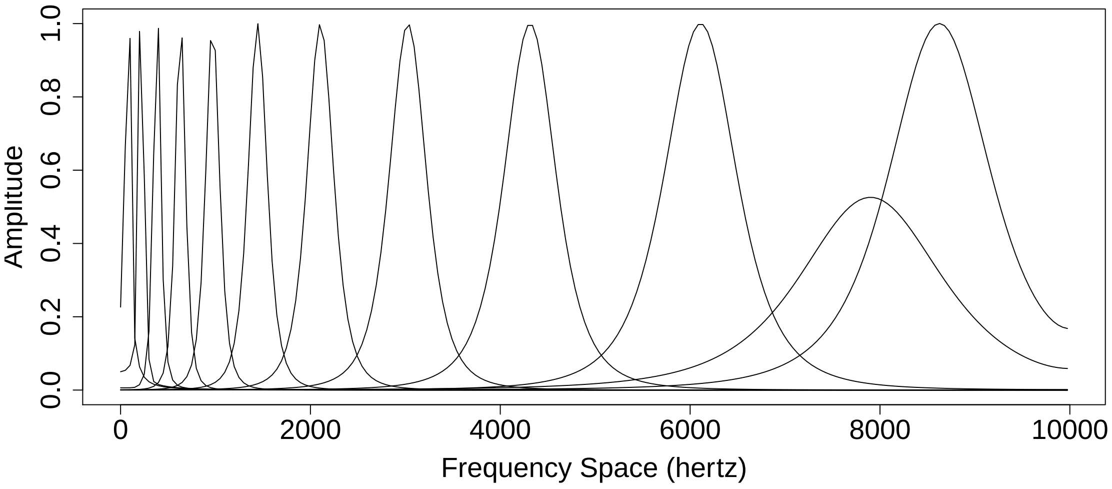


Figure 3.8: Subset of the gammatone filter bank $\mathcal{G}_{\mathsf{ERB}(C)}$ with 36 channels, which is taken as a model of the basilar membrane.

Typically, a gammatone is given as an impulse response in the time domain, where b is a bandwidth, n_γ is the filter order, a_γ is the amplitude, and φ is the phase:

$$\gamma(t) = a_\gamma t^{n_\gamma - 1} e^{-2\pi b t} \cos((2\pi f)t + \varphi).$$

We take $\varphi = 0$, $a_\gamma = 1$, and $n_\gamma = 4$ to be fixed throughout. Given an n-channel ERB filter bank $\mathcal{E}$, we derive an n-channel *gammatone filter bank* $\mathcal{G}_{\mathsf{ERB}(C)}$ whose filters are the frequencies responses of $\gamma(t)$ with $b = \mathsf{ERB}(f_r)$ for each $\mathcal{B}_{f_r} \in \mathcal{E}$. A subset of a 36-channel gammatone filter bank is depicted in Figure 3.8. Gammatone filter banks are physiologically appropriate (de Boer, 1973; Aertsen and Johannesma, 1980; Carney and Yin, 1988), psychoacoustically useful (Patterson and Moore, 1986; Glasberg and Moore, 1990), and computationally convenient for modeling the frequency sensitivity of the basilar membrane.

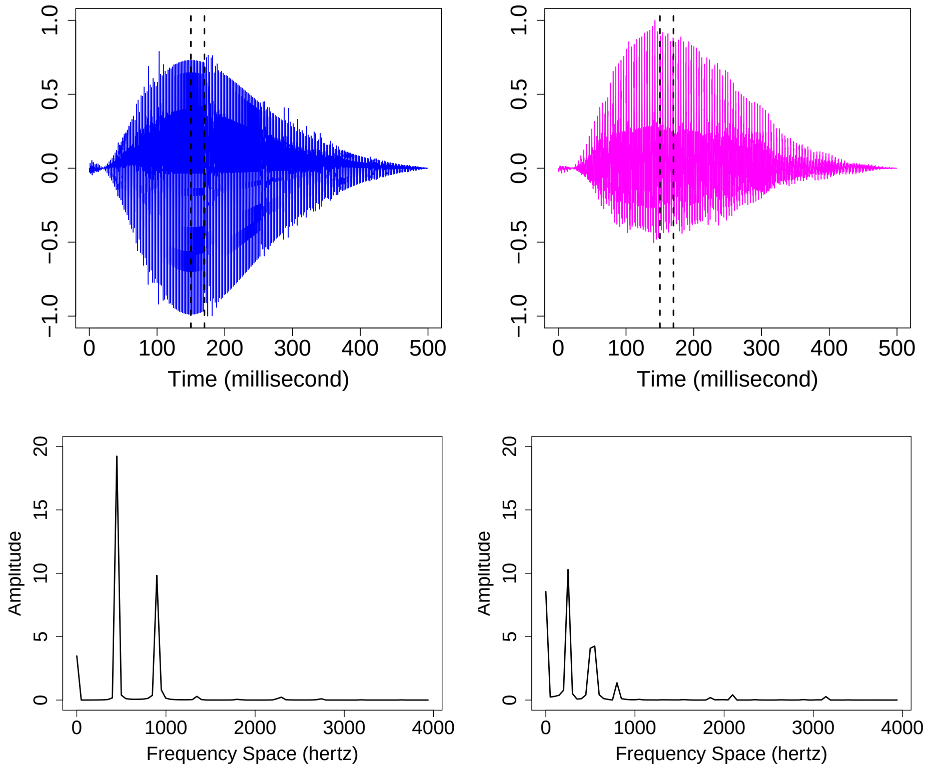


Figure 3.9: Time (top) and frequency (bottom) domain representations of infant (left) and adult (right) Vlab neutral vocal tract productions $s(0.5)$ and $s(10)$. Cursors on the upper graph demarcate the 20ms analysis window.

Now that we have a model of the basilar membrane we can specify its effect on vowel signals. We exemplify using the vowel signals $s(0.5)$ and $s(10)$, depicted in the time (top) and frequency (bottom) domains in Figure 3.9. The vowel signals are "produced" by the Vlab articulatory synthesizer model of the "neutral vocal tract" configuration for an infant (left) and adult caretaker (right). Details on Vlab vowel signal synthesis are specified in Section 4.2.3. Given a vowel signal s, a *time slice* from s is simply a segment of s over some connected interval of the time domain of s. A *spectral representation* of a vowel signal s is taken to be a multitaper spectrum of a time slice from s (see Reidy, 2013).

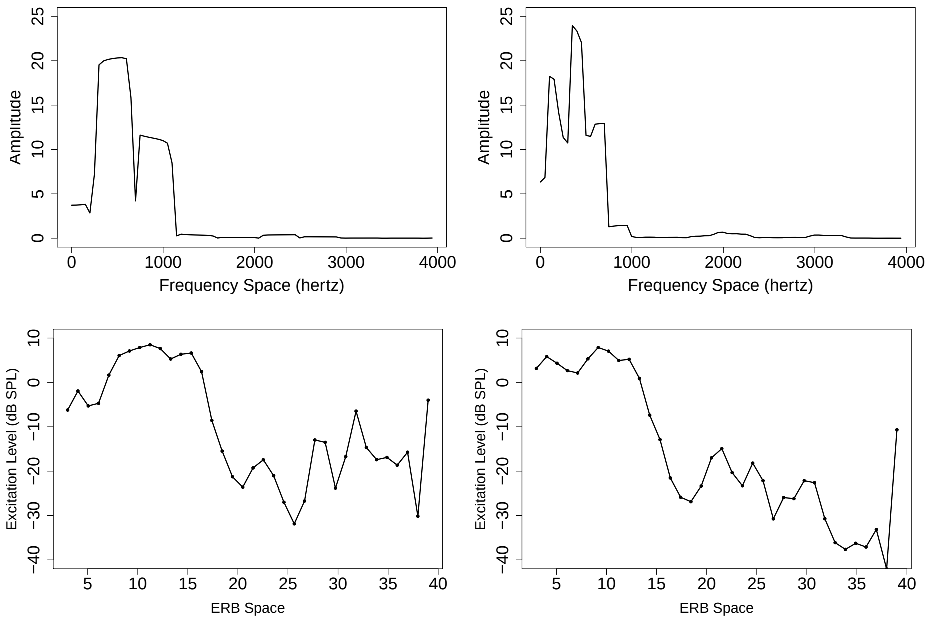


Figure 3.10: Multitaper spectral (top) and excitation pattern (bottom) representations of infant (left) and adult (right) Vlab neutral vocal tract productions $s(0.5)$ and $s(10)$.

The standard frequency domain representations of $s(0.5)$ and $s(10)$ in Figure 3.9 (bottom), called "periodograms," are single-taper spectral representations derived from the time slices between the dashed lines in Figure 3.9 (top). In general, periodograms are multitaper spectra using one taper. The spectral representations of $s(0.5)$ and $s(10)$ in Figure 3.10, derived from the same time slices, are multitaper spectra using eight tapers. In the remainder of this chapter all spectral representations are obtained in this fashion.

In this dissertation, we use a fixed gammatone filter bank $\mathcal{G}_{\mathsf{ERB}(C)}$ with 36 channels as a model of the basilar membrane, a subset of which is depicted in Figure 3.8. A filter bank is *applied* to a spectral representation of s through mulitplication of the representation by each filter, and then summing the resulting functions. The $\log$ output of a gammatone filter bank $\mathcal{G}_{\mathsf{ERB}(C)}$ applied to a spectral representation of a vowel signal s is called an *excitation* *pattern* for s under $\mathcal{G}_{\mathsf{ERB}(C)}$. Excitation pattern representations of vowel signals $s(0.5)$ and $s^(10)$ are shown in Figure 3.10 (bottom). The y-axis values of the 36 dots in the bottom figures constitute the 36-component *excitation vectors* for $s(0.5)$ and $s(10)$, respectively. In the remainder of this chapter all excitation vectors are obtained in this fashion.

Given a formant vector $\mathbf{f}^k \in \textsc{mvs}(a)$, we denote its corresponding excitation vector as $\mathbf{e}^k$. More generally, for each maximal vowel space $\textsc{mvs}(a)$, define the corresponding *maximal auditory space* as follows: $\textsc{mauds}(a) = \{\mathbf{e}^k \mid \mathbf{f}^k \in \textsc{mvs}(a)\}$. Maximal auditory spaces are assumed to be embedded within an *auditory reference frame*, taken to be the Euclidean space $\mathbb{R}^{36}$. The next section focuses on the cognitive organization of maximal auditory spaces into manifolds, and their alignment based on caretaker responses.

3.3.2 Auditory Manifolds, Pairings, and Structural Computations

Given our models of caretakers, we now turn to the infant side of vowel category transfer. Within our vocal learning environment, infant agents, or simply infants, are modeled in terms of cognitive structures used to organize auditory representations and caretaker responses. The main kind of structure used is called a *manifold*, as discussed in Chapter 2 and defined mathematically in Section 2.3. Familiarity with the definitions is assumed throughout the remainder of this chapter. The other kind of structure is called a *pairing*, which is simply a set of ordered pairs. Both kinds of structures are discussed below.

Given a maximal auditory space $\textsc{mauds}(a)$, an *auditory manifold over* $\textsc{mauds}(a)$, denoted $M(a) = (V(a), E(a), w(a))$, is simply a weighted graph derived from $\textsc{mauds}(a)$. We assume an indexing on $V(a)$ where the vertex $v_k \in V(a)$ corresponds to $\mathbf{e}^k \in \textsc{mauds}(a)$. For simplicity, weight functions are assumed to be constant, assigning a value of 1 to each edge of an auditory manifold, unless otherwise stated. Auditory manifolds model an infant's cognitive organization of both internalized acoustic signals, and auditory representations derived from an internal model (Wolpert et al., 1995). They form the basis of our model of an intramodal auditory vowel normalization.

Given a response pairing $T(a_0, a_1)$ over $\textsc{mvs}(a_0)$ and $\textsc{mvs}(a_1)$ for $\ell \in \textsc{lang}$ with category transfer function $C(T(a_0, a_1))$, we define an *auditory pairing* $I(T(a_0, a_1))$ as a set of ordered pairs $(\mathbf{e}^j, \mathbf{e}^k)$ where $(\mathbf{f}^j, \mathbf{f}^k) \in T(a_0, a_1)$. Auditory pairings model an infant's internalization of the response pairings separated out by a caretaker during turn-taking vocal exchanges. They are also the locus for vowel category acquisition and the basis for the perceptual magnet effect within the model. In this connection, we need the following definitions.

We define a *socio-auditory weighting* over $I(T(a_0, a_1))$ to be a nonnegative function

$$S(C(T(a_0, a_1))) : I(T(a_0, a_1)) \to \mathbb{R}.$$

Each auditory pair $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$ derives from a pair $(\mathbf{f}^j, \mathbf{f}^k) \in I(T(a_0, a_1))$, whence $(\mathtt{c}, g) = C(T(a_0, a_1))(\mathbf{f}^j, \mathbf{f}^k)$, The *socio-auditory weight* $\iota(g)$ assigned to $(\mathbf{e}^j, \mathbf{e}^k)$ is a function of the transfer weight g. An auditory pairing with a socio-auditory weighting is called a *socio-auditory pairing*. Socio-auditory pairings model an infant's internalization of the acoustic and social signals provided by a caretaker during turn-taking vocal exchanges.

We define a *pre-categorical equivalence* over $I(T(a_0, a_1))$ as an equivalence relation over the pairs in $I(T(a_0, a_1))$. We also define a *category transfer interpretation* over $I(T(a_0, a_1))$ for a language ℓ to be a function

$$I(C(T(a_0, a_1))) : I(T(a_0, a_1)) \to C_\ell.$$

Each category transfer function $C(T(a_0, a_1))$ yields a *simple category transfer interpretation* over $I(T(a_0, a_1))$ whereby each auditory pair $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$ is assigned the category c where $(\mathsf{c}, g) = C(T(a_0, a_1))(\mathbf{f}^j, \mathbf{f}^k)$. An auditory pairing with a category transfer interpretation is called a *categorical auditory pairing*. A socio-auditory pairing with a category transfer interpretation is called a *socio-categorical auditory pairing*. Each category transfer interpretation for ℓ over $I(T(a_0, a_1))$ yields a pre-categorical equivalence over $I(T(a_0, a_1))$ in the obvious way. In this dissertation, we restrict our attention to pre-categorical equivalences derived from category transfer interpretations. However, this is a simplifying assumption.

We turn now to computations over manifolds. Let $M(a_0)$ and $M(a_1)$ be auditory manifolds, and let $I(T(a_0, a_1))$ be a socio-auditory pairing, which is assumed to be an alignment for MAUDS(a_0) and MAUDS(a_1). The socio-auditory weighting on $I(T(a_0, a_1))$ yields a weight function on the alignment relation derived from $I(T(a_0, a_1))$, which is used to form a combined weighted graph $M(a_0, a_1)$ from $M(a_0)$ and $M(a_1)$, called an *auditory commensuration manifold over* $M(a_0)$ *and* $M(a_1)$ *derived from* $I(T(a_0, a_1))$, or simply a *commensuration manifold*. The computation involving the combination of the manifolds $M(a_0)$ and $M(a_1)$ is called *auditory normalization*. That is, auditory normalization is a binary operation on manifolds, and can be viewed as a structure co-opting generative computation. To emphasize the importance of socio-auditory pairing, we notate the auditory normalization of manifolds $M(a_0)$ and $M(a_1)$ via a socio-auditory weighting $S(C(T(a_0, a_1)))$ as a mapping over triples:

$$\textsc{audNorm} : (M(a_0), M(a_1), S(C(T(a_0, a_1)))) \mapsto M(a_0, a_1).$$

The notation highlights the necessity of each of the cognitive structures $M(a_0)$, $M(a_1)$, and $I(T(a_0, a_1))$ in yielding the commensuration manifold $M(a_0, a_1)$. Auditory commensuration manifolds provide the basis for vowel category acquisition as well as the perceptual magnet effect, and we treat each in turn.

Let $M(a_0, a_1)$ be a commensuration manifold derived from $I(T(a_0, a_1))$, and consider the simple category transfer interpretation $I(C(T(a_0, a_1))$ over $I(T(a_0, a_1))$ for ℓ. Let $\textsc{mauds}(a_0, a_1) = \textsc{mauds}(a_0) + \textsc{mauds}(a_1)$ (i.e., their disjoint union), and let $L_M(a_0, a_1)$ be the graph Laplacian for $M(a_0, a_1)$. The graph Laplacian $L_M(a_0, a_1)$ provides the means for (i) extending the pre-categorical equivalence over $I(T(a_0, a_1))$ to all of $\textsc{mauds}(a_0, a_1)$, and (ii) deriving "warped representations" of the representations in $\textsc{mauds}(a_0, a_1)$. The former is achieved through the use of "manifold regularization" (Belkin et al., 2004, 2006), which spreads the pre-categorical equivalence to the whole of the commensuration manifold $M(a_0, a_1)$. A number of regularization algorithms are available, though, as a heuristic, we use the plearn algorithm available at `http://www.cse.ohio-state.edu/~mbelkin/algorithms/Laplacian.tar`. We notate the computation of equivalence via a category transfer interpretation $I(C(T(a_0, a_1)))$ as a mapping over ordered pairs:

$$\textsc{audEquiv} : (M(a_0, a_1), I(C(T(a_0, a_1)))) \mapsto \mathsf{equiv}(\textsc{mauds}(a_0, a_1)).$$

The equivalence relation $\mathsf{equiv}(\textsc{mauds}(a_0, a_1))$ is called a *categorical equivalence over* $\textsc{mauds}(a_0, a_1)$, or simply, a *categorical equivalence*. The notation highlights the necessity of each of the cognitive structures $M(a_0, a_1)$ and $I(C(T(a_0, a_1)))$ in yielding the equivalence over $\textsc{mauds}(a_0, a_1)$.

Concerning the derivation of warped representations, we need a few definitions. Given an auditory manifold $M(a)$ over $\textsc{mauds}(a)$, an *m-dimensional auditory warping of* the space $\textsc{mauds}(a)$, denoted $\textsc{warp}(a)$, is an m-dimensional eigenmap derived from $M(a)$. The jth row of $\textsc{warp}(a)$, denoted $\mathbf{w}^j$ is the *warped representation* of the excitation vector $\mathbf{e}^j \in \textsc{mauds}(a)$. Auditory warping also applies to auditory commensuration manifolds. Let $\textsc{warp}(a_0, a_1)$ be the m-dimensional eigenmap derived from $M(a_0, a_1)$, and let $\textsc{warp}(a_0)$ be the m-dimensional eigenmap of $\textsc{mauds}(a_0)$ with respect to $\textsc{warp}(a_0, a_1)$, and $\textsc{warp}(a_1)$ the m-dimensional eigenmap of $\textsc{mauds}(a_1)$ with respect to $\textsc{warp}(a_0, a_1)$. The jth row of $\textsc{warp}(a_0)$, denoted $\mathbf{w}^j$ is the *warped representation* of the excitation vector $\mathbf{e}^j \in \textsc{mauds}(a_0)$. Similarly, the kth row of $\textsc{warp}(a_1)$, denoted $\mathbf{w}^k$ is the warped representation of the excitation vector $\mathbf{e}^k \in \textsc{mauds}(a_1)$. We take Laplacian eigenmapping to be a operation on auditory manifolds called *auditory warping*, which is denoted as follows:

$$\textsc{aud}\mathrm{W}\textsc{arp} : (M(a_0, a_1), S(C(T(a_0, a_1))) \mapsto \textsc{warp}(a_0, a_1).$$

The notation reflects the fact that the warped representations in $\textsc{warp}(a_0, a_1)$ derive from the commensuration manifold $M(a_0, a_1)$. The warped representations exist within a *warped reference frame*.

3.3.3 Infant Structures

Within our vocal learning environment, infant agents, or simply infants, are modeled as structures $(M, I(T))$ where M is a set of manifolds, and $I(T)$ is a set of socio-auditory pairings over the manifolds in M, together with category transfer interpretations.

Definition 3.2 (Infant Agent). Given a caretaker $c^{\ell}_{a_0} = (Q, T)$, an *infant of age* a_1 *of* $c^{\ell}_{a_0}$ is a structure $i^{\ell}_{a_1} = (M, I(T))$ where M is a set of manifolds $M(a)$ over $\textsc{mauds}(a)$ for each $a \in \{a_0, a_1\}$. The age a_1 identifies $\textsc{mauds}(a_1)$ as the infant's own maximal auditory space. The set $I(T)$ is accordingly composed of socio-categorical auditory pairings $I(T(a, a_1))$, assumed to be alignments for $\textsc{mauds}(a_0)$ and $\textsc{mauds}(a_1)$, along with their corresponding category transfer interpretations $I(C(T(a_0, a_1)))$. We assume that the set of manifolds M inherits the binary operation $\mathsf{audNorm}$, and the operations $\mathsf{audEquiv}$ and $\mathsf{audWarp}$.

The basic idea behind the formulation is that the infant is internalizing the attempt at interaction made by the caretaker in their own attempt to establish commensuration across the auditory manifolds. Hence the auditory normalization computation the infant carries out is based on response pairings derived from the caretaker. In the next section, we show how formulating infants in this fashion provides a potential basis for a developmental model of vowel category acquisition and the perceptual magnet effect.

3.4 Vocal Learning Environment

In order to illustrate the definitions and concepts put forward in this chapter, we create a simple vocal learning environment. We provide visualization of each stage of model creation (when possible), along with interpretation and discussion of the modeling output. To begin with, the MVSs used in this section are of the form $\textsc{mvs}(a)$ for $a \in \textsc{vlabAges}$, and $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$ are shown in Figure 3.11.

For each $a \in \textsc{vlabAges}$, we create $\textsc{mvs}(a)$, as well as $\textsc{mauds}(a)$. For each language community $\ell \in \textsc{lang}$, and for each subject s^{ℓ}_{τ}, we create a caretaker $\ell^{\tau}_{10} = (Q^{\ell}_{\tau}, T^{\ell}_{\tau})$, modifying the caretaker notion slightly in order to emphasize the subject and language community the caretaker model derives from. We use age 10 for the caretaker since the age 10 prototypes were rated most like those of a young female adult by subjects who provided the vowel category judgements and goodness ratings (see Plummer et al., 2013b, for comparisons of age and gender ratings across language communities).

For each $\mathsf{c} \in C_{\ell}$ and $a \in \textsc{vlabAges}$, we take $Q^{\ell}_{\tau}(\mathsf{c}, a) = P'_{\mathsf{c}}(s^{\ell}_{\tau}, 10, a)$ (with vowel category extension parameter $\alpha = 0.5$) to be a VCRS in Q^{ℓ}_{τ}. Moreover, for each $\mathsf{c} \in C_{\ell}$

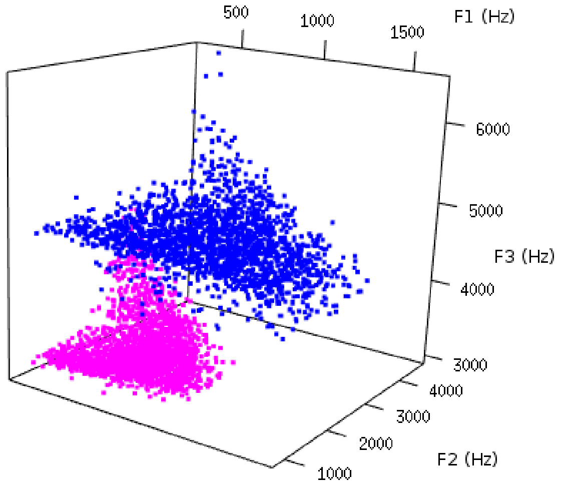


Figure 3.11: Approximated Vlab maximal vowel spaces $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$. The MVSs are depicted within a three-dimensional acoustic reference frame to emphasize their general lack of overlap.

and $a \in \textsc{vlabAges}$, let $b_\upsilon(\texttt{c}, a)$ be the set of υ formant vectors in $\textsc{mvs}(a)$ with the highest VCRS values under $Q^\ell_\tau(\texttt{c}, a)$, indexed from 1 (highest VCRS value) to υ (lowest VCRS value). Each response pairing $T^\ell_\tau(10, a)$ in T^ℓ_τ is a set of ordered pairs $(\mathbf{f}^j, \mathbf{f}^k)_i$, where $\mathbf{f}^j$ is the ith formant vector in $b_\upsilon(\texttt{c}, 10)$, and $\mathbf{f}^j$ the ith formant vector in $b_\upsilon(\texttt{c}, a)$, for each $\texttt{c} \in C_\ell$. Moreover, each response pairing $T^\ell_\tau(10, a)$ has a corresponding category transfer function $C(T^\ell_\tau(10, a))$.

For each caretaker $\ell^\tau_{10} = (Q^\ell_\tau, T^\ell_\tau)$, we create an infant $\ell^\tau_{0.5} = (M^\ell_\tau, I(T^\ell_\tau))$, where M contains the auditory manifolds $M(0.5)$ and $M(10)$ over $\textsc{mauds}(0.5)$ and $\textsc{mauds}(10)$, respectively. $M(0.5)$ and $M(10)$ are constructed using a k-nearest-neighbors computation, where $k = 20$, over $\textsc{mauds}(0.5)$ and $\textsc{mauds}(10)$. The set $I(T^{\ell}_{\tau})$ contains the socio-auditory pairing $I(T^{\ell}_{\tau}(10, 0.5)) = \{(\mathbf{e}^j, \mathbf{e}^k) \mid (\mathbf{f}^j, \mathbf{f}^k) \in T^{\ell}_{\tau}(10, 0.5)\}$, whose weight function is a positive constant set to 10. It is assumed that $I(T^{\ell}_{\tau}(10, 0.5))$ is also a categorical-auditory pairing with category transfer interpretation $I(C(T^{\ell}_{\tau}(10, 0.5)))$. These nearest-neighbors and socio-auditory weigthing parameter values are meant to yield comprehensible heuristics exemplifying the vocal learning environment. These parameters are subject to investigation, though quantitative evaluation of their role in modeling the acquisition of vowel normalization is beyond the scope of this dissertation.

3.4.1 Demonstrations

In order to make the modeling approach concrete, we demonstrate the "acquisition" of vowel normalization within the vocal learning environment for caretakers J^{20}_{10} and G^{12}_{10}, and their corresponding infants $J^{20}_{0.5}$ and $G^{12}_{0.5}$, focusing on the corner vowels $\{\mathsf{i},\mathsf{a},\mathsf{u}\}$. The figures corresponding to the demonstrations are presented at the end of the chapter. A larger sample of figures for infants and caretakers is presented in Appendix B.

Demonstration 3.4.1. We focus first on J^{20}_{10} and $J^{20}_{0.5}$, stepping through the entire acquisition procedure. We list the steps below, presenting the corresponding figures afterwards.

External Signals: The VCRSs $Q^{J}_{20}(\mathsf{c}, 0.5)$ and $Q^{J}_{20}(\mathsf{c}, 10)$ (where $\mathsf{c} \in \{\mathsf{i},\mathsf{a},\mathsf{u}\}$) for caretaker J^{20}_{10} are depicted in Figure 3.12. These VCRSs for J^{20}_{10} yield the response pairing $T^{J}_{20}(10, 0.5)$, which is depicted in Figure 3.7, along with its category transfer function $C(T^{J}_{20}(10, 0.5))$. The language- and dyad-specificiy of acquisition begin at this stage the procedure.

Internalization: The infant $J^{20}_{0.5}$ internalizes the response pairing $T^{J}_{20}(10, 0.5)$ along with its category transfer function $C(T^{J}_{20}(10, 0.5))$, yielding the socio-auditory pairing

$I(T_{20}^{J}(10,0.5))$ and the categorical transfer interpretation $I(C(T_{20}^{J}(10,0.5)))$. The necessary internalization of external vocal and social signals takes places at this stage.

Internal Computations: Having internalized $I(T_{20}^{J}(10,0.5))$, along with its socio-auditory weighting $S(C(T_{20}^{J}(10,0.5)))$, the infant computes the auditory normalization

$$\textsc{audNorm} : (M(10), M(0.5), S(C(T_{20}^{J}(10,0.5)))) \mapsto M(10,0.5)$$

over the "self" auditory manifold $M(0.5)$, and the "caretaker" auditory manifold $M(10)$. Importantly, normalization makes use of the socio-auditory weights only, and not the categories from $S(C(T_{20}^{J}(10,0.5)))$. The output $M(10,0.5)$ is a structure providing the means to (i) compute the "warping" of perception, and (ii) categorize the representations in $\textsc{mauds}(10,0.5))$.

Concerning (i), the socio-auditory weighting $S(C(T_{20}^{J}(10,0.5)))$, yields the following warping:

$$\textsc{audWarp} : (M(10,0.5), S(C(T_{20}^{J}(10,0.5))) \mapsto \textsc{warp}(10,0.5).$$

The warped representations in $\textsc{warp}(10,0.5)$ are depicted in Figure 3.13 in terms of their first three components. Concerning (ii), having also internalized the category transfer interpretation $I(C(T_{20}^{J}(10,0.5)))$, the infant computes the auditory equivalence:

$$\textsc{audEquiv} : (M(10,0.5), I(C(T_{20}^{J}(10,0.5)))) \mapsto \mathsf{equiv}(\textsc{mauds}(10,0.5)).$$

The equivalence relation over $\textsc{mauds}$(10,0.5) is depicted over the maximal vowel spaces $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$ in Figure 3.14. **QEF**

Demonstration 3.4.2. We next focus on G^{12}_{10} and $G^{12}_{0.5}$, again stepping through the entire acquisition procedure. We list the steps below, presenting the corresponding figures afterwards.

External Signals: The VCRSs $Q^G_{12}(\mathsf{c}, 0.5)$ and $Q^g_{12}(\mathsf{c}, 10)$ (where $\mathsf{c} \in \{\mathsf{i{,}a{,}u}\}$) for caretaker G^{12}_{10} are depicted in Figure 3.15. These VCRSs for G^{12}_{10} yield the response pairing $T^G_{12}(10, 0.5)$, which is depicted in Figure 3.16, along with its category transfer function $C(T^G_{12}(10, 0.5))$.

Internalization: The infant $G^{12}_{0.5}$ internalizes the response pairing $T^G_{12}(10, 0.5)$ along with its category transfer function $C(T^G_{12}(10, 0.5))$, yielding the socio-auditory pairing $I(T^G_{12}(10, 0.5))$ and the categorical transfer interpretation $I(C(T^G_{12}(10, 0.5)))$.

Internal Computations: Having internalized $I(T^G_{12}(10, 0.5))$, along with its socio-auditory weighting $S(C(T^G_{12}(10, 0.5)))$, the infant computes the auditory normalization

$$\textsc{audNorm} : (M(10), M(0.5), S(C(T^G_{12}(10, 0.5)))) \mapsto M(10, 0.5)$$

over the "self" auditory manifold $M(0.5)$, and the "caretaker" auditory manifold $M(10)$. The output $M(10, 0.5)$ yields a structure providing the means to compute the "warping" of perception, as well as categorize the representations in $\textsc{mauds}(10, 0.5))$.

Concerning the former, the socio-auditory weighting $S(C(T^G_{12}(10, 0.5)))$, yields the following warping:

$$\textsc{audWarp} : (M(10, 0.5), S(C(T^G_{12}(10, 0.5))) \mapsto \textsc{warp}(10, 0.5).$$

The warped representations in $\textsc{warp}(10, 0.5)$ are depicted in Figure 3.17 in terms of their first three components.

Concerning the latter, having also internalized the category transfer interpretation $I(C(T^G_{12}(10, 0.5)))$, the infant computes the auditory equivalence:

$$\textsc{audEquiv} : (M(10, 0.5), I(C(T^G_{12}(10, 0.5)))) \mapsto \mathsf{equiv}(\textsc{mauds}(10, 0.5)).$$

The equivalence relation over MAUDS(10,0.5) is depicted over $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$ in Figure 3.18. **QEF**

3.4.2 Discussion

We begin with a few points about specific aspects of Demonstrations 3.4.1 and 3.4.2 before proceding to more general discussion, using the former to illustrate. The same points apply to the entire vocal learning environment, though we mainly limit the discussion to the two demonstrations.

i) We conceive of the response pairing $T^J_{20}(10, 0.5)$ as a model of pairs derived from infant-caretaker interaction, though the derivation procedure is left unspecified. Since the pairing operation is commutative, it may be the case that the caretaker has responded to an infant vocalization or vice versa. That is, the order of vocal turn-taking is not an imposition of the model, though the addition of ordering may itself be an interesting line of inquiry.

ii) We have modeled the internalization of the category transfer function $C(T^J_{20}(10, 0.5))$, with the highest level of fidelity, e.g., every pair in $C(T^J_{20}(10, 0.5))$ has a corresponding pair in $I(T^J_{20}(10, 0.5))$. Yet, this in no way suggests that infants internalize every vocal interaction with their caretakers. The internalization of $C(T^J_{20}(10, 0.5))$ may easily be made to reflect the fact that infants do not internalize every such interaction.

iii) The infant internalizes the response pairing $T^J_{20}(10, 0.5)$ along with its category transfer function $C(T^J_{20}(10, 0.5))$, yielding the socio-auditory pairing $I(T^J_{20}(10, 0.5))$ and the categorical transfer interpretation $I(C(T^J_{20}(10, 0.5)))$. However, the auditory normalization computation makes use of the socio-auditory weights only, and not the categories from $S(C(T^J_{20}(10, 0.5)))$.

iv) The output structure $M(10, 0.5)$ providing the means to (i) categorize the representations in MAUDS$(10, 0.5))$, and (ii) compute the "warping" of perception, is a graph structure in its own right, independent of the representations used to construct it. In this sense, auditory normalization is not a computation over auditory, or other psychophysical representations, but a cognitive operation over cognitive structures.

We now turn to the output of Demonstration 3.4.1. The representations in WARP$(10, 0.5)$ yielded by the Laplacian eigenmapping AUDWARP$(M(10, 0.5), S(C(T^J_{20}(10, 0.5)))$ are depicted in Figure 3.13 in terms of their first three components. In computational practice, however, these representations have m-many components, where m is the number of excitation vectors involved in the auditory normalization computation (4000 in this version of the vocal learning environment). The "features" corresponding to the second and third components of the warped representations yielded by auditory normalization are predicted to provide the basis for an internal phonological system. The warped representations in WARP(10,0.5), are depicted in terms of these two components in Figure 3.13 (left), and Figure 3.17 (left).

Examination of Figures 3.13 and 3.17 suggests that the feature corresponding to the first component of each warped representation reflects the socio-auditory argument of the normalization computation. On this basis, we suggest a re-conceptualization of the perceptual magnet effect in terms of these warped representations, which adheres to the interpretations of (at least) their first three components. The re-conceptualization includes the spatial positioning of these representations within the warped reference frame, and the manifolds that are computed over these representations. The latter adds a structural aspect to the magnet effect, and the merit of this addition is discussed along with categorization below.

Regarding the former aspect of re-conceptualization, the auditory representations in the socio-auditory pairing $I(T^J_{20}(10, 0.5))$ serve as "perceptual magnets" for the remaining auditory representations in $\textsc{mauds}(10, 0.5)$, which are not components of pairs in $I(T^J_{20}(10, 0.5))$. Accordingly, auditory representations in $\textsc{mauds}(10, 0.5)$ close to representations in pairs in $I(T^J_{20}(10, 0.5))$ in terms of the metric over the auditory reference frame are expected to be even closer in terms of the distance between their corresponding warped representations in WARP(10,0.5) over the warped reference frame. Conversely, auditory representations in $\textsc{mauds}(10, 0.5)$ far from representations in pairs in $I(T^J_{20}(10, 0.5))$ in terms of the metric over the auditory reference frame are expected to be even farther in terms of the distance between their corresponding warped representations in WARP(10,0.5) over the warped reference frame. The warped representations in Figure 3.13 show the magnet effect beginning to take shape following normalization.

The capability of the model to capture the language-specificity of the magnet effect as conceptualized in terms of representations in WARP(10,0.5) is revealed by contrasting Figures 3.13 and 3.17. The same auditory manifolds are used in both Demonstration 3.4.1 and Demonstration 3.4.2 by infants $J^{20}_{0.5}$ and $G^{12}_{0.5}$, respectively, as models of self, and of caretakers J^{20}_{10} and G^{12}_{10}, respectively. The only difference between these demonstrations is in the category transfer functions $C(T^J_{20}(10, 0.5))$ and $C(T^G_{12}(10, 0.5))$, and their respective socio-auditory pairings $I(T^J_{20}(10, 0.5))$ and $I(T^G_{12}(10, 0.5))$. Notice that the warped representations corresponding to u in Figure 3.13 are more toward the inside of the point cloud than those in Figure 3.17.

Before moving on, it is important to state that, aside from language-specificity, Figures 3.13 and 3.17, together with those in Appendix B, also suggest that the normalization computation is specific to infant-caretaker dyads. In this connection, our model makes the clear prediction that the acquisition of vowel normalization is an individual property, being based on signals internalized through dyad-specific vocal interactions, and individualized manifold representations of the self and others. In other words, the model predicts that there can be individual differences in "phonological grammar" within a community, of the sort that sociolinguists study, as well as individual differences in "phonetic ability" of the sort that speech language pathologists study.

Concerning categorization, the equivalence relation $\mathsf{equiv}(\textsc{mauds}(10, 0.5))$ yielded by the computation $\textsc{audEquiv}(M(10, 0.5), I(C(T^J_{20}(10, 0.5))))$ is depicted in Figure 3.14 over representations in $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$. Similarly, the equivalence relation yielded by the computation $\textsc{audEquiv}(M(10, 0.5), I(C(T^G_{12}(10, 0.5))))$ is depicted in Figure 3.18. over representations in $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$. Figure 3.14 shows that $J^{20}_{0.5}$'s internalization of u reflects a higher F2 "location" within the acoustic reference relative to $G^{12}_{0.5}$'s, which is what we would predict, given the cross-language difference between /u/ being an unrounded vowel that is not as back as /o/ in most speaking styles for most Japanese speakers but a very back, very rounded vowel for Greek speakers. This suggests that the response pairing constructed by J^{20}_{10} during vocal exchange may result in a successful category transfer to $J^{20}_{0.5}$, with Figure 3.18 indicating the same for G^{12}_{10} and $G^{12}_{0.5}$.

In both demonstations, overall categorization is very coarse and the category boundaries bleary. Returning to the latter part of our re-conceptualization of the perceptual magnet effect – that involving manifolds computed over warped representations – we suggest that manifolds created over the warped representations in WARP(10,0.5) provide the infrastructure that facilitates a more refined and well-deliniated spread of categorization information. Put in these terms, we suggest that manifolds created over warped representations may be a better basis for categorization than manifolds over the auditory representations that warped representations are derived from, and in the sequel to this dissertation, we test this hypothesis. However, in either case, vowel category acquisition is carried out based on the language- and dyad-specificity of the acquisition of vowel normalization, whence the language- and dyad-specificity carries over to vowel category acquisition. If true, the individualistic nature of vowel category acquisition may percolate up the "ladder of abstraction," into phonological and syntactic acquistion, and likely beyond.

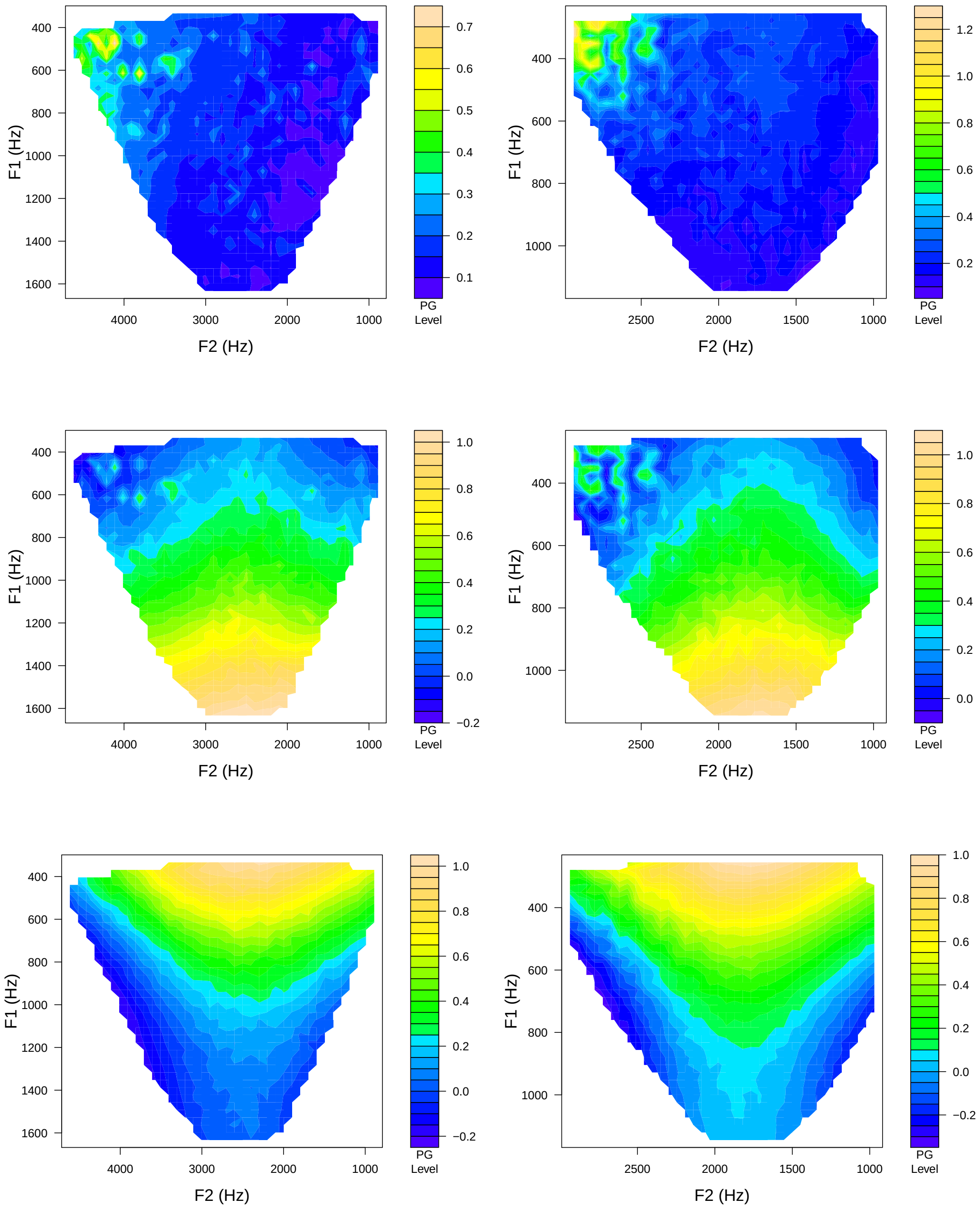


Figure 3.12: Vowel category response surfaces $Q^J_{20}(\mathsf{c}, 0.5)$ (left) and $Q^J_{20}(\mathsf{c}, 10)$ (right) for caretaker J^{20}_{10} derived from subject s^J_{20} ($\mathsf{c} \in \{$i,a,u$\}$).

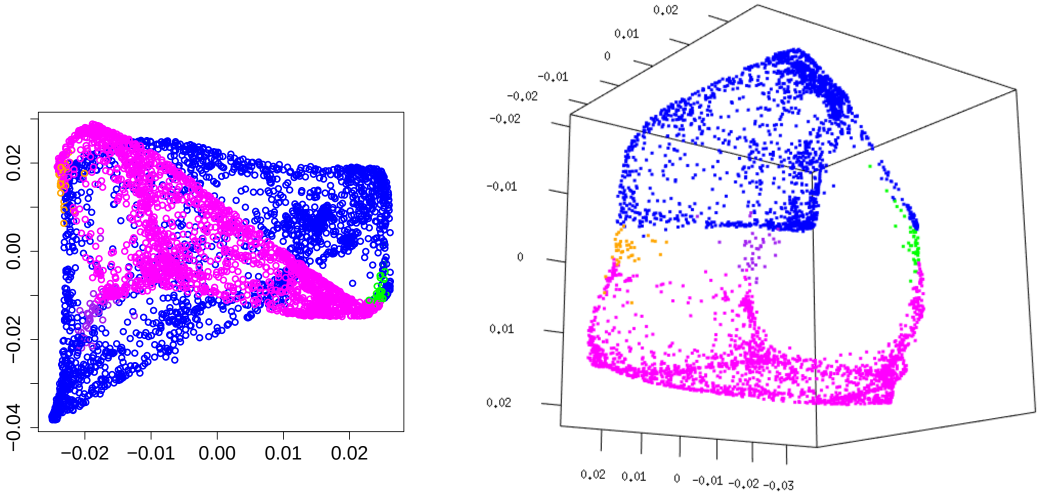


Figure 3.13: The warping yielded by socio-auditory weighting $S(C(T^J_{20}(10, 0.5)))$. In the two-component depiction (left) the u representations are in the lower left corner.

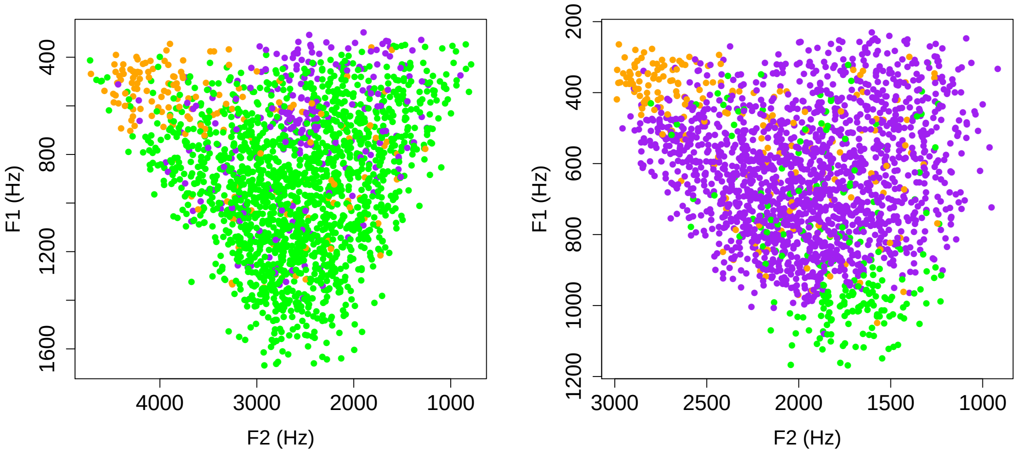


Figure 3.14: The auditory equivalence derived from the category transfer interpretation $I(C(T^J_{20}(10, 0.5)))$. The equivalence relation is depicted over the representations on MVS(10) and MVS(0.5).

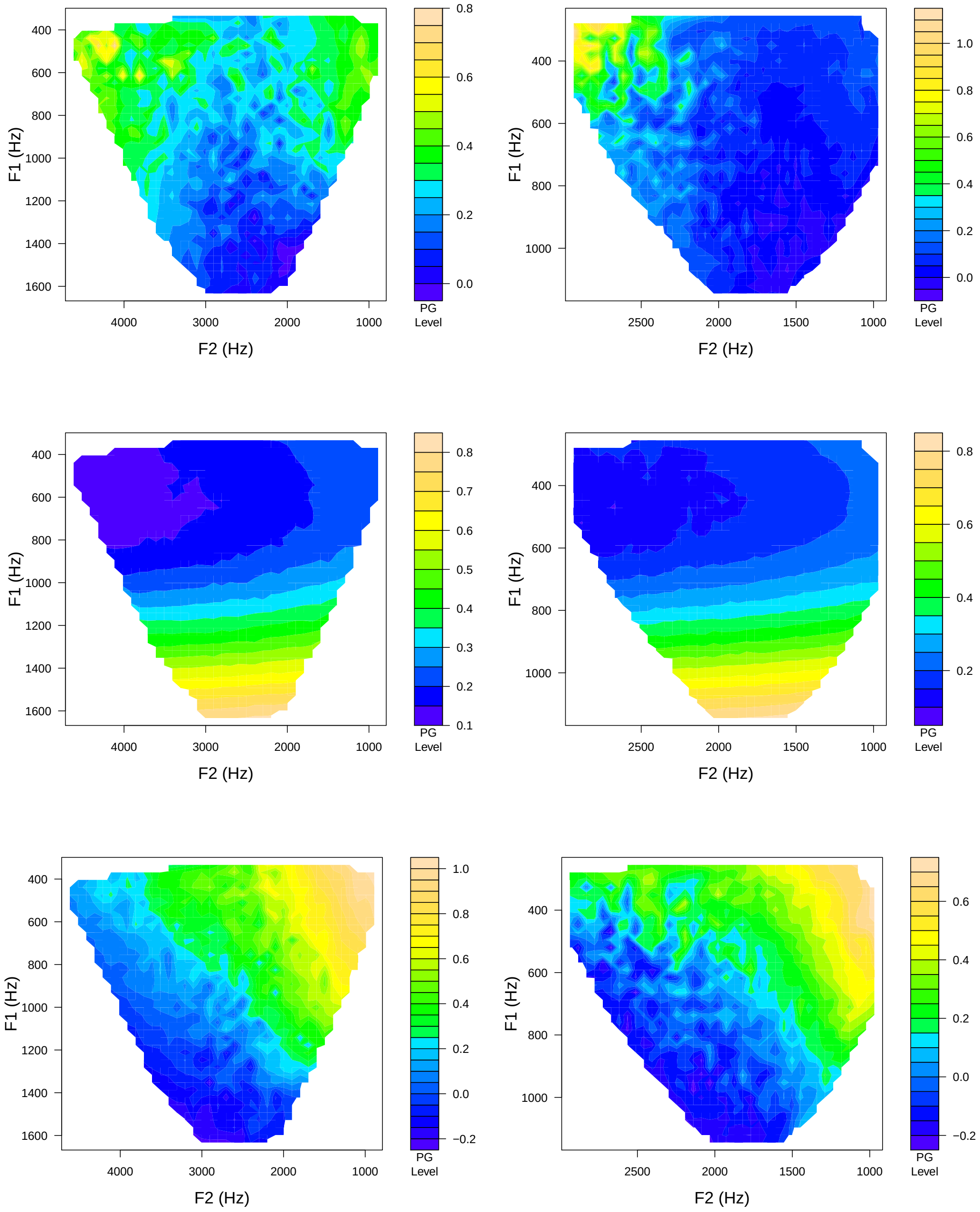


Figure 3.15: Vowel category response surfaces $Q^G_{12}(\mathsf{c}, 0.5)$ (left) and $Q^G_{12}(\mathsf{c}, 10)$ (right) for caretaker G^{12}_{10} derived from subject s^G_{12} ($\mathsf{c} \in \{\mathsf{i,a,u}\}$).

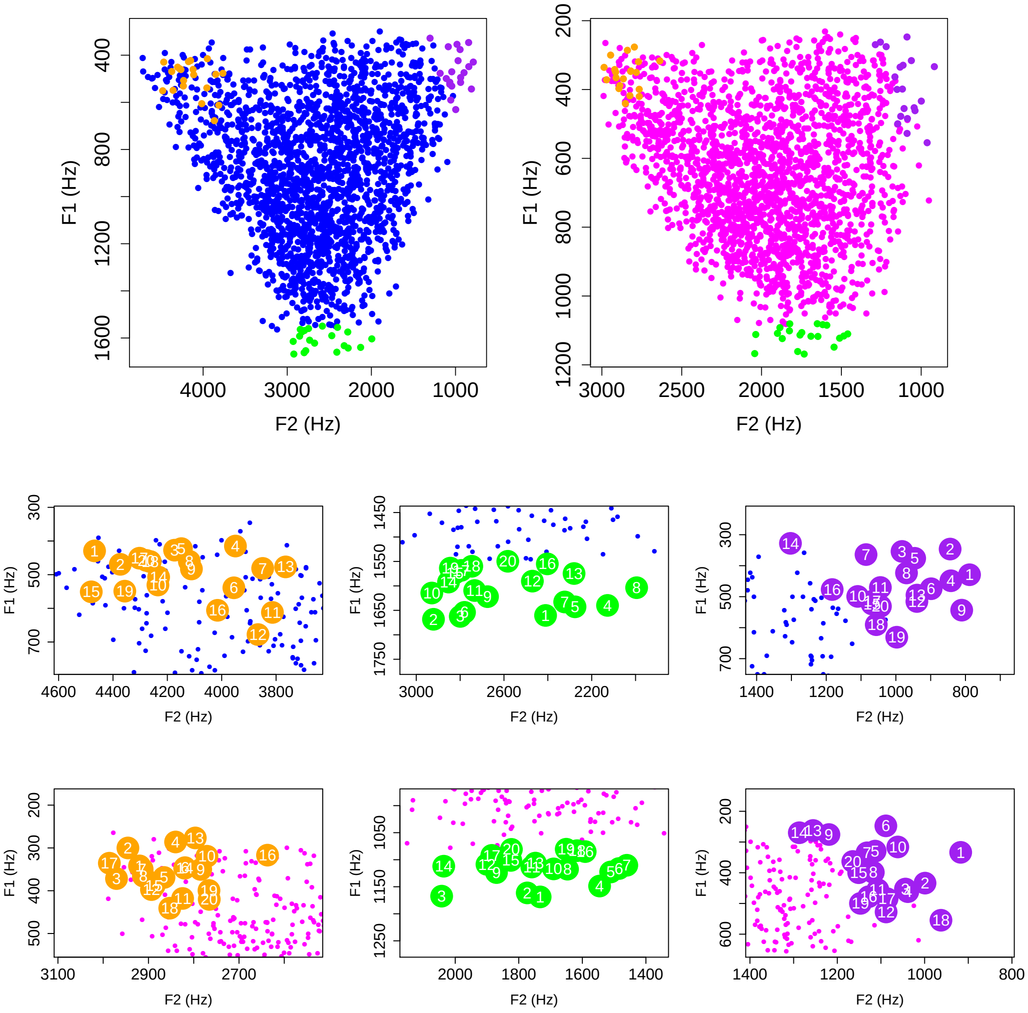


Figure 3.16: A response pairing (middle and bottom rows) over MVS(0.5) (top, left) and MVS(10) (top, right) derived from $Q^G_{12}(\mathsf{c}, 0.5)$ and $Q^G_{12}(\mathsf{c}, G, 10)$ for subject s^G_{12} ($\mathsf{c} \in \{$i,a,u$\}$).

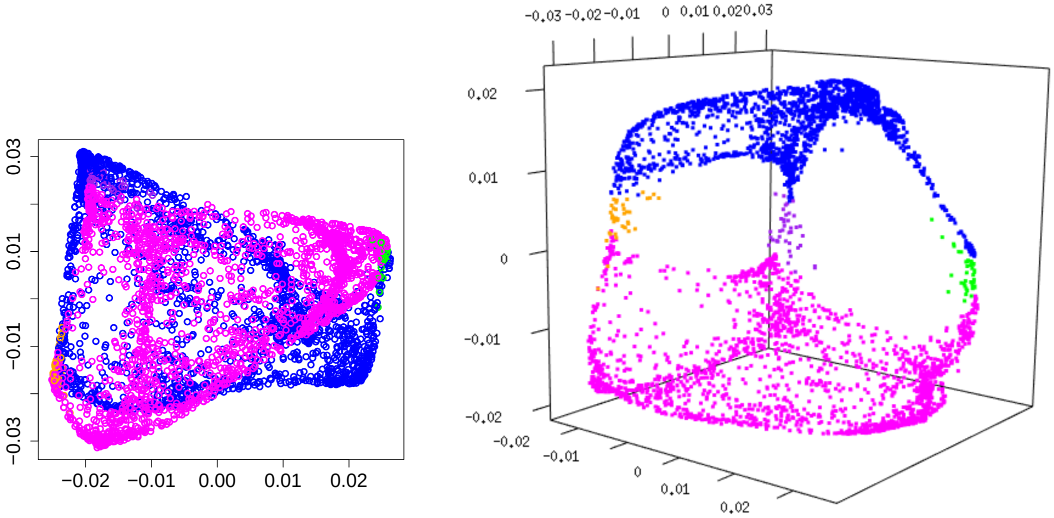


Figure 3.17: The warping yielded by socio-auditory weighting $S(C(T^G_{12}(10, 0.5)))$. In the two-component depiction (left) the u representations are in the upper left corner.

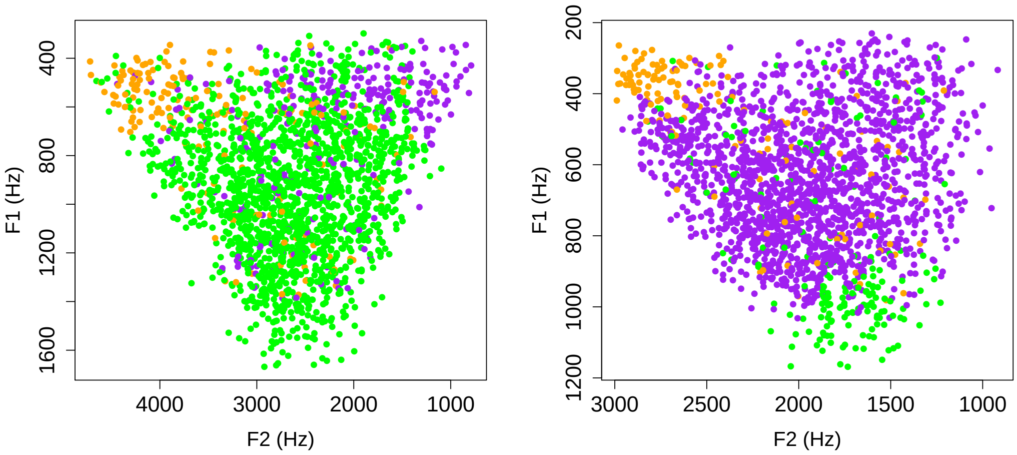


Figure 3.18: The auditory equivalence derived from the category transfer interpretation $I(C(T^G_{12}(10, 0.5)))$. The equivalence relation is depicted over representations in MVS(10) and MVS(0.5).

CHAPTER 4: MULTIFOLD REPRESENTATION AND INTERMODAL ASPECTS

In this chapter, we extend the conceptual foundation for vocal learning and its relation to the acquisition of vowel normalization presented in Chapter 3. The extension has two principal facets: (i) the incorporation of articulatory representations within the model, as well as relations between articulatory and auditory representations that yield intermodal representations, and (ii) the incorporation of "multifold" auditory representations within the vocal learning model. We take articulatory representations as our point of departure, beginning with a review of models of articulatory synthesis, and then shifting toward their more recent appropriation within cognitive models of articulation. We review the use of these representations within "task dynamics" models of speech production formulated in the late 1980s by Elliot Saltzman and Kevin Munhall, along with the subsequent "auditory target" model for the acquisition of articulatory-auditory mappings during spoken language acquisition developed by Frank Guenther in the mid 1990s. The latter model takes vowel normalization to be a fixed computation localized within a single modality, a view common to most modeling architectures.

We again take a sharp break from the standard conceptualization of vowel normalization in favor of one in which it is a computation over "intermodal manifolds" derived from an infant's early intermodal computations over auditory and articulatory representations. We focus on articulatory representations derived from an articulatory synthesis paradigm whose primary aim is the simulation of human speech production through construction of models of the human vocal tract based on its developmental biology (Sections 4.2.1 and 4.2.2). We then present the extension of the "static" auditory representations described in Chapter 3 to multifold representations that serve as a basis for more elaborate representations of vowel signals (Section 4.2.3).

To adduce the intermodal conceptualization of vowel normalization, we extend the framework put forward in Chapter 3 to include a methodology for modeling the use of the internalization of a relationship between an infant's interpretation of their own vocalizations and differentiated caretaker responses to those vocalizations in computing equivalence classes over intermodal representations (Section 4.3). The methodology is again expressed through the creation of a vocal learning environment consisting of caretaker and infant agents constructed from cross-linguistic perceptual categorization results. The potential of the approach is demonstrated via its computational structural output (Section 4.4).

4.1 Articulation and Vocal Learning

The construction of "mechanisms" that simulate the articulatory and other motor behaviors of living beings is a concept that dates back well into antiquity, represented prominently in literature by Daedalus's prolific array of creations. Dating the practice of constructing such objects is more challenging, due to the lack of lasting evidence that might authenticate them, and a tendency toward artifice. Nevertheless, the construction of such mechanisms has long been viewed as a viable route to scientific understanding of the phenomena that inspire them. The phonetician G. Oscar Russell (1928, Chapter 1), for example, provides a historical review of the progress in modeling up to the early 19th century. According to Russell, articulatory synthesis models have existed in some legitimate form since the 16th century. These early synthesis models mostly made use of mechanical contraptions that roughly simulated the physical characteristics of the vocal tract, yet some of them nevertheless achieved significant success in illuminating certain aspects of the speech production process. For example, Russell's (1928) own (in)famous work was based in part on that of two 19th century modelers, Wilfred Willis and Charles Wheatstone, whose ideas Russell identifies as having been "adduced from some type of actual scientific proof," and having attained "some degree of acceptance" (p. 21).

Willis was interested in explaining differences in vowel quality, and hypothesized that "a voiced vowel is made up of at least two tones: that produced by the vocal cavity or cavities, and that resulting from the vibration of the vocal cords" (Russell, 1928, p. 21). To test this, he conducted experiments using reeds of various sizes and pipes of various lengths, intended as analogs of the vocal chords and vocal cavities, respectively. Based on the results, Willis (1830) extrapolated his "cavity tone theory" of vowel quality, wherein each vowel is characterized by the tone of some corresponding vocal cavity form, independent of the tone generated by the vocal chords. Attempting to extend Willis's "unisonant resonance" cavity tone theory, Wheatstone (1837) constructed a "speaking machine," based on a design by Wolfgang von Kempelen, and used it in experiments that led him to a multiple resonance theory of vowel quality later adapted by von Helmholtz (1863). Wheatstone's major theoretical contribution was the idea that the vocal chords generated a complex tone, with the vocal cavity augmenting certain harmonics of the tone, in addition to its fundamental frequency.

Mechanical synthesis models persisted into the early 20th century, though eventually gave way to electrical models. One of the earliest, created by Stewart (1922), was simply a "functional copy of the vocal organs" whose operation was based solely "upon the production of audio-frequency oscillations in electrical circuits" (p. 311). More elaborate devices followed, such as H. W. Dudley's Vocoder and the subsequent Voder, developed at Bell Labs by H. W. Dudley and R. R. Riesz, and displayed at the 1939 World's Fair, which was able to produce voiced and unvoiced speech sounds by passing "energy" and random noise, respectively, through a keyboard-operated resonance device. Soon after, the Haskins Laboratories Pattern Playback Machine (Cooper et al., 1951) appeared, which was capable of "reconvert[ing] spectrograms into sound, either in their original form or after modification" (p. 319). The Pattern Playback Machine along with Fant's (1953) OVE synthesizer provided the basis for substantial reciprocation of progress in articulatory synthesis and phonetic research during the 1960s (see Klatt, 1987; Beckman, 1997, for further detail).

Electrical synthesizers eventually gave way to software-based models as computer programming developed into an inexpensive and highly portable means of model exchange. Moreover, the "symbolic" developments in phonological analysis largely due to Chomsky and Halle (1968) dovetailed with the software-based modeling methods, leading to a new period of reciprocal progress in the 1970s and early '80s. For example, Carlson and Granström's (1975) "phonetically oriented programming language for rule description of speech" made use of context-sensitive grammar rules defined over several levels of speech sound representations in the generation of synthesized speech sounds. Hertz's (1982) Speech Research System (SRS) similarly "uses four kinds of rules that apply in succession to convert a text string into sound: text-modification, conversion, feature-modification, and parameter rules" (p. 1155). The representations used in the transition from text to speech use concepts from both contemporary phonological advances and phonetic knowledge accumulated over several preceding decades.

From the mid 1980s onward linguistically-oriented software-based synthesis shifted toward "natural" models of the human vocal tract, some of which were derived directly from vocal tract image data using the rapidly advancing computational statistical methods developed in the preceding decades. Maeda's (1990) articulatory synthesis model, discussed in Section 4.2.1, is based on a statistical modeling approach that organizes midsagittal vocal tract tracings using numeric interpretations of Lindblom and Sundberg (1971) "articulatory blocks." The blocks are scaled and additively combined to generate new midsagittal vocal tract shapes which feed into a signal generation algorithm. The model has since been extended by Boë and Maeda (1998) to generate midsagittal vocal tract shapes over an age set ranging from infancy to early adulthood. These age-varying synthesis models continue the reciprocal relationship with theory, as they have come to play a significant role in the debate on the evolution of spoken language (see Boë et al., 2002; de Boer and Fitch, 2010). Boë and Maeda's (1998) age-varying model continues to be extended, with one of the more modern versions serving as the articulatory model used in this dissertation (see Section 4.2.2).

Advances in computational geometry and the mathematical biosciences in the 1970s led to the advent of "finite element" synthesis (e.g., Wilhelms-Tricarico, 1995) during the 1980s and '90s, which explicitly and directly models the structures of the articulatory system, and the relations between them. The ArtiSynth model put forward by Fels et al. (2006) is a "set of models, along with constraints which control their interactions, which can be combined to form an integrated model" of the vocal tract (p. 2). The models themselves "can represent specific anatomical structures, such as a muscle-activated jaw..., the tongue..., the airway cavity itself..., or acoustical production entities" and can be implemented as "particle/spring/rigid body systems..., spline-based curves and surfaces, and PCA-driven point clouds or meshes" (p. 2). The interaction between models "is achieved using constraint components, which provide a general mechanism for specifying interactions between models" and "enforce themselves by modifying the input and/or output variables of the models which they are interconnecting" (p. 3). The ArtiSynth model is an open-source project which continues to be modified and refined as more is learned about the anatomy and physiology of the structures being modeled. At present, the finite element modeling approach does not extend to infant vocal tracts, thus we have not attempted to incorporate such models within our own approach, yet, the general approach seems promising.

The construction of more sophisticated models was accompanied by greater attention to their planned operation and control. Although motor control of the body had long been understood to be a property of the brain and central nervous system, the complexity of that control had scarcely come into focus by the beginning of the 20th century. As study of the anatomy and physiology of the brain progressed in the early 20th century, along with advances in the theory of complex dynamical systems, cohesive conceptualizations of motor planning and control of bodies began to take shape, leading to new applications and advances. By the 1950s, articulatory synthesis models became conceptually integrated within planning and control theory, focusing attention on the nature of the representation of articulation within the mind (see Grimme et al., 2011, for a review). More broadly, the organization and relation of representations for planning and control is a key issue in the design of simulations and models of goal-oriented human actions, ranging from reaching tasks to speech articulation. The typical organizational scheme involves specifying a "configuration space" whose representations correspond to the structural components of a system and their positions relative to one another, and a "task space" whose representations correspond to the potential "end-effector" positions of the system.

To exemplify, the consider the "robot manipulator" depicted in Figure 4.1 composed of "n-links interconnected by joints into an open kinematic chain" (Spong, 1996, p. 1340). The "joint variables, $q_1, \ldots, q_n$, are the relative angles between the links" where "q_i is the angle between link $i-1$ and i," and $q_i \in [0, 2\pi)$. A vector $q = (q_1, \ldots, q_n)$ whose components are values for the n joint angles is called a "configuration," and the "set of all possible configurations is called configuration space or joint space," which in this case is the "n-dimensional torus $\mathcal{T}^n = S^1 \times \cdots \times S^1$, where S^1 is the unit circle." The position of the robot manipulator can be described in terms of configurations within the configuration space. The motion of the manipulator may be described in terms of sequences of configurations. Now, suppose the task of the robot manipulator is to grasp an object in the environment with its "end-effector" (link 6), depicted as a "$\sqcap$" attached to link 5 in Figure 4.1. The "space of all positions and orientations (called poses) of the end-effector" is called the "task space," which may be characterized as follows:

> "If a coordinate frame, called the base frame or world frame is established at the base of the robot, and a second frame, called the end-effector frame or task frame, is attached to the end-effector, then the end-effector position is given by a vector $x \in \mathbb{R}^3$ specifying the coordinates of the origin of the task frame in the base frame, and the end-effector orientation is given by a 3×3 [special orthogonal] matrix R.. . .The set of all such orientation matrices forms the special orthogonal group $SO(3)$. The task space is then isomorphic to the special Euclidean group, $SE(3) = \mathbb{R}^3 \times SO(3)$." Spong (1996, p. 1340)

Motor planning, which involves the generation of desired end-effector paths, may also be carried out in the task space and related to configuration space planning and control. In the example above, the relation takes the form

$$f : \mathcal{T}^n \to SE(3)$$

called a *forward kinematic transformation*. An inverse of f is called an *inverse kinematic transformation*. Specification of a forward kinematic transformation also yields derivative

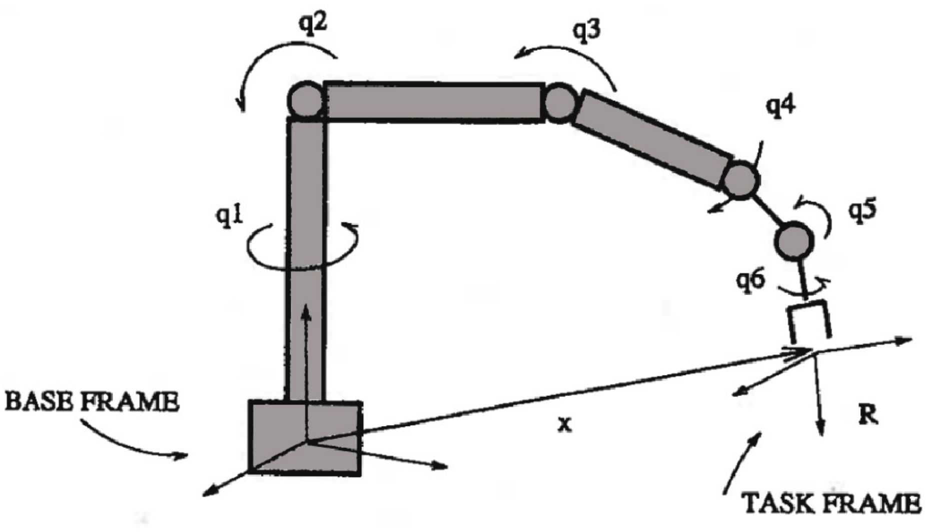


Figure 4.1: From Spong (1996), page 1340. A robot manipulator composed of n-links interconnected by joints into an open kinematic chain. The joint space variables $q_1, \ldots, q_n$, are the relative angles between the links. The end-effector (link 6) is depicted as a "⊓" attached to link 5.

relations that capture other aspects of trajectory planning, e.g., *velocity kinematic transformations* $J_f(q)$ between joint space velocities $\dot{q}$ and end-effector velocities $J_f(q)(\dot{q})$, and their pseudoinverses.

In a review of approaches to limb and speech motor control, Grimme et al. (2011) characterize two main lines in the history of conceptualization of the goals of motor control within the task space framework. The division is based on the assignment of priority to a planning domain, which may be the task space or the configuration space. On the "intrinsic" approach, priority is given to an effector-specific configuration space, while on the "extrinsic" approach, priority is given to the task space, where planning may occur independent of effector-system consideration. Grimme et al. (2011) extend the intrinsic-extrinsic dichotomy to speech production, basing the division on the assignment of priority to one of the sensory modalities within which speech production goals are assumed to be situated.

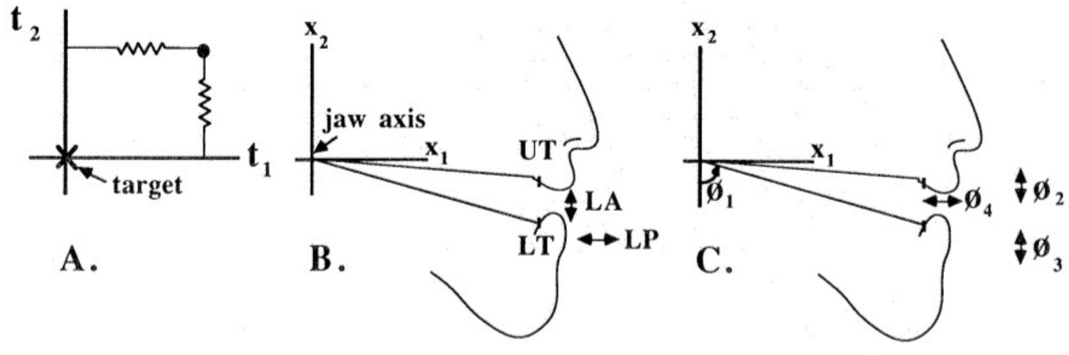


Figure 4.2: Bilabial task from Saltzman (1986). A. Task space $(\tilde{t})$. Closed circle denotes current system configuration. Squiggles denote each axis' dynamics in lumped forms; B. Jaw space $(\tilde{x})$. Local tract variables (LP = lip protrusion, LA = lip aperture) are expressed in jaw coordinates. UT and LT denote positions of upper and lower front teeth, respectively; C. Model articulator space $(\tilde{\phi})$. ϕ's denote articulatory variables.

Roughly, the division amounts to whether task spaces are conceptualized in terms of articulation or audition. The "intrinsic" conceptualization is articulatory-centric, drawing on the Motor Theory of speech perception (e.g., Liberman et al., 1967; Liberman and Mattingly, 1985), Fowler's (1986) Direct Realism, and later developments in articulatory phonology (Browman and Goldstein, 1990a,b). The "extrinsic" conceptualization is auditory-centric, drawing on studies of acoustic control of articulation (e.g., Perkell et al., 1993; Savariaux et al., 1995), and acoustic-based theories of perception (e.g., Stevens, 1972, 1989). To illustrate the division, we briefly review two key task space models of motor behavior exhibited by the speech production system.

In the mid 1980s, Saltzman and Kelso (1987) put forward a "task-dynamic" approach to motor planning and control wherein goals are situated within a task space characterized by the functional aspects of the motor action being carried out. With respect to speech production, Saltzman (1986) exemplifies the approach with a task space taken to be a two-dimensional Euclidean space whose axes correspond to articulatory constriction parameters. Figure 4.2 (A.) depicts a task frame in which the target constriction is at the origin, and the current constriction is denoted by a closed circle. The task dynamics are represented as equations of motion within the task space, and are kinematically transformed into jaw space trajectories that yield effector-specific motion control of the tongue, jaw, and lips for speech production (Figure 4.2 B). The resulting jaw space trajectories are kinematically transformed into a four-dimensional model articulator space representing "moving" components of the tongue, jaw, and lips (Figure 4.2 C). The task-dynamic approach is argued to be able to capture the articulatory system's ability to make rapid adjustments in response to small perturbations that occur during speech production, along with a number of other motor planning aspects of speech production (Saltzman and Munhall, 1989; Saltzman, 1995).

In early 1990s, an alternative to the articulatory-centric approach began to take shape, in which task spaces are defined over acoustic/auditory degrees of freedom (see Maeda, 1991). Guenther (1995) and Guenther et al. (1998) put forward a model focusing on the relationship

$$f : V \to P$$

between an "articulatory" reference frame V defined by the parameters of Maeda's (1990) VLAM, and an "auditory perceptual" reference frame P defined using log ratios over formant values: $\log(F2)/\log(F1)$, $\log(F3)/\log(F2)$, $\log(F2)/\log(F1)$, and $\log(F1)/SR$, where SR is a "sensory reference," calculated using the geometric average of all values of F0 for a given talker (Miller, 1989). Feedback-based estimations of pseudoinverses of the velocity kinematic transformation $J_f(\theta)$ are taken as models of the acquisition of a cross-modal auditory-perceptual-articulatory mapping ultimately used to relate trajectories in the auditory-perceptual frame to trajectories in the articulatory frame.

Aside from the nature of the task space, Guenther's (1995) model brought attention to the role of relations between articulatory and auditory representations in during language acquisition. Properties of the vocal tract are structurally linked to acoustic cues that factor into speech perception and production, particularly with respect to vowels (see Honda, 1996; Perkell, 1996). Moreover, relations between articulatory and auditory representations appear to be present early in development (Kuhl and Meltzoff, 1982), and likely play a substantial role in spoken language acquisition (Kuhl and Meltzoff, 1996). Within planning and control modeling, Perrier (2005) argues for a multisensory approach which incorporates a "modality hierarchy" in which auditory representations are given priority, while Schwartz et al.'s (2012) Perception-for-Action-Control theory emphasizes role of relations between articulatory and auditory representations in speech perception.

In Chapter 3 an infant's interpretations of social agents in the vocal learning environment were characterized as auditory manifolds embedded within the auditory reference frame. Although valuable insight may be gained from a "unisensory" basis of investigation into spoken language acquisition, it is likely the case that adopting a "multisensory" conceptualization involving relations between articulatory and auditory representations, inter alia, should also be investigated. To the point, comparison of models involving only subsets of sensory modalities may shed light on spoken language acquisition in cases of abberational development, as disussed in Section 5.2.2. In this chapter, we expand the conceptualization put forward in Chapter 3 so that an infant's interpretation of a social agent in the vocal learning environment is constructed over structures relating the infant's articulatory representations to the auditory representations of that agent. These "sensorimotor" structures evolve into "intermodal" structures as cognitive computation develops. In the remainder of this section, we lay out the foundation of our approach to the modeling of the acquisition of sensorimotor and intermodal structures, and the role they play in vowel normalization. In order to formulate these structures, we first need to formulate structures from both the sensory and motor modalities that we are concerned with. We have covered auditory structures in detail in Chapter 3, thus we begin with articulatory structures.

Recall the configuration space $\mathcal{T}^n$ for the robot-manipulator depicted in Figure 4.1. The space $\mathcal{T}^n$ is characterized as an n-dimensional torus, though it is also an n-dimensional manifold (in the topological sense). Moreover, the task space $SE(3)$ is a Lie group, and hence a differentiable manifold. Indeed, manifolds have been an integral part of the specifications of configuration and task spaces since the inception of the planning and control modeling. To illustrate, recall the basic example from Chapter 2 involving a ball attached to a rod fixed to a swivel such that the rod can swivel freely both horizontally and vertically. The configuration space for the rod is the torus $\mathcal{T}^2$, while the task space is the two-dimensional surface of the sphere centered at the swivel, with radius equal to the length of the rod, both of which are manifolds. Aside from these global shapes, other planning and control structures have been characterized as manifolds, including the "uncontrolled manifold" for a given end-effector pose, roughly defined as the set of configurations that yield that pose. The name is

> "based on the idea that the central nervous system controls less precisely or even not at all, which configuration within that manifold is being realized, while control perpendicular to that manifold is used to achieve the task of keeping or moving the [end-effector] to its desired location" (Schöner et al., 2008, p. 24),

highlighting the fact that “[s]peech articulatory movements are not fixed programs that unfold invariably” (p. 23). The uncontrolled manifold is argued to be a key concept in formulating motor equivalence (Saltzman et al., 2006; Schöner et al., 2008). Moreover, the existence of such manifolds suggests “that task-level goals structure the coupling among the components of the articulatory apparatus” (p. 10 Grimme et al., 2011). If true, it follows that a simple “inversion” approach to multisensory matching is untenable, even if language-specific.

More broadly, manifold-based conceptual frameworks are finding greater application in the modeling of anatomical and physiological articulatory phenomena (e.g., Jansen and Niyogi, 2006; Seo et al., 2010, 2011; Ma and Fu, 2012). In light of the the growing applicability of manifold-based foundations, we adopt the following extension of our conceptual basis:

> Cognitive organization of articulatory representations is a prerequisite for the acquisition of vowel normalization, and is a significant component in an infant’s development of a cohesive representation of the self and of their caretakers, in the sense discussed in Section 1.2.2. Specifically, these components are articulatory manifolds.

We now turn to relations over structures within the auditory and articulatory frames. We take as our point of departure Davenport’s (1977) brief yet useful history on the nature of cross-modal inquiry over the last few centuries. Davenport (1977) writes that “[t]he concept of the interrelatedness of the senses...has had a long philosophical history extending at least as far back as Locke and Berkeley; yet the clinical and experimental investigations into this concept are relatively recent in origin” (p. 74). The fact that clinical work and experimental investigations were, at the time, just getting under way likely accounts for the two underlying assumptions of the field prior to 1970. As Davenport explains, “[o]ne held that cross-modal perception is uniquely human, and the other that it is dependent on language" (ibid). Davenport continues, "[a] half-dozen experiments performed since 1970, however, have drastically altered these assumptions about cross-modal perception," suggesting that it is not unique to humans and does not dependent on language. Expanding on the latter point, Davenport et al. (1973) maintain that the ability to "abstract information across different sense modalities," present in both apes and humans, "could well be the rudimentary basis which is essential for the emergence of language" (p. 21). Furthermore, in a report to the New York Academy of Sciences Davenport (1976) writes the following on the relevance of cross-modal perception to the "origin and evolution of speech and language":

> First, it appears that multimodal information extraction of environmental information is likely to result in more veridical perception, and may facilitate cognitive functioning. Second,...cross-modal perception requires the derivation of modality-free information, a "representation." That an organism can have the same representations, concepts or percepts, regardless of the method of peripheral reception, confers great advantage on that animal in coping with the demands of living. (p. 147)

The main product of cross-modal perception identified by Davenport is an "abstract" and "modality-free" representation of multimodal sensory information. While the existence of such representations is largely accepted at present, the exact nature of their "derivation" during early infancy is still a matter of debate, as is the scope of "veridical perception." We first review the conceptual history of the former, building up discussion of the latter.

Piaget (1936) characterized the derivation of the relations of sensory information during development as "integration," wherein initially separate sensory modalities gradually coalesce to form an integrated "multisensory" system. With respect to developmental studies, this approach "set the research agenda, with developmental psychologists seeking either to support [t]his description, or more recently, to reject it" (Smith, 1994, p. x). Indeed, by the

1960s, work on “amodal” sensory information – i.e., information that is in some way invariant across the different modalities – served as a basis for the “differentiation” approach to multisensory relations (Gibson, 1966, 1969), wherein an infant’s sensory perception is initially unified, and increasingly differentiated during the course of development. By the end of the 1980s, differentiation had become the mainstream approach, with a number of different theories creating a rather broad platform for research into the nature of multimodal perception (see Lewkowicz and Lickliter, 1994, for further details). Still, despite the dichotomous presentation, both integration- and differentiation-based theories have in common that multisensory perception is “a progressive process that results in the improvement of early emerging multisensory perceptual abilities and the proliferation of new ones with development and increasing experience” (Lewkowicz and Ghazanfar, 2009, p. 470).

During the 1980s, an alternative approach to multisensory perception began to take shape based on the “Intensity Hypothesis” (see Lewkowicz and Turkewitz, 1980) whose “basic premise...is that during the first few months of life infants’ responsiveness to stimulation in all modalities is dominated by the quantitative aspects of stimulation,” while over time “responsiveness to the qualitative aspects of stimulation emerges” (Lewkowicz, 1994, pp. 166-7). The approach assumes that “both integration and differentiation processes are involved in the developmental process that leads up to the emergence of these more advanced response mechanisms” (p. 167). Importantly, focus is shifted more to the developmental aspects of multisensory perception, which brings into view the underappreciated fact that

> “the changing functional properties of the sensory systems, the differential effects of experience, and the processes of differentiation and integration all contribute to the development of specific intersensory functions through an intricate process of interaction among all of these factors.” (p. 169)

In light of this, “predictions based on end point (i.e., adult) performance are inappropriate and can lead to false predictions” (ibid). This formulation, if even remotely on the right track, casts significant doubt on the merit of distributional learning models that depend on similitude between adult and infant cognitive capacities, particularly those based on statistical learning mechanisms (e.g., McMurray et al., 2009; Rasilo et al., 2013; Hörnstein, 2013), and define success in terms of adult categorization behavior.

Considerations of this kind have evinced aspects outside of the typical scope of multisensory research which are relevant to the concept of “veridical perception.” To illustrate, Lewkowicz and Ghazanfar’s (2009) focus on “perceptual regression” brings into view a “multisensory perceptual narrowing” that extends the conceptualization of perceptual warping discussed in Chapter 3. Specifically, “infant sensitivity to multisensory relations is [predicted to be] initially broad and then narrows with development” (p. 472), which is adduced by two studies in which

> “4-, 6-, 8- and 10-month-old infants were presented with side-by-side movies of the faces of a rhesus monkey producing a coo call on one side and a grunt call on the other side, and their preferences were measured for each of these visual calls in silence and then in the presence of one of the corresponding audible vocalizations....Consistent with perceptual narrowing, 4- and 6-month-old infants matched the visual and audible calls by looking longer at the visible call in the presence of the corresponding vocalization than in its absense....In contrast, 8- and 10-month-old infants exhibited no evidence of multisensory matching.” (ibid)

The results suggest that the 4- and 6-month-old infants are at a developmental stage wherein multisensory matching is a prominent component in cognitive functioning, while the 8- and 10-month-old infants may have progressed beyond a stage in which multisensory matching plays less of a role due to the emergence of more abstract representations. If true, the multisensory pairing of auditory and visual representations may constitute a significant, differentiated developmental stage that buttresses concept and category formation.

From this audio-visual basis, we extrapolate two key precepts. The first is that infants carry out multisensory matching over auditory and articulatory representations early in spoken language acquisition. If true, the sensorimotor pairing of articulatory and auditory representations may play a significant role in language acquisition long before vowel categories begin to form. Moreover, the pairings may provide the means for the creation of representations of social agents. With respect to our conceptual basis, we assume that:

- Infants carry out multisensory matching over articulatory and auditory representations, and use a pairing structure to store the matched representations.
- Auditory and articulatory manifolds are aligned using the pairings to yield sensorimotor manifolds, which constitute the basis of an infant's development of a cohesive representation of the self and of their caretakers in the sense discussed in Section 1.2.2. In the simplest case, the infant aligns a manifold over their articulatory representations with a manifold over their auditory representations, yielding a model self, while also aligning a manifold over their articulatory representations with a manifold over the auditory representations of the caretaker, yielding a model of the caretaker.

The latter assumption has proven more tendentious than others that constitute our conceptual basis, perhaps because it is a departure from a more "common sense" conceptualization in favor of adhering to principles and theoretical coherence, as well as capturing emerging experimental data. In principle, our conceptual foundation assumes an infant's creation both of a model self and of a model caretaker, and theoretical symmetry suggests that the creation procedure is the same in each case, barring abberation (e.g., abnormal development). Recent experimental evidence adducing the existence of "mirror neurons"

– roughly characterized as neurons that fire both while producing and while perceiving action (Rizzolatti et al., 1996; Kohler et al., 2002) – have been hypothesized to play a key role in speech perception and production, and broader cognitive phenomena related to spoken language acquistion (see *Brain and Language* Volume 112, Issue 1). Mirror neuron activity has been incorporated within Guenther et al.'s (2006) neural model of speech acquisition via the speech sound map, which activates during both speech production and perception. This neural evidence and modeling suggests a cognitive structural link between perception and production during the acquistion of vowel normalization, which we model in terms of manifold alignment. Although potentially perceived as a departure from common sense, we feel this conceptualization is promising.

The former assumption is rather broad in scope, and is consistent with a number of potential conceptualizations of the development of multisensory matching during early infancy. Thus, before proceeding we discuss what we take to be the conceptual foundation of the modeling approach we put forward in Section 4.3. We assume that early babbling jumpstarts the development of the formation of pairing structures that the infant uses to link articulatory manifolds over their own articulatory representations to auditory manifolds over their own auditory representations. The development of pairing structures is then co-opted for relating models of the self to models of others in two distinct ways. The first involves the use of pairings for relating preliminary models of the self to models of others, e.g., the formation of socio-auditory pairings used to align auditory manifolds. The second involves the use of pairing structures in the formation of more elaborated models of others that include an imputed mirroring of the infant's own internal auditory-articulatory pairings, as described in the latter assumption above. In the model formulation in Section 4.3, we are assuming that co-opting has already taken place, and that socio-auditory pairings yield the sensorimotor pairings that guide the alignment of articulatory and auditory manifolds as infants create models of the social agents in their environment. However, this is an assumption that faciliates modeling at this stage of investigation, and is not an irrevocable aspect of the modeling framework. Indeed, sensorimotor alignment is likely to require a more flexible formulation.

The second key precept is that intermodal representations exist, and are distinct from sensorimotor pairings. This precept is in line with Meltzoff and Kuhl's (1994) conceptualization of intermodal representation, defined as "a higher order phonetic representation of speech that acts as a mediator between nonidentical information in the two [or more] modalities" (p. 360). Intermodal representations are included in Guenther et al.'s (2006) neural model of acquisition as posited speech sound map representations mediating between perceptual and motor representations, and the frame they inhabit is assumed be fixed. Given our conceptual basis, sensorimotor manifolds yield transformations from the auditory and articulatory reference frames to "intermodal" reference frames, creating representations which exhibit language-specific multisensory narrowing (Section 4.3.1). The intermodal reference frames themselves are predicted to be language-specific in terms of the number of axes required for category acquisition. Moreover, intermodal representations are subject to the same computations as auditory and articulatory representations. With respect to our conceptual basis, we assume that:

- Cognitive organization of intermodal representations is a prerequisite for the acquisition of vowel normalization, and is a significant component in an infant's development of a cohesive representation of the self and of their caretakers, in the sense discussed in Section 1.2.2. Specifically, these cohesive representations are intermodal manifolds.

- Infants carry out pairing over intermodal representations, and the pairing of representations is provided by vocal exchange with a caretaker.
- Infants use this pairing to derive mappings between intermodal manifolds representing a self and their caretakers. This computation is called *intermodal vowel normalization*. In the simplest case, an intermodal manifold representing a self is "aligned" with an intermodal manifold representing a single caretaker.

From this basis, we again take the view that vowel category acquisition is an emergent phenomenon which arises from a vowel normalization computation.

The conceptual formulation of the acquisition of vowel normalization put forward in this chapter and throughout this dissertation is meant to distinguish the acquisition of normalization as a metaphysical entity that is separable from the broader suite of vocal learning phenomena. This separation is strengthened by a technical formulation emphasizing the reference frames, structures, and computations involved in spoken language acquisition, which highlight the role of normalization as a generative procedure over manifolds, rather than a statistical summary achieved through the "practice" of learning, which rather makes use of the frames, structures, and computations. The technical formulation is mostly in line with the general task-dynamic approach to planning and control, which focuses attention on "skill acquisition," formulated as "establishing a one-to-one correspondence between the functional characteristics of the skill and the dynamical regime underlying the performance of that skill" (Saltzman and Kelso, 1987, p. 86), as distinguished from "skill learning," which roughly involves the algorithmic determination of task-specific configuration trajectories. The corresponding extension of the model to include sensorimotor and intermodal phenomena only fortifies the argument in favor of serious consideration of the frames, structures, and computations, independent of the token summaries they may yield.

In the vocal learning environment formulated in Chapter 3, each infant agent in the vocal learning environment makes use of an auditory reference frame to create auditory manifolds that constitute representations of the social agents relevant to their acquisition of spoken language. In the simplest case, the social agents are the infant and a single caretaker. The auditory manifolds M_I and M_C representing these agents are then normalized using a socio-auditory pairing to yield a commensuration manifold that provides the means for the computation of warped representations, and vowel category transfer, inter alia. In the remainder of this chapter, we modify and extend the framework for modeling the acquisition of vowel normalization put forward in Chapter 3 to include "articulatory" and "intermodal" aspects of vocal learning. Specifically, we extend the following components.

Internal Computation: We take cognitive computation to also involve: i) the creation of manifolds over articulatory representations, derived either from infant productions or from internal self-signaling, ii) the creation of a pairing of auditory representations of infant productions with articulatory representations of infant productions, and a pairing of auditory representations of caretaker responses with articulatory representations of infant productions, both derived from turn-taking vocal exchanges, along with the assignment to each pair an interpretation of the social signal imposed on by the caretaker on their response, iii) the alignment of auditory and articulatory manifolds using these pairings, which yields sensorimotor manifolds, and thus transformations from the auditory and articulatory reference frames to an intermodal frame, iv) the creation of manifolds and pairings over intermodal representations, the latter computed from auditory representations derived from turn-taking vocal exchanges, v) the alignment of intermodal manifolds using these pairings, which yields commensuration manifolds. In this fashion, the spaces that are ultimately used for sensorimotor coordination and vowel categorization are language-specific, and moreover, dyad-specific, rather than fixed across languages.

Behavior: The resulting behavior, e.g., categorical perception of vowel signals, the perceptual magnet effect, multisensory narrowing etc., is based on emphasis of regions in the aligned manifolds made salient as a result of the internal computations. Importantly, the categorical behavior is a derivative aspect of the internal computation. The main output of the computations is a conceptual structure allowing for general equivalence computations between the infant and conspecifics.

The extended vocal learning environment again consists of models of caretaker agents representing five different language communities (American English, Cantonese, Greek, Japanese, and Korean) derived from vowel category perception experiments, and models of infant agents that "vocally interact" with their caretakers. The basic representational extension is the inclusion of articulatory and multifold auditory representations (Section 4.2). We then characterize the internalization of these signals and the internal computations over them (Sections 4.3.1 and 4.3.2), demonstrating how they yield the external behavior (Section 4.4). The main computations carried out within the extended vocal learning environment are summarized in Appendix A (Figures A.4 through A.8 depict the computations described in the following sections).

4.2 Articulatory and Multifold Representations

The human articulatory system is a complex, multipurpose biological system involving rich interaction between many anatomical structures above the transverse plane. The articulatory system is typically divided into two subsystems: the phonatory system, composed of the diaphragm, lungs, and trachea, and the supralaryngeal vocal tract system. In this dissertation, we are primarily concerned with the supralaryngeal vocal tract system, and in this section, we present a model of the speech capabilities of the human articulatory system localized within this system.

We begin this section with a review of the development of the components of the linear articulatory modeling paradigm (Lindblom and Sundberg, 1971; Maeda, 1990, 1991; Boë and Maeda, 1998), stepping through the key aspects (Section 4.2.1). We then discuss an age-varying extension of the approach called the *Variable Linear Articulatory Model* (VLAM Boë and Maeda, 1998), and a recent implementation of the model called the Vlab (Section 4.2.2). The settings of the Vlab parameters, called "articulatory configurations," are taken to be articulatory representations an infant organizes during the acquisition of vowel normalization. We conclude with the formulation of "multifold representations" over the vowel signals synthesized by the Vlab (Section 4.2.3).

4.2.1 Basis of Variable Linear Articulatory Modeling

We begin with the basic framework for the creation of an articulatory synthesizer laid out by Heinz and Stevens (1965), which involves the following steps: (i) obtaining X-ray tracings of the vocal tract that include all relevant details, such as stable landmarks, for quantitative measurement; (ii) devising a reference coordinate system for quantitative measurement that can be precisely defined in terms of stable landmarks readily observable on the tracings; (iii) devising a set of transformations based on a knowledge of structural anatomy which allows one to infer the acoustically relevant cross-sectional area from the measurements; (iv) calculating an acoustic spectrum from a specification of the cross-sectional area as a function of distance along the vocal-tract midline.

Given a set of midsagittal X-ray tracings (e.g., the one depicted in Figure 4.3, left) it is clear that the vocal-tract contains portions that are rectangle-like and portions that are circle-like, suggesting the combination of a rectangular coordinate system and a polar coordinate system to achieve a more useful coordinate system for vocal-tract cross-sectional area calculation. A *semi-polar coordinate system* for vocal-tract measurement, depicted in Figure 4.3 (right), is constructed in the following way. Adhering to the desideratum concerning stable landmarks within the X-ray data: (i) place points at the posterior nasal spine (PNS), the anterior inferior corner of the second cervical vertebra (CV2), and the tip of the

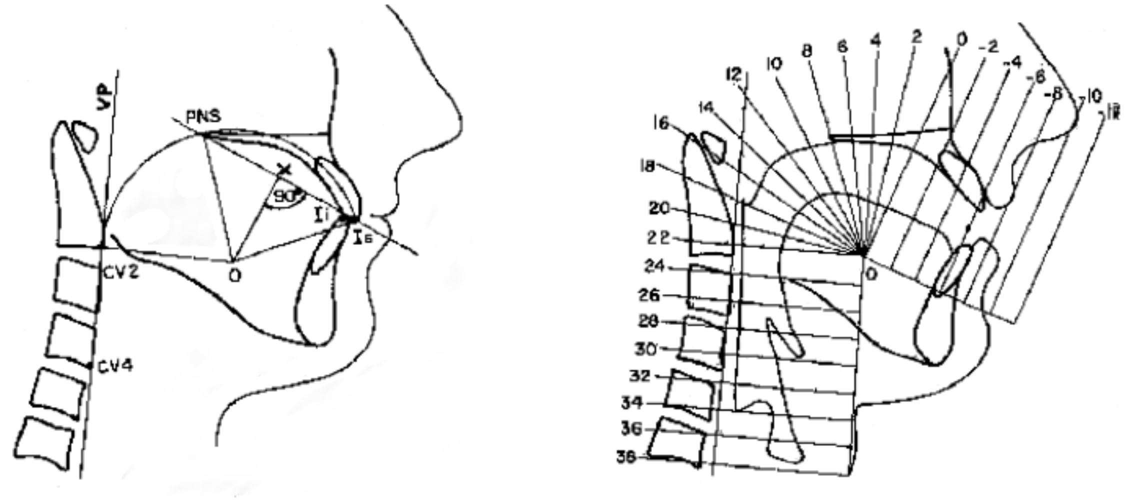


Figure 4.3: Images from Heinz and Stevens (1965). Locations of points on vocal-tract tracing (left) used to construct a semi-polar coordinate system (right).

crown of the most anterior maxillary incisor (Is); (ii) select as the origin O, the point that is equidistant from these three points; (iii) select a point at the anterior interior corner of the fourth cervical vertebra (CV4), and draw a line VP through CV2 and CV4; (iv) let X be the point midway between PNS and Is, and draw the ray from O that passes through X. (v) draw a ray A from O, in the direction of the lips, that is parallel with the line segment (PNS)(IS); (vi) similarly, draw a ray V from O, in the direction of the larynx, that is parallel with the line VP. Once the reference points and rays are in place, (i) a standard polar coordinate system is used to yield the coordinates of points the area of the plane that lies between OX and O(CV2); (ii) situate a quadrant of a standard rectangular coordinate system on the plane such that the origin of the quadrant is at the crux of OX and A; (iii) situate a quadrant of a standard rectangular coordinate system on the plane such that the origin of the quadrant is at the crux of O(CV2) and V. A precise assignment of coordinates depends

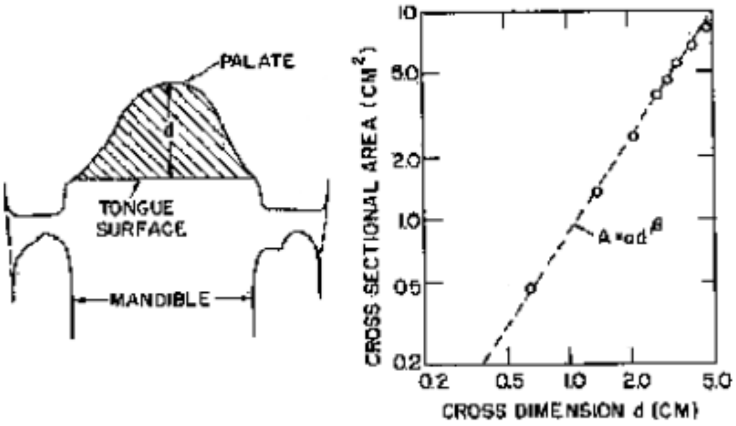


Figure 4.4: Image from Heinz and Stevens (1965). The area transformation is given as $a[n] = \alpha_n g(d[n])^{\beta_n}$, where g is a function, and α_n and β_n are constants, all determined empirically for each n.

on how fine-grained a set of measurements one is interested in, that is determination of a unit.

Given a midsagittal vocal-tract tracing with a semi-polar coordinate system, let t_u denote the function whose graph is the upper tracing line of the vocal-tract, and t_l the function whose graph is the lower tracing line of the vocal-tract. We sample f_u and f_l at integer multiples of the unit of our coordinate system, and let $t_u[n]$ and $t_l[n]$ denote the finitely many samples from each continuous function. Let $d[n]$ denote the linear distance between $t_u[n]$ and $t_l[n]$, called a *cross-sectional distance*. The scalar $d[n]$ gives us a way to compute the relevant maximum distance from the lower tracing to the upper tracing which delimits the computation of cross-sectional area of the vocal-tract lying on the plane determined by the line "$x = n$". The area transformation is given as $a[n] = \alpha_n g(d[n])^{\beta_n}$ (see Figure 4.4, right), where g is a function, and α_n and β_n are constants, all determined empirically for each n. Given an array $a[n]$ of cross-sectional areas, a *transfer function* is derived using the

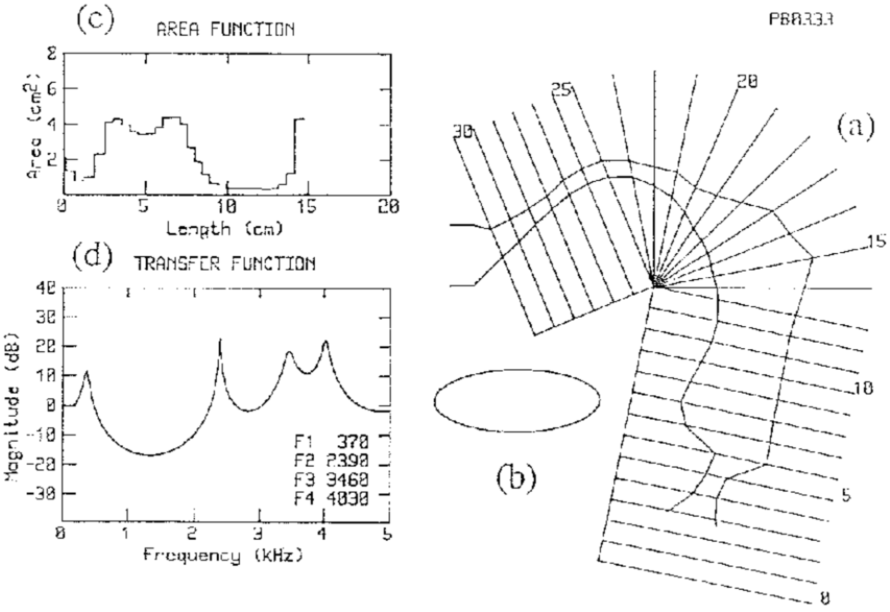
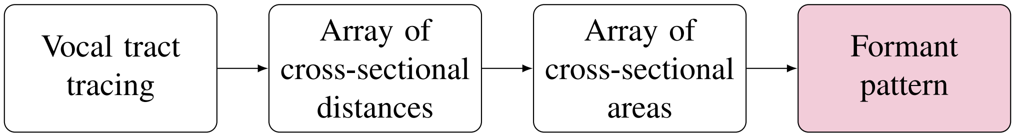


Figure 4.5: Images from Maeda (1990). (a) A midsagittal vocal-tract tracing and semi-polar coordinate system. (b) A depiction of lip rounding. (c) A cross-sectional area function indicating the cross-sectional area for the given vocal-tract tracing over distance from the glottis. (d) A transfer function derived from the cross-sectional area function indicating the resonance properties of the given vocal-tract tracing.

method specified in Badin and Fant (1984), providing the resonances, or *formant pattern* of the given midsagittal vocal-tract shape. A midsagittal vocal-tract tracing with a semi-polar coordinate system and corresponding cross-sectional area function and transfer function are shown in Figure 4.5. Putting it all together, we have the following flow of information in simulating the articulatory-acoustic relationship from a given vocal-tract tracing:


Vocal tract tracing
Array of cross-sectional distances
Array of cross-sectional areas
Formant pattern

Now, suppose we have m many X-ray tracings for a given speaker, and suppose we assign coordinates to each such that there are n samples taken from their upper and lower vocal-tract functions. For each of the m-many X-ray tracing, we then have n cross-sectional distances. Let $d_j[k]$ denote the kth cross-sectional distance for the jth tracing, and call each array

$$d_j = [d_j[1], d_j[2], \ldots, d_j[n]]^T$$

a *cross-sectional distance vector*. We can arrange the cross-sectional distance vectors in an $n \times m$ matrix:

$$D = [d_1, d_2, \ldots, d_m] = \begin{pmatrix} d_1[1] & d_2[1] & \cdots & d_m[1] \\ d_1[2] & d_2[2] & \cdots & d_m[2] \\ \vdots & \vdots & \ddots & \vdots \\ d_1[n] & d_2[n] & \cdots & d_m[n] \end{pmatrix}.$$

A reasonable next step is to impose a bit of theory on the cross-sectional distance data wrought from the tracings to make it easier to interpret and control computationally. Toward this goal, Maeda (1991) assumes that "during speech production, the complex activities of the articulatory organs are organized into a limited number of independently controllable functional blocks. Let us call these blocks 'elementary articulators' and their actions 'elementary gestures"' (p.7). Based on Lindblom and Sundberg's (1971) model, the elementary articulators are assumed to be the jaw, the larynx, the tongue body, dorsum, and tip (or apex), and lip height and protrusion. Maeda (1991) further assumes "that the influence of each elementary gesture on the deviation of the tongue from its neutral shape is proportional to a parameter representing a strength of that gesture, and that the influences from the different articulators can be added up to produce the final tongue shape" (p. 7). Specifically, it is assumed that each of the columns in D can be treated as values of a random vector in $\mathbb{R}^n$, or more simply as values taken by n real random variables, and the value of each random variable is a linear combination of seven parameters, one for each elementary articulator, each of whose values corresponding to the influence of an elementary gesture on tongue position.

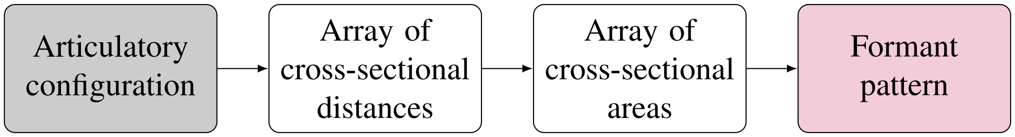
Given our matrix of cross-sectional distances D as input, a factor analysis outputs two matrices:

$$\Lambda_D = \begin{pmatrix} \lambda_{1,1} & \cdots & \lambda_{1,\omega} \\ \vdots & \ddots & \vdots \\ \lambda_{n,1} & \cdots & \lambda_{n,\omega} \end{pmatrix} \quad F_D = \begin{pmatrix} f_{1,1} & \cdots & f_{1,m} \\ \vdots & \ddots & \vdots \\ f_{\omega,1} & \cdots & f_{\omega,m} \end{pmatrix}$$

satisfying (in addition to some constraints) the decomposition equation $D = \Lambda_D F_D + \epsilon$ where ϵ is a matrix containing error values of the decomposition (and the mean of the random vector yielding D, normally subtracted from D in the above equation, is suppressed). The matrices Λ_D and F_D decompose D into "factor loadings" (the entries of Λ_D) and ω many unobservable random variables, called "factors," realized here as the columns in F_D. The jth column of F_D represents a "state in factor space" that corresponds to the cross-sectional distance vector d_j, and the vocal-tract shape that d_j represents. Note that Λ_D and F_D are unique only up to orthogonal transformation. Ideally, the factors of D are interpretable as elementary articulators, and their values as elementary gestures, yet, this is not always the case, especially since rotations (orthogonal transformations) of F_D are also solutions to the decomposition equation. Maeda (1979) uses "arbitrary" factor analysis to improve the interpretability of factors as components of elementary articulators. A factor analysis on D whose factors are interpretable as elementary articulators is called an *elementary articulatory factor model*. The factors themselves are called *elementary articulators*, and the values they take on are called *elementary gestures*.

Let $A_D = (\Lambda_D, F_D)$ be an elementary articulatory factor model on a distance matrix D. Since the columns in F_D can be viewed as tuples of elementary gestures, we can extrapolate a space $F_A \subseteq \mathbb{R}^7$ that contains tuples whose components are possible elementary gestures.

The points in F_A are called *articulatory configurations*, and F_A is called a *configuration space*. Since the dimensions of the ambient space $\mathbb{R}^7$ correspond to elementary articulators, we refer to $\mathbb{R}^7$ as an *articulatory space*, which is fixed as the *articulatory reference frame* throughout this dissertation. The matrix Λ_D is then re-interpreted as a linear transformation Λ_A from the configuration space F_A to the distance vector space $\mathbb{R}^n$. The linear transformation Λ_A is called a *linear articulatory transformation*, and the ordered pair (Λ_A, F_A) is called a *linear articulatory model* (LAM). Given a LAM, we have the following sequence of transformations simulating the articulatory-acoustic relationship:


Articulatory configuration
Array of cross-sectional distances
Array of cross-sectional areas
Formant pattern

The sequence of transformations described above can be composed into a relation over the set of articulatory configurations and a set of formant patterns. In the remainder of this dissertation, we focus on relations between articulatory configurations and formant patterns, and the vowel signals they yield. The modeling of vocal learning carried out in Chapter 3 made use of auditory representations derived from vowel signals yielded by articulatory configurations, yet the articulatory configurations themselves were not explicitly modeled. Moreover, the auditory representations were "static," in the sense that they captured only component-wise structural relations, as opposed to broader vowel-internal structural relations. In the coming sections, we expand our model of vocal learning by explicitly incorporating articulatory configurations and "multifold" auditory representations of the vowel signals they yield.

4.2.2 Articulatory Configurations as Representations

The relation over the set of articulatory configurations and a set of formant patterns described by a LAM models vocal tracts that fall within a small age range. In order to model vocal tracts that fall outside of this range, this relation must be made modifiable in some way. Boë and Maeda (1998) put forward a "growth model" extension of the LAM approach described in the previous section, called the *variable linear articulatory model*, or VLAM. According to Boë (1999), the VLAM modifies "the longitudinal dimension of the vocal tract according to two scaling factors: one for the anterior part of the vocal tract and the other for the pharynx" (p. 2502). For example, the "infant" (blue) and "adult" (magenta) vocal tracts depicted in Figure 4.6 (left) result from different anterior and pharyngeal scaling values. The maximal vowel spaces these two scale settings are capable of yielding are pictured in Figure 4.6 (right). More generally, this non-uniform scaling "permits the evolution of the vocal tract shape to be simulated, month by month and year by year" (ibid). Thus, rather than a single relation between a set of articulatory configurations and formant patterns modeling a small range of vocal tracts, the VLAM provides a collection of such relations modeling a far wider range of vocal tract ages in terms of lengths and structures. Boë et al. (2002) argue that this approach is capable of modeling vocal tracts of lengths ranging from those of adults to young infants, as well as vocal tract structures of other species (e.g., Neanderthals).

It should be mentioned that the aptness of VLAM vocal tracts to certain modeling activities has been challenged, e.g., by de Boer and Fitch (2010), particularly with respect to infant vocal tracts. Specifically, the criticism centers on the "logical circularity" of the use of adult production data in constructing models which are then used as evidence for the production capabilities of infants. Although we do not disagree with the criticism, per se,

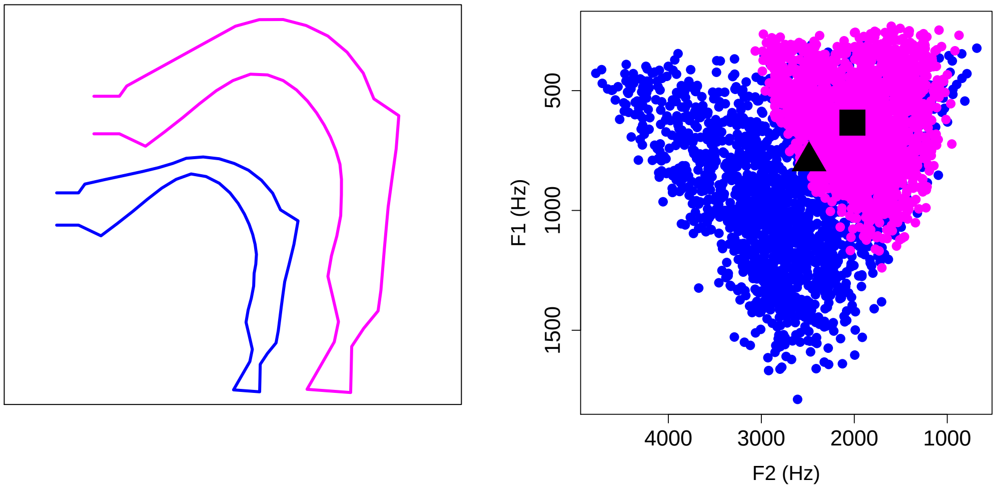


Figure 4.6: (left) Vlab midsagittal representations of a neutral articulatory configuration for ages 0.5 and 10 in years, along with the corresponding densely-sampled MVS within formant space (right). The triangle and square indicate the formant vectors corresponding to the neutral infant and adult midsagittal representations, respectively.

we appeal to the position expressed in the Preface of this dissertation. That is, modeling of the articulations of infants begins with the unrealistic assumption that the infant articulatory system is a (nonuniformly) scaled-down version of the articulatory system of an adult, and progress can be made in this fashion, even as we seek insight into the differences between infant and adult articulatory systems, and how an infant overcomes the differences in spoken language acquisition. In other words, we view the VLAM as a scaffold, rather than a dais, for further work.

VLAM-based vocal tract modeling continues to be modified and refined as more is learned about infant articulation. In this dissertation we use a recent version called the “Vlab” (Boë et al., 2010), which is based on more complete data on vocal tract growth, including that of young children. The seven elementary articulators are those mentioned

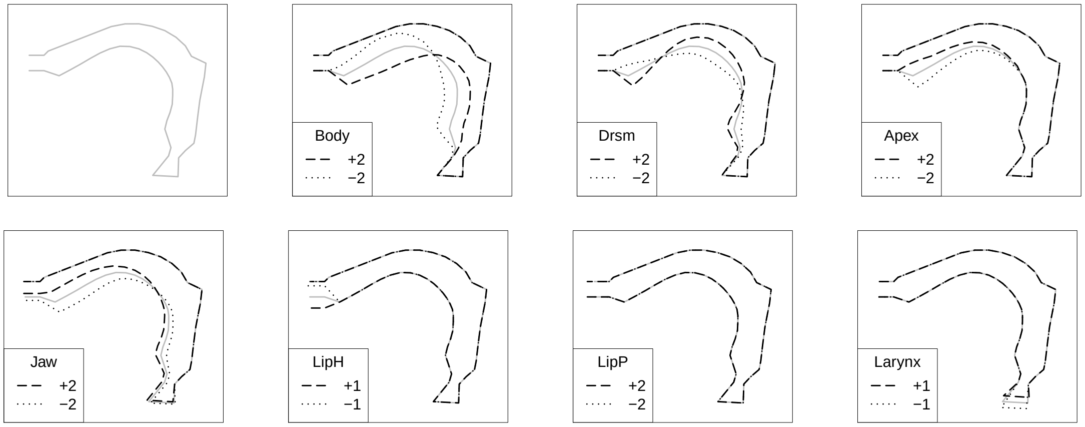


Figure 4.7: The age 10 neutral articulatory configuration (top, left), and seven elementary gestures: (top, left to right) the tongue body, dorsum, and tip (or apex), (bottom, left to right) jaw raising/lowering, lip height and protrusion, and the larynx. In the case of lip protrusion, the influence of the articulator results in either extended (+2) or retracted (-2) lips, relative to the neutral tract.

in the previous section: the jaw, the tongue body, dorsum, and tip (or apex), lip height and protrusion, and the larynx. The range of elementary gestures for the first six articulators is roughly $[-3, 3]$, and $[-1, 1]$ for the larynx. The influence of the elementary gestures on tongue shape is depicted in Figure 4.7 for the Vlab set at age 10. The top, left image depicts the vocal tract yielded by the *neutral articulatory configuration* $(0, 0, 0, 0, 0, 0, 0)$. For each elementary articulator, the corresponding image depicts vocal tracts yielded by articulatory configurations whose only contributing factor is that articulator. In the case of lip protrusion, the influence of the articulator results in either extended (+2) or retracted (-2) lips, relative to the neutral tract. The Vlab's nonuniform scaling is exhibited in Figure 4.6. The "infant" (blue) and "adult" (magenta) vocal tract representations depicted (left) are output by the Vlab set at ages 0.5 and 10 years, respectively, for the neutral configuration. The maximal vowel spaces $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$ are represented in Figure 4.6 (right). Below is a list of configurations for corner vowels for each of these ages, with the corresponding vocal tract and MVS representations depicted in Figure 4.8.

	Jaw	Body	Drsm	Apex	LipP	LipH	Larynx
neutral	0	0	0	0	0	0	0
infant /i/	1	-2	-2	-3	-2	2	0
infant /a/	-3	2	-1	-3	-3	3	0
infant /u/	2	2	3	-3	0	-1	0
adult /i/	2	-3	-2	-3	-3	1	0
adult /a/	-3	2	-1	-3	-3	3	0
adult /u/	1	3	3	0	3	-1	0

Recall that VLABAGES is the set of all possible age settings for the Vlab. Notice that the configuration for the infant /a/ is identical to the configuration for the adult /a/. More generally, the configuration spaces for each age in VLABAGES may overlap substantially, yet, as this example illustrates, what differs across ages is the relation between configurations and formant patterns. Thus the following conceptual organization is meaningful. For each age $a \in$ VLABAGES, the set of all articulatory configurations for a that do not result in occlusion of the oral cavity is called a *maximal articulatory space for age a*, denoted MARS(a). Each articulatory configuration within MARS(a) is identified with an *articulatory vector* whose components are the values of the elementary articulator components for that articulatory configuration. For each $a \in$ VLABAGES, each articulatory vector $\mathbf{a}^k \in$ MARS(a) yields a formant vector $\mathbf{f}^k$ in the maximal vowel space MVS(a).

4.2.3 Multifold Representations of Vowels

In this section, we broadly describe the Vlab's vowel synthesis procedure and the steps involved in going from articulatory synthesis to the auditory modeling described in Section 3.3.1. We illustrate the procedure by extending the example from Section 3.3.1, taking as our starting point the neutral configuration $(0, 0, 0, 0, 0, 0, 0)$, which yields the vowel signals $s(0.5)$ and $s(10)$ depicted in Figure 3.9 (top) and Figure 4.9 (bottom) for Vlab ages

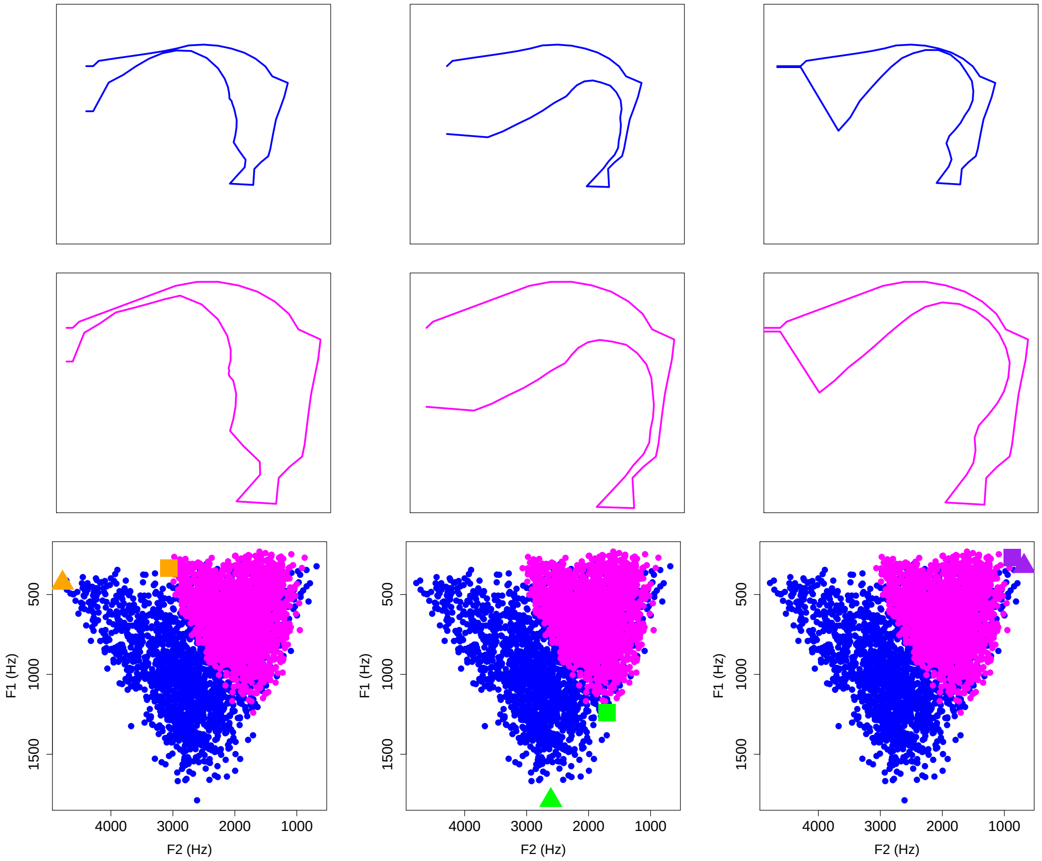


Figure 4.8: The top row contains age 0.5 Vlab midsagittal representations yielded by articulatory configurations for the corner vowels /i/ (left), /a/ (mid), and /u/ (right). Similarly, the middle row contains age 10 Vlab midsagittal representations yielded by articulatory configurations for the corner vowels /i/ (left), /a/ (mid), and /u/ (right). The triangles and squares within each MVS indicate the formant vectors corresponding to the infant and adult midsagittal representations, respectively.

0.5 and 10. In addition to an articulatory configuration, the synthesis procedure requires a signal duration value, and a fundamental frequency trajectory, both of which are independently specifiable. Throughout this dissertation, we fix the duration for all synthesized vowels at 500ms. The fundamental frequency trajectory varies with modeling age, though is held constant within each age. The trajectory computation for an age a is a smoothing

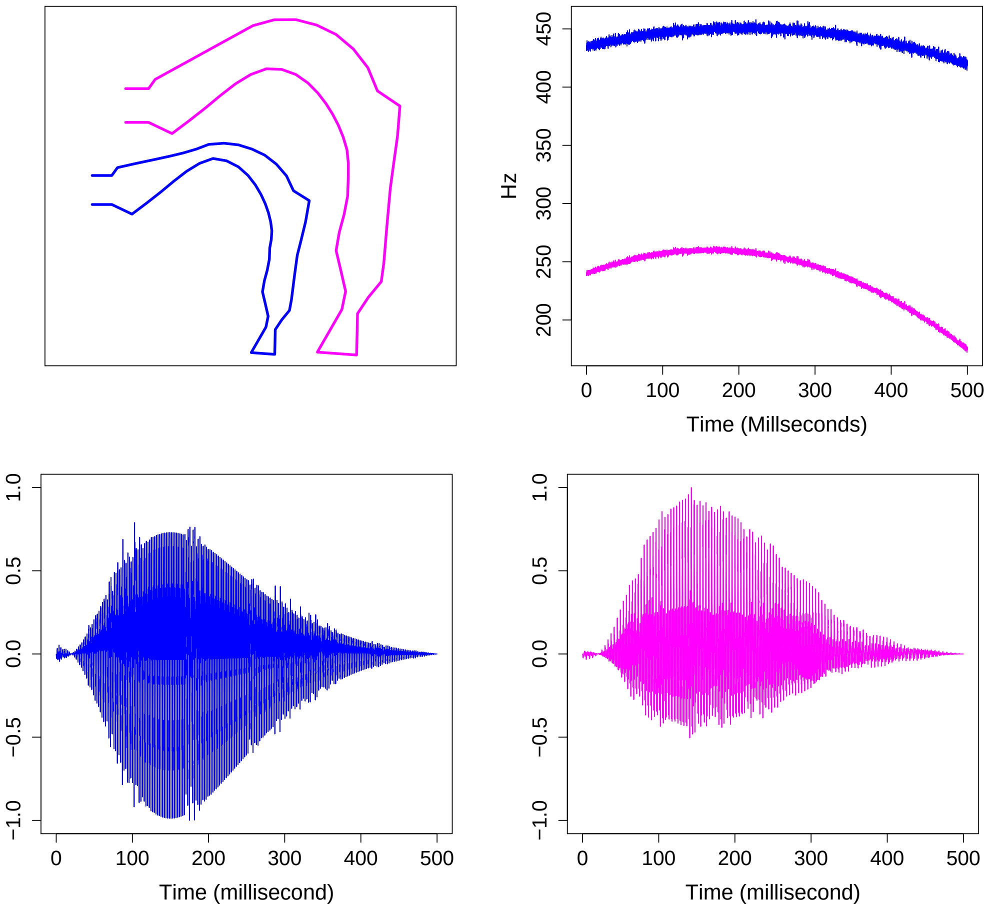


Figure 4.9: The vocal tract representations (top, left), fundamental frequency trajectories (top, right), and synthesized vowel signals corresponding to the neutral configuration for ages 0.5 (bottom, left) and 10 (bottom, right)

spline interpolation over 3 fundamental frequency values which are represented as a *fundamental frequency vector* $f_0(a) = (f_0^1(a), f_0^2(a), f_0^3(a))$. We take $f_0(0.5) = (435, 450, 420)$ and $f_0(10) = (240, 260, 175)$. The midsagittal vocal tract representations, fundamental frequency trajectories, and synthesized vowel signals corresponding to the neutral configuration for ages 0.5 and 10 are depicted in Figure 4.9.

In Section 3.3.1, the vowel signals $s(0.5)$ and $s(10)$ were used to illustrate the computation of excitation pattern representations from their spectral representations (see Figure 3.10). The spectral representations were taken to be multitaper spectra (see Reidy, 2013) computed over a single time slice from each signal. The use of one time slice from a vowel signal in constructing spectral and excitation pattern representations yields “static” information about that signal, i.e., information about representation-internal relations, e.g., the the vowel-internal relation between F1 and F2. In this section, we move toward “multifold” representations of vowel signals that yield vowel-internal information across spectral and excitation pattern representations. We limit our consideration of multifold representations to vowel signals themselves, to the exclusion of “multifold articulatory representations,” which involve relations across articulatory representations. We illustrate multifolds with auditory multifolds that use more than one time slice from a vowel signal. Given a vowel signal s, a *time slice sequence of order n* from s is simply n-many time slices from s. Given a time slice sequence of order n from s, an *n-fold spectral representation* for a vowel signal s is taken to be the n-many multitaper spectra of the n-many time slices from s.

Recall that we use a fixed gammatone filter bank $\mathcal{G}_{\mathsf{ERB}(C)}$ with 36 channels as a model of the basilar membrane. A filter bank is *applied* to an n-fold spectral representation of s through application of the filter to each of the n-many representations of s. The $\log$ output of a gammatone filter bank $\mathcal{G}_{\mathsf{ERB}(C)}$ applied to an n-fold spectral representation for a vowel signal s is called an *n-fold excitation pattern* for s under $\mathcal{G}_{\mathsf{ERB}(C)}$. The dashed lines in Figure 4.10 (top) represent time slice sequences of order 3 over the vowel signals $s(0.5)$ and $s(10)$. The corresponding 3-fold excitation patterns are pictured in Figure 4.10

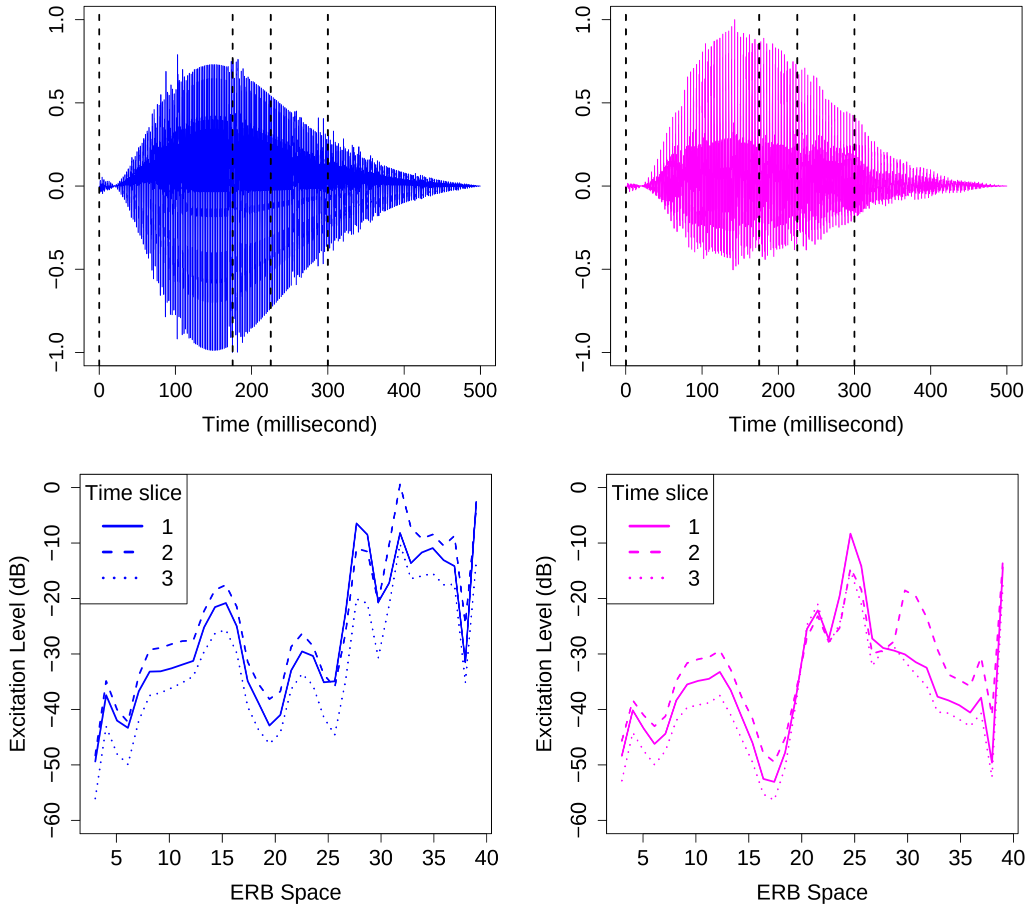


Figure 4.10: (top) Vowel signals $s(0.5)$ (left) and $s(10)$ (right) with time slice sequences of order 3, and the corresponding 3-fold excitation patterns (bottom). Cursors on the upper graph demarcate the analysis windows.

(bottom). We define an *n-fold excitation vector* in the obvious way. In the remainder of this chapter all n-fold excitation vectors are obtained in this fashion.

Given a formant vector $\mathbf{f}^k \in \textsc{mvs}(a)$, we denote its corresponding n-fold excitation vector as $\{\mathbf{e}^{k_i} \mid 1 \leq i \leq n\}$. Although the subscripting on the index k reflects the fact that the excitation vectors $\mathbf{e}^{k_i}$ derive from the vowel signal corresponding to the formant vector $\mathbf{f}^k$, it is assumed that the $\mathbf{e}^{k_i}$ are independent of each other. Given this assumption, for each maximal vowel space $\textsc{mvs}(a)$, we define the corresponding *n-fold maximal auditory space* as follows: $\textsc{mauds}_n(a) = \bigcup\{\mathbf{e}^{k_i} \mid \mathbf{f}^k \in \textsc{mvs}(a) \text{ and } 1 \leq i \leq n\}$. Note that each maximal auditory space $\textsc{mauds}(a)$ is a 1-fold maximal auditory space. In the remainder of this dissertation, we fix the order of n-fold representations at $n \in \{1, 3\}$. We refer to $\textsc{mauds}(a) = \textsc{mauds}_1(a)$ as a *static maximal auditory space*. The subscript on $\textsc{mauds}(a)$ is omitted when the distinction between static and general n-fold is not relevant.

Since each $\textsc{mauds}(a)$ is embedded within the auditory reference frame $\mathbb{R}^{36}$, it is a straightforward exercise to carry over the concepts defined for maximal auditory spaces to n-fold maximal auditory spaces. Given an age a, an *n-fold auditory manifold* over $\textsc{mauds}(a)$, denoted $M(a) = (V(a), E(a), w(a))$, is simply a weighted graph derived from $\textsc{mauds}_n(a)$. We assume an indexing on $V(a)$ where the vertex $v_{k_i} \in V(a)$ corresponds to $\mathbf{e}^{k_i} \in \textsc{mauds}_n(a)$. Given a response pairing $T(a_0, a_1)$ over $\textsc{mvs}(a_0)$ and $\textsc{mvs}(a_1)$ for $\ell \in \textsc{lang}$ with category transfer function $C(T(a_0, a_1))$, an *auditory pairing* $I(T(a_0, a_1))$ is a set of ordered pairs $(\mathbf{e}^{j_i}, \mathbf{e}^{k_i})$ where $(\mathbf{f}^j, \mathbf{f}^k) \in T(a_0, a_1)$. A socio-auditory weighting over $I(T(a_0, a_1))$ is a nonnegative function

$$S(C(T(a_0, a_1))) : I(T(a_0, a_1)) \to \mathbb{R}.$$

For each auditory pair $(\mathbf{e}^{j_i}, \mathbf{e}^{k_i}) \in I(T(a_0, a_1))$, we have $(\mathtt{c}, g) = C(T(a_0, a_1))(\mathbf{f}^j, \mathbf{f}^k)$, and the socio-auditory weight $\iota(g)$ assigned to the auditory pair is a function of the transfer weight g. A category transfer interpretation over $I(T(a_0, a_1))$ for a language ℓ is a function

$$I(C(T(a_0, a_1))) : I(T(a_0, a_1)) \to C_\ell.$$

Each category transfer function $C(T(a_0, a_1))$ yields a simple category transfer interpretation over $I(T(a_0, a_1))$ whereby each auditory pair $(\mathbf{e}^{j_i}, \mathbf{e}^{k_i}) \in I(T(a_0, a_1))$ is assigned the category $\mathtt{c}$ where $(\mathtt{c}, g) = C(T(a_0, a_1))(\mathbf{f}^j, \mathbf{f}^k)$.

The primary advantage of defining n-fold maximal auditory spaces in this manner is the linear increase in number of representations that enter into auditory pairings. This facilitates the normalization computation, making it more robust, as will be shown in Section 4.4. In the next section, we turn to a far broader extension of the modeling framework.

4.3 Intermodal Structures

In this section, we continue to extend our model of infant agents and the structures they use for organizing vocal interaction with caretakers and the internal computations carried out over their representations of acoustic and social signals put forward in Section 3.3. The main aspects of the extension are presented in the remainder of this section, beginning with the modeling of the organization of articulatory representations using manifolds, and the creation of intermodal representations through the alignment of articulatory and auditory manifolds based on caretaker response pairings. These representations are then organized using manifolds, which are again aligned based on caretaker responses. We take the latter computation to be an "intermodal" vowel normalization, which expands on the conceptualization put forward in the previous chapter. In the remainder of this section, we formulate the approach, and follow with a demonstration of the concepts in the next section.

4.3.1 Articulatory Manifolds and Intermodal Computations

Recall that within our vocal learning environment, infants are modeled in terms of cognitive structures used to organize auditory representations and caretaker responses. The main kind of structure used is a manifold, as defined in Section 2.3 and discussed in Chapters 2 and 3. The other kind of structure used is a pairing, which is simply a set of ordered pairs, as discussed in Chapter 3. Both kinds of structures play a key role in the extended model of infants put forward in this section, and familiarity with the definitions is assumed throughout the remainder of this chapter.

Given a maximal articulatory space $\textsc{mars}(a)$, an *articulatory manifold over* $\textsc{mars}(a)$, denoted $A(a) = (V(a), E(a), w(a))$, is simply a weighted graph derived from $\textsc{mars}(a)$. We assume an indexing on $V(a)$ where the vertex $v_k \in V(a)$ corresponds to $\mathbf{a}^k \in \textsc{mars}(a)$. For simplicity, weight functions are assumed to be constant, assigning a value of 1 to each edge of an articulatory manifold, unless otherwise stated. Articulatory manifolds model an infant's cognitive organization of their articulatory representations, and together with multifold auditory manifolds, form the basis of our intermodal model of vowel normalization.

In the remainder of this section, we restrict our focus to static maximal auditory spaces. The formulation that follows is meant to extend the role of auditory pairings derived from caretaker response pairings by relating each auditory representation in an auditory pairing to an articulatory representation, yielding "sensorimotor pairings" to be used in manifold alignment. The restriction to static auditory manifolds is due to the vagueness that comes with linking n-fold auditory representations with articulatory representations, whose clarification is beyond the scope of the present work. At present, we define a *sensorimotor pairing for age* a as set of ordered pairs $(\mathbf{a}^k, \mathbf{e}^j)$ where $\mathbf{a}^k \in \textsc{mars}(a)$ and $\mathbf{e}^j \in \textsc{mauds}(a')$.

Let $I(T(a_0, a_1))$ be an auditory pairing derived from a response pairing $T(a_0, a_1)$ over $\textsc{mvs}(a_0)$ and $\textsc{mvs}(a_1)$ for $\ell \in \textsc{lang}$ with category transfer function $C(T(a_0, a_1))$. The auditory pairing $I(T(a_0, a_1))$ provides the basis for two sensorimotor pairings relative to the maximal articulatory set $\textsc{mars}(a_1)$, which are derived as follows. Let $\pi_1(I(T(a_0, a_1))) = \{\mathbf{e}^j \mid (\mathbf{e}^k, \mathbf{e}^j) \in I(T(a_0, a_1))$. The *sensorimotor pairing for age* a_1 *relative to* $\textsc{mars}(a_1)$ derived from $I(T(a_0, a_1))$, denoted $I_{a_1}(T(a_0, a_1))$, is the sensorimotor pairing whose pairs are of the form $(\mathbf{a}^k, \mathbf{e}^k)$ where $\mathbf{e}^k \in \pi_1(I(T(a_0, a_1)))$ and $\mathbf{a}^k \in \textsc{mars}(a_1)$ is the articulatory vector that yields the formant vector $\mathbf{f}^k \in \textsc{mvs}(a_1)$, whose vowel signal yields $\mathbf{e}^k$. The *sensorimotor pairing for age* a_0 *relative to* $\textsc{mars}(a_1)$ derived from $I(T(a_0, a_1))$, denoted $I^{a_0}(T(a_0, a_1))$, is the sensorimotor pairing $\{(\mathbf{a}^k, \mathbf{e}^j) \mid (\mathbf{a}^k, \mathbf{e}^k) \in I_{a_1}(T(a_0, a_1))\}$. The set $I^{a_0}_{a_1}(T(a_0, a_1))$ containing these sensorimotor pairings is called the *sensorimotor pairing structure relative to* $\textsc{mars}(a_1)$ derived from $I(T(a_0, a_1))$.

Let $S(C(T(a_0, a_1)))$ be a socio-auditory weighting over $I(T(a_0, a_1))$, and moreover let $I^{a_0}_{a_1}(T(a_0, a_1))$ be a sensorimotor pairing structure relative to $\textsc{mars}(a_1)$ derived from $I(T(a_0, a_1))$. Without loss of generality, consider $I_{a_1}(T(a_0, a_1))$, the sensorimotor pairing for age a_1 relative to $\textsc{mars}(a_1)$. We define a *socio-sensorimotor weighting* over $I_{a_1}(T(a_0, a_1))$ derived from $S(C(T(a_0, a_1)))$ to be a nonnegative function

$$S_{a_1}(C(T(a_0, a_1))) : I_{a_1}(T(a_0, a_1)) \to \mathbb{R}.$$

For each auditory pair $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$, we have $\iota(g) = S(C(T(a_0, a_1))(\mathbf{e}^j, \mathbf{e}^k)$ where g is the transfer weight assigned to $(\mathbf{f}^j, \mathbf{f}^k)$. The *socio-sensorimotor weight* $\delta(\iota(g)))$ assigned to a sensorimotor pair $(\mathbf{a}^k, \mathbf{e}^k) \in I_{a_1}(T(a_0, a_1))$ is a function of the socio-auditory weight $\iota(g)$. A sensorimotor pairing with a socio-sensorimotor weighting is called a *socio-sensorimotor pairing*. The same definitions apply to $I^{a_0}(T(a_0, a_1))$, the sensorimotor pairing for age a_0 relative to $\textsc{mars}(a_1)$. Specifically, a *socio-sensorimotor weighting* over $I^{a_0}(T(a_0, a_1))$ derived from $S(C(T(a_0, a_1)))$ is a nonnegative function

$$S^{a_0}(C(T(a_0, a_1))) : I^{a_0}(T(a_0, a_1)) \to \mathbb{R}.$$

It is assumed that the socio-sensorimotor weight assigned to a sensorimotor pair $(\mathbf{a}^k, \mathbf{e}^j) \in I^{a_0}(T(a_0, a_1))$ is the socio-sensorimotor weight $\delta(\iota(g)))$ assigned to the sensorimotor pair $(\mathbf{a}^k, \mathbf{e}^k) \in I_{a_1}(T(a_0, a_1))$, whence $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$. The set $S^{a_0}_{a_1}(C(T(a_0, a_1)))$

containing these socio-sensorimotor weightings is called the *socio-sensorimotor weighting structure relative to* $\textsc{mars}(a_1)$ derived from $S(C(T(a_0, a_1)))$. A sensorimotor pairing structure together with a socio-sensorimotor weighting structure is called a *socio-sensorimotor pairing structure*. Socio-sensorimotor pairing structures, like socio-auditory pairings, model an infant's internalization of the acoustic and social signals provided by a caretaker during turn-taking vocal exchanges.

We turn now to computations over manifolds. Let $M(a_0)$ and $M(a_1)$ be auditory manifolds, $A(a_1)$ an articulatory manifold, and let $I(T(a_0, a_1))$ be a socio-auditory pairing, which is assumed to be an alignment for $\textsc{mauds}(a_0)$ and $\textsc{mauds}(a_1)$. The sensorimotor pairings $I_{a_1}(T(a_0, a_1))$ and $I^{a_0}(T(a_0, a_1))$ derived from $I(T(a_0, a_1))$ yield alignments for $\textsc{mauds}(a_1)$ and $\textsc{mars}(a_1)$, and $\textsc{mauds}(a_0)$ and $\textsc{mars}(a_1)$, respectively. Let $S^{a_0}_{a_1}(C(T(a_0, a_1)))$ be a socio-sensorimotor pairing structure derived from the socio-auditory weighting $S(C(T(a_0, a_1)))$. The socio-sensorimotor weightings $S_{a_1}(C(T(a_0, a_1)))$ and $S^{a_0}(C(T(a_0, a_1)))$ yield weight functions on the respective alignment relations derived from $I_{a_1}(T(a_0, a_1))$ and $I^{a_0}(T(a_0, a_1))$. These weight functions are used to form the combined weighted graphs $H(a, a_1)$, from $A(a_1)$ and $M(a)$, where $a \in \{a_0, a_1\}$, called *sensorimotor manifolds over $A(a_1)$ and $M(a)$ derived from $I^{a_0}_{a_1}(T(a_0, a_1))$*, or simply *sensorimotor manifolds*. The computation involving the combination of the manifolds $A(a_1)$ and $M(a)$ is called *sensorimotor alignment*. To emphasize the importance of socio-sensorimotor pairing structures, we notate the sensorimotor alignment of manifolds $A(a_1)$ and $M(a)$ via a socio-sensorimotor weighting $S^{a_0}_{a_1}(C(T(a_0, a_1)))$ as a mapping over triples:

$$\textsc{smAlig} : (A(a_1), M(a), S^{a_0}_{a_1}(C(T(a_0, a_1)))) \mapsto H(a, a_1).$$

The notation highlights the necessity of each of the cognitive structures $M(a)$, $A(a_1)$ and $I_{a_1}^{a_0}(T(a_0, a_1))$ in yielding the aligned manifold $H(a, a_1)$. Sensorimotor manifolds provide the basis for intermodal vowel normalization acquisition, which we treat in the next section.

We take "intermodal" representations of auditory representations to be the output of a Laplacian eigenmapping over auditory and articulatory representations. Let $a \in \{a_0, a_1\}$ and let IMNARROW(a, a_1) be the m-dimensional eigenmap derived from $H(a, a_1)$. Let MIMS(a) be the m-dimensional eigenmap of MAUDS(a) with respect to IMNARROW(a, a_1). The jth row of MIMS(a), denoted $\mathbf{c}^j$, is the *intermodal representation* of the excitation vector $\mathbf{e}^j \in$ MAUDS(a). We notate the intermodal computation as follows:

$$\text{IMNARROWING} : (H(a, a_1), S_{a_1}^{a_0}(C(T(a_0, a_1)))) \mapsto \text{IMNARROW}(a, a_1).$$

Note that IMNARROW(a, a_1), by definition, contains the representations in MIMS(a). The set MIMS(a) of intermodal representations with respect to IMNARROW(a, a_1) is called a *maximal intermodal space for age* a, which exists within an *intermodal reference frame*.

Given a maximal intermodal space MIMS(a), an *intermodal manifold over* MIMS(a), denoted $K(a) = (V(a), E(a), w(a))$, is simply a weighted graph derived from MARS(a). We assume an indexing on $V(a)$ where the vertex $v_k \in V(a)$ corresponds to $\mathbf{c}^k \in$ MIMS(a). For simplicity, weight functions are assumed to be constant, assigning a value of 1 to each edge of an intermodal manifold, unless otherwise stated. Intermodal manifolds model an infant's organization of articulatory and auditory representations into a model of "self" and a model of "other," i.e., their caretaker. As in Chapter 3, the distinction between the models of self and other is based on differentiated auditory representations, however, in the intermodal case these models include articulatory representations.

4.3.2 Intermodal Manifolds and Vowel Normalization

We now introduce the structures and computations that model the acquisition of intermodal vowel normalization. The structures are, again, manifolds and pairings, the main difference being the spaces they organize, which are taken to be maximal intermodal spaces $\text{MIMS}(a)$. We begin by defining an *intermodal pairing for ages* a_0 and a_1 as set of ordered pairs $(\mathbf{c}^j, \mathbf{c}^k)$ where $\mathbf{c}^j \in \text{MIMS}(a_0)$ and $\mathbf{c}^k \in \text{MIMS}(a_1)$. The sensorimotor pairings $I_{a_1}^{a_0}(T(a_0, a_1))$ relative to $\text{MARS}(a_1)$ derived from $I(T(a_0, a_1))$ yield an intermodal pairing for ages a_0 and a_1 in the following way. Given $(\mathbf{a}^k, \mathbf{e}^j) \in I^{a_0}(T(a_0, a_1))$ and $(\mathbf{a}^k, \mathbf{e}^k) \in I_{a_1}(T(a_0, a_1))$, form the pair $(\mathbf{c}^j, \mathbf{c}^k)$ where $\mathbf{c}^j \in \text{MIMS}(a_0)$ is the intermodal representation for $\mathbf{e}^j$, and $\mathbf{c}^k \in \text{MIMS}(a_1)$ is the intermodal representation for $\mathbf{e}^k$. The intermodal pairing is denoted $J(T(a_0, a_1))$. Note that each intermodal pair $(\mathbf{c}^j, \mathbf{c}^k) \in J(T(a_0, a_1))$ corresponds to the auditory pair $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$.

Let $S(C(T(a_0, a_1)))$ be a socio-auditory weighting over $I(T(a_0, a_1))$. We define a *socio-intermodal weighting* over $J(T(a_0, a_1))$ derived from $S(C(T(a_0, a_1)))$ to be a non-negative function

$$S_J(C(T(a_0, a_1))) : J(T(a_0, a_1)) \to \mathbb{R},$$

For each auditory pair $(\mathbf{e}^j, \mathbf{e}^k) \in I(T(a_0, a_1))$, we have $\iota(g) = S(C(T(a_0, a_1))(\mathbf{e}^j, \mathbf{e}^k)$. The *intermodal weight* $\kappa(\iota(g))$ assigned to the intermodal pair $(\mathbf{c}^j, \mathbf{c}^k) \in J(T(a_0, a_1))$ is a function of the socio-auditory weight $\iota(g)$. An intermodal pairing with a socio-intermodal weighting is called a *socio-intermodal pairing*. Socio-intermodal pairings, like socio-auditory pairings, model an infant's internalization of the acoustic and social signals provided by a caretaker during turn-taking vocal exchanges.

The remaining definitions in this section mirror those from Section 3.3.2. We define a *pre-categorical equivalence* over $J(T(a_0, a_1))$ as an equivalence relation over the pairs in $J(T(a_0, a_1))$. We also define a *category transfer interpretation* over $J(T(a_0, a_1))$ for a language ℓ to be a function

$$J(C(T(a_0, a_1))) : J(T(a_0, a_1)) \to C_\ell.$$

Each category transfer interpretation over $I(C(T(a_0, a_1))$ yields a *simple category transfer interpretation* over $J(T(a_0, a_1))$ whereby each intermodal pair $(\mathbf{c}^j, \mathbf{c}^k) \in J(T(a_0, a_1))$ is assigned the category c where $\mathsf{c} = I(C(T(a_0, a_1)))(\mathbf{e}^j, \mathbf{e}^k)$. An intermodal pairing with a category transfer interpretation is called a *categorical intermodal pairing*. A socio-intermodal pairing with a category transfer interpretation is called a *socio-categorical intermodal pairing*. Each category transfer interpretation for ℓ over $J(T(a_0, a_1))$ yields a pre-categorical equivalence over $J(T(a_0, a_1))$ in the obvious way. We restrict our attention to pre-categorical equivalences derived from category transfer interpretations. However, this is a simplifying assumption.

We turn now to computations over intermodal manifolds. Let $K(a_0)$ and $K(a_1)$ be intermodal manifolds, and let $J(T(a_0, a_1))$ be a socio-categorical intermodal pairing, which is assumed to be an alignment for $\textsc{mims}(a_0)$ and $\textsc{mims}(a_1)$. The socio-intermodal weighting on the pairing $J(T(a_0, a_1))$ yields a weight function on the alignment relation derived from $J(T(a_0, a_1))$, which is used to form a combined weighted graph $K(a_0, a_1)$ from $K(a_0)$ and $K(a_1)$, called an *intermodal commensuration manifold over $K(a_0)$ and $K(a_1)$ derived from $J(T(a_0, a_1))$*, or simply a *commensuration manifold*. The computation involving the combination of the manifolds $K(a_0)$ and $K(a_1)$ is called *intermodal normalization*. That is, intermodal normalization is a binary operation on manifolds, and can be viewed as a structure co-opting generative computation. To emphasize the importance of socio-intermodal pairing, we notate the intermodal normalization of manifolds $K(a_0)$ and $K(a_1)$

via a socio-intermodal weighting $S_J(C(T(a_0, a_1)))$ as a mapping over triples:

$$\textsc{imNorm} : (K(a_0), K(a_1), S_J(C(T(a_0, a_1)))) \mapsto K(a_0, a_1).$$

The notation highlights the necessity of each of the cognitive structures $K(a_0)$, $K(a_1)$ and $J(T(a_0, a_1))$ in yielding the aligned manifold $K(a_0, a_1)$. Within the extended vocal learning model, commensuration manifolds may provide the basis for vowel category acquisition as well as the perceptual magnet effect.

Let $K(a_0, a_1)$ be a commensuration manifold derived from the pairing $J(T(a_0, a_1))$, and let $J(C(T(a_0, a_1))$ be the category transfer interpretation over $J(T(a_0, a_1))$ for ℓ. Moreover, let $\textsc{mims}(a_0, a_1) = \textsc{mims}(a_0) + \textsc{mims}(a_1)$ (i.e., their disjoint union). Let $L(a_0, a_1)$ be the graph Laplacian for $K(a_0, a_1)$. The graph Laplacian $K(a_0, a_1)$ provides the means for (i) deriving "commensuration representations" of the representations in $\textsc{mims}(a_0, a_1)$, and (ii) extending the pre-categorical equivalence over $J(T(a_0, a_1))$ to all of $\textsc{mims}(a_0, a_1)$. The former is achieved through the use of manifold regularization (Belkin et al., 2004, 2006), which spreads the pre-categorical equivalence to the whole of the aligned manifold $M(a_0, a_1)$. We use the plearn algorithm available at `http://www.cse.ohio-state.edu/~mbelkin/algorithms/Laplacian.tar`. We notate the computation of equivalence via a category transfer interpretation $J(C(T(a_0, a_1)))$ as a mapping over ordered pairs:

$$\textsc{imEquiv} : (K(a_0, a_1), J(C(T(a_0, a_1)))) \mapsto \mathsf{equiv}(\textsc{mims}(a_0, a_1)).$$

The equivalence relation $\mathsf{equiv}(\textsc{mims}(a_0, a_1))$ is called a *categorical equivalence over* $\textsc{mims}(a_0, a_1)$, or simply, a *categorical equivalence*. The notation highlights the necessity of each of the cognitive structures $K(a_0, a_1)$ and $J(C(T(a_0, a_1)))$ in yielding the equivalence over $\textsc{mims}(a_0, a_1)$.

Concerning the derivation of commensurate representations, we need a few definitions. Given an intermodal manifold $K(a)$ over MIMS(a), an *m-dimensional intermodal warping of* MIMS(a), denoted IMWARP(a), is an m-dimensional eigenmapping derived from $K(a)$. The jth row of IMWARP(a), denoted $\mathbf{z}^j$ is the *warped representation* of the intermodal vector $\mathbf{c}^j \in$ MIMS(a). Similarly, intermodal warping applies to commensuration manifolds. Let IMWARP(a_0, a_1) be the m-dimensional eigenmap derived from $K(a_0, a_1)$, and IMWARP(a_0) the m-dimensional eigenmapping of MIMS(a_0) with respect to IMWARP(a_0, a_1), and IMWARP(a_1) the m-dimensional eigenmapping of MIMS(a_1) with respect to IMWARP(a_0, a_1). The jth row of IMWARP(a_0), denoted $\mathbf{z}^j$ is the *warped representation* of the intermodal vector $\mathbf{c}^j \in$ MIMS(a_0). Similarly, the kth row of IMWARP(a_1), denoted $\mathbf{z}^k$ is the commensuration representation of the intermodal vector $\mathbf{c}^k \in$ MIMS(a_1). We take Laplacian eigenmapping to be an operation on intermodal manifolds called *intermodal warping*, which is denoted as follows:

$$\text{IMWARP} : (K(a_0, a_1), S_J(C(T(a_0, a_1))) \mapsto \text{IMWARP}(a_0, a_1).$$

The notation reflects the fact that the warped representations in IMWARP(a_0, a_1) derive from the commensuration manifold $M(a_0, a_1)$. The warped representations exist within a *warped reference frame*.

4.3.3 Infant Structures

Recall that within our vocal learning environment as established in Chapter 3, infant agents, or simply infants, are modeled as structures $(M, I(T))$ where M is a set of manifolds, and $I(T)$ is a set of socio-auditory pairings over the manifolds in M, together with category transfer interpretations. The extensions put forward in this chapter are incorporated as follows.

Definition 4.1 (Infant Agent)**.** Given a caretaker $c^{\ell}_{a_0} = (Q, T)$, an *infant of age* a_1 *of* $c^{\ell}_{a_0}$ is a structure $i^{\ell}_{a_1} = (M, I(T))$ where M is a set of manifolds $M(a)$ over $\textsc{mauds}(a)$ for each $a \in \{a_0, a_1\}$. The age a_1 identifies $\textsc{mauds}(a_1)$ as the infant's own maximal auditory space. The set $I(T)$ is accordingly composed of socio-auditory pairings $I(T(a, a_1))$, assumed to be alignments for $\textsc{mauds}(a_0)$ and $\textsc{mauds}(a_1)$, along with their corresponding category transfer interpretations $I(C(T(a_0, a_1)))$. The socio-auditory pairings and their corresponding category transfer interpretations yield socio-intermodal pairings that are alignments for $\textsc{mauds}(a_1)$ and $\textsc{mars}(a_1)$, and $\textsc{mauds}(a_0)$ and $\textsc{mars}(a_1)$. They also yield socio-intermodal pairings that are alignments for $\textsc{mims}(a_0)$ and $\textsc{mims}(a_1)$. We assume that the set of manifolds M inherits the binary operations audNorm, smAlig, and imNorm (all of which are instances of the same operation), the operations audEquiv, imEquiv (both of which are instances of the same operation), and the operations audWarp, imNarrowing, and imWarp (all of which are instances of the same operation).

The basic idea behind the formulation is that the infant is internalizing the attempt at vocal interaction made by the caretaker in their own attempt to establish commensuration across manifold representations of the self and caretaker. Hence the normalization computations the infant carries out are based on response pairings derived from the caretaker. In the next section, we again show how formulating infants in this fashion provides a potential basis for a developmental model of vowel category acquisition and the perceptual magnet effect.

4.4 Vocal Learning Environment Redux

In order to illustrate the definitions and concepts put forward in this chapter, we again create a simple vocal learning environment. We provide visualization of each stage of model creation (when possible), along with interpretation and discussion of the modeling output. We again use the MVSs $\textsc{mvs}(a)$ for $a \in \textsc{vlabAges}$, and $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$ as shown in Figure 4.11.

For each language community $\ell \in \textsc{lang}$, and for each subject s^{ℓ}_{τ}, we create a caretaker $\ell^{\tau}_{10} = (Q^{\ell}_{\tau}, T^{\ell}_{\tau})$, modifying the caretaker notion slightly in order to emphasize the subject and language community the caretaker model derives from. For each $\mathsf{c} \in C_{\ell}$ and $a \in \textsc{vlabAges}$, we take $Q^{\ell}_{\tau}(\mathsf{c}, a) = P'_{\mathsf{c}}(s^{\ell}_{\tau}, 10, a)$ (with vowel category extension parameter $\alpha = 0.5$) to be a VCRS in Q^{ℓ}_{τ}. Moreover, for each $\mathsf{c} \in C_{\ell}$ and $a \in \textsc{vlabAges}$, let $b_{\upsilon}(\mathsf{c}, a)$ be the set of υ formant vectors in $\textsc{mvs}(a)$ with the highest VCRS values under $Q^{\ell}_{\tau}(\mathsf{c}, a)$, indexed from 1 (highest VCRS value) to υ (lowest VCRS value). Each response pairing $T^{\ell}_{\tau}(10, a)$ in T^{ℓ}_{τ} is a set of ordered pairs $(\mathbf{f}^{j}, \mathbf{f}^{k})_{i}$, where $\mathbf{f}^{j}$ is the ith formant vector in $b_{\upsilon}(\mathsf{c}, 10)$, and $\mathbf{f}^{j}$ the ith formant vector in $b_{\upsilon}(\mathsf{c}, a)$, for each $\mathsf{c} \in C_{\ell}$. Moreover, each response pairing $T^{\ell}_{\tau}(10, a)$ has a corresponding category transfer function $C(T^{\ell}_{\tau}(10, a))$.

For each caretaker $\ell^{\tau}_{10} = (Q^{\ell}_{\tau}, T^{\ell}_{\tau})$, we create an infant $\ell^{\tau}_{0.5} = (M^{\ell}_{\tau}, I(T^{\ell}_{\tau}))$, where M contains the auditory manifolds $M(0.5)$ and $M(10)$ over $\textsc{mauds}(0.5)$ and $\textsc{mauds}(10)$, respectively. The vocal learning environment is extended to include the 3-fold auditory manifolds $M_3(0.5)$ and $M_3(10)$ over $\textsc{mauds}_3(0.5)$ and $\textsc{mauds}_3(10)$, as well as the articulatory manifolds $A(0.5)$ and $A(10)$ over $\textsc{mars}(0.5)$ and $\textsc{mars}(10)$. Each of these manifolds is constructed using a k-nearest-neighbors computation. The set $I(T^{\ell}_{\tau})$ contains the socio-categorical auditory pairing $I(T^{\ell}_{\tau}(10, 0.5)) = \{(\mathbf{e}^{j}, \mathbf{e}^{k}) \mid (\mathbf{f}^{j}, \mathbf{f}^{k}) \in T^{\ell}_{\tau}(10, 0.5)\}$, whose weight function is a positive constant set to 20. It is assumed that $I(T^{\ell}_{\tau}(10, 0.5))$ has a simple category transfer interpretation $I(C(T^{\ell}_{\tau}(10, 0.5)))$. Furthermore, it is assumed that $I(T^{\ell}_{\tau}(10, 0.5))$ and $I(C(T^{\ell}_{\tau}(10, 0.5)))$ have the 3-fold counterparts $I_3(T^{\ell}_{\tau}(10, 0.5)) = \{(\mathbf{e}^{j_i}, \mathbf{e}^{k_i}) \mid (\mathbf{e}^{j}, \mathbf{e}^{k}) \in I(T^{\ell}_{\tau}(10, 0.5))$ and $1 \leq i \leq 3\}$, and $I_3(C(T^{\ell}_{\tau}(10, 0.5)))$.

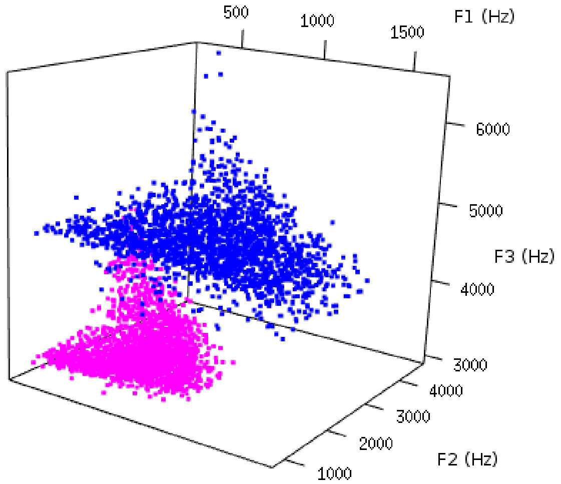


Figure 4.11: Approximated Vlab maximal vowel spaces $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$. The MVSs are depicted within a three-dimensional acoustic reference frame to emphasize their general lack of overlap. We have repeated Figure 3.11 to emphasize that the basis for computing the response surfaces that feed into a caretaker's response pairings is the same as in Chapter 3, even though in this chapter these pairings are the basis for a broader range of auditory representations.

4.4.1 Demonstrations

In order to make the extension of the modeling approach concrete, we demonstrate the "acquisition" procedure for the 3-fold version of auditory normalization and intermodal vowel normalization within the vocal learning environment for caretakers J^{20}_{10} and G^{12}_{10}, and their corresponding infants $J^{20}_{0.5}$ and $G^{12}_{0.5}$, focusing on the corner vowels $\{\mathsf{i},\mathsf{a},\mathsf{u}\}$. The first pair of demonstrations are 3-fold counterparts to the pair of demonstrations presented in Chapter 3. The second pair are intermodal counterparts. The figures corresponding to the demonstrations are presented at the end of the chapter.

Demonstration 4.4.1. We focus first on J^{20}_{10} and $J^{20}_{0.5}$, stepping through the entire acquisition procedure. We list the steps below, presenting the corresponding figures afterwards.

External Signals: The VCRSs $Q^J_{20}(\mathsf{c}, 0.5)$ and $Q^J_{20}(\mathsf{c}, 10)$ (where $\mathsf{c} \in \{\mathsf{i,a,u}\}$) for caretaker J^{20}_{10} are depicted in Figure 3.12. These VCRSs for J^{20}_{10} yield the response pairing $T^J_{20}(10, 0.5)$, which is depicted in Figure 3.7, along with its category transfer function $C(T^J_{20}(10, 0.5))$. The language- and dyad-specificity of acquisition begin at this stage the procedure.

Internalization: The infant $J^{20}_{0.5}$ internalizes the response pairing $T^J_{20}(10, 0.5)$ along with its category transfer function $C(T^J_{20}(10, 0.5))$, yielding the socio-auditory pairing $I_3(T^J_{20}(10, 0.5))$ and the simple categorical transfer interpretation $I_3(C(T^J_{20}(10, 0.5)))$. The necessary internalization of external vocal and social signals takes places at this stage.

Internal Computations: Having internalized $I_3(T^J_{20}(10, 0.5))$, along with its socio-auditory weighting $S_3(C(T^J_{20}(10, 0.5)))$, the infant computes the auditory normalization

$$\textsc{audNorm} : (M_3(10), M_3(0.5), S_3(C(T^J_{20}(10, 0.5)))) \mapsto M_3(10, 0.5).$$

over the "self" auditory manifold $M_3(0.5)$, and the "caretaker" auditory manifold $M_3(10)$. The output $M_3(10, 0.5)$ yields the means compute the "warping" of perception. The socio-auditory weighting $S_3(C(T^J_{20}(10, 0.5)))$, yields the following warping:

$$\textsc{audWarp} : (M_3(10, 0.5), S_3(C(T^J_{20}(10, 0.5))) \mapsto \textsc{warp}(10, 0.5).$$

The warped representations in $\textsc{warp}(10, 0.5)$ are depicted in Figure 4.12 (top), in terms of their first three components. **QEF**

Demonstration 4.4.2. We next focus on G^{12}_{10} and $G^{12}_{0.5}$, again stepping through the entire acquisition procedure. We list the steps below, presenting the corresponding figures afterwards.

External Signals: The VCRSs $Q^G_{12}(\mathsf{c}, 0.5)$ and $Q^g_{12}(\mathsf{c}, 10)$ (where $\mathsf{c} \in \{\mathsf{i,a,u}\}$) for caretaker G^{12}_{10} are depicted in Figure 3.15. These VCRSs for G^{12}_{10} yield the response pairing $T^G_{12}(10, 0.5)$, which is depicted in Figure 3.16, along with its category transfer function $C(T^G_{12}(10, 0.5))$.

Internalization: The infant $G^{12}_{0.5}$ internalizes the response pairing $T^G_{12}(10, 0.5)$ along with its category transfer function $C(T^G_{12}(10, 0.5))$, yielding the socio-auditory pairing $I_3(T^G_{12}(10, 0.5))$ and the categorical transfer interpretation $I_3(C(T^G_{12}(10, 0.5)))$.

Internal Computations: Having internalized $I_3(T^G_{12}(10, 0.5))$, along with its socio-auditory weighting $S_3(C(T^G_{12}(10, 0.5)))$, the infant computes the auditory normalization

$$\textsc{audNorm} : (M_3(10), M_3(0.5), S_3(C(T^G_{12}(10, 0.5)))) \mapsto M_3(10, 0.5).$$

over the "self" auditory manifold $M_3(0.5)$, and the "caretaker" auditory manifold $M_3(10)$. The output $M_3(10, 0.5)$ yields the means for the "warping" of auditory perception. The socio-auditory weighting $S_3(C(T^G_{12}(10, 0.5)))$, yields the following warping:

$$\textsc{audWarp} : (M_3(10, 0.5), S_3(C(T^G_{12}(10, 0.5))) \mapsto \textsc{warp}(10, 0.5).$$

The warped representations in $\textsc{warp}(10, 0.5)$ are depicted in Figure 4.12 (bottom), in terms of their first three components. **QEF**

Demonstration 4.4.3. We return to J^{20}_{10} and $J^{20}_{0.5}$, stepping through the intermodal acquisition procedure, taking sensorimotor pairing as our point of departure, as the external signals are the same as the multifold example above. We list the steps below, presenting the corresponding figures afterwards.

Internal Computations: Having internalized $I(T^J_{20}(10, 0.5))$, along with its socio-auditory weighting $S(C(T^J_{20}(10, 0.5)))$, the infant creates the socio-sensorimotor pairing structure relative to $\textsc{mars}(0.5)$, which includes the sensorimotor pairings $I_{0.5}(T^J_{20}(10, 0.5))$ and $I^{10}(T^J_{20}(10, 0.5))$, and the socio-sensorimotor weightings $S_{0.5}(C(T^J_{20}(10, 0.5)))$ and $S^{10}(C(T^J_{20}(10, 0.5)))$. The infant uses these pairings to compute the sensorimotor manifolds

$$\textsc{smAlig} : (A(0.5), M(a), S^{10}_{0.5}(C(T^J_{20}(10, 0.5)))) \mapsto H(a, 0.5),$$

over the auditory manifold $M(a)$, and the articulatory manifold $A(0.5)$, for $a \in \{0.5, 10\}$. The output $H(a, 0.5)$ provides the means to compute intermodal representations, which are yielded by the following intermodal mappings:

$$\textsc{imNarrowing} : (H(a, 0.5), S^{10}_{0.5}(C(T^J_{20}(10, 0.5))) \mapsto \textsc{imNarrow}(a, 0.5).$$

The intermodal representations yielded by the mappings are depicted in Figure 4.13 in terms of their first three components. The maximal intermodal spaces $\textsc{mims}(0.5)$ and $\textsc{mims}(10)$ yield the intermodal manifolds $K(0.5)$ and $K(10)$.

The infant also creates the socio-categorical intermodal pairing $J(T^J_{20}(10, 0.5))$ with socio-intermodal weighting $S_J(C(T^J_{20}(10, 0.5)))$ and category transfer interpretation $J(C(T^J_{20}(10, 0.5)))$. The weighting $S_J(C(T^J_{20}(10, 0.5)))$ is used to compute the intermodal normalization

$$\textsc{imNorm} : (K(10), K(0.5), S_J(C(T^J_{20}(10, 0.5)))) \mapsto K(10, 0.5).$$

The output $K(10, 0.5)$ yields the means to compute the "warping" of perception, and categorize the representations in $\textsc{mims}(10, 0.5))$.

The socio-auditory weighting $S_J(C(T^G_{12}(10,0.5)))$, yields the following warping:

$$\textsc{imWarp} : (K(10,0.5), S_J(C(T^G_{12}(10,0.5))) \mapsto \textsc{imWarp}(10,0.5).$$

The warped representations in $\textsc{imWarp}(10,0.5)$ are depicted in Figure 4.14 in terms of their first three components. Having also internalized the category transfer interpretation $J(C(T^G_{12}(10,0.5)))$, the infant computes the intermodal equivalence:

$$\textsc{imEquiv} : (K(10,0.5), J(C(T^G_{12}(10,0.5)))) \mapsto \mathsf{equiv}(\textsc{mims}(10,0.5)).$$

The equivalence relation over $\textsc{mims}$(10,0.5) is depicted over $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$ in Figure 4.17 (top). **QEF**

Demonstration 4.4.4. We return to G^{12}_{10} and $G^{12}_{0.5}$, stepping through the intermodal acquisition procedure, taking sensorimotor pairing as our point of departure, as the external signals are the same as the multifold example above. We list the steps below, presenting the corresponding figures afterwards.

Internal Computations: Having internalized $I(T^G_{12}(10,0.5))$, along with its socio-auditory weighting $S(C(T^G_{12}(10,0.5)))$, the infant creates the socio-sensorimotor pairing structure relative to $\textsc{mars}(0.5)$, which includes the sensorimotor pairings $I_{0.5}(T^G_{12}(10,0.5))$ and $I^{10}(T^G_{12}(10,0.5))$, and the socio-sensorimotor weightings $S_{0.5}(C(T^G_{12}(10,0.5)))$ and $S^{10}(C(T^G_{12}(10,0.5)))$. The infant uses these pairings to compute the sensorimotor manifolds

$$\textsc{smAlig} : (A(0.5), M(a), S^{10}_{0.5}(C(T^G_{12}(10,0.5)))) \mapsto H(a,0.5),$$

over the auditory manifold $M(a)$, and the articulatory manifold $A(0.5)$, for $a \in \{0.5, 10\}$. The output $H(a, 0.5)$ provides the means to compute intermodal representations, which are yielded the following intermodal mappings:

$$\textsc{imNarrowing} : (H(a, 0.5), S^{10}_{0.5}(C(T^G_{12}(10, 0.5))) \mapsto \textsc{imNarrow}(a, 0.5).$$

The intermodal representations yielded by the mapping are depicted in Figure 4.15 in terms of their first three components. The maximal intermodal spaces $\textsc{mims}(0.5)$ and $\textsc{mims}(10)$ yield the intermodal manifolds $K(0.5)$ and $K(10)$.

The infant also creates the socio-categorical intermodal pairing $J(T^G_{12}(10, 0.5))$ with socio-intermodal weighting $S_J(C(T^G_{12}(10, 0.5)))$ and category transfer interpretation $J(C(T^G_{12}(10, 0.5)))$. The weighting $S_J(C(T^G_{12}(10, 0.5)))$ is used to compute the intermodal normalization

$$\textsc{imNorm} : (K(10), K(0.5), S_J(C(T^G_{12}(10, 0.5)))) \mapsto K(10, 0.5).$$

The output manifold $K(10, 0.5)$ yields the means to categorize the representations in $\textsc{mims}(10, 0.5))$, as well as the “warping” of perception.

The socio-auditory weighting $S_J(C(T^G_{12}(10, 0.5)))$, yields the following warping:

$$\textsc{imWarp} : (K(10, 0.5), S_J(C(T^G_{12}(10, 0.5))) \mapsto \textsc{imWarp}(10, 0.5).$$

The warped representations in $\textsc{imWarp}(10, 0.5)$ are depicted in Figure 4.16.

Having also internalized the category transfer interpretation $J(C(T^G_{12}(10, 0.5)))$, the infant computes the intermodal equivalence:

$$\textsc{imEquiv} : (K(10, 0.5), J(C(T^G_{12}(10, 0.5)))) \mapsto \mathsf{equiv}(\textsc{mims}(10, 0.5)).$$

The equivalence relation over MIMS(10,0.5) is depicted over $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$ in Figure 4.17 (bottom). **QEF**

4.4.2 Discussion

We begin with a few points about specific aspects of Demonstrations 4.4.1 and 4.4.3, and 4.4.2 and 4.4.4, which involve the caretaker J_{10}^{20} and infant $J_{0.5}^{20}$, and the caretaker G_{10}^{12} and infant $G_{0.5}^{12}$, respectively. The same points apply to the entire vocal learning environment, though we mainly limit the discussion to these two demonstrations. Before proceeding, we repeat several points discussed in Secton 3.4.2.

i) We conceive of the response pairing $T_{20}^{J}(10, 0.5)$ as a model of pairs derived from infant-caretaker interaction, though the derivation procedure is left unspecified. Since the pairing operation is commutative, it may be the case that the caretaker has responded to an infant vocalization or vice versa. That is, the order of vocal turn-taking is not an imposition of the model, though the addition of ordering may itself be an interesting line of inquiry.

ii) We have modeled the internalization of the category transfer function $C(T_{20}^{J}(10, 0.5))$, with the highest level of fidelity, e.g., every pair in $C(T_{20}^{J}(10, 0.5))$ has a corresponding pair in $I(T_{20}^{J}(10, 0.5))$, or set of pairs in $I_3(T_{20}^{J}(10, 0.5))$. Yet, this in no way suggests that infants internalize every vocal interaction with their caretakers. The internalization of $C(T_{20}^{J}(10, 0.5))$ may easily be made to reflect the fact that infants do not internalize every such interaction.

iii) The infant internalizes the response pairing $T_{20}^{J}(10, 0.5)$ along with its category transfer function $C(T_{20}^{J}(10, 0.5))$, yielding the socio-auditory pairings $I(T_{20}^{J}(10, 0.5))$ and $I_3(T_{20}^{J}(10, 0.5))$, and the categorical transfer interpretations $I(C(T_{20}^{J}(10, 0.5)))$ and

$I_3(C(T_{20}^J(10, 0.5)))$. However, the sensorimotor alignment and intermodal normalization computations derive from the socio-auditory weights only, and not the categories from $S(C(T_{20}^J(10, 0.5)))$ and $S_3(C(T_{20}^J(10, 0.5)))$.

Concerning Demonstrations 4.4.1 and 4.4.2, the warped representations contained in $\textsc{warp}(10, 0.5)$ yielded by the Laplacian eigenmappings

$$\textsc{aud}\textsc{Warp}(M(10, 0.5), S(C(T_{20}^J(10, 0.5))) \quad \textsc{aud}\textsc{Warp}(M(10, 0.5), S(C(T_{12}^G(10, 0.5)))$$

are depicted in Figure 4.12 in terms of their first three components. The key difference between these demonstrations and those in Secton 3.4.2 is their use of 3-fold excitation vectors. In practice, the use of 3-fold excitation vectors increases the number of auditory pairs that guide auditory normalization, and the number of auditory representations over which auditory manifolds are formed. This increase provides more data points for the alignment algorithm to work with, mitigating sparsity problems at the cost of slowing down algorithmic performance. In principle, the use of 3-fold excitation vectors focuses attention on the internalization of vowel signals by the auditory system. We selected n-fold excitation vectors as the starting point for reasoning about the internalization since a finite set over n elements is the simplest structure that can be assumed. Indeed, it is important to note that sequential information is not encoded within the n-fold vectors, as incorporating order results in a more complex structure. In the sequal to this dissertation, we take n-fold vectors as a point of departure for further investigation of the structural encoding of vowel signals by the auditory system.

We now turn to Demonstrations 4.4.3 and 4.4.4. Without loss of generality, we focus on the former, beginning with an unpacking of the sensorimotor alignment computation. Having internalized the socio-auditory pairing $I(T^J_{20}(10, 0.5))$, along with its socio-auditory weighting $S(C(T^J_{20}(10, 0.5)))$, the infant creates the socio-sensorimotor pairing structure relative to $\textsc{mars}(0.5)$, which includes the sensorimotor pairings $I_{0.5}(T^J_{20}(10, 0.5))$ and $I^{10}(T^J_{20}(10, 0.5))$, and the socio-sensorimotor weightings $S_{0.5}(C(T^J_{20}(10, 0.5)))$ and $S^{10}(C(T^J_{20}(10, 0.5)))$.

The infant uses the pairing $I_{0.5}(T^J_{20}(10, 0.5))$ in the sensorimotor alignment computation

$$\textsc{smAlig} : (A(0.5), M(0.5), S^{10}_{0.5}(C(T^J_{20}(10, 0.5)))) \mapsto H(0.5, 0.5),$$

which yields the sensorimotor manifold $H(0.5, 0.5)$. This sensorimotor manifold is the basis for the infant's model of self. Indeed, $H(0.5, 0.5)$ is used to compute intermodal representations of the self:

$$\textsc{imNarrowing} : (H(0.5, 0.5), S^{10}_{0.5}(C(T^J_{20}(10, 0.5))) \mapsto \textsc{imNarrow}(0.5, 0.5).$$

The intermodal representations in $\textsc{mims}(0.5)$ derived from $\textsc{imNarrow}(0.5, 0.5)$ are depicted in Figure 4.13 (left) in terms of their first three components. In similar fashion, the infant uses the pairing $I^{10}(T^J_{20}(10, 0.5))$ in the sensorimotor alignment computation

$$\textsc{smAlig} : (A(0.5), M(10), S^{10}_{0.5}(C(T^J_{20}(10, 0.5)))) \mapsto H(10, 0.5),$$

which yields the sensorimotor manifold $H(10, 0.5)$. This sensorimotor manifold is the basis of the infant's model of their caretaker. Indeed, $H(10, 0.5)$ is used to compute intermodal representations of the caretaker:

$$\textsc{imNarrowing} : (H(10, 0.5), S^{10}_{0.5}(C(T^J_{20}(10, 0.5))) \mapsto \textsc{imNarrow}(10, 0.5).$$

The intermodal representations in $\textsc{mims}(10)$ derived from $\textsc{imNarrow}(10, 0.5)$ are depicted in Figure 4.13 (right) in terms of their first three components.

In computational practice, the representations in both $\textsc{mims}(0.5)$ and $\textsc{mims}(10)$ have m-many components, where m is the number of excitation and articulatory vectors involved in the sensorimotor alignment computation (4000 in this version of the vocal learning environment). Interpretation of these components is less clear than in the case of auditory warping, though the representations in Figure 4.13 exhibit a similar kind of warping based on the contiguity between the auditory and articulatory manifolds created by the sensorimotor pairings, akin to the warping of auditory representations based on the contiguity between auditory manifolds created by auditory pairings. Moreover, a number of open questions remain as to which components of intermodal representations are used as input for futher computations, e.g., the creation of intermodal manifolds. It may be the case that a different number of components are used in forming intermodal manifolds for different social agents, as well as different compnents across different social agents. In these demonstrations, we keep to the simplest approach, adopting uniformity across agents, but this is only a modeling assumption.

Representations in the maximal intermodal spaces $\textsc{mims}(0.5)$ and $\textsc{mims}(10)$ are thus restricted to their first three components when the intermodal manifolds $K(0.5)$ and $K(10)$ are constructed. The manifolds $K(0.5)$ and $K(10)$ are then aligned to form the commensuration manifold $K(10, 0.5)$ during intermodal normalization, providing the means to derive the warped representations in $\textsc{imWarp}(10, 0.5)$ depicted in Figures 4.14 in terms of their first three components. The commensuration manifold $K(10, 0.5)$ also provides the means to compute the equivalence relation

$$\textsc{imEquiv}(K(10, 0.5), J(C(T^J_{20}(10, 0.5))))$$

depicted in Figure 4.17 (top) over representations in the maximal vowel spaces $\text{MVS}(0.5)$ and $\text{MVS}(10)$.

The sensorimotor alignment computations in Demonstration 4.4.4 are identical to those in Demonstration 4.4.3, save for the use of the socio-auditory pairing $I(T_{12}^{G}(10, 0.5))$, along with its socio-auditory weighting $S(C(T_{12}^{G}(10, 0.5)))$. In this case, the commensuration manifold $K(10, 0.5)$ provides the means to derive the warped representations in $\text{IMWARP}(10, 0.5)$ depicted in Figures 4.16 in terms of their first three components. The commensuration manifold $K(10, 0.5)$ also provides the means to compute the equivalence relation

$$\text{IMEQUIV}(K(10, 0.5), J(C(T_{12}^{G}(10, 0.5))))$$

depicted in Figure 4.17 (bottom) over representations in $\text{MVS}(0.5)$ and $\text{MVS}(10)$.

Our interpretation of the output of Demonstrations 4.4.3 and 4.4.4 is simply carried forward from Secton 3.4.2. Specifically, examination of Figures 4.14, 4.16, and 4.17 support the introduction of a structural component to the perceptual magnet effect, along with the language- and dyad-specific nature of intermodal normalization, and vowel category acquisition.

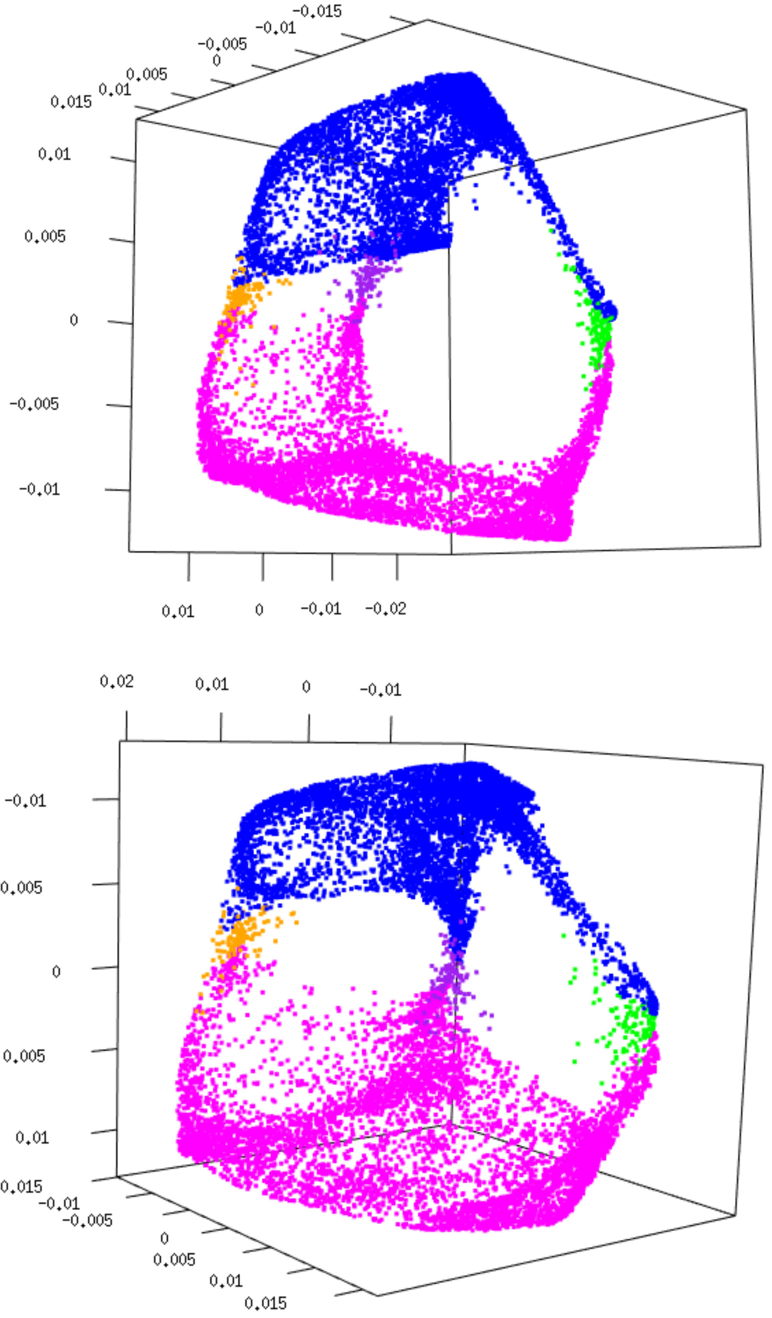


Figure 4.12: The warpings yielded by socio-auditory weightings $S_3(C(T^J_{20}(10, 0.5)))$ (top) and $S_3(C(T^G_{12}(10, 0.5)))$ (bottom).

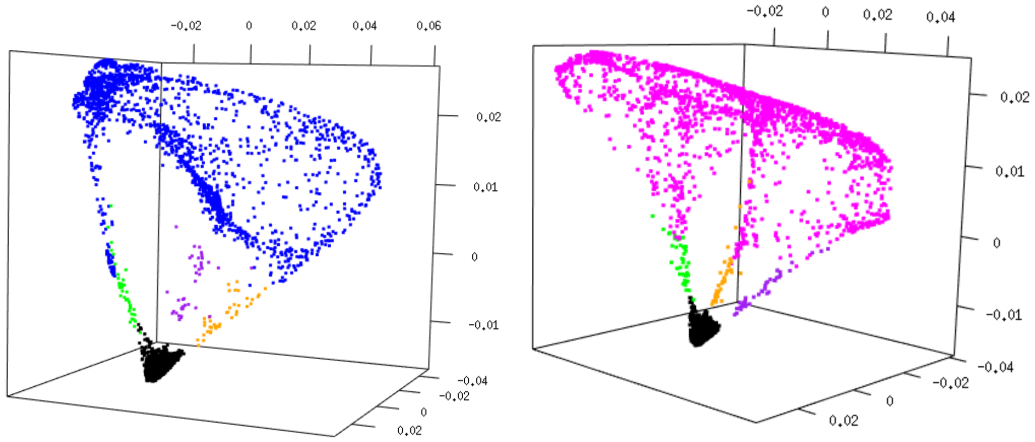


Figure 4.13: The intermodal representations yielded by socio-sensorimotor weighting $S^{10}_{0.5}(C(T^J_{20}(10, 0.5)))$.

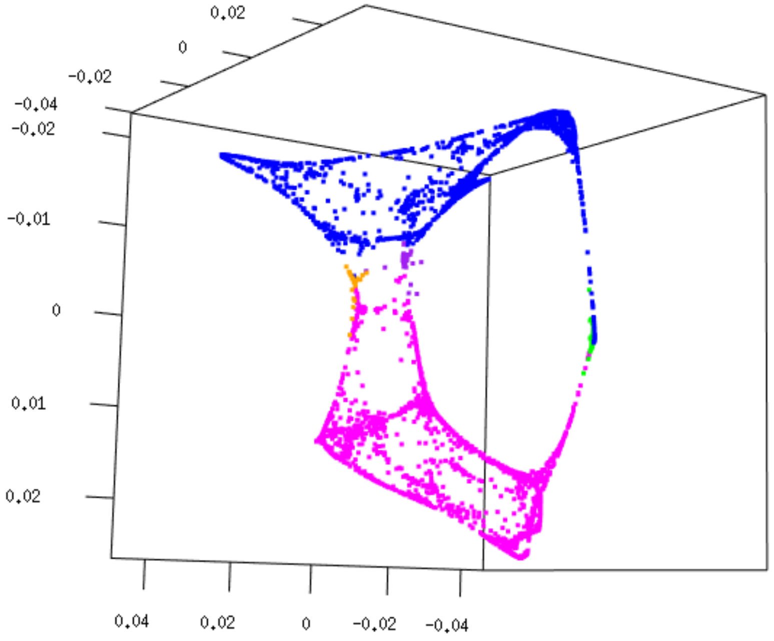


Figure 4.14: The warping yielded by socio-intermodal weighting $S(C(T^J_{20}(10, 0.5)))$.

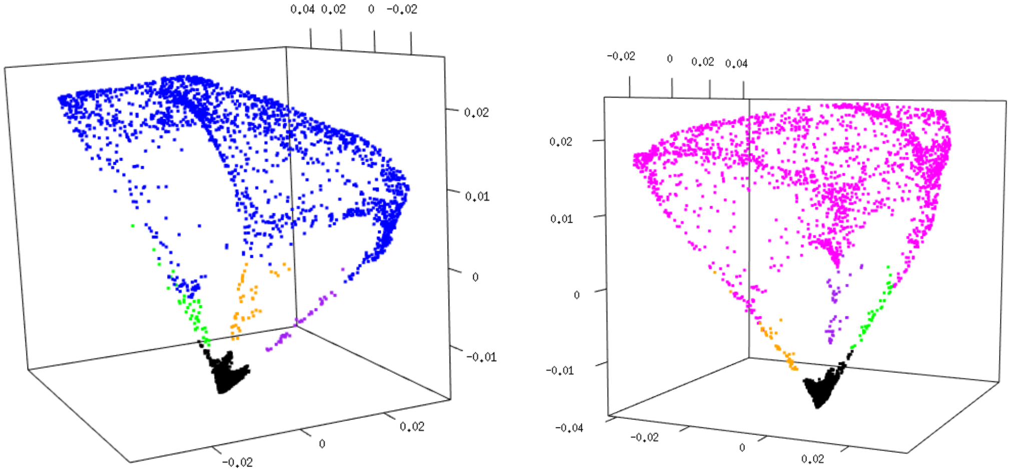


Figure 4.15: The intermodal representations yielded by socio-sensorimotor weighting $S^{10}_{0.5}(C(T^G_{12}(10, 0.5)))$.

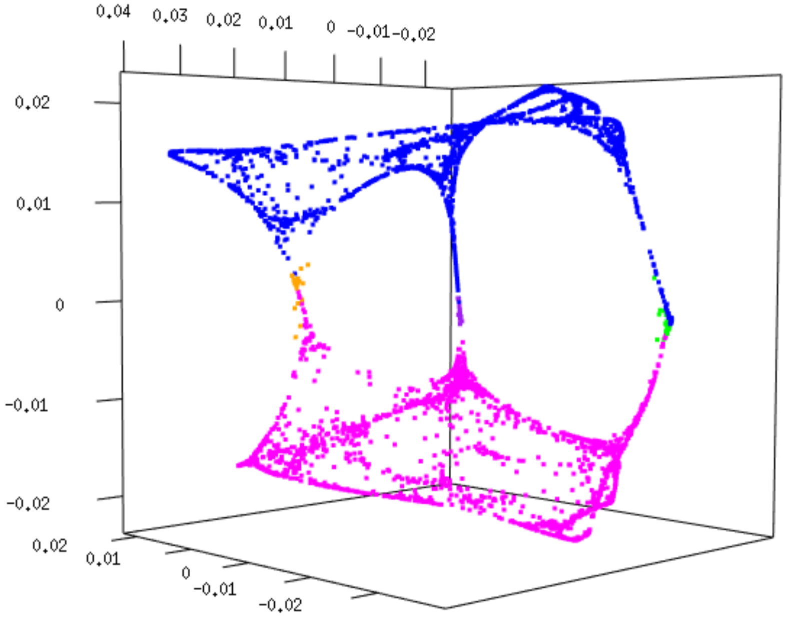


Figure 4.16: The warping yielded by socio-intermodal weighting $S(C(T^G_{12}(10, 0.5)))$.

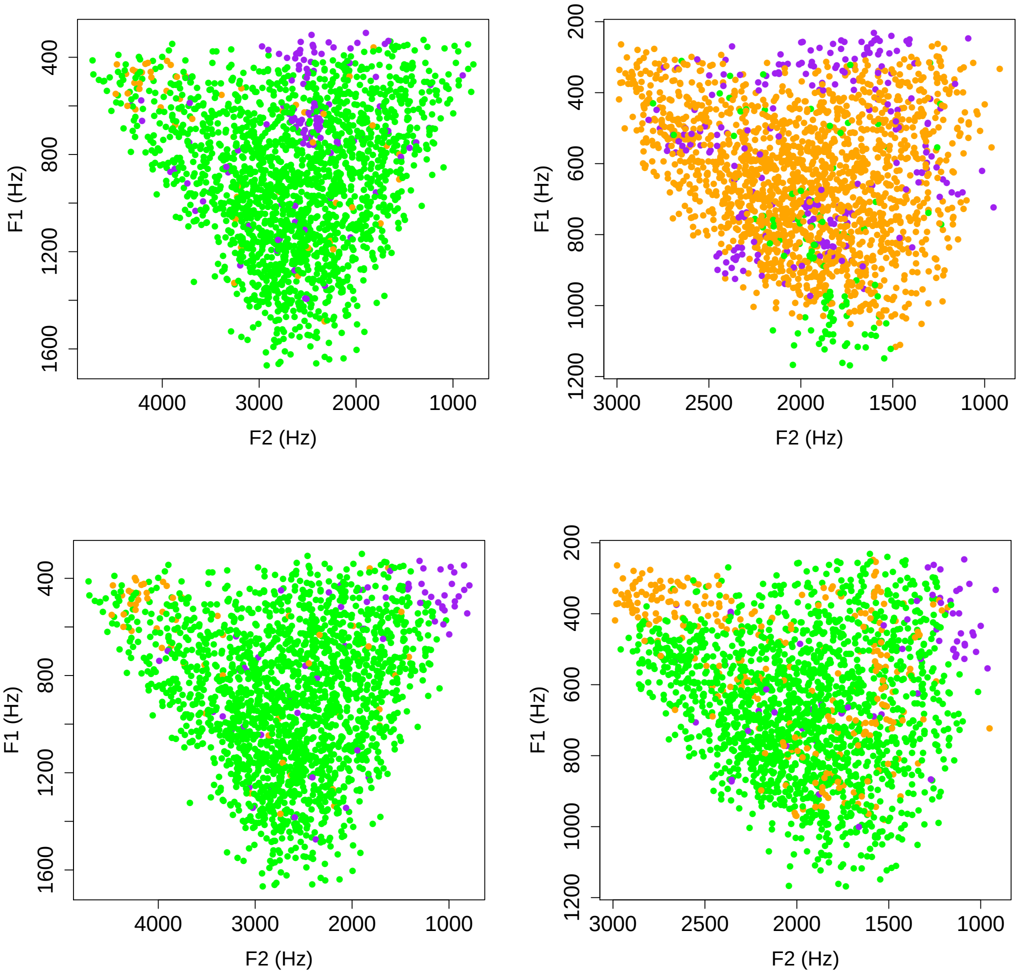


Figure 4.17: The intermodal equivalences derived from the category transfer interpretations $J(C(T^J_{20}(10, 0.5)))$ (top) and $J(C(T^G_{12}(10, 0.5)))$ (bottom). The equivalence relations are depicted over representations in $\textsc{mvs}(10)$ and $\textsc{mvs}(0.5)$.

CHAPTER 5: GENERAL DISCUSSION

In this closing chapter, we present a colloquial retrospection on the modeling approach put forward in earlier chapters, and a discussion of the approach's future prospects and potential for prediction. The retrospection serves to summarize the formulations and demonstrations, following the presentation sequence of earlier chapters, and accordingly assumes some basic familiarity with their terms and concepts. The future directions are treated very broadly, and are based on the main tenets of the modeling philosophy described in the preface. Accordingly, a number of metaphysical and epistemological stances are repeated and given greater emphasis in hopes that the points will be clearer after working through their implications for modeling vowel normalization and category acquisition.

5.1 Retrospection

The main modeling contribution of this dissertation is the creation of a virtual environment for vocal learning, called throughout the "vocal learning environment." The environment consists of models of caretaker agents representing five different language communities (American English, Cantonese, Greek, Japanese, and Korean) derived from vowel category perception experiments, and models of infant agents that "vocally interact" with their caretakers. Aspects of the vocal learning environment are presented in detail in Chapters 3 and 4. In this section we summarize the modeling approach that is put forward in this dissertation through formulation and demonstration of the vocal learning environment.

The summary is presented according to the component-wise phenomenal division that has carried through the dissertation.

EXTERNAL SIGNALS: We take the external signals involved to be the acoustic signals produced by an infant and a single caretaker, together with social signals that correspond to the acoustic signals produced by the caretaker, as well as the infant. The acoustic signals are modeled using the output of an age-varying articulatory synthesizer (Sections 4.2.1 and 4.2.2), where the signals are characterized in terms of their formant frequencies. The collection of all potential formant patterns for a given age, characterized in terms of their first three formant frequencies, is called a *maximal vowel space* for that age (Section 3.2). For each adult caretaker in the vocal learning environment, and for each maximal vowel space age, the caretaker's language-specific vowel category knowledge is modeled as a set of functions from the maximal vowel space to *goodness values*. For each vowel category in the caretaker's language, the goodness value assigned to a formant pattern within the maximal vowel space models how good the caretakers feels that formant pattern is as an example of the vowel category. These functions, called *vowel category response surfaces*, locate regions over the maximal vowel space that contain good examples of their language's vowel categories (Section 3.2.2). The social signals that caretakers communicate to infants during vocal exchanges are modeled as functions over vowel category response surface values (Section 3.2.3).

INTERNALIZATION: We take the internalization of signals to involve the creation of auditory representations over the acoustic signals derived from the infant and caretaker productions, as well as the creation of representations derived from interpretation of the caretaker's social signals. The creation of auditory representations from acoustic signals by the auditory system is modeled as the transduction of vowel signals through a series of transformations, each of which models a component of the auditory system. The output of the series of transformations is an auditory representation called an *excitation pattern*, which reflects the amount of signal energy recorded by a linear sequence of regions along the basilar membrane (Section 3.3.1, while auditory representations are treated more generally as “multifold” representations in Section 4.2.3). The internalization of the social signals that caretakers communicate to infants during vocal exchanges is modeled as a function over those social signals whose values are called *socio-auditory weight* (Section 3.3.2).

INTERNAL COMPUTATION AND BEHAVIOR: We take the stock of internal computations to include the following: (i) a manifold formation computation that forms manifolds over auditory and articulatory representations, or representations derived from these through other internal computations, (ii) a pairing computation that forms pairing structures over auditory and articulatory representations, or representations derived from these through other internal computations, (iii) an alignment computation that acts as a binary operation on manifolds, i.e., which takes two manifolds and yields a new one, which models auditory normalization, sensorimotor alignment, and intermodal normalization, (iv) a transformation computation that maps representations in a given set of reference frames to representations in a new reference frame based on the alignment of manifolds embedded within the initial set of frames.

Modeling begins with the instantiation of internal computations as auditory computations. The acquisition of a cognitive structure that provides an infant with the means for representing auditory (sensory) information, which is assumed to be vowel-like, coming from both the infant and the caretaker is modeled in terms of manifold formation (Chapter 2) over auditory representations. Manifolds formed over auditory representations are called *auditory manifolds*. The acquisition of a cognitive structure which an infant uses to interpret interactions with their caretaker is modeled as a pairing operation over infant and caretaker auditory representations derived from turn-taking vocal exchanges. Each pair in these *auditory pairings* is assigned the socio-auditory weight corresponding to the social signal imposed by the caretaker on the caretaker's response to the infant's production.

The acquisition of an *auditory normalization* computation yielding equivalences between representations of qualitatively similar vowels that may differ absolutely in representation due to speaker variation is modeled as manifold alignment (Chapter 2, Section 3.3.2). The alignment of auditory manifolds yields a transformation that models perceptual warping. The transformation is characterized as a Laplacian eigenmapping from the auditory reference frame to a "warped" representation frame (Section 3.3.2). The alignment also yields the means for categorization of auditory representations based on category transfer information derived from vocal exchanges.

In addition to auditory computations, the vocal learning model includes articulatory, sensorimotor, and intermodal computations. To begin, the acquisition of a cognitive structure that provides an infant with the means for representing articulatory (motor) information derived from the infant's own vowel productions is modeled in terms of manifold formation over articulatory representations (as characterized in Section 4.2). Manifolds formed over articulatory representations are called *articulatory manifolds*. The auditory pairings used to align auditory manifolds are also be used to align articulatory manifolds with auditory manifolds. Specifically, the auditory pairings yield two pairings, the first of which is a pairing over auditory representations of infant productions with articulatory representations of those productions, and a pairing over auditory representations of caretaker responses to the infant productions with articulatory representations of the infant productions. Each pair in these *sensorimotor pairings* is assigned a *socio-sensorimotor weight* derived from the socio-auditory weight corresponding to the social signal imposed by the caretaker on the caretaker's response to the infant's production. The acquisition of a cognitive structure that provides an infant with the means to relate auditory structures and articulatory structures is also modeled as manifold alignment (Section 4.3.1). The resulting *sensorimotor manifolds* yield transformations that models multisensory narrowing. The transformations are characterized as Laplacian eigenmappings from the auditory and articulatory reference frames to an "intermodal" representation frame (Section 4.3.1).

The intermodal representations are used to form *intermodal manifolds* which serve as the infant's cognitive representations of the infant and the caretaker. The auditory pairings used to align auditory manifolds, and to align articulatory manifolds with auditory manifolds, are also used to align intermodal manifolds. Specifically, the auditory pairings yield a single *intermodal pairing* over intermodal representations of infant productions with intermodal representations of caretaker responses to the infant productions. Each pair in the intermodal pairing is assigned a *socio-intermodal weight* derived from the socio-auditory weight corresponding to the social signal imposed by the caretaker on the caretaker's response to the infant's production.

The acquisition of an *intermodal normalization* computation yielding equivalences between representations of qualitatively similar vowels that may differ absolutely in representation due to speaker variation is again modeled as manifold alignment (Section 4.3.2). The alignment of intermodal manifolds yields a transformation that models perceptual warping. The transformation is characterized as a Laplacian eigenmapping from the intermodal reference frame to a "warped" representation frame (Section 4.3.2). The alignment also yields the means for categorization of intermodal representations based on category transfer information derived from vocal exchanges.

5.2 Future Directions

The retrospection above also facilitates discussion of potential modifications, improvements, and expansions to the components of the vocal learning environment, as well as broader correlates of the modeling approach and conceptual organization. In this concluding section, we list suggestions for further work within the vocal learning environment, structured in terms of its components, and follow with a brief treatment of its purview.

5.2.1 Modifying the Vocal Learning Environment Model

EXTERNAL SIGNALS: Focus on external signals is mainly limited to their categorization and corresponding goodness levels. There are several immediate extensions that may be considered.

- The first involves modifying the acoustic parameters of the signals used in the vocal learning environment. Importantly, duration was fixed across all vowel signals, which led to an artificially simple method for auditory internalization. In order to address practical concerns, the modeling framework needs to be able to accommodate vowel signals of varying duration.

- The use of multiple time slices may provide the means for i) representing the effects of the changing amplitude and varying f0 patterns which mimic the different "melodies" of infant-directed speech, ii) allowing for changes in spectral structure resulting from changing articulatory configurations over time, and iii) developing a way of modeling the internalization of the sequential or dynamic nature of the unfolding changes over time.

- Another potential extension is the incorporation of other kinds of external signals, with the most obvious being visual signals of others' speech productions. The introduction of a third sensory modality is easily handled within our modeling and conceptual framework, and visual representations lend themselves nicely to the manifold approach.

- The incorporation of visual signals provides the means to include visual social signals. Specifically, visual cues to a caretaker's goodness ratings may help the infant differentiate between feedback indicating a high quality "adult-like" infant production and a lower quality, more "infant-like" infant production. Visual cues of this nature may be modeled using the transfer functions and socio-auditory weights that guide the infant's acquisition of vowel normalization.

- The last extension we mention pertains to social categories. Perceptual categorization experiments provide the goodness ratings that form the basis of the vowel category response surfaces used to model caretaker responses to infant vocalizations. Subjects in these experiments also provided age and gender ratings of the stimuli, revealing rich, language-specific interpretations with sociolinguistic underpinnings (Plummer et al., 2013b). These preliminary results suggest that caretaker feedback is far more nuanced and structurally complex than has typically been assumed. This in turn suggests that the socio-auditory weightings that influence normalization are complex, and may have many contributing factors that influence each other in a variety of ways. The model of caretaker response pairings and transfer functions put forward in this dissertation is flexible enough to encompass factoring, which immediately provides means to model more complex caretaker feedback.

INTERNALIZATION: Modeling of internalization is limited to two components: internalization of social signals and acoustic signals. Extension of the modeling of external signals likely necessitates extension of the modeling of internalization, e.g., the inclusion of visual signals necessitates modeling of the internalization of visual signals. In the short-term, practicable concerns my lead to the rapid incorporation of representation of acoustic properties such as duration. More broadly, understanding of the anatomy and physiology of the infant auditory system is only just beginning to take shape (see Zemlin, 1998, Chapters 6 and 7). Knowledge of the adult auditory system itself is only slightly more advanced, and is in large part derived from research on other mammals. The relatively basic auditory modeling used in this dissertation is meant primarily to focus attention on the fact that internalization of acoustic signals by the auditory system is an important part of speech perception. Importantly, more inclusive models may reveal more of complexity involved in the acquistion of normalization, while highliting the merit of our model. For example, incorporating an auditory reference frame whose representations also encode organism-internal information, e.g., that from bone conduction in addition to signal from the eardrum, allows for modeling of more complex sensory phenomena during acquisition, potentially providing new avenues of investigation in auditory learning. Similarly, the ontogeny of the internalization of social signals and its interaction with other functional interpretations of vocal exchange is also understudied (Hsu et al., 2013), and the simplified model used in this dissertation is meant only as a first step. Advances in understanding and modeling of the ontogeny of the auditory system and the interpretation of vocal exchange may be incorporated as more is learned.

INTERNAL COMPUTATION AND BEHAVIOR: The most immediately actionable modifications to the modeling approach lie within the stock of internal computations. A number of different kinds of manifold constructions and alignment paradigms exist (Wang, 2010; Ma and Fu, 2012), which provide an assortment of structural models potentially applicable to the acquisition of vowel normalization. Moreover, the treatment of perceptual warping and multisensory narrowing in terms of Laplacian eigenmapping is one way among many in which they may be formulated within the more general graph-based alignment paradigm. Computational investigation of these variations may be the quickest route to deliverables in this area of research.

5.2.2 Broader Issues

We turn now to two key points brought to light in formulating the vocal learning environment. The first concerns the generative conceptualization of multisensory matching, modeled as sensorimotor pairing and alignment. The manifolds involved in sensorimotor alignment are entities from different reference frames, and since the alignment computation does not discriminate with respect to the reference frames that manifolds are constructed over, it may relate manifolds across *any* two (or more) sensory modalities. This immediately raises the question of how auditory and articulatory (and moreover visual) manifolds come to be arguments for the alignment computation during acquisition, rather than say, tactile manifolds, especially in light of successful language acquisition in the absence of sensory input from one or more of these “typical” modalities. One possible answer is that Perrier’s (2005) notion of a hierarchical organization of the modalities in which audition is given higher priority than other modalities makes auditory structures a default type for normalization, with other manifolds types filling that role in the absence of auditory sensory input, as in the case of congenital deafness. Another potential factor, proffered throughout this dissertation, is that the basis of acquisition is in the relation of representations of the self to representations of others. On this view, the modalities that factor into the alignment computation are determined by their "referential value" with respect to social agents. In principle, the same general design could adequately handle the use of visual manifolds for the acquisition of a signed language by deaf infants from Deaf caretakers. Structures over modalities that factor into the organization of sensory experience underscore the complexity of acquisition, likely placing its core aspects outside of the scope of distributional learning.

The second point concerns the scope of the generative conceptualization of normalization. The main point of this dissertation is that normalization is a generative procedure involving the relation of representations. Although this position is in some sense at odds with the largely reductive and summary-based position of most computational modeling, it seems to be virtual truism. Looking toward the human sciences, rather than the computational sciences, adduces the generative position, and may be more conceptually advantageous, if not numerically expedient. The use of reference frames in modeling psychological concepts goes back to Lewin (1936), while Goffman's (1974) "frame analysis," which encompasses relations and operations over frames – called "frame alignment" – has had substantial influence on sociological theory. More recently, the general approach has been applied to literary historical analysis. For example, Rose (2001) appeals to Goffman's (1974) conceptualization of frames in order to describe how readers in Industrial age Britain interpreted the literature and art they were exposed to. Specifically, readers are taken to "build up a repertoire of interpretive strategies," in which case "the value of a liberal education lies not so much in acquiring facts or absorbing 'eternal truths,' but in discovering new ways to interpret the world" (p. 7). The focus is on the intake of other points of view, rather than discovery of a "true signal."

These examples indicate that the interpretive power of the generative approach is quite broad. The modeling approach put forward in this dissertation is an attempt at formulating these interpretations, and a first step toward a basis for quantitative analysis and prediction. The nature of the relations between sociality and multisensory integration is scarcely understood, with theory and modeling only beginning to take shape, either at the neural level of description (e.g., Oberman and Ramachandran, 2007, 2008), or the socio-cognitive (e.g., Ozturk et al., 2013). If the acquisition of vowel normalization is indeed an operation over cogintive representations of social agents in an infant's vocal learning environment, and vowel normalization influences phonological acquisition, then the models put forward in this dissertation may be applied to the quantitative investigation of these relations. In the sequel to this dissertation, we begin to formulate quantitative measures of the acquisition of vowel normalization, which may aid in this pursuit.

APPENDIX A: ACQUISITION COMPUTATIONS

In this appendix, we present graphical representations of the computations carried out within the vocal learning environment, based on the following heuristic involving a single caretaker (age 10) and a single infant (age 0.5). The key aspects of the heuristic are as follows.

The respose pairing and category transfer function:

$T(10, 0.5)$	$C(T(10, 0.5))$
$(\mathbf{f}^1, \mathbf{f}^7)$	(i, g_1)
$(\mathbf{f}^3, \mathbf{f}^9)$	(u, g_2)
$(\mathbf{f}^5, \mathbf{f}^{11})$	(a, g_3)

The socio-categorical auditory pairing:

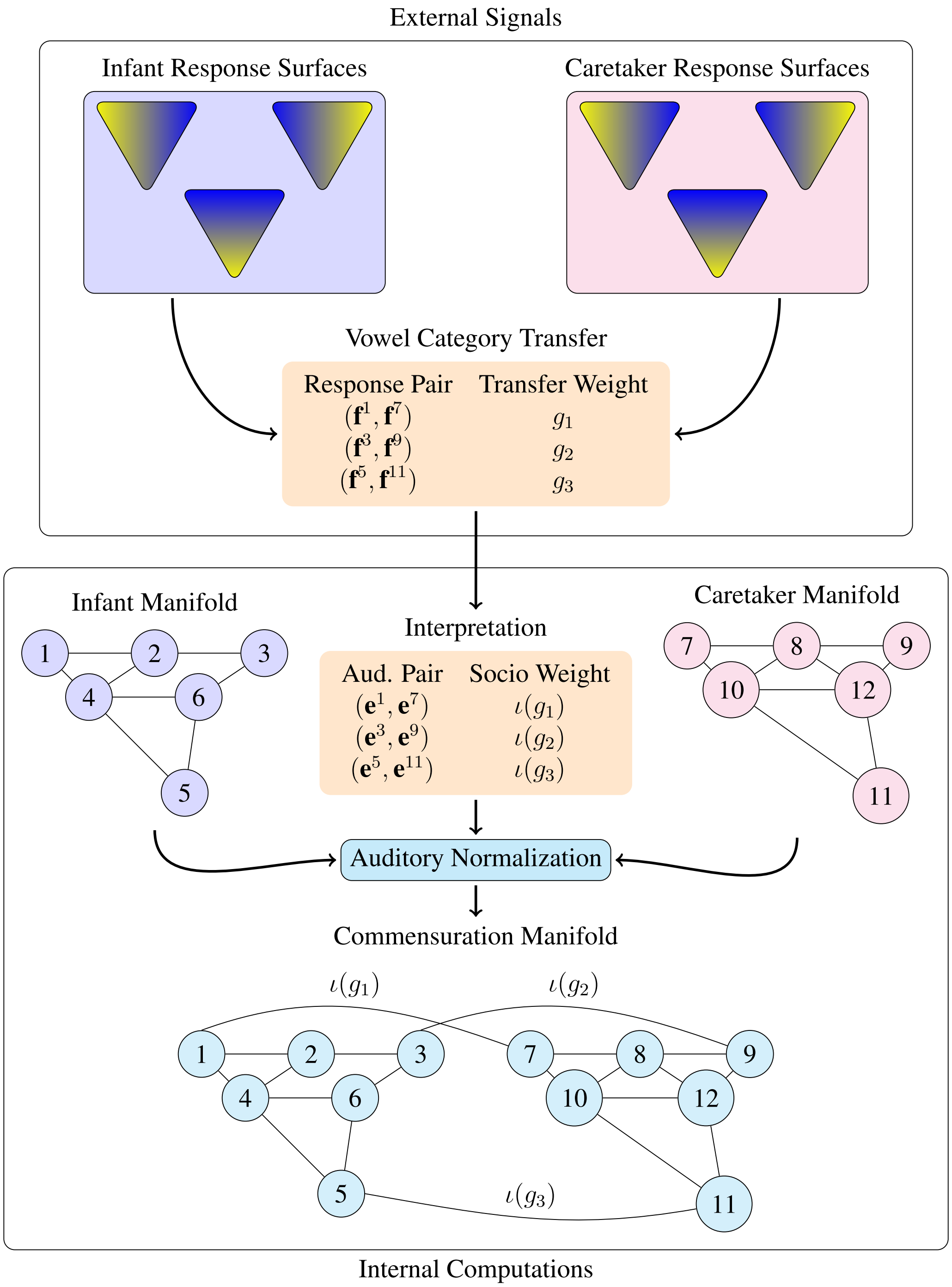
$I(T(10, 0.5))$	$S(C(T(10, 0.5)))$	$I(C(T(10, 0.5)))$
$(\mathbf{e}^1, \mathbf{e}^7)$	$\iota(g_1)$	i
$(\mathbf{e}^3, \mathbf{e}^9)$	$\iota(g_2)$	u
$(\mathbf{e}^5, \mathbf{e}^{11})$	$\iota(g_3)$	a

The socio-sensorimotor pairing structure relative to MARS(0.5):

$I_{0.5}(T(10, 0.5))$	$S_{0.5}(C(T(10, 0.5)))$	$I^{10}(T(10, 0.5))$	$S^{10}(C(T(10, 0.5)))$
$(\mathbf{a}^1, \mathbf{e}^1)$	$\delta(\iota(g_1))$	$(\mathbf{a}^1, \mathbf{e}^6)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^3)$	$\delta(\iota(g_2))$	$(\mathbf{a}^3, \mathbf{e}^9)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^5)$	$\delta(\iota(g_3))$	$(\mathbf{a}^5, \mathbf{e}^{11})$	$\delta(\iota(g_3))$

The socio-categorical intermodal pairing:

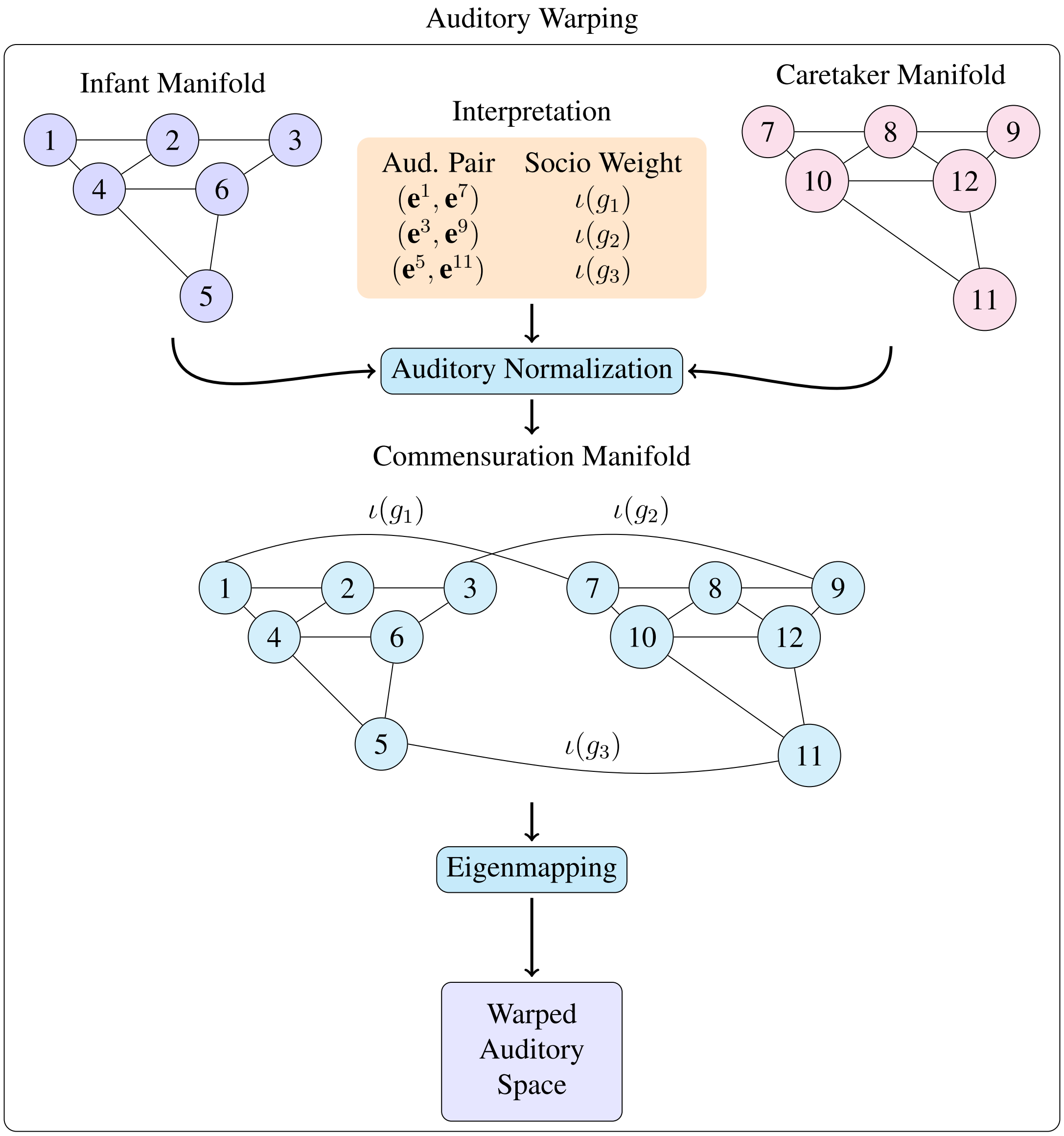
$J(T(10, 0.5))$	$S_J(C(T(10, 0.5)))$	$J(C(T(10, 0.5)))$
$(\mathbf{c}^1, \mathbf{c}^7)$	$\kappa(\iota(g_1))$	i
$(\mathbf{c}^3, \mathbf{c}^9)$	$\kappa(\iota(g_2))$	u
$(\mathbf{c}^5, \mathbf{c}^{11})$	$\kappa(\iota(g_3))$	a



Response Pair	Transfer Weight
$(\mathbf{f}^1, \mathbf{f}^7)$	g_1
$(\mathbf{f}^3, \mathbf{f}^9)$	g_2
$(\mathbf{f}^5, \mathbf{f}^{11})$	g_3

Aud. Pair	Socio Weight
$(\mathbf{e}^1, \mathbf{e}^7)$	$\iota(g_1)$
$(\mathbf{e}^3, \mathbf{e}^9)$	$\iota(g_2)$
$(\mathbf{e}^5, \mathbf{e}^{11})$	$\iota(g_3)$

Figure A.1: System architecture for auditory normalization.



Aud. Pair	Socio Weight
$(\mathbf{e}^1, \mathbf{e}^7)$	$\iota(g_1)$
$(\mathbf{e}^3, \mathbf{e}^9)$	$\iota(g_2)$
$(\mathbf{e}^5, \mathbf{e}^{11})$	$\iota(g_3)$

Figure A.2: System architecture for auditory warping.

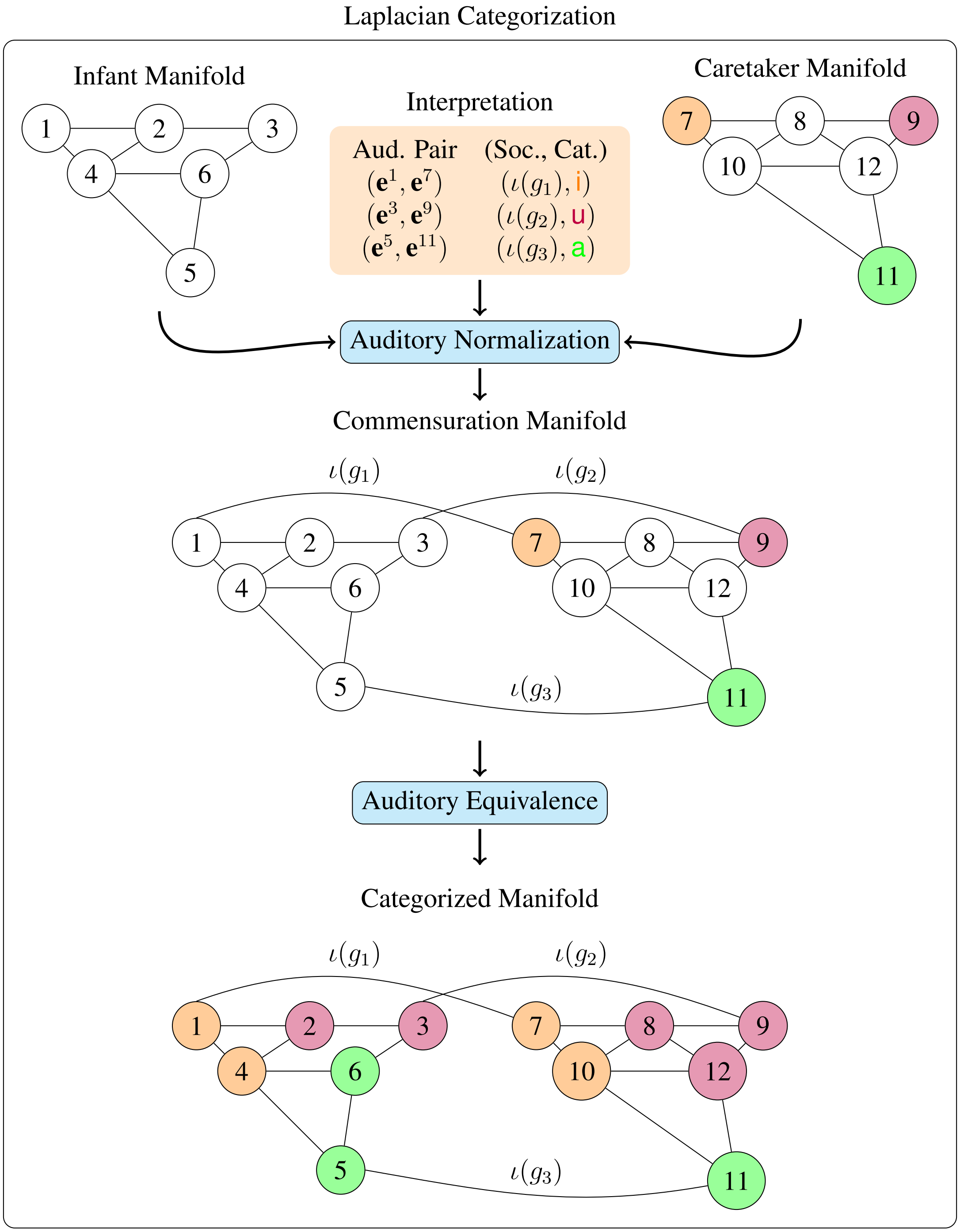
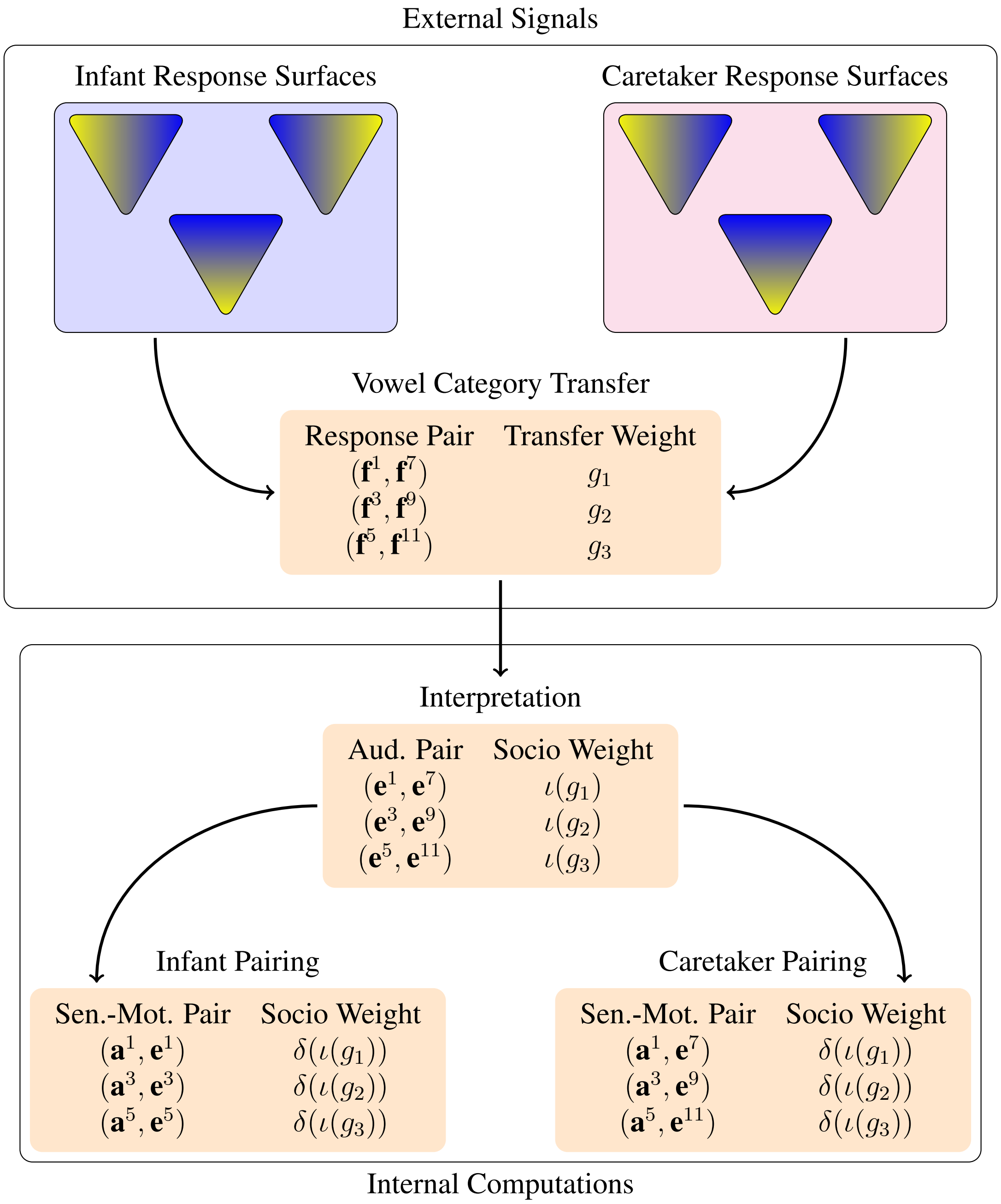


Figure A.3: System architecture for vowel categorization.



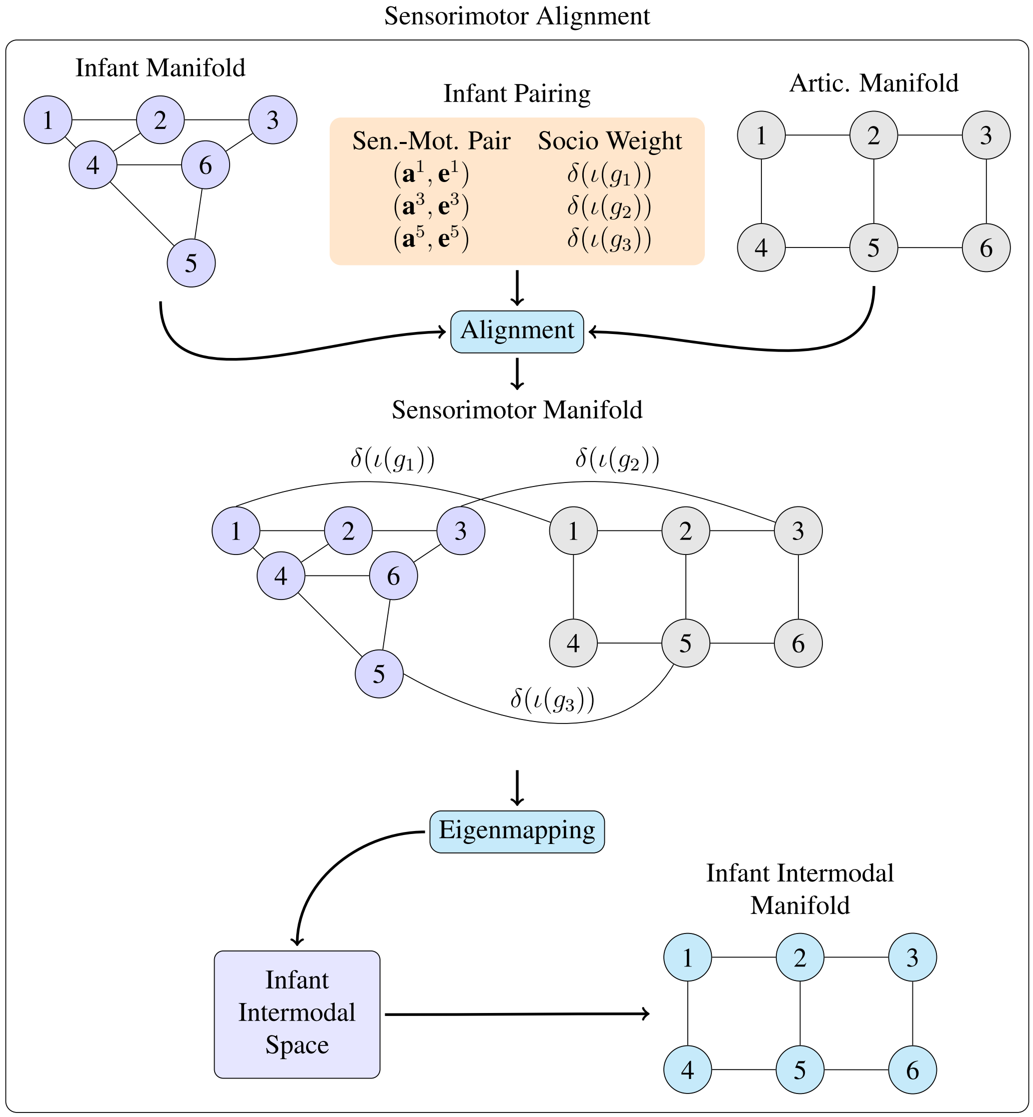
Response Pair	Transfer Weight
$(\mathbf{f}^1, \mathbf{f}^7)$	g_1
$(\mathbf{f}^3, \mathbf{f}^9)$	g_2
$(\mathbf{f}^5, \mathbf{f}^{11})$	g_3

Aud. Pair	Socio Weight
$(\mathbf{e}^1, \mathbf{e}^7)$	$\iota(g_1)$
$(\mathbf{e}^3, \mathbf{e}^9)$	$\iota(g_2)$
$(\mathbf{e}^5, \mathbf{e}^{11})$	$\iota(g_3)$

Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^1)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^3)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^5)$	$\delta(\iota(g_3))$

Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^7)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^9)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^{11})$	$\delta(\iota(g_3))$

Figure A.4: System architecture for sensorimotor pairing.



Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^1)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^3)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^5)$	$\delta(\iota(g_3))$

Figure A.5: System architecture for sensorimotor alignment.

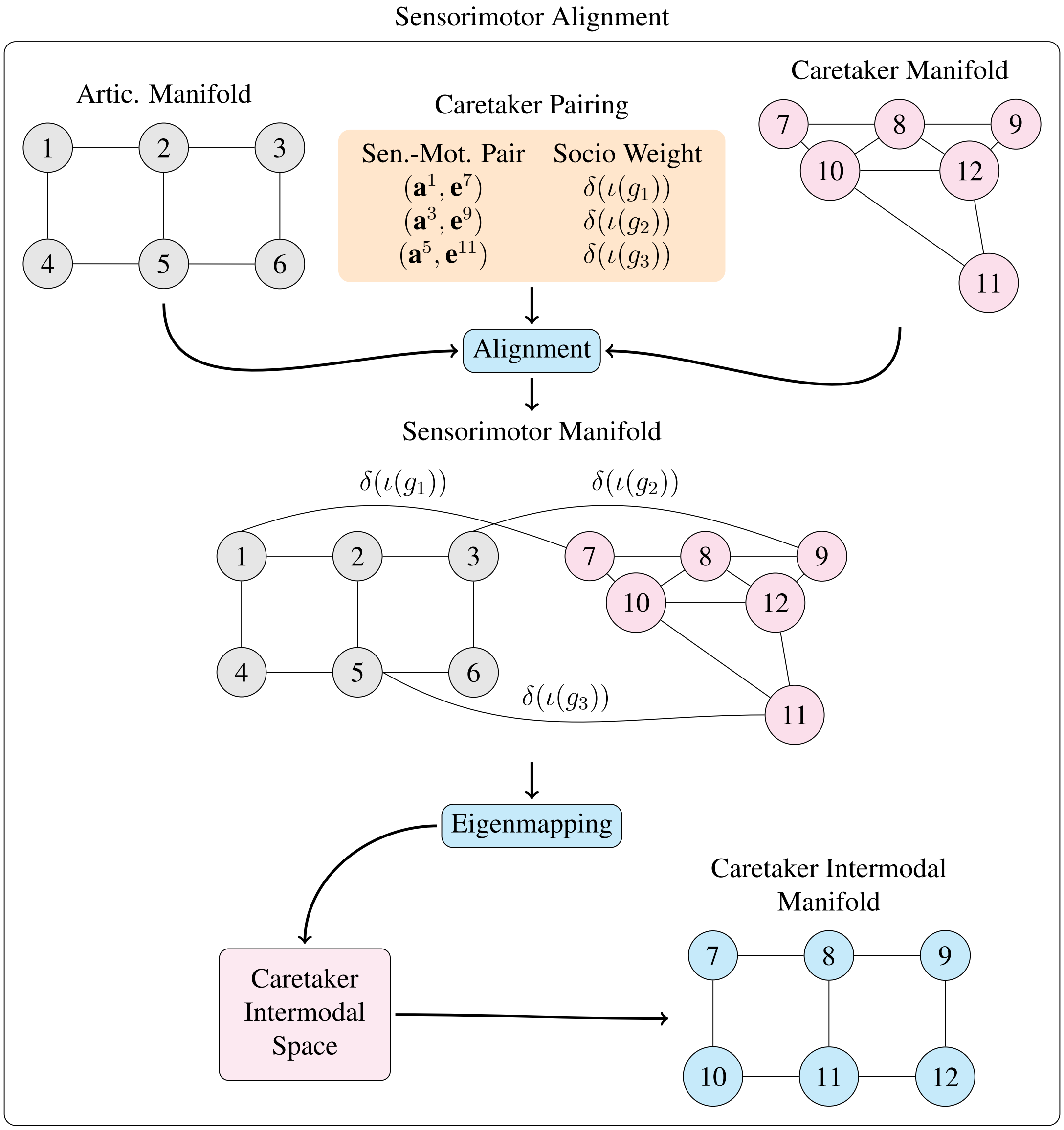


Figure A.6: System architecture for sensorimotor alignment.

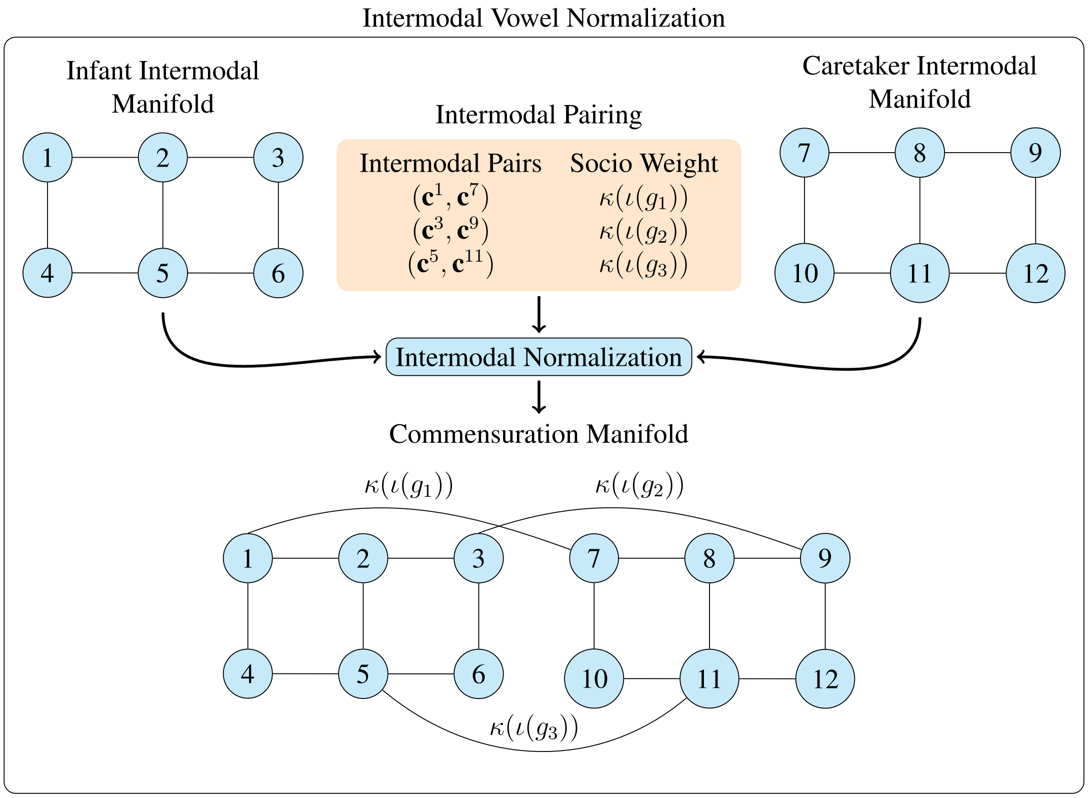


Figure A.7: System architecture for intermodal normalization.

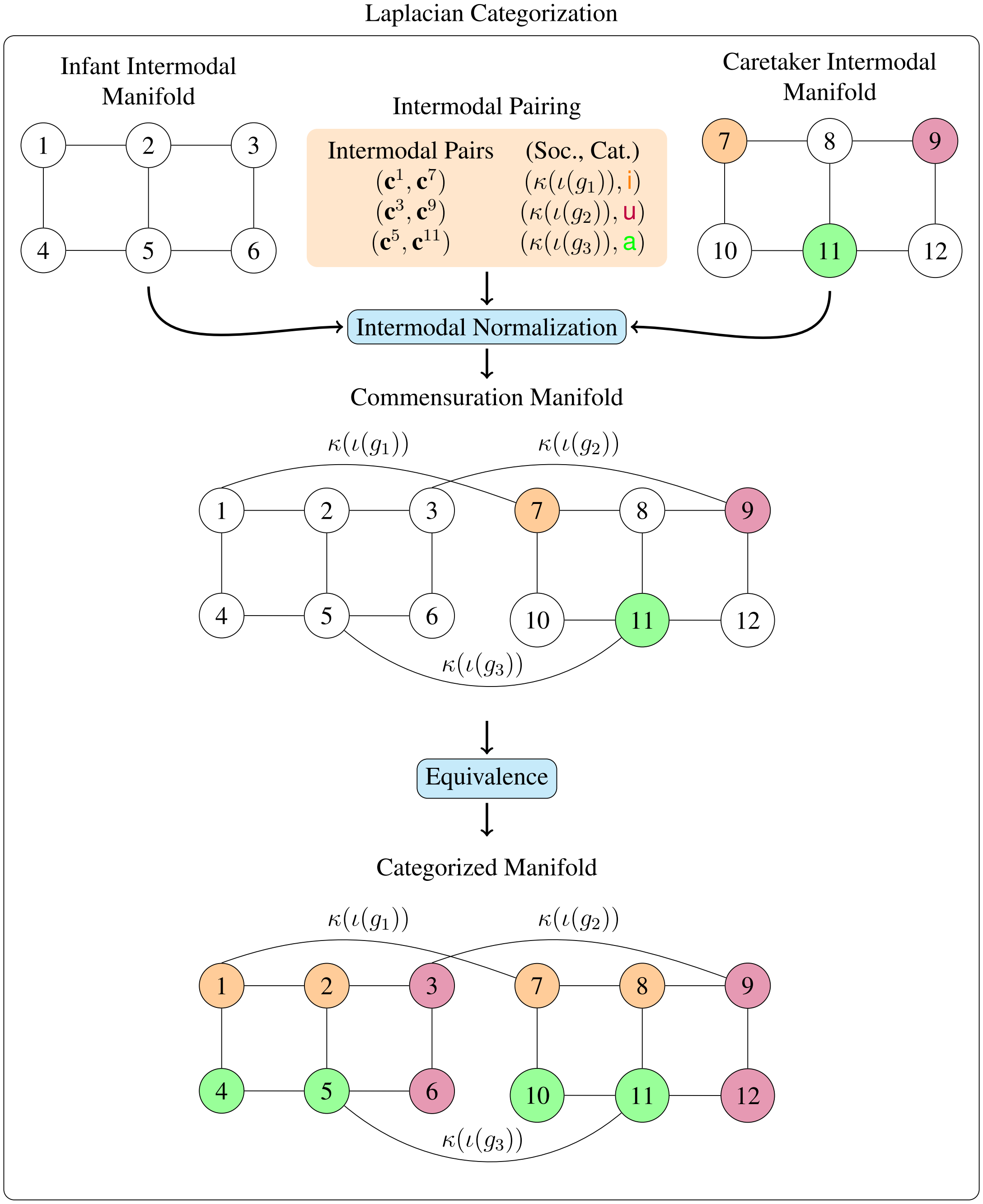

Laplacian Categorization
Infant Intermodal Manifold
1
2
3
4
5
6
Intermodal Pairing
Intermodal Pairs (Soc., Cat.)
$(\mathbf{c}^1, \mathbf{c}^7)$ $(\kappa(\iota(g_1)), \text{i})$
$(\mathbf{c}^3, \mathbf{c}^9)$ $(\kappa(\iota(g_2)), \text{u})$
$(\mathbf{c}^5, \mathbf{c}^{11})$ $(\kappa(\iota(g_3)), \text{a})$
Caretaker Intermodal Manifold
7
8
9
10
11
12
Intermodal Normalization
Commensuration Manifold
$\kappa(\iota(g_1))$
$\kappa(\iota(g_2))$
$\kappa(\iota(g_3))$
Equivalence
Categorized Manifold
$\kappa(\iota(g_1))$
$\kappa(\iota(g_2))$
$\kappa(\iota(g_3))$

Figure A.8: System architecture for vowel categorization.

APPENDIX B: ALIGNMENT FIGURES

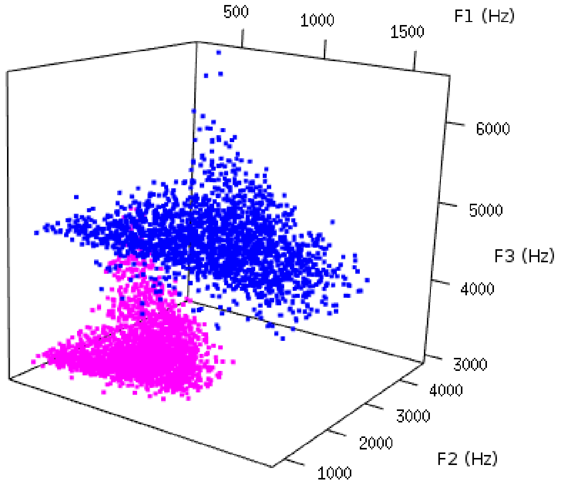


Figure B.1: Approximated Vlab maximal vowel spaces $\textsc{mvs}(0.5)$ and $\textsc{mvs}(10)$. The MVSs are depicted within a three-dimensional acoustic reference frame to emphasize their general lack of overlap. We have repeated Figures 3.11 and 3.11 to emphasize that the basis for computing the response surfaces that feed into a caretaker's response pairings is the same as in Chapters 3 and 4.

In this appendix, we provide several auditory normalization computations that highlight points made in discussion of Demonstrations 3.4.1 and 3.4.2. Specifically, we present vowel category response surfaces, and the warped representations they yield, derived from subjects s^J_{13}, s^G_4, s^E_5, and s^K_{17}. All modeling aspects of the extensions are identical to those in Demonstrations 3.4.1 and 3.4.2 save for the vowel category response surfaces and the socio-auditory pairings they yield.

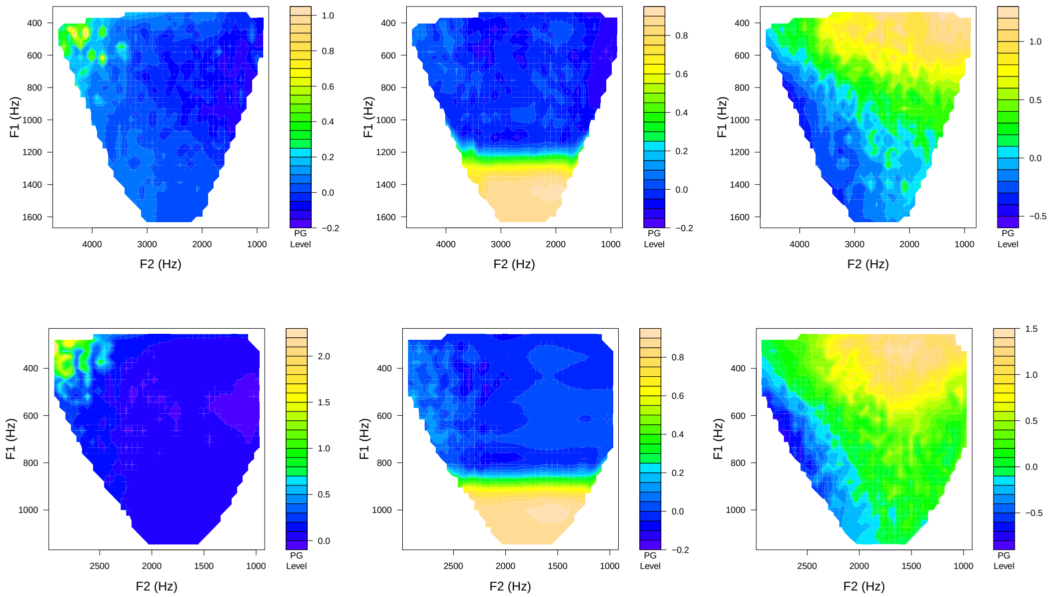


Figure B.2: Vowel category response surfaces $Q^J_{13}(\mathsf{c}, 0.5)$ (top) and $Q^J_{13}(\mathsf{c}, 10)$ (bottom) for caretaker J^{13}_{10} derived from subject s^J_{13} ($\mathsf{c} \in \{\mathsf{i,a,u}\}$).

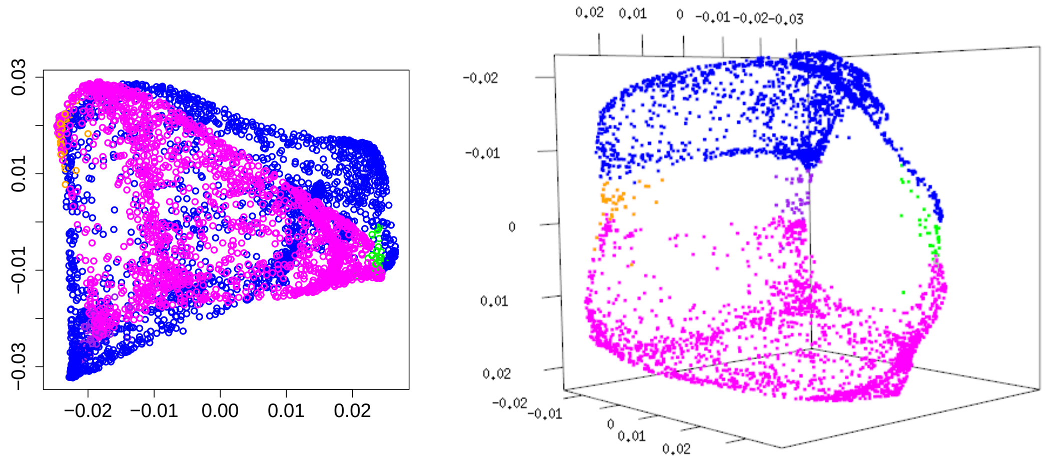


Figure B.3: The warping yielded by socio-auditory weighting $S(C(T^J_{13}(10, 0.5)))$. In the two-component depiction (left) the u representations are in the lower left corner.

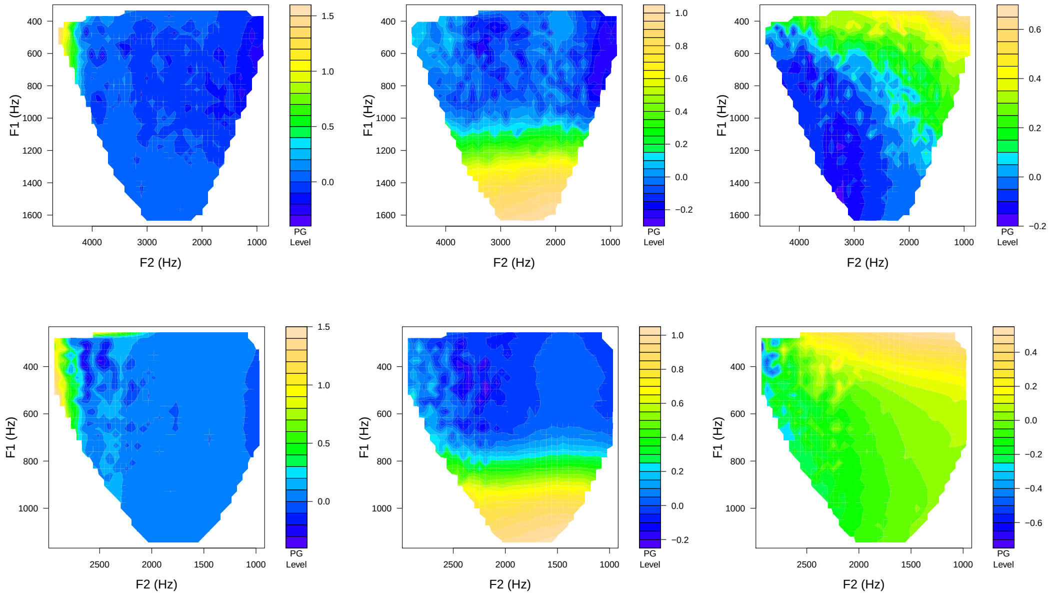


Figure B.4: Vowel category response surfaces $Q_4^G(\mathsf{c}, 0.5)$ (top) and $Q_4^G(\mathsf{c}, 10)$ (bottom) for caretaker G_{10}^4 derived from subject s_4^G ($\mathsf{c} \in \{\mathsf{i,a,u}\}$).

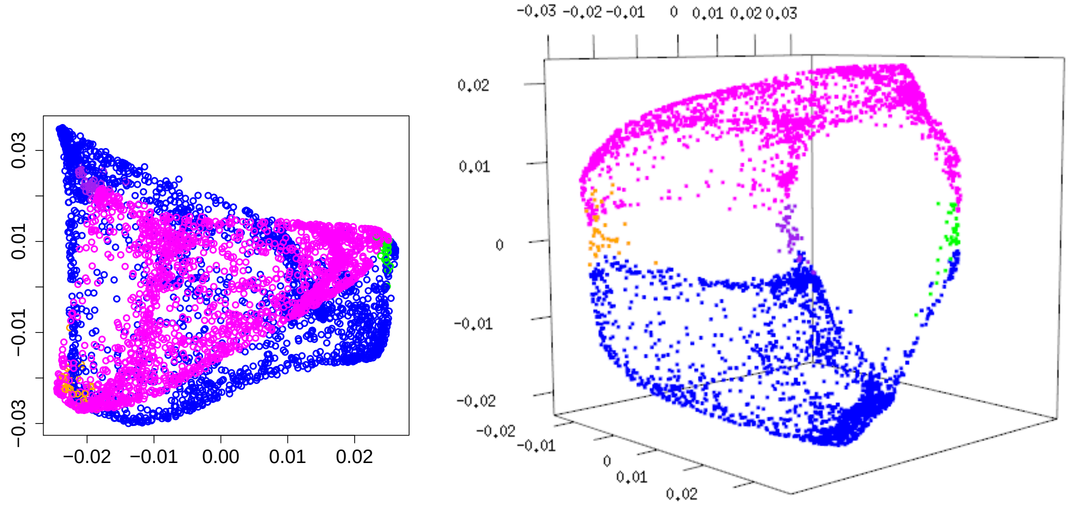


Figure B.5: The warping yielded by socio-auditory weighting $S(C(T_4^G(10, 0.5)))$. In the two-component depiction (left) the u representations are in the upper left corner.

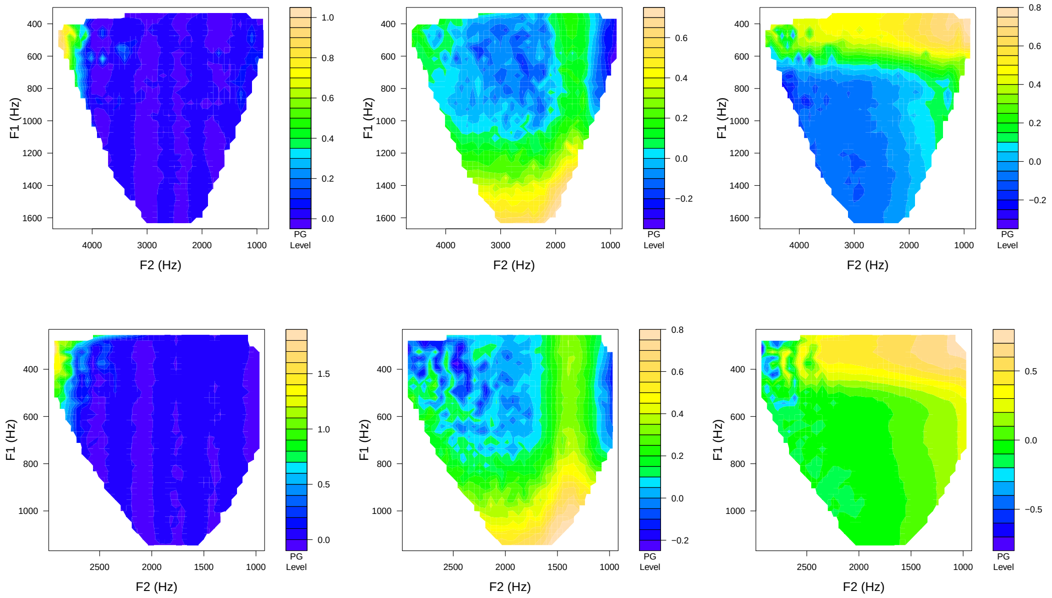


Figure B.6: Vowel category response surfaces $Q_5^E(\mathsf{c}, 0.5)$ (top) and $Q_5^E(\mathsf{c}, 10)$ (bottom) for caretaker E_{10}^5 derived from subject s_5^E ($\mathsf{c} \in \{\mathsf{i},\mathsf{a},\mathsf{u}\}$).

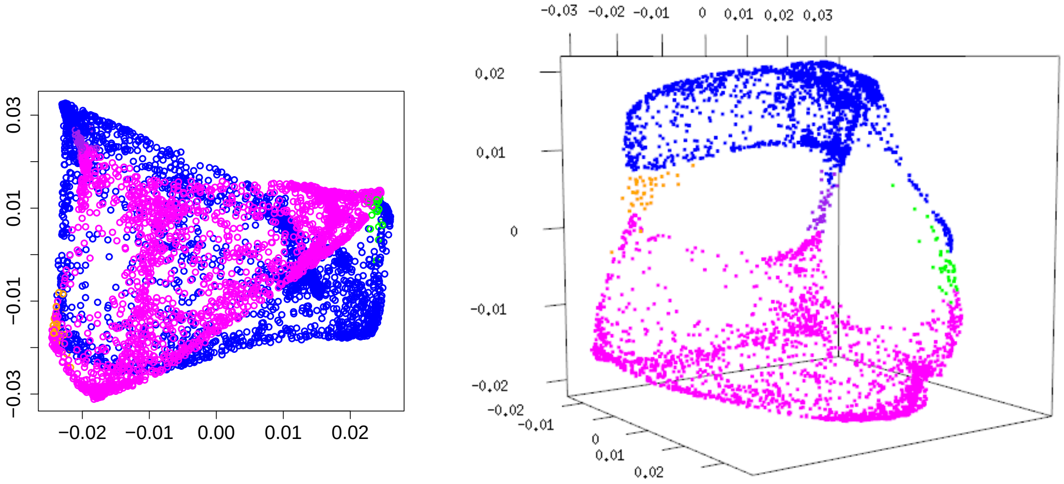


Figure B.7: The warping yielded by socio-auditory weighting $S(C(T_5^E(10, 0.5)))$. In the two-component depiction (left) the u representations are in the upper left corner.

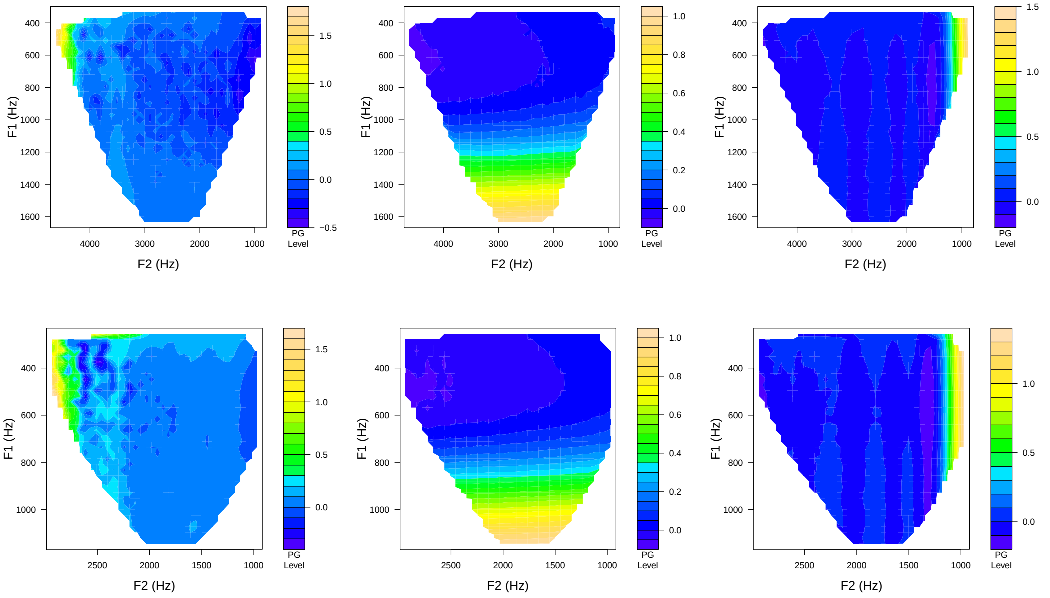


Figure B.8: Vowel category response surfaces $Q^K_{17}(\mathsf{c}, 0.5)$ (top) and $Q^K_{17}(\mathsf{c}, 10)$ (bottom) for caretaker K^{17}_{10} derived from subject s^K_{17} ($\mathsf{c} \in \{\mathsf{i},\mathsf{a},\mathsf{u}\}$).

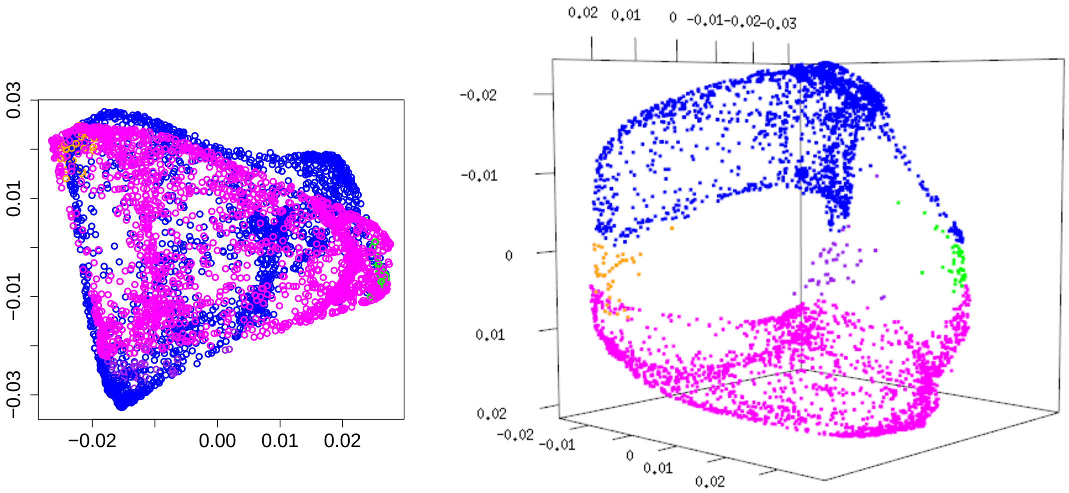


Figure B.9: The warping yielded by socio-auditory weighting $S(C(T^K_{17}(10, 0.5)))$. In the two-component depiction (left) the u representations are in the lower left corner.

APPENDIX C: BASIC MATHEMATICS

Without going into foundational detail, we will assume that there are things called *sets*, and we assume with the usual stock of sets from basic mathematics, in particalar, the set of natural numbers, denoted by $\mathbb{N}$, the set of real numbers, denoted by $\mathbb{R}$, and the set of complex numbers, denoted by $\mathbb{C}$. Moreover, we assume a relationship between sets called *membership*. Given two sets A and B, it holds that either A is a member of B, which we notate by $A \in B$, or A is not a member of B, which we notate by $A \notin B$. The sets that are members of a given set A, are also called *elements* of A. For example, each natural number n is an element of $\mathbb{N}$, and thus we write $n \in \mathbb{N}$, while most real numbers, such as π, are not, whence we write $\pi \notin \mathbb{N}$. A set A is a *subset* of a set B, denoted by $A \subseteq B$, if every element of A is also an element of B. For example, every natural number $n \in \mathbb{N}$ is a real number, and so $n \in \mathbb{R}$, whence $\mathbb{N} \subseteq \mathbb{R}$. Given two sets A and B, the *union of* **A** *and* **B**, denoted $A \cup B$, is the set of all sets that are elements of A or elements of B. The *intersection of* **A** *and* **B**, denoted $A \cap B$, is the set of all sets that are elements of A and elements of B. The *set difference of* **A** *and* **B**, denoted $A - B$, is the set of all sets that are elements of A, and not elements of B.

Given sets A and B, we call the set (A, B) the *ordered pair of* **A** *and* **B**, and A and B are called the *first component* and second component of the pair, respectively. We use ordered pairs to construct two important “cartesian” objects.

Given any two sets A and B, the *cartesian product of* **A** *and* **B**, denoted $A \times B$, is the set of all ordered pairs whose first component is an element of A and whose second component is an element of B. The set A is called the *first factor* of $A \times B$, while B is called the *second factor*. We can take the cartesian product repeatedly to form more complex sets. For example, given a set C we can form the product $(A \times B) \times C$ of *ordered triples* whose elements are of the form $((a, b), c)$ where $a \in A$, $b \in B$, and $c \in C$. It is easy to show that the parenthetical grouping interal to an ordered triple can be suppressed, thus we drop the parentheses from them and the product term hereafter. The definitions and concepts can be extended to ordered quadruples, quintuples, etc., in an obvious way. Given $n \in \mathbb{N}$ and sets $A_1, A_2, \ldots, A_n$, the set $(A_1, A_2, \ldots, A_n)$ is called an *ordered n-tuple*. Moreover, given $n \in \mathbb{N}$ and sets $A_1, A_2, \ldots, A_n$, their *n-fold cartesian product* is the set

$$A_1 \times A_2 \times \cdots \times A_n =_{def} \{(a_1, a_2, \ldots, a_n) \mid a_1 \in A_1, a_2 \in A_2 \ldots a_n \in A_n\}.$$

For any set A, a *cartesian power of A* is a cartesian product all of whose factors are A. For example, the *first cartesian power of A*, denoted A^1, is just A. The *cartesian square of A*, written A^2, is $A \times A$. More generally, for a natural number $n > 2$, the *n-th cartesian power of A*, denoted A^n, is the n-fold cartesian product all of whose factors are A.

The *cartesian coproduct*, or more commonly *disjoint union* of A and B, written $A + B$ is the set of ordered pairs (x, y) such that either $x = 0$ and $y \in A$, or $x = 1$ and $y \in B$. A and B are called the *cofactors* of $A + B$. We can take the disjoint union repeatedly to form more complex sets. For example, given a set C we can form the dijoint union $(A + B) + C$. It is easy to show that the parenthetical grouping can be suppressed, thus we drop the parentheses from them and the disjoint union hereafter. Given $n \in \mathbb{N}$ and sets $A_1, A_2, \ldots, A_n$, their *n-fold disjoint union* is the set $A_1 + A_2 + \cdots + A_n$. For any set A, a *cartesian copower of A* is a cartesian coproduct all of whose cofactors are A.

We return to the cartesian product which is used to define a number of important objects. Given two sets A and B, a *relation R between A and B* is simply a subset of the cartesian product of A and B. That is, $R \subseteq A \times B$. To be more precise, R is a *binary relation*. We define a *ternary relation among the sets A, B, and C* to be a subset of the threefold cartesian product $A \times B \times C$; thus a ternary relation is a set of ordered triples. For $n \in \mathbb{N}$ greater than 3, we define an *n-ary relation* are defined in the obvious way.

Let R be a binary relation on a set A. We say that R is *irreflexive* if $(a, a) \notin R$ for all $a \in A$. We say that R is *symmetric* if $(a, b) \in R$ implies that $(b, a) \in R$ for all $a, b \in A$. The *symmetric closure* of R is the relation $R \cup \{(b, a) \mid (a, b) \in R\}$. An *adjacency relation on a set A* is a binary relation on A that is irreflexive and symmetric. A *(simple, undirected) graph* is an ordered pair (V, E), where V is a set, and E is a set of two-element subsets of V. Given a graph $G = (V, E)$, the set V is called the *vertex set of G*, and its elements *vertices of G*, while the set E is called the *edge set of G*, and its elements the *edges of G*. The edge set E induces an adjacency relation on V.

A relation f between A and B is called a *function from A to B* if and only if for every $a \in A$, there exists a unique $b \in B$ such that $(a, b) \in f$. In this case we write $f : A \to B$ and refer to A as the *domain* of f, denoted $dom(A)$, and to B as the *codomain* of f, denoted $cod(f)$. For each $a \in dom(f)$, the unique b such that $(a, b) \in f$ is called the *value of f at a*, written $f(a)$. We also say that f maps a to b, written $f : a \mapsto b$. Given two sets A and B, the *characteristic function of B in A* is the function that maps each $x \in A$ to 1 if $x \in B$, and to 0 if $x \in A - B$.

For any set A, and $n \in \mathbb{N}$, a function A^n to A is called an *n-ary operation on A*. A *unary operation on A* is simply a function from A to itself. A function from the natural number 1 to A is called a *nullary operation*, which we use to conceptualize constants from the set A. A *binary operation on A* is a function from $A \times A$ to A. For example, addition of real numbers is a binary operation over $\mathbb{R}$, denoted $+_{\mathbb{R}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and the ordered pair

$(\mathbb{R}, +_{\mathbb{R}})$ denotes the "space" of real numbers with addition. Similarly, addition of complex numbers is a binary operation over $\mathbb{C}$, yielding the space $(\mathbb{C}, +_{\mathbb{C}})$. We can complicate our spaces by endowing them with further operations, like multiplication, which is also a binary operation (when defined appropriately) over both $\mathbb{R}$ and $\mathbb{C}$. Let $1_{\mathbb{R}} : 1 \to \mathbb{R}$ denote the nullary operation whose value is the real number 1, and $0_{\mathbb{R}} : 1 \to \mathbb{R}$ denote the nullary operation whose value is the real number 0. The space $(\mathbb{R}, +_{\mathbb{R}}, \cdot_{\mathbb{R}}, 0_{\mathbb{R}}, 1_{\mathbb{R}})$ is a special algebraic object called a "field," with additive identity $0_{\mathbb{R}}$ and multiplicative identity $1_{\mathbb{R}}$. The space $(\mathbb{C}, +_{\mathbb{C}}, \cdot_{\mathbb{C}}, 0_{\mathbb{C}}, 1_{\mathbb{C}})$ is also a field with additive identity $0_{\mathbb{C}}$ (a nullary operation taking on the value $(0, 0)$) and multiplicative identity $1_{\mathbb{C}}$ (a nullary operation taking on the value $(1, 0)$). We refer to these spaces as simply the "real field" $\mathbb{R}$ and the "complex field" $\mathbb{C}$. We let F denote an unspecified field in the general definitions hereafter, though in this dissertation we make use of the real and complex fields only.

Definition C.1. A *vector space* over a field F (e.g., $\mathbb{R}$ or $\mathbb{C}$) is an ordered triple $(V, +, \cdot)$ with binary operations $+ : V \times V \to V$, called *vector addition*, and $\cdot : F \times V \to V$, called *scalar multiplication*, such that

- $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$,
- $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in V$,
- There exists an element $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$ for all $\mathbf{u} \in V$,
- For any $\mathbf{u} \in V$, there exists an element $\mathbf{v} \in V$ such that $\mathbf{u} + \mathbf{v} = \mathbf{0}$,
- $a \cdot (b \cdot \mathbf{u}) = (a \cdot b) \cdot \mathbf{u}$ for all $a, b \in F$ and $\mathbf{u} \in V$,
- $1 \cdot \mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$ (1 is the multiplicative identity in F),
- $a \cdot (\mathbf{u} + \mathbf{v}) = (a \cdot \mathbf{u}) + (a \cdot \mathbf{v})$ for all $a \in F$ and $\mathbf{u}, \mathbf{v} \in V$,
- $(a + b) \cdot \mathbf{u} = (a \cdot \mathbf{u}) + (b \cdot \mathbf{u})$ for all $a, b \in F$ and $\mathbf{u} \in V$.

The elements of V are called *vectors*, and the element $\mathbf{0} \in V$ is called the *zero vector* of V. We typically denote the triple $(V, +, \cdot)$ simply by V. The elements of F are called *scalars*. If the field F in the above definition is the real field, then V is called a *real vector space*, and if F is the complex field, V is called a *complex vector space*.

In the remainder of this section we drop the convention of denoting vectors using boldface. A simple familiar example is the real vector space $(\mathbb{R}, +, \cdot)$. In this case, vector addition and multiplication are identical to addition and multiplication in the field $\mathbb{R}$. Another simple example is the complex vector space $(\mathbb{C}, +, \cdot)$, where $+$ and $\cdot$ are complex addition and multiplication. It is important to note that given a vector space $(V, +, \cdot)$ over a field F, vector addition $+$ and scalar multiplication $\cdot$ may differ from the addition and multiplication operations over F. A simple nontrivial example is the real vector space $(\mathbb{R}^2, +, \cdot)$. Vector addition is defined as follows: $(a, b) + (c, d) = (a + c, b + d)$, and scalar multiplication is defined as follows: $\alpha \cdot (a, b) = (\alpha a, \alpha b)$ for all $\alpha \in \mathbb{R}$.

Given a natural number $n > 2$, $\mathbb{R}^n$ is a real vector space with the obvious generalizations of the definitions of vector addition and scalar multiplication given above. These vector spaces are canonical examples of *coordinate spaces*, and we refer to them as *real coordinate spaces*. Similarly, $\mathbb{C}^n$ is a complex vector space with the same definitions of vector addition and scalar multiplication given above, called *complex coordinate spaces*.

Definition C.2. A complex vector space H is called an *inner product space* if for each ordered pair of vectors (x, y), where $x, y \in H$, there is associated a complex number $\langle x, y\rangle$ called the *inner product* of x and y, such that for all $x, y, z \in H$:

- $\langle y, x\rangle = \overline{\langle x, y\rangle}$.
- $\langle x + z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$
- $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$ (for scalar α).

- $\langle x, x\rangle \geq 0$.
- $\langle x, x\rangle = 0$ iff $x = 0$ (the zero vector).

Let F^n be a coordinate space, where F is either the complex or real field. The inner product on F^n is the *dot product*, defined as follows for all $x, y \in F^n$:

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.

Let H be an inner product space. For all $x \in H$, we define the *norm of x*, denoted $||x||$ to be the non-negative square root of $\langle x, x\rangle$. That is, $||x||^2 = \langle x, x\rangle$. We can define the distance between vectors x and y to be $||x - y||$. Specifically, the operation $d(x, y) = ||x - y||$ acts as a *metric* on H, i.e., (i) $d(x, y) = 0$ if and only if $x = y$, (ii) $d(x, y) = d(y, x)$, and (iii) $d(x, z) \leq d(x, y) + d(y, z)$. Thus, the vector space H together with $|| \cdot ||$ forms a *metric space*.

Definition C.3. A *metric space* is an ordered pair (H, d) where H is a set and d is a metric on H.

Let F^n be a coordinate space, where F is either the complex or real field. The operation $d(x, y) = ||x - y||$ based on the dot product acts as a *metric* on F^n, hence the coordinate spaces together with their dot products form metric spaces. We use the real coordinate spaces construed as metric spaces in this fashion as our models of reference frames.

BIBLIOGRAPHY

Adank, P., Smits, R., and van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5):3099–3107. 13, 36

Aertsen, A. M. H. J. and Johannesma, P. I. M. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38(4):223–234. 111

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In Fant, G. and Tatham, M. A. A., editors, *Auditory Analysis and Perception of Speech*, pages 103–113. Academic Press, London. 36

Ames, H. and Grossberg, S. (2008). Speaker normalization using cortical strip maps: a neural model for steady-state vowel categorization. *Journal of the Acoustical Society of America*, 124(6):3918–3936. 13, 38

Ananthakrishnan, G. and Salvi, G. (2011). Using imitation to learn infant-adult acoustic mappings. In *Proceedings of INTERSPEECH 2011*, pages 765–768. 17, 38

Badin, P. and Fant, G. (1984). Notes on vocal tract computations. *KTH Speech Transmission Laboratory – Quarterly Progress & Status Report*, 2-3:53–108. 159

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication*, 22(2-3):251–267. 17

Bates, E. and Elman, J. (1996). Learning rediscovered. *Science*, 274(5294):1849–1850. 85, 86

Beck, J. M. (1996). Organic variation of the vocal apparatus. In Hardcastle, W. J. and Laver, J., editors, *Handbook of Phonetic Sciences*, pages 256–297. Blackwell, Cambridge, England. 94

Beckman, M. E. (1997). Speech models and speech synthesis. In van Santen, J. P., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 185–209. Springer-Verslag, New York. 137

Belkin, M., Matveeva, I., and Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In Shawe-Taylor, J. and Singer, Y., editors, *COLT*, volume 3120 of *Lecture Notes in Computer Science*, pages 624–638. Springer. 117, 179

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396. 11, 64, 68

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(2):2399–2434. 117, 179

Bleile, K. M., Stark, R. E., and McGowan, J. S. (1993). Speech development in a child after decannulation: further evidence that babbling facilitates later speech development. *Clinical Linguistics and Phonetics*, 7(4):319–337. 12, 77

Bloom, K. and Lo, E. (1990). Adult perceptions of vocalizing infants. *Infant Behavior and Development*, 13(2):209–219. 18

Bloom, K., Russell, A., and Wassenberg, K. (1987). Turn-taking affects the quality of infant vocalizations. *Journal of Child Language*, 14(2):211–227. 16, 81

Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis. In *Proceedings of the International Congress Phon. Sci*, volume 3, pages 2501–2504. 163

Boë, L.-J., Heim, J.-L., Honda, K., and Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, 30(3):465–484. 138, 163

Boë, L.-J. and Maeda, S. (1998). Modélization de la croissance du conduit vocal. Éspace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse. In *Journée d'Études Linguistiques: "La Voyelle dans Tous ces États"*, pages 98–105. Nantes, France. 94, 138, 155, 156, 163

Boë, L.-J., Perrier, P., and Girin, B. (2010). Vlab. Software package. Université du Québec à Montréal. 164

Boë, L.-J., Perrier, P., Guérin, B., and Schwartz, J.-L. (1989). Maximal vowel space. In *EUROSPEECH 09*, pages 281–284, Paris, France. 94

Browman, C. and Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18(3):299–320. 142

Browman, C. and Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology: Vol. 1. Between the Grammar and Physics of Speech*, pages 341–376. Cambridge University Press, Cambridge, England. 142

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510. 93, 99

Callan, D. E., Kent, R., Guenther, F., and Vorperian, H. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech and Hearing Research*, 43(3):721–736. 6, 7, 9, 11, 12, 38, 42

Carlson, R. and Granström, B. (1975). A phonetically oriented programming language for rule description of speech. In Fant, G., editor, *Speech Communication*, volume 2, pages 245–253. Almqvist & Wiksell Stockholm. 137

Carney, L. H. and Yin, T. C. T. (1988). Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *Journal of Neurophysiology*, 60(5):1653–1677. 111

Catchpole, C. K. and Slater, P. J. B. (1995). *Bird song: Biological Themes and Variations*. Cambridge University Press, New York. 77

Chiba, T. and Kajiyama, M. (1941). *The Vowel, its Nature and Structure*. Tokyo-Kaiseikan Publishing Company, Ltd., Tokyo. 4, 34, 35, 36

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row. 137

Chung, F. R. K. (1997). *Spectral Graph Theory*. Regional Conference Series in Mathematics. American Mathematical Society. Number 92. 52, 66

Clopper, C. (2009). Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass*, 3(6):1430–1442. 36

Cooper, F. S., Liberman, A. M., and Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences of the United States of America*, 37(5):318. 137

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. Griffin Press: Adelaide, South Australia. Printed for The Limited Editions Club, 1971. 15

Davenport, R. K. (1976). Cross modal perception in apes. *Annals of the New York Academy of Sciences*, 280(1):143–149. 42, 147

Davenport, R. K. (1977). Cross-modal perception: a basis for language? In Rumbaugh, D. M., editor, *Language Learning by a Chimpanzee*, pages 73–83. Academic Press. 146

Davenport, R. K., Rogers, C. M., and Russell, I. S. (1973). Cross modal perception in apes. *Neuropsychologia*, 11(1):21–28. 147

Davis, B. L. and MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, 38(6):1199–1211. 12

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465. 17

de Boer, B. and Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, 18(1):36–47. 138, 163

de Boer, B. and Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4):129–134. 89

de Boer, B. and Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, 18(2):141–154. xxii de Boer, E. (1973). On the principle of specific coding. *Journal of Dynamic Systems, Measurement, and Control*, 95:265–273. 111

Deck, K. A. and Trautwein, W. (1964). Ionic currents in cardiac excitation. *Pflügers Archive*, 280:65–80. xxvi

Doupe, A. J. and Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, 22(1):567–631. 69, 75, 91

Edelman, G. W. (1987). *Neural Darwinism*. Basic Books, New York. 17

Eimas, P. D. (1975a). Auditory and phonetic coding of the cues for speech: Discrimination of the /r-l/ distinction by young infants. *Perception & Psychophysics*, 18(5):341–347. 83

Eimas, P. D. (1975b). Speech perception in early infancy. In Cohen, L. and Salapatek, P., editors, *Infant Perception, Vol. 2: From Sensation to Cognition*, pages 193–321, New York. Academic. 83

Fant, G. (1953). Speech communication research. *Royal Swedish Academy of Engineering Sciences*, 2:331–337. 137

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752–782. 6

Fels, S., Vogt, F., Van Den Doel, K., Lloyd, J., Stavness, I., and Vatikiotis-Bateson, E. (2006). ArtiSynth: A biomechanical simulation platform for the vocal tract and upper airway. In *International Seminar on Speech Production, Ubatuba, Brazil*. 138

Fitch, W. T. (2004). Evolving honest communication systems: Kin selection and ”mother tongues”. In Oller, D. K. and Griebel, U., editors, *Evolution of Communication Systems: A Comparative Approach*, pages 275–296. MIT Press., Cambridge, Massachusetts. 19

Fowler, C. A. (1986). An event approach to the study of speech perception. *Journal of Phonetics*, 14:3–28. 142

Fox, R. and Jacewicz, J. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *Journal of the Acoustical Society of America*, 126:2603–2618. 10

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823. 93

Fromkin, V., Krashen, S., Curtis, S., Rigler, D., and Rigler, M. (1974). The development of language in Genie: a case of language acquisition beyond the “critical period”. *Brain and Language*, 1:81–107. 76

Gallese, V. (2001). The ‘shared manifold’ hypothesis. from mirror neurons to empathy. *Journal of Consciousness Studies*, 8(5-7):33–50. 21, 43

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development*. Appleton-Century-Crofts, New York. 148

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston. 148

Girolametto, L., Weitzman, E., Wiigs, M., and Pearce, P. S. (1999). The relationship between maternal language measures and language development in toddlers with expressive vocabulary delays. *American Journal of Speech-Language Pathology*, 8(4):364. 82

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise methods. *Hearing Research*, 47:103–138. 111

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press. 207

Goldstein, M. H., King, A. P., and West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences*, 100(13):8030–8035. 81

Goldstein, M. H. and Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, 19(5):515–523. 16, 44

Goldstein, M. H., Schwade, J. A., and Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child development*, 80(3):636–644. 81

Goldstein, U. G. (1980). *An articulatory model of the vocal tract of the growing child*. PhD thesis, Massachusetts Institute of Technology. 94

Gould, S. J. (1997a). Darwinian fundamentalism. *New York Review of Books*, 44(10):34–37. (June 12). xxii

Gould, S. J. (1997b). Evolution: The pleasures of pluralism. *New York Review of Books*, 44(11):47–52. (June 26). xxii

Grimme, B., Fuchs, S., Perrier, P., and Schöner, G. (2011). Limb versus speech motor control: A conceptual review. *Motor control*, 15(1):5–33. 139, 141, 146

Gros-Louis, J., West, M. J., Goldstein, M. H., and King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30(5):112–119. 16, 44, 81, 82

Gu, C. (2002). *Smoothing spline ANOVA models*. Springer. 93

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102:594–621. 6, 9, 10, 11, 42, 143, 144

Guenther, F. H. (2003). Neural control of speech movements. In Schiller, N. O. and Meyer, A., editors, *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarites*, pages 209–239. Walter de Gruyter. 28, 29

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96:280–301. 6, 7, 9, 11, 12, 17, 91, 151, 152

Guenther, F. H. and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *J. of the Acoustical Society of America*, 100:1111–1121. 5, 6, 87

Guenther, F. H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633. 6, 143

Guenther, F. H. and Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25:408–422. 7, 38, 42

Halle, M. and Stevens, K. N. (1962). Speech recognition: A model and a program for research. *Information Theory, IRE Transactions on*, 8(2):155–159. 35

Ham, J., Lee, D. D., and Saul, L. K. (2005). Semisupervised alignment of manifolds. In Ghahramani, Z. and Cowell, R., editors, *Proc. of the Ann. Conf. on Uncertainty in AI*, volume 10, pages 120–127. 11, 47, 64, 68

Harvey, W. (1628). *Exercitatio Anatomica De Motu Cordis Et Sanguinis In Animalibus*. xxvii

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, New York. 93, 99

Heintz, I., Beckman, M., Fosler-Lussier, E., and Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. In *INTERSPEECH 09*, pages 688–691, Brighton, UK. 17, 38

Heinz, J. M. and Stevens, K. N. (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proceedings of the 5th International Conference on Acoustics*, page A44. xvi, 156, 157, 158

Hertz, S. R. (1982). From text to speech with SRS. *The Journal of the Acoustical Society of America*, 72(4):1155–1170. 137

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In Sankoff, D., editor, *Linguistic Variation: Models and Methods*, pages 161–171. New York: Academic. 13, 36

Hockett, C. F. (1965). Sound change. *Language*, 41(2):185–204. 24, 39

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544. xxvi

Holliday, J. J., Beckman, M. E., and Mays, C. (2010). Did you say susi or shushi? Measuring the emergence of robust fricative contrasts in English- and Japanese-acquiring children. In *Proceedings of INTERSPEECH 2010*, pages 1886–1889. 107

Honda, K. (1996). Organization of tongue articulation for vowels. *Journal of Phonetics*, 24:39–52. 6, 144

Hörnstein, J. (2013). *Developmental approach to early language learning in humanoid robots*. PhD thesis, Universidade Técnica de Lisboa Instituto Superior Técnico. 6, 9, 18, 91, 149

Howard, I. S. and Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, 15:85–117. 9, 17, 43, 45

Hsu, H. C., Iyer, S. N., and Fogel, A. (2013). Effects of social games on infant vocalizations. *Journal of Child Language*, 41(1):1–23. 21, 205

Immelmann, K. (1969). Song development in the zebra finch and other estrildid finches. In Hinde, R. A., editor, *Bird Vocalizations*. Cambridge Univ. Press, London. 76

Ishihara, H., Yoshikawa, Y., Miura, K., and Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *Autonomous Mental Development, IEEE Transactions on*, 1(4):217 –225. 13, 17, 38

Iverson, P. and Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *The Journal of the Acoustical Society of America*, 99(2):1130–1140. 84

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, 62(4):874–886. 85, 86

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. M.I.T. Press, Cambridge, Massachusetts. 35

Jansen, A. and Niyogi, P. (2006). Intrinsic fourier analysis on the manifold of speech sounds. In *in IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 241–244. 11, 32, 146

Jansen, A. and Niyogi, P. (2007). Semi-supervised learning of speech sounds. In *Proceedings of INTERSPEECH 2007*. 11, 32

Jespersen, O. (1922). *Language: Its Nature, Development, and Origin*. W. W. Norton & Company, New York. 25

Johnson, K. (1990a). Contrast and normalization in vowel perception. *Journal of Phonetics*, 18:229–254. 14, 20, 37, 39

Johnson, K. (1990b). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88(2):642–654. 14

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson and Mullennix, editors, *Talker Variability in Speech Processing*, pages 145–165. San Diego: American Press. 37

Johnson, K. (2005). Speaker normalization in speech perception. In Remez, R. E. and Pisoni, D. B., editors, *The Handbook of Speech Perception*, pages 363–389. Blackwell. 14, 39

Joos, M. (1948). Acoustic phonetics. *Language*, 24(2):5–136. 1, 34, 35, 36, 39, 40

Kallay, J. and Holliday, J. J. (2012). Using spectral measures to differentiate Mandarin and Korean sibilant fricatives. In *Proceedings of INTERSPEECH 2012*. 107

Kamen, R. S. and Watson, B. C. (1991). Effects of long-term tracheostomy on spectral characteristics of vowel productions. *Journal of Speech and Hearing Research*, 34:1057–1065. 6, 77

Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793. 137

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., and Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297(5582):846–848. 151

Kohn, M. E. and Farrington, C. (2012). Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. *Journal of the Acoustical Society of America*, 131:2237–2248. 14, 39

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69. 5

Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift für Tierpsychologie*, 22(7):770–783. 77

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, 66(6):1668–1679. 12, 37, 40

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, 6:263–285. 12, 37, 40

Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the protoypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50:93–107. 5, 84, 85

Kuhl, P. K. (2000). A new view of language acquisition. *PNAS*, 97(22):11850–11857. 82, 84

Kuhl, P. K. (2007). Is speech ‘gated’ by the social brain? *Developmental Science*, 10(1):110–120. 82, 84, 85, 87, 90, 91

Kuhl, P. K. and Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, 218(4577):1138–1141. 8, 42, 144

Kuhl, P. K. and Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, 100(4):2425–2438. 12, 37, 144

Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101. 90

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608. 85

Labov, W. (1963). The social motivation of a sound change. *Word*, 19:273–309. 35

Labov, W. (1966). *The Social Stratification of English in New York City*. Cambridge University Press. 35

Labov, W. (1972). *Sociolinguistic Patterns*. Univ. Penn. Press. 35

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the XIIIth International Congress on Phonetic Sciences*, volume 2, pages 140–147. 87

Ladefoged, P. and Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1):98–104. 14, 35, 36, 39

Lake, B. M., Vallabha, G. K., and McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, 1(1):35–43. 6

Lane, H. L. (1976). *The Wild Boy of Aveyron*. Harvard University Press, Cambridge MA. 76

Lasky, R. E., Syrdal-Lasky, A., and Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 20(2):215–225. 83

Leake, Chauncey, D. (1928). *Anatomical Studies of the Motion of the Heart and Blood: The Leake Translation*. Charles C. Thomas, Springfield, Illinois. xxviii

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105(3):1455–1468. 12

Lewin, K. (1936). Topological psychology. *Mac Graw Hill*. 28, 207

Lewkowicz, D. J. (1994). Development of intersensory perception in human infants. In Lewkowicz, D. J. and Lickliter, R., editors, *The Development of Intersensory Perception: Comparative Perspectives*, pages 165–203. Lawrence Erlbaum Associates, Inc. 148

Lewkowicz, D. J. and Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences*, 103(17):6771–6774. 86

Lewkowicz, D. J. and Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in cognitive sciences*, 13(11):470–478. 9, 148, 149

Lewkowicz, D. J. and Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, 109(5):1431–1436. 86

Lewkowicz, D. J. and Lickliter, R., editors (1994). *The Development of Intersensory Perception: Comparative Perspectives*. Lawrence Erlbaum Associates, Inc. 8, 148

Lewkowicz, D. J. and Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology*, 16(6):597–607. 148

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6):431. 142

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36. 142

Lindblom, J. and Sundberg, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50:1166–179. 94, 138, 155, 160

Lloyd, R. J. (1890). *Some Researches into the Nature of Vowel-Sound*. Turner and Dunnett, Liverpool. 33, 35

Locke, J. L. and Pearson, D. M. (1990). Linguistic significance of babbling: evidence from a tracheostomized infant. *Journal of Child Language*, 17:1–16. 78

Ma, Y. and Fu, Y. (2012). *Manifold Learning Theory and Applications*. CRC Press. 11, 47, 64, 146, 206

MacKain, K. S. (1983). Speaking without a tongue. *Journal of the National Student Speech Language Hearing Association*, 11:46–71. 78

Maeda, S. (1979). Une modèle articulatoire de la tongue avec des composantes linéaires. *10émes JEP, GALF*, pages 152–164. 161

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. and Marchal, A., editors, *Speech Production and Speech Modeling*, pages 131–149. The Netherlands: Kluwer Academic Publishers. xvi, 6, 94, 138, 143, 155, 159

Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, 19:321–331. 94, 143, 155, 160

Masataka, N. (1993). Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three- and four-month-old Japanese infants. *Journal of Child Language*, 20:303–312. 80, 81

Masataka, N. (2003). *The Onset of Language*. Cambridge University Press, Cambridge, UK. 16, 17, 19, 44, 80

Masataka, N. and Biben, M. (1987). Temporal rules regulating affiliative vocal exchanges of squirrel monkeys. *Behaviour*, 101:311–319. 79

Massaro, D. W. and Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, 2:15–35. 96

Maye, J., Werker, J. F., and Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111. 85, 88

McCowan, B. and Reiss, D. (1997). Vocal learning in captive bottlenose dolphins: A comparison with humans and nonhuman animals. In Snowdon, C. T. and Hausberger, M., editors, *Social Influences on Vocal Development*, pages 178–207. Cambridge University Press. 75

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *Speech, Language and the Law*, 13:89–126. 10

McMurray, B., Aslin, R. N., and Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12:369–378. 88, 89, 90, 91, 149

Mead, G. H. (1909). Social psychology as counterpart to physiological psychology. *Psychological Bulletin*, 6(12):401–408. 19

Meltzoff, A. (2007). The ‘like me’ framework for recognizing and becoming an intentional agent. *Acta Psychologica*, 124:26–43. 20, 43

Meltzoff, A. N. and Kuhl, P. K. (1994). Faces and speech: Iintermodal processing of biologically relevant signals in infants and adults. In Lewkowicz, D. J. and Lickliter, R., editors, *The Development of Intersensory Perception: Comparative Perspectives*, pages 335–369. Lawrence Erlbaum Associates, Inc. 8, 152

Ménard, L. and Boë, L.-J. (2000). Exploring vowel production strategies from infant to adult by means of articulatory inversion of formant data. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 465–468. Beijing, China. 95

Ménard, L., Schwartz, J.-L., and Boë, L.-J. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *Journal of the Acoustical Society of America*, 111(4):1892–1905. 37, 94, 95

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85:2114–2134. 13, 143

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, 50(1-3):271–284. 93, 96

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and cognitive processes*, 12(5/6):865–869. 93, 96

Miller, M. B. and Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Reviews in Microbiology*, 55(1):165–199. 70

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, 18(5):331–340. 83

Moore, B. C. J. and Glasberg, B. G. (1996). A revision of Zwicker's loudness model. *Acta Acoustica*, 82:335–345. 5

Moore, B. C. J., Glasberg, B. G., and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240. 107

Munson, B., Ménard, L., Beckman, M. E., Edwards, J., and Chung, H. (2010). Sensorimotor maps and vowel development in English, Greek, and Korean: A cross-linguistic perceptual categorizaton study (A). *Journal of the Acoustical Society of America*, 127:2018. 93, 95, 96

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85(5):2088–2113. 10

Nearey, T. M. and Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80:1297–1308. 10

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, 51:574–585. 10

Niyogi, P. (2004). Towards a computational model of human speech perception. In *Proceedings of the Conference on Sound to Sense, MIT (In Honor of Ken Stevens' 80th birthday)*. 32

Noble, D. (1962). A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pacemaker potentials. *Journal of Physiology*, 160:317–352. xxvi

Noble, D. (2002). Modelling the heart: insights, failures, and progress. *BioEssays*, 24(2):1155–1163. xxvi

Oberman, L. M. and Ramachandran, V. S. (2007). The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin*, 133:310–327. 208

Oberman, L. M. and Ramachandran, V. S. (2008). Preliminary evidence for deficits in multisensory integration in autism spectrum disorders: The mirror neuron hypothesis. *Social Neuroscience*, 3:348–355. 208

Ohms, V. R., Gill, A., Van Heijningen, C. A. A., Cate, C. T., and Beckers, G. J. L. (2010). Zebra finches exhibit speaker-independent phonetic perception of human speech. In *Proceedings of the Royal Society B: Biological Sciences*, volume 277, pages 1003–1009. 74

Oller, D. K. and Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development*, 59:441–449. 12

Oudeyer, P.-Y. (2001). The origins of syllable systems: An operational model. In *Proc. 23rd Ann. Conf. of the Cognitive Science Society*, volume 23, pages 744–749. 17

Oudeyer, P.-Y. (2002). Phonemic coding might result from sensory-motor coupling dynamics. In Hallam, B., Floreano, D., Hallam, J., Hayes, G., and Meyer, J.-A., editors, *Proc. 7th Int'l Conf. on the Simulation of Adaptive Behavior*, pages 406–416. MIT Press. 7, 9, 11, 17, 42

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233:435–449. xxii

Ozturk, O., Krehm, M., and Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114:173–186. 208

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59:640–654. 109

Patterson, R. D. and Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In Moore, B. C. J., editor, *Frequency Selectivity in Hearing*, pages 123–177. Academic Press, London. 111

Perkell, J. S. (1996). Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modelling. *Journal of Phonetics*, 24:3–22. 6, 144

Perkell, J. S., Matthies, M. L., Svirsky, M. A., and Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel [u]: A pilot ”motor equivalence” study. *The Journal of the Acoustical Society of America*, 93(3):2948–2961. 142

Perrachione, T. K., Tufo, S. N. D., and Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science*, 333:595. 20, 43

Perrier, P. (2005). Control and representations in speech production. *ZAS Papers in Lingustics*, 40:109–132. 144, 206

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184. 4, 27, 28, 29, 30, 32, 35

Piaget, J. (1936). *The Origins of Intelligence in Children*. W. W. Norton & Company Inc., New York, NY, USA. Translated by Margaret Cook, 1963. 7, 147

Piaget, J. (1945). *Play, Dreams and Imitation in Childhood*. W. W. Norton & Company Inc., New York, NY, USA. Translated by C. Gattegno and F. M. Hodgson, 1962. 15

Plummer, A. R. (2012a). Aligning manifolds to model the earliest phonological abstraction in infant caretaker vocal imitation. In *13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, Portland, OR. 11

Plummer, A. R. (2012b). Manifold alignment, vocal imitation, and the perceptual magnet effect. In *The Annual International Child Phonology Conference (ICPC 2012)*, Minneapolis, MN. 32, 97

Plummer, A. R., Beckman, M. E., Belkin, M., Fosler-Lussier, E., and Munson, B. (2010). Learning speaker normalization using semisupervised manifold alignment. In *Proceedings of INTERSPEECH 2010*, Tokyo. 11

Plummer, A. R., Ménard, L., Munson, B., and Beckman, M. E. (2013a). Comparing vowel category response surfaces over age-varying maximal vowel spaces within and across language communities. In *Proceedings of INTERSPEECH 2013*. 93, 103

Plummer, A. R., Munson, B., Ménard, L., and Beckman, M. E. (2013b). Examining the relationship between the interpretation of age and gender across languages. In *21st International Congress on Acoustics, 165th Meeting of the Acoustical Society of America, 52nd Meeting of the Canadian Acoustical Society (ICA 2013)*, Montréal, CA. 119, 204

Rasilo, H., Räsänen, O., and Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, 55(9):909–931. 6, 9, 18, 91, 149

Reidy, P. (2013). An introduction to random processes for the spectral analysis of speech data. *OSU Working Papers in Linguistics*, 60:67–116. 107, 112, 169

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141. 151

Rose, J. (2001). *The Intellectual Life of the British Working Classes*. Yale University Press. 207

Rosenberg, S. (1997). *The Laplacian on a Riemannian manifold: an introduction to analysis on manifolds*. Cambridge University Press. 64, 66

Russell, G. O. (1928). *The Vowel*. The Ohio State University Press. 4, 30, 135, 136

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928. 85, 88, 91

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In Heuer, H. and Fromm, C., editors, *Experimental Brain Research Series 15*, pages 129–144. New York: Springer-Verlag. xvi, 9, 29, 142, 143

Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In Port, R. F. and van Gelder, T., editors, *Mind as motion: Explorations in the dynamics of cognition*, pages 149–173. MIT Press, Cambridge, MA. 9, 29, 143

Saltzman, E. and Kelso, J. A. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, 94(1):84–106. 142, 153

Saltzman, E., Kubo, M., and Tsao, C. C. (2006). Controlled variables, the uncontrolled manifold, and the task-dynamic model of speech production. In Divenyi, P., Greenberg, S., and Meyer, G., editors, *Dynamics of Speech Production and Perception*, pages 21–31. 146

Saltzman, E. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382. 9, 29, 143

Savariaux, C., Perrier, P., and Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, 98(5):2428–2442. 142

Schöner, G., Martin, V., Reimann, H., and Scholz, J. (2008). Motor equivalence and the uncontrolled manifold. In *Proceedings of the International Seminar on Speech Production (ISSP 2008) in Strassbourg*, pages 23–28. 145, 146

Schwartz, J.-L., Basirat, A., Ménard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336–354. 144

Schwartz, J.-L., Boë, L.-J., and Abry, C. (2007). Linking dispersion-focalization theory and the maximum utilization of the available distinctive features principle in a perception-for-action-control theory. In Sole, M.-J., Beddor, P. S., and Ohala, M., editors, *Experimental Approaches to Phonology*, pages 104–124. Oxford University Press. 94

Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997a). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25:255–286. 95

Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997b). Major trends in vowel system inventories. *Journal of Phonetics*, 25:233–253. 95

Seo, S., Chung, M. K., and Vorperian, H. K. (2010). Heat kernel smoothing using Laplace-Beltrami eigenfunctions. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 505–512. Springer. 146

Seo, S., Chung, M. K., Whyms, B. J., and Vorperian, H. K. (2011). Mandible shape modeling using the second eigenfunction of the Laplace-Beltrami operator. In *SPIE Medical Imaging*, pages 79620Z–79620Z. International Society for Optics and Photonics. 146

Seung, H. S. and Lee, D. D. (2000). The manifold ways of perception. *Science*, 290(5500):2268–2269. 11

Shaw, E. A. G. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *Journal of the Acoustical Society of America*, 56(6):1848–1861. 107

Smith, D., Patterson, R., Turner, R., Kawahara, H., and Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, 117:305–318. 37

Smith, L. B. (1994). Foreword. In Lewkowicz, D. J. and Lickliter, R., editors, *The Development of Intersensory Perception: Comparative Perspectives*, pages ix–xix. Lawrence Erlbaum Associates, Inc. 147

Smith, R. (1997). *The Norton History of the Human Sciences*. W. W. Norton & Company. 19

Spong, M. W. (1996). Motion control of robot manipulators. In Levine, W. S., editor, *The Control Handbook*, pages 1339–1351. CRC Press. xvi, 140, 141

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 32(1):47–55. 35

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In David, E. E. and Denes, P. B., editors, *Human Communication: A Unified View*, pages 51–66. McGraw-Hill. 142

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17:3–45. 142

Stevens, S. S., Volkman, J., and Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190. 5

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. *Nature*, 110:311–312. 136

Stoel-Gammon, C. and Otomo, K. (1986). Babbling development of hearing-impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders*, 51:33–41. 11

Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature*, 259:39–41. 83

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, 28:12–23. 13, 38

Tamis-LeMonda, C. S., Bornstein, M. H., and Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development*, 72(3):748–767. 82

Thelen, E. and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, London. 17

Thorpe, W. H. (1958). The learning of song patterns by birds, with especial reference to the song of the chaffinch, *Fringilla coelebs*. *Ibis*, 100(4):535–570. 76

Tinbergen, N. (1953). *Social Behaviour in Animals*. London: Methuen. xxi

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, 20:410–433. xxi, xxiv

Tonndorf, J., editor (1981). *Physiological Acoustics*. Benchmark Papers in Acoustics. Hutchinson Ross Publishing Company. 107

Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Languages and Cognitive Processes*, 26(7):952–981. 6

Vallée, N., Boë, L.-J., and Payan, Y. (1995). Vowel prototypes for UPSID's 33 phonemes. In *Proceedings of ICPhS 2*, pages 424–427. Stockholm. 95

von Békésy, G. (1947). The variations of phase along the basilar membrane with sinusoidal vibrations. *Journal of the Acoustical Society of America*, 19:452–460. 108

von Helmholtz, H. (1863). *On the Sensations of Tone*. David McKay, New York. trans. A. J. Ellis, 1912. 136

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics. 93

Wang, C. (2010). *A Geometric Framework For Transfer Learning Using Manifold Alignment*. PhD thesis, University of Mass. Amherst. 11, 206

Waters, C. M. and Bassler, B. L. (2005). Quorum sensing: cell-to-cell communication in bacteria. *Annual Review of Cell & Developmental Biology*, 21:319–346. xiv, 70, 71, 73

Weeks, J. R. (2002). *The Shape of Space*. Marcel Dekker, Inc., 2nd edition. 31

Westermann, G. and Miranda, E. R. (2002). Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research*, 31(4):367–375. 7, 9, 11, 42

Westermann, G. and Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, 89:393–400. 7, 9, 11, 42

Wheatstone, C. (1837). Reed-organ pipes, speaking machines, etc. *London and Westminster Review*. 136

Whitney, W. D. (1875). *The Life and Growth of Language*. H. S. King & Co. 19

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press, USA. 97

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *The Journal of the Acoustical Society of America*, 97:3085–3098. 138

Willis, R. (1830). On vowel sounds, and on reed-organ pipes. *Transactions of the Cambridge Philosophical Society*, 3:231–268. 136

Winston, P. (2011). Keynote Panel – The Golden Age – A Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience. MIT Symposium on Brains, Minds, and Machines. Video available at: http://mit150.mit.edu/symposia/brains-minds-machines. xxiii

Winters, S. J., Levi, S. V., and Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*, 123(6):4524–4538. 43

Wollock, J. (1997). *The Noblest Animate Motion: Speech, physiology and medicine in pre-Cartesian linguistic thought*. John Benjamins Publishing Company. 20

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882. 43, 115

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC. 93, 99, 100

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114. 99, 100

Zahorian, S. A. and Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94:1966–1982. 10

Zemlin, W. R. (1998). *Speech and Hearing Science: Anatomy and Physiology*. Allyn & Bacon, fourth edition. 205

Zuidema, W. and de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, 37(2):125–144. xxii

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248. 5