University of California

Los Angeles

Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Linguistics

by

Ying Lin

2005

The dissertation of Ying Lin is approved.

Ying-Nian Wu

Abeer Alwan

Bruce Hayes

Edward Stabler, Committee Co-chair

Patricia Keating, Committee Co-chair

University of California, Los Angeles

2005

To my parents:

Lin Jin-Sheng and Lin Hui-Zhen

with love and gratitude

Table of Contents

List of Figures

List of Tables

Acknowledgments

I was very fortunate to have met many nice people who helped me during my five years at UCLA.

First, my deepest gratitude goes to my co-chairs: Pat Keating and Ed Stabler. Pat Keating always patiently watched over my growth in the graduate school, provided timely advice at the important moments, and showed her confidence in me at times of confusion and frustration. The topic of the dissertation originated from a summer project with her. Without her warm encouragement and cheerful support, I would not have continued to pursue the same question. Ed Stabler has been my mentor since my first year at UCLA, and I felt much reassured when he kindly agreed to serve as co-chair of the committee. His emphasis on clarity and his commitment to scientific goals have been constant sources of inspiration. I thank him for his generous support during the various stages of my graduate study, as well as for giving me the freedom and encouragement to pursue my own interests.

Influences of my other committee members are also evident, starting from the classes I took from them. Bruce Hayes helped me develop my interest in phonological acquisition in his seminar. Every meeting in his office was an adventure filled with inspiring questions and conversations. I owe Yingnian Wu most of what I know about statistics. His view on scientific modeling will also stay with me. I also thank Abeer Alwan for teaching me about speech, and expressing interest in my work.

My thanks also extend to my other teachers in the Department of Linguistics at UCLA, including: Colin Wilson, Sun-Ah Jun, Ed Keenan, Hilda Koopman, Tim Stowell, Kie Zuraw, Marcus Kracht and Daniel Büring. Outside the depart-

ment, I am also grateful to Dr. Charles Taylor for giving me the opportunity to work in his lab.

Graduate school could have been a much less pleasant experience without friends in the phonetics lab and in the department (too many to name here): Tim Arbisi-Kelm, Leston Buell, Ivano Caponigro, Melissa Epstein, Christina Esposito, John Foreman, Henk Harkema, Alex Hurst, Ben Keil, Sameer Khan, Haiyong Liu, Kuniko Nielsen, Katya Petrova, Rebecca Scarborough, Shabnam Shademan, Marcus Smith, Mariam Sy, Harold Torrence, Stefano Vegnaduzzo and Jie Zhang. They have helped me in many aspects of my life.

I would also like to thank my teachers at Beijing University who introduced me to linguistics. They are Professor Wang Hong-Jun, Professor Yuan Yu-Lin and Dr. Gao Xiao-Hong. It would not have been possible for me to pursue a Ph.D. in the U.S. without their encouragement and help.

Finally, my thanks go to my partner Hisayo Suzuki for her love, support and patience. Thank you for all the waiting.

Vita

1979	Born, Fuzhou, Fujian Province, China
2000	B.S. in Information Science, Minor in Computer Software, Beijing University, Beijing
2005	M.A. in Mathematics, UCLA, Los Angeles
2001-2004	Research Assistant Department of Linguistics and Department of Ecology and Evolutionary Biology, UCLA
2002-2004	Teaching Assistant Department of Linguistics and Department of Asian Languages and Cultures, UCLA

Publications and Presentations

Ying Lin. Learning Stochastic OT grammars: A Bayesian approach with Data Augmentation and Gibbs sampling. In *Proceedings of The 43rd Annual Meeting of the Association for Computational Linguistics, 2005.*

Ying Lin. Two perspectives on Malagasy reduplication: derivational and Optimality Theoretic analyses. To appear in *The 12th meeting of the Austronesian Formal Linguistics Association.*

Ying Lin. Learning phonetic features from waveforms. In *UCLA Working Papers in Phonetics*, Los Angeles, 2004.

Gregory M. Kobele, Jason Riggle, Travis Collier, Yoosook Lee, Ying Lin, Yuan Yao, Charles Taylor, and Edward P. Stabler. Grounding as learning. In S. Kirby, editor, *Proceedings of the Workshop on Language Evolution, ESSLLI '03*, Vienna, 2003.

Edward Stabler, Travis Collier, Gregory Kobele, Yoosook Lee, Ying Lin, Jason Riggle, Yuan Yuao, and Charles Taylor. The learning and emergence of mildly context-sensitive languages. In *Proceedings of the European Conference of Artificial Life*, 2003.

Learning Stochastic OT grammars, Ying Lin, *The 79th Annual Meeting of the Linguistic Society of America,* Jan. 2005

Learning segments from waveforms, Ying Lin, *The 138th Annual Meeting of the Acoustical Society of America*, Nov. 2004

Acoustic adaptation in frog calls from French Guyana, Yuan Yao and Ying Lin, *The 138th Annual Meeting of the Acoustical Society of America*, Nov. 2004

Unsupervised learning of broad phonetic classes using a statistical mixture model, Ying Lin, *The 137th Annual Meeting of the Acoustical Society of America*, May 2004

Abstract of the Dissertation

Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition

by

Ying Lin

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2005

Professor Patricia Keating, Co-chair

Professor Edward Stabler, Co-chair

Although the study of various language structures is essential to linguistics, the field owes the world the answer to a very fundamental question: *where do the structures come from?* For example, phonology – the theory of sound structures – does not offer an account of how children identify sounds from words they hear. However, synthesizing recent developments in two previously unrelated fields – infant speech perception and speech recognition – has made it possible to attack this problem for the first time.

This dissertation builds a computational model in an attempt to show how features and segments, the basic elements of phonological structure, could be learned from acoustic speech data. Within a statistical learning framework, phonological acquisition proceeds from a "holistic" representation of lexical items to a compositional one. Features are learned as partitions between natural classes of sounds, while segments are learned from waveforms through iterative learning of categories and hidden structures. Results from experiments conducted on audio

speech databases suggest that the model is able to group acoustically similar sounds with respect to natural classes, and that segment-sized units based on natural classes can arise from unsupervised learning of the acoustic data from isolated words.

Topics in this dissertation are arranged to introduce new components in an incremental manner: Chapter 2 introduces the main modeling tool – the statistical mixture model. Chapter 3 applies the mixture model to the task of segment clustering. Chapter 4 extends the basic clustering framework by adding the segmentation of words, and later phonotactics, into the learning framework. Chapter 5 considers the problem of learning with a lexical model and its implications. Chapter 6 summarizes the work presented, and discusses the implications of the model for empirical research and possible directions for future work.

CHAPTER 1

Introduction

1.1 The building blocks of language

A fundamental characteristic of the human language system is that it is *compositional*: we have an evident ability to construct an infinite variety of new utterances by reusing parts to build hierarchical structures. Such a hierarchical structure is usually conceived as consisting of many levels, each studied by a separate area of linguistics. At the very bottom level of the hierarchy, the objects of interest are the sounds of language that combine to form *phonological structures.* Traditionally, phonology studies how sounds combine to form meaningful units, while phonetics mainly studies the production, perception and acoustic properties of these sounds.

The assumption that a language's sound system is based on individual sounds, or segments, has had a long history in the descriptive work on a variety of languages around the world. Much of such work predated Saussure's *Premier Cours de Linguistique Générale* (1907) and formed the foundation of descriptive linguistics in the first half of twentieth century. As a present-day linguist undertakes the task of transcribing sounds of a language, the productivity of his work would depend on the phonetic alphabet[1] to be used in transcription, which in turn is informed by a segmental analysis of the language being studied. Therefore it is

[1] For example, the standard International Phonetic Alphabet.

not an exaggeration to say that the segment has been the *de facto* building block in linguistic description.

Since the work of Jakobson, Fant and Halle, (1952), linguistic analysis of phonological structure has been based on *distinctive features.* A binary distinctive feature distinguishes two natural classes of sounds. For example, [p], [t] and [k] share the feature [-voice], while [b], [d] and [g] share the feature [+voice]. Distinctive features are designed to be a system that would accommodate sounds in all human languages. Within this system, phonemes, or segments, are given a secondary status – they are regarded as “bundles” of features. Although not treated as fundamental, segments are indispensable for phonological analyses, since many important generalizations still depend on treating segments as basic units. In the past few decades, many research topics in generative phonology have grown from the assumption that features and segments are the basic units of the phonological grammar in a speaker’s mind.

Although perhaps few researchers would question the analytic value of features and segments, not everyone would accept that a speaker uses knowledge of features and segments with regard to her own language. This disagreement is often referred to as the debate on the “psychological reality” of features and segments. In the field of phonetics, it has been known for some time that there are no obvious boundaries between segments in connected speech (Pike, 1943; Firth, 1948) because articulators move continually from one target to another (Liberman et al., 1967). As commonly taught in introductory phonetics, features and segments are units of speech at an abstract level of description, and the mapping between these units and acoustic signals depends on individual languages. Such a division of labor allows phonology and phonetics to focus on their respective research topics, but also begs the apparent question:

Where do features and segments come from? (1.1)

As an attempt to address the interface between phonology and phonetics, there have been several attempts to relate abstract phonological categories to phonetic details of speech. Consider, for example, some examples of recent work on "phonetics-based phonology":

> *A vowel with F1=700Hz must be perceived as some category.*
> (Boersma, 1998)

> *An inventory of vowels must maximize contrasts.*
> (Flemming, 1995)

Phonetics-based phonology is appealing because of its claim that formal phonological analysis needs to be sensitive to phonetic details, but it does not differ from its predecessors in formal phonology with regard to one important point: all phonological analyses assume a prior stage of analysis, in which the waveform or the speech production process is temporally mapped to phonological units. Therefore neither the traditional phonology nor the more recent phonetics-based phonology have addressed question (1.1).

In a broader context, the issue of building blocks is not only relevant to generative grammar, but also to virtually all studies of language learning/acquisition. The reason is that by definition, any higher-level structure is formed by sequences of sound units. Take, for example, the study of infants' acquisition of phonotactics (Jusczyk, Luce, and Charles-Luce, 1994). The research question was formulated as whether infants are sensitive to sequential patterns of sounds in their ambient language. Needless to say, without deciding on a set of units, such "sequential patterns" will not be well-defined. As another example, computational models of word discovery from speech (Cartwright and Brent, 1994) invariably assumes the input data is coded by strings of phonetic symbols. A more sophisticated version of this type of approach appears in de Marcken's dissertation (1995), where he performs unsupervised lexical learning from audio recordings. Yet this experiment does not differ substantially from the previous ones, since a phoneme speech recognizer is used as a pre-processor to generate strings of phonetic symbols. Moreover, many other computational learning models have used a segmental encoding of speech as input data, including the induction of phonological rules (Gildea and Jurafsky, 1996), and the mapping between morphological paradigms (Albright and Hayes, 2003), among others.

1.2 The puzzle of early phonological acquisition

As discussed above, several streams of research, spanning a number of different fields, have gone quite far with the assumption that segmental representation is a basic level of linguistic description. For their purposes, such an assumption helps establish a proper representation of the data and helps them focus on their questions of interest. However, these works do not provide clues about how the segmental assumption itself may be replaced by more elementary principles, which is the focus of question (1.1) raised above.

Due to the complex nature of the problem, it is worthwhile to make the question more specific and consider two versions of (1.1). The first version involves the macroscopic *evolution* of language as a system of communication and may be phrased as follows:

What are the possible factors or mechanisms that would influence the *organization of the sound systems of human language?*

This question does not lie within the scope of this dissertation. What the current work is concerned with, instead, is evolution in a microscopic timescale – the period in which the baby becomes acquainted with the sound system of its native language. In particular, we are interested in the puzzle of *early phonological acquisition*. As will be discussed further in Section 1.5, I chose to focus on the stage of early phonological acquisition, because this is the period in which the child's sound system is shaped most dramatically by her environment. Therefore it may provide the most insights into the issue in which we are interested. From this narrower perspective, (1.1) can be rephrased as follows:

How do children arrive at the abstract features and segments, or more generally, the atomic units of phonological representation?

1.3 The importance of methodology

Of course, the above question stands out as a puzzle only if we assume that newborns do not yet know everything. Conceivably, one may argue that linguists need not be concerned with the question "where do the units come from", since features and segments may be part of the knowledge that every human language learner is equipped with from birth. Yet a brief reflection would suggest that this argument sounds implausible – each of us grows up in an ambient language, and we are aware that different languages' sound systems have different units, at least at the level of detail. Hence the puzzle cannot be simply dismissed as a non-question.

Taking one step back, it is also conceivable to consider shortcuts to the puzzle of early phonological acquisition. For example, if features are universal (Chomsky and Halle, 1968) and every language learner has access to all of them, perhaps the children's task is characterized as realizing which ones are distinctive and how they are implemented in the target language. If this is correct, then it is most valuable for children to collect minimal pairs of words (Jakobson, 1941). However, substantial numbers of minimal pairs come so late that it would be difficult for children to develop phonology until they have acquired a vocabulary of significant size (Luce, 1986; Charles-Luce and Luce, 1990). The assumption of a universal feature set simplifies the problem of learning, but at the cost that phonological acquisition must be delayed until enough words are learned. Since it seems unnatural to assume that young children to have no knowledge of phonology, such a shortcut also seems unsatisfactory.

The motivation for the current dissertation is that traditional linguistic analysis has not yet applied appropriate tools for resolving the puzzle of early phonological acquisition. In other words, it is helpful to be aware of other methodologies that may also contribute to our understanding of the puzzle, from neighboring fields of formal linguistics. For example, experimental methodologies for studying infants have developed significantly in the last 30 years, producing some very influential work in phonological acquisition. This work encourages a general interpretation that at least some aspects of early phonological representation are influenced by the learning experience. As I will discuss in Section 1.5, it is the empirical work on phonological acquisition that sets the stage for the current thesis.

The methodology of the current dissertation project is modeling. My aim is to build a computational model to show how early phonological representation can be learned from acoustic speech data. In other words, I try to understand the following question: how can children "crack the speech code," discover the smallest building blocks of language, and gradually put them into use when they learn new words? Such an effort will also give linguists a different way of answering questions like, "*Where do the symbols X, Y and Z come from in your theory?*" Such questions are common from people who are not familiar with the field, yet the methodology of theoretical linguistics has not been able to provide a direct answer.

Before describing our model, I will clarify the stance taken in the present thesis, and then turn to a review of recent findings in phonological acquisition that provide empirical support for the position taken.

1.4 Our focus on speech perception

Most studies of phonological development can be roughly divided into two categories: the first records what children say, the second designs experiments to test what they perceive. The relation between speech production and perception is a complex issue (Liberman et al., 1967; Fowler, 1986), as is their relation in the process of language development (Locke, 1983; Jusczyk, 1997). Although some would argue that phonological acquisition cannot succeed without a perception-production loop, the relation between speech production and perception is not within the scope of this thesis. Instead, the current dissertation chooses to focus on only one side of the story, by assuming that at least *some* understanding of phonological acquisition can be achieved by modeling the child's conception of adult language. This key assumption is a fairly standard one that has been taken by many others. As Hayes (2004) argues:

> *... it is also important to consider acquisition from another point of view, focusing on the child's internalized conception of the adult*

> *language. ... the limiting case is the existence of language-particular phonological knowledge in children who cannot say anything at all.*

In other words, it is of no interest for us to argue whether phonological acquisition can be achieved completely through speech perception alone, such as in the case of aphonic/cannulated children (Locke, 1983). Rather, the interesting lesson that we might learn is how one source of information – acoustic signals – can be exploited for the purpose of identifying features and segments.

The focus on perception distinguishes our work from those based on neural networks (Guenther, 1995; Plaut and Kello, 1999). This distinction reflects differences in modeling philosophy and will be discussed further at the end of the chapter. Because of our theoretical focus, the literature review below will predominantly cover studies of speech perception.

1.5 Past results on the development of child phonology

Linguists have long noticed that developing phonological structure is a major cognitive advance (Hockett, 1960). For example, Jackendoff (2002) remarked (p. 244):

> *[The innovation of phonological structure] requires us to think of the system of vocalizations as combinatorial, in that the concatenation of inherently meaningless phonological units leads to an intrinsically unlimited class of words ... It is a way of systematizing existing vocabulary items and being about to create new ones.*

However, the logical prerequisite for innovating phonological structure is establishing the atoms of the structure, a process that is far less obvious than often assumed. But since the pioneering work of Eimas et al. (1971), the field of developmental psychology has greatly advanced our understanding of this process. I will present a brief review of the main results that provide empirical ground for my model. For more details, the reader is referred to more comprehensive reviews such as Gerken (1994) and Kuhl (2004).

1.5.1 Learning sound categories

The first step in going from sound waves to phonological structures is learning sound categories. Much has been discovered about how this step is achieved by infants during the period between birth and 3-6 months. For example, exposure to an ambient language starts early (Mehler et al., 1988; Lecanuet and Granier-Deferre, 1993) and is quantitatively significant (Weijer, 2002). Young infants are found to distinguish different categories of sounds much like adults (Eimas et al., 1971). They even do better than adults at hearing distinctions in a foreign language (Lasky, Syldal-Lasky, and Klein, 1975; Streeter, 1976). But they gradually change from being a universal perceiver to a language-specific one, i.e. they lose the sensitivity to distinguish sounds that are not distinctive in the ambient language (Werker and Tees, 1984; Werker and Tees, 1992). It is widely assumed that the underlying mechanism is the ability to learn patterns from distributions of sounds (Kuhl, 1993; Maye, Werker, and Gerken, 2002).

1.5.2 Recognizing words

Aided by the development of other cognitive processes (Jusczyk et al., 1990; Tomasello, 2003), babies of 7-8 months exhibit an impressive ability to detect words (Jusczyk and Aslin, 1995; Jusczyk and Hohne, 1997). The nature of inputs to infants' word learning ability is still under debate, but a number of studies have attested to the great influence of single-word utterances (Ninio, 1993; Brent and Siskind, 2001; Weijer, 2001), short utterances (Ratner, 1996) and perceptually salient words (Peters, 1983) in utterances. The special prosodic structure of child-oriented speech (Fernald and Kuhl, 1987; Werker, Pegg, and McLeod, 1994) also seems to play a role. One of the mechanisms suggested for word segmentation is termed "statistical learning": infants are sensitive to patterns of sounds in the input language (Saffran, Aslin, and Newport, 1996). These patterns span many aspects of phonetics and phonology, such as: distinctions between native and non-native sounds (Jusczyk et al., 1993), phonotactics (Mattys and Jusczyk, 2001), coarticulation (Johnson and Jusczyk, 2001) and prosodic cues (Jusczyk, Cutler, and Redanz, 1993).

1.5.3 Early childhood: Lack of evidence for segmental structures

The general consensus in the field of child language is that early lexical representation/processing is not based on sub-lexical units, but on a set of under-analyzed word-sized atoms. In addition to calling it "holistic", researchers have used other terms like "phonological idioms" (Leopold, 1947), "frozen forms" (Ingram, 1979), "unanalyzed wholes" (Walley, 1993) and "speech formulas" (Peters, 1983). Arguments for this view come from many sources. For example, Jusczyk et al. (1993)'s experiment showed that after infants were familiarized with a CV stimuli set {[bo],[bi],[ba]}, they did not show more preference for a new stimulus [bu] and that for [du]. Similar conclusions have been drawn from other experimental studies, such as Nittrouer, Studdert-Kennedy and McGowan (1989). On the other hand, if we examine children's early productions, they are not organized with units smaller than words, but with ones larger than words (Vihman, 1996). Evidence of this kind led developmentalists to conclude that infants' discriminative ability is mostly based on the use of "global", rather than "local" information in the acoustic signals. This "holistic" view also helps explain other phenomena: compared to new-born infants' surprising ability to detect subtle differences, older children often failed to learn minimal pairs of nonsense words (Garnica, 1973; Stager and Werker, 1997) – a surprising result. In particular, Pater, Stager and Werker (2004) used minimal pairs that differ in two features, such as [pin] and [din]. They found that 14-month old infants failed to learn the minimal pairs of words in a task that is designed to test their word learning abilities, even though they were able to distinguish the same pairs of words in a discriminative task. In contrast, 18-month old infants appeared to be capable of both. Many have offered similar interpretations of this type of finding. For example, Anisfeld (1984) commented that infants are not responding to the stimuli "analytically". The characterization of the difference between child and adult phonology, expressed through the contrast between words like "holistic" and "global" versus "analytic" and "local", reflects a common belief that segmental structure is not inherent to child phonology, but develops over time.

1.5.4 Emergentist ideas and other competing candidates for sub-lexical units

Although many arguments have been made against sub-lexical structure in early development, experts generally agree that children do organize their lexicon with units smaller than words (Jusczyk, 1986; Gerken, 1994; Kuhl, 2004) at some point in their childhood (Gerken, Murphy, and Aslin, 1995). Moreover, this reorganization occurs before the age of reading instruction (Walley, 1993). Such a conclusion is natural to expect, since in order to reach the stage of "infinite uses of finite means", some type of composition must be present in their sound system.

The obvious gap between the aforementioned "holistic" hypothesis and compositionality has given rise to a trend of thought that treats the sub-lexical structure as an *emergent* phenomenon. For example, Lindblom (1992; 2000) considers the possibility of taking gestures (Browman and Goldstein, 1989) as the basic units of phonological organization, and argues that they can arise as a result of compression of gestural scores. Since Lindblom's work directly relates to the emergence of segments, but does not formulate and test a model, we will postpone the discussion of it until 6.3.

Units other than segments have also been considered as possibilities. For example, many researchers have suggested infants use syllable-like units in an early stage of phonological acquisition (Jusczyk, 1986). Yet they are reluctant to argue for syllable-based representation as a necessary intermediate stage, partly due to the difficult of defining syllables on the speech stream. An exception is MacNeilage(1998)'s association between syllables and cycles of mandibular movements and his idea that syllable-based "frames" form the first stage of phonological acquisition.

1.5.5 The lexicon and the development of phonological structure

The idea that the lexicon influences the development of phonological structure came from Ferguson and Farwell (1975) and has gained sympathy among a long list of researchers, e.g. (Charles-Luce and Luce, 1990; Fowler, 1991; Jusczyk, 1994; Metsala and Walley, 1998; Beckman and Edwards, 2000). For example,

Walley (1993) states:

> *The phonemic segment emerges first as an implicit, perceptual unit by virtue of vocabulary growth and only later as a more accessible, cognitive unit.*

The transition from a holistic to a compositional representation, i.e. the "whole-to-part" shift, has been thought to start at the time when a "burst" is observed in children's vocabulary growth (Ferguson and Farwell, 1975; Benedict, 1979; Menyuk, Menn, and Silber, 1986)[2]. The arguments made for a transition come from studies of babbling (Locke, 1983), segmental duration (Edwards, Beckman, and Munson, 2004), performance in gating experiments (Walley and Ancimer, 1990), similarity judgments, mispronunciation detection (Walley and Metsala, 1990) and sound-based word associations (Vihman, 1996). It was also observed that lexical restructuring is not an across-the-board change, but rather a gradual and word-specific process that lasts into middle childhood (Anisfeld, 1984; Fowler, 1991). Even for children in middle childhood, studies of spoken word recognition show that they still engage in more holistic lexical processing than older listeners (Walley, 1993).

There are several conceptual models that have been proposed for this transition in the literature. One was offered by Jusczyk (1986; 1992; 1993). In his *WRAPSA*[3] model, Jusczyk imagined that the lexicon is structured with many "neighborhoods". During development, the increasing density of a lexical neighborhood (Luce, 1986) causes a pressure towards the emergence of sub-lexical units. Details of the restructuring process may depend on the familiarity of the

[2] However, more recent studies showed that it is a more gradual process in general (Bates and Goodman, 1997).

[3] WRAPSA stands for "Word Recognition And Phonetic Structure Acquisition."

words (Metsala and Walley, 1998), lexical neighborhood densities (Garlock, Walley, and Metsala, 2001) and gradient phonotactics (Storkel, 2001). For example, it was suggested in the *Lexical Restructuring Model* (Metsala and Walley, 1998) that words in dense neighborhoods and those learned earlier are more prone to segmental restructuring.

As a consequence of phonological development, children's lexicons are structured in a way different from those of adults. For example, the beginning of a word is more important to adults (Marslen-Wilson, 1987), but children do not show this preference (Cole and Perfetti, 1980; Gerken, Murphy, and Aslin, 1995).

1.5.6 Segmental representation in adults

Despite the arguments against segments as the primary unit of speech (e.g. Browman and Goldstein, (1989)), few in the field would deny the status of segments as one level of phonological representation for adults. Evidence for the use of segments has been gathered through experimental studies of speech perception (Fletcher, 1953; Allen, 1994; Nearey, 1997; Nearey, 2001). For example, results reviewed by Allen (1994) show that the correct identification of nonsense CVC syllables in noise can be very accurately predicted from the marginal correct identification of their constituent phonemes, suggesting that the recognition of segment-sized parts leads to the identification of nonsense syllables.

In addition, support for segments also comes from studies of reading. Awareness of segments may arise as early as in toddlers (Menn, 1983) and keeps developing throughout childhood[4] (Fowler, 1991). Learning sound-to-letter mappings has been shown to be necessary for a reader of a phonemic orthography, and the

[4] The specific age at which children become aware of the segments is a topic of debate (Gerken, Murphy, and Aslin, 1995).

young pre-reader's phonological awareness is a predictor of their later reading ability (Liberman et al., 1977). The effect of phonology in reading has led to proposals that combine orthographic with segmental representations (Harm and Seidenberg, 1999). Although much caution should be taken in setting phonological awareness equal to the existence of segments, these studies provided some indirect support for segments as units of speech, at least for adults.

1.5.7 Summary: A whole-to-part process

The above sections presented a review of relevant studies on how phonological representation develops in children, along the timeline from birth to later in childhood. Summarizing others' views, there has been a growing consensus in the past 30 years that phonological structures are learned gradually through development. Conceptual models have also appeared to characterize current thinking in the field: at least in the early stages, the discovery of phonological structure is seen as a *consequence*, instead of a *prerequisite*, of word learning. Although the arrow of causality is also thought to go the other way[5], it is the "whole-to-part" direction that has drawn significant attention in the literature.

However, none of the proposals are specific enough to be quantitative, and this results in a lack of a clear understanding of early phonological acquisition. To take one example, the learning of phonetic categories from waveforms is often ascribed to some "general pattern-finding" abilities. Yet in the literature, nothing specific enough has been said so that we can actually see what kind of "pattern-finding" can help children identify atoms of their language. As another example, the child's lexicon is often conceived as various "lexical neighborhoods" formed by words that are "similar" to each other. But without explaining in what space

[5] For example, Jusczyk (1997) suggests that discovery of segmental structure also leads to a dramatic expansion of vocabulary size.

words are distributed, and what makes words similar, our understanding of the lexical restructuring process will stay at an intuitive level.

Due to the difficulty of assessing human behavior, especially in infants, it is unlikely that experimental studies alone can produce answers to the questions raised above. Hence I intend to bring in a different perspective – computational modeling. But unlike previous models, here one cannot assume speech is represented as strings of symbols. Instead, this is the research question – one must directly deal with speech signals to show how strings of symbols can be derived. Building such a model is not a trivial task. It turns out that a statistical framework, supplemented with tools from speech engineering, is most appropriate for the current purpose. The rest of the chapter describes such a framework.

1.6 Statistical learning of phonological structures: a Pattern-Theoretic approach

The pattern finding skill mentioned above has gradually been known as "statistical learning". As noted in 1.5.1 and 1.5.2, many psycholinguists consider statistical learning a supporting mechanism for phonological acquisition. But in the world of mathematical sciences, the word "statistical learning" also stands for a booming subject that has received much attention and produced its own conferences and standard texts (Mitchell, 1997; Vapnik, 2000; Duda, Hart, and Stork, 2000; Hastie, Tibshirani, and Friedman, 2002). However, not much effort has been spent on connecting these two directions of research. The main reason why this is not yet happening is a major difference in research goals. The technical work on statistical learning has been focused on pattern recognition, i.e., machines that will map patterns to pre-determined outputs (e.g. $\{0, 1\}$-values in classification or probabilities in inference). Perhaps the best example of pattern recognition is the engineering work on speech; speech recognition has made much progress in the 80's and 90's, but most of the work on sub-word units has been directed towards improving recognition accuracy[6]. Nevertheless, the scientific interest in human language learning has always been on the *structures of patterns*. Although many tools in the current models are borrowed from speech recognition, the interest of the present study is not in engineering, but in identifying phonological structures from patterns of acoustic signals.

Before discussing how the current model utilizes statistical learning, the first question I would like to clarify is: "How should statistics be used to answer our question?" This question has a more general version regarding the role of statistics in cognitive modeling. In recent years, a response has been provided by applied mathematicians working on Pattern Theory (Grenander, 1996; Mumford, 2002), influencing research in computational modeling of vision. According to Grenander (1996), the goal of Pattern Theory is to:

> *... create a theory of patterns that can be used systematically to codify knowledge at least in many situations. This is similar to knowledge engineering in artificial intelligence, where computer scientists try to create representations that are possible to exploit by implementing them on the computer. Our approach differs in that it will be based on an algebraic/probabilistic foundation that has been developed for exactly this purpose. The algebraic component will express rules or regularity, the probabilistic one the variability, thus combining the two opposing themes mentioned above.*

[6] Most commonly, the sub-word units are context-dependent segments (e.g. triphones), while the effort has been spent on the tradeoff between a large number of units and a limited amount of training data.

This is exactly the approach we want for the phonological acquisition task. The algebraic component in our model is the phonological representations of lexical items, and the probabilistic component deals with the mapping between representation and acoustic signals. Given this setting, I would also like to state explicitly what is meant by "learning phonological structure": choosing a symbolic representation according to probabilistic criteria. Assumptions about the nature of the representation must be made before we know from what we are choosing. However, the choice is not based on considerations of whether this will lead to a theoretically insightful explanation of phonological systems. Instead, it is determined by probability functions whose values depend on data from the environment. The learning of phonological structures will be formalized mathematically as optimization problems, and the testing of the model will be performed with several computationally intensive algorithms.

Although the pattern-theoretic approach intends to model certain aspects of language acquisition, it is not our intention to claim that the equations and algorithms are actually what is computed by the brain. To clarify this point, it is illuminating to consider Marr (1982)'s three levels of understanding complex information processing systems: a computational level, an algorithmic level and an implementation level. Unlike in other types of models such as neural networks, it is the computational level that the current thesis intends to investigate.

Four assumptions serve as cornerstones of the model. First, I assume that a word-learning process gives infants enough exposure (training data) for learning "holistic" (or acoustic) representations of words or phrases, justified by the empirical findings reviewed in 1.5.2. One could also start without this assumption, but would then have the additional problem of dealing with learning words[7]. Since

[7] As mentioned in Section 1.1, such attempts have been made in unsupervised lexical acqui-

other cognitive processes also facilitate word learning (Jusczyk, 1997; Tomasello, 2003) and we are mainly interested in phonological structure, assuming that some holistic representations are given is a necessary simplification that preserves the main characteristics of the problem.

The second assumption is that children are able to detect acoustic changes within the signals they hear, and group piece-wise similar acoustic signals as sound categories. Although we do not know whether in fact it is a general skill to learn sound categories from signals and treat signal streams as sequences of categories (Kuhl, 2004), such a skill is the bias that is built into our model that restricts the space of phonological representations.

The third assumption is that the representation learning problem is solved by *iterative learning*. As also noted in work on grammar learning (Tesar and Smolensky, 1996), the problem of learning *hidden structure* is pervasive in language learning. For example, there is a fairly good understanding about how infants acquire vowel categories from isolated stimuli (Kuhl, 1993; Lacerda, 1998; Maye, Werker, and Gerken, 2002), but not much is known about how infants acquire vowels from speech streams and further acquire phonotactics (Jusczyk and Aslin, 1995). The missing link in the picture is due to the fact that segmental structure is "hidden": there are few breaks in the acoustic signal.

The current solution conceives an iterative task: once it is known how words can be broken up into sounds, the sound categories can be improved; conversely, a better grasp of sound categories in turn leads to better segmentation and improves knowledge of phonotactics. This view suggests a different perspective in the debate about whether category learning precedes or follows word learning (Werker, 2003). In my opinion, segmentation is better seen not as a *task*, but

sition models like (Brent and Cartwright, 1996; de Marcken, 1996).

as a *computational process* that is actively engaged in phonological acquisition. Similar views are also presented by Edwards, Beckman and Munson (2004):

> *The relationship between knowledge of the phonological grammar and processing of phonological patterns is a symbiotic one. Knowledge feeds on processing, and processing feeds on knowledge.*

In terms of computation, the iterative learning strategy used in the current study has its roots in statistics (Baum, 1972; Dempster, Laird, and Rubin, 1977; Tanner and Wong, 1987) and is in essence very similar to one that has been applied successfully in speech recognition (Rabiner and Juang, 1993) and computer vision (Tu and Zhu, 2002). In a rather different tradition, another variant of such a strategy can also be seen in speech models based on neural networks, such as the use of feedback in network training (Guenther, 1995).

The last assumption is that symbolic structures need to be augmented with exemplars to allow enough flexibility in modelling speech. Exemplar-based models (Johnson, 1997b; Pierrehumbert, 2001; Pierrehumbert, 2003) are based on the intuition that symbolic representations need to be enriched with details from real-world signals. These details often create ambiguities in mapping the signals to symbolic structures. For example, a given speech sound may be ambiguous between two categories (Hillenbrand et al., 1995), and a given word may be segmented in several ways (Glass and Zue, 1988). This intuition is also shared in the current study, but embedded in a different formalization – statistical mixture models.

A mixture model treats the variation among data as generated by several independent sources, and has a long history in statistics (Wu, 1996). Compared to exemplar-based models (Johnson, 1997b), mixture models can be seen as adding a level of generalization to simply storing large numbers of exemplars[8]. Specifically, we use a mixture model at the level of phonological categories, and another mixture model for multiple segmentations of each utterance. Learning problems with mixture models are solved with iterative algorithms, thus addressing the issue of how symbolic and probabilistic components interact during learning. As a result of iterative learning, different representations for the same lexical entry also form a natural mixture model, thereby reconciling traditional symbolic views and recent "episodic" views of the lexicon (Goldinger, 1997).

1.7 Overview of the model

Based on the assumptions outlined above, I give a brief overview of the project undertaken in the current dissertation. The model takes acoustic speech data for utterances as input, and outputs phonological representations as well as phonotactic knowledge derived from the lexicon. For illustrative purposes, four components are identified in order to highlight their connections with previous work in the literature: *segmentation*, *phonological development*, *lexicon building*, and *lexical restructuring*. Most of the computation is carried out between the first three components, which will be elaborated on in Chapter 3, 4 and 5, respectively. The relationship between those components is shown in Figure 1.1:

[8] Raw acoustical traces, for example. The difficulty in taking an approach that lets the data "speak for themselves" comes from a fundamental problem in non-parametric statistics, which will be discussed further in 6.3.

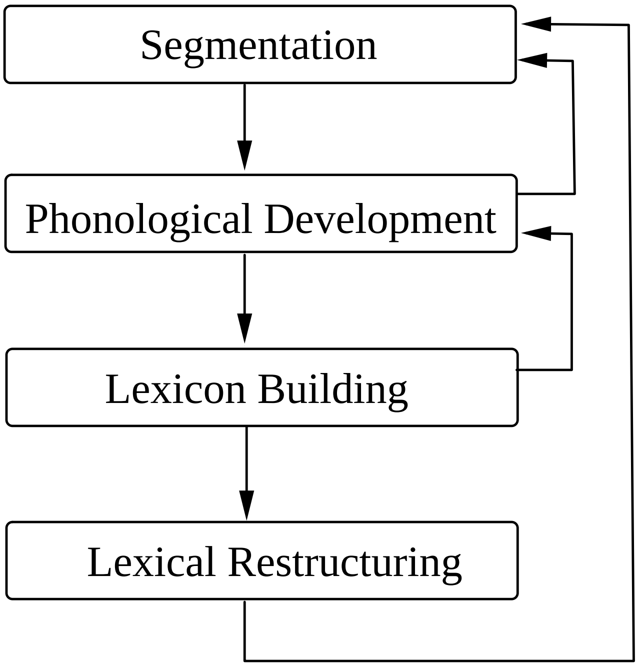

Figure 1.1: Structure of the model

1.7.1 Segmentation

Segmentation refers to the process by which acoustic signals and lexical representations are used to obtain a set of instances for each phonological unit. We will be using two kinds of segmentations. *Acoustic segmentation* is based on discontinuities in acoustic signals, or "landmarks" (Stevens, 1998). Algorithms for acoustic segmentation have appeared in the speech recognition literature (Svendson and Soong, 1987; Glass, 1988; Bacchiani, 1999). With dynamic programming, these algorithms try to identify piece-wise stationary regions in the signal, thereby locating all the possible segment boundaries. In the current model, we will use acoustic segmentation to produce instances of initial categories and provide a starting point for iterative learning.

In *Model-based segmentation*, the goal is to map different parts of the acoustic signal to a sequence of unit models. There are two prerequisites for such a mapping: First, we need a hypothesis space to search for the segmentation. Second, we also need a criterion for optimality in order to decide which segmentations are optimal. With these prerequisites, the procedure for finding the segmentation is posed as a statistical inference problem: inferring the value of the hidden variables from the data and the model parameters. Since we use Hidden Markov Models (HMM) as models of categories, each segmentation can be obtained using *Viterbi decoding* (Rabiner, 1989), within a search space encoded a finite-state network. Since phonotactic and lexical knowledge are encoded in such a space, they will influence the result of segmentation.

1.7.2 Phonological development

Given an exemplar-based view of phonological structures, phonological knowledge is encoded as parameters in the mixture model. Phonological development refers to updating these parameters, i.e. the learning of phonological categories and sequential patterns in lexical representations. Technically, the problem of learning categories is often referred to as *clustering* (Linde, Buzo, and Gray, 1980; McLachlan and Basford, 1988). The speech recognition literature that is most relevant to the current project is the work on *Acoustic Sub-Word Units* (ASWU) (Svendsen et al., 1989; Lee et al., 1989; Bacchiani and Ostendorf, 1999; Riccardi, 2000; Singh, Raj, and Stern, 2002). However, since the original purpose of ASWU is to optimize recognition performance rather than derive meaningful units, I implement the clustering step in ASWU systems using a mixture of HMMs, which was previously applied in clustering problems in other domains (Li and Biswas, 2002; Alon et al., 2003).

When the segmental boundaries are given, features are discovered in a sequence by repeatedly partitioning the clusters, starting from gross distinctions such as the one between obstruents and sonorants. Subtler distinctions are discovered later and lead to finer natural classes of sounds, along lines compatible with previous proposals (Pierrehumbert, 2003). The current model of feature discovery assumes the definition of feature to be distinctions between natural classes (Ladefoged, 2001) and shares the view that features are language-specific rather than universal (Kaisse, 2000; Mielke, 2004). When the segmental boundaries are not given, our model is applied in the task of learning from waveforms. In the iterative learning procedure, the knowledge of categories and phonotactics benefits, and is benefited by, inferring hidden structures. The algorithm based on HMMs is in essence similar to *Viterbi training* (Jelinek, 1976; Juang and Rabiner, 1990) and converges to a local maximum in the data likelihood.

1.7.3 Lexicon building

The task of lexicon building refers to using the representations obtained from phonological development to build an input/receptive lexicon, in the sense of (Menn, 1992). Each entry of the lexicon may consist of multiple exemplars with weights, and each lexical exemplar is a sequence of phonological category models. The lexicon is the crucial source of information for accumulating knowledge of phonotactics, indicated by a small loop in Figure 1.1. But in implementation, the lexicon is also used in the segmentation steps and actively participates in the iterative learning process.

This step is similar to the lexicon building task in ASWU systems (Paliwal, 1990) in that the mixture of lexical exemplars can be seen as a way of doing pronunciation modeling (Bacchiani, 1999). But there is a major difference between the current model and the work in speech recognition: instead of deriving a large number of data-driven units that do not have any linguistic significance (Singh, Raj, and Stern, 2000), we focus on a small number of units that intend to capture broad natural classes. The assumption that the early lexicon may consist of broad categories has appeared in other places, for example, Menn (1992), Lacerda and Sundberg (2004). Along an unrelated line of research, consequences of using broad classes in lexical representation have also been explored in speech recognition research, such as in studies of large lexicons based on broad classes (Shipman and Zue, 1982; Carter, 1987; Vernooij, Bloothooft, and van Holsteijn, 1989; Cole and Hou, 1988).

1.7.4 Lexical restructuring

A common problem in mathematical modeling is deciding the number of parameters. This is the problem of *model selection*. Given a restricted set of representations, we need a model selection criterion and a search strategy in the hypothesis model space. In our case, the dimensionality of the model depends on the number of unit categories in the phonetic inventory, since lexical representation and phonotactics are generated as solutions to the optimization problem once the phonetic inventory is fixed. Ideally, the model selection criterion and search strategy should be motivated by the insights from child language studies. In particular, the "lexical restructuring" proposal (Metsala and Walley, 1998) gives hints on how a model should be selected: the lexicon as a whole determines what phonological units should be used in lexical representation. Central to such an effort is a formulation of lexical distance and neighborhoods. In accordance with the view of the lexicon as consisting of exemplars, two ways of quantitatively assessing lexical distance will be discussed: one is an extension of simple edit-distance to the mixture model; the other is *Kullback-Leibler divergence*, a concept from

Information Theory (Cover and Thomas, 1991) intended to quantify the global distances between the acoustic models. Though a full-fledged demonstration of lexical restructuring lies beyond the scope of this dissertation, some preliminary steps will be outlined towards a formal characterization of the lexicon-driven idea.

In summary, the model describes the following picture for the acquisition of sub-lexical structures: children learn categories of sounds from speech streams, in an iterative manner. The result of such learning is a set of atomic probabilistic models, as well as symbolic representation of lexical items. As the number of lexical entries increases, distinctions between lexical items gradually diminish, thereby favoring a refined inventory of categories and corresponding lexical representations. Although identifying phonological units and forming lexical representations are often thought of as two separate problems, a coherent story becomes possible only by treating them *together* in a joint modeling framework.

1.8 Organization of the dissertation

Topics in this dissertation are arranged in an incremental manner. Chapter 2 introduces the main modeling tool – the statistical mixture model. Chapter 3 applies the mixture model to the task of segment clustering. Chapter 4 extends the basic clustering framework by adding the segmentation of words, and later phonotactics, into the learning framework. Chapter 5 considers the problem of learning with a lexical model. The implications of the lexical model for several topics will also be discussed, such as lexical distance, lexical neighborhood, model refinement and model selection. Chapter 6 summarizes the work presented, and discusses the implications and possible directions for future work. Materials in each chapter are roughly divided as follows: the main text focuses on definition of the problem, general discussion, and in some cases algorithms; while the technical details, such as the equations and occasional proofs, are deferred to the appendix.

CHAPTER 2

Mixture models and their application to clustering

The main question that we would like to address in this chapter is the following: how can phonological categories be learned from given instances of speech sounds? In technical contexts, the same question arises under the name of *clustering*: grouping data into a few subsets based on some measure of affinity or similarity. Although many clustering methods exist in the literature, the basic method used in the present study is a parametric approach using mixture models. In other words, we assume that the data is generated by a model, whereas the task of learning is characterized as finding parameters of the model that achieve a reasonably good fit to the data.

2.1 Mixture models

Mixture models have a long history in statistics and are often applied to situations where the set of data being studied consists of multiple subsets, but the subset memberships of data samples are not observable. For example, all the instances of vowels in a language are a collection of instances of vowel phonemes. But in a spoken language, the labels for each vowel are not available for the language learner. The task of the learner, however, is to acquire knowledge of the phonemes as well as infer the category memberships of these sound instances.

In this thesis, we are only concerned with *finite mixture models*, where the number of sub-populations is known and fixed. The degree to which the model (denoted by θ hereafter) predicts a given set of data (denoted by y) is characterized by a probability function, called the *likelihood* function. Given a finite mixture model, the likelihood function $p(y|\theta)$[1] can be written in the following form [2]:

$$p(y|\theta) = \lambda_1(\theta)f_1(y|\theta) + \lambda_2(\theta)f_2(y|\theta_2) + \cdots + \lambda_M(\theta)f_M(y|\theta) \tag{2.1}$$

θ denotes the collection of unknown parameters in the model. The number of mixture components is given by M, a fixed integer. The distribution functions $f_i(y|\theta), i = 1, \cdots, M$ each characterize a component of the mixture. They are usually chosen from some common families of distributions. When $f_i, i = 1, \cdots, M$ are chosen from the same family, an alternative notation for the mixture distribution is often written as $p(y|\theta_i), i = 1, \cdots, M$, where $\theta = (\theta_1, \cdots, \theta_M)$, i.e. a vector form representing the model parameters as consisting of several subsets of parameters. The probability that a sample is drawn from component m is $\lambda_m(\theta)$, which depends on θ and is also subject to the constraint $\sum_m \lambda_m(\theta) = 1$. Alternatively, one may view λ_m as a *prior probability*, which determines how much of the data is accounted for by the m-th component in the mixture. As an example of finite mixture models, consider the stimuli used in the Voice Onset Time (VOT) training/perception experiment in Maye and Gerken (2000):

[1] Read as "the probability of y given θ".

[2] The notation follows Wu (1996).

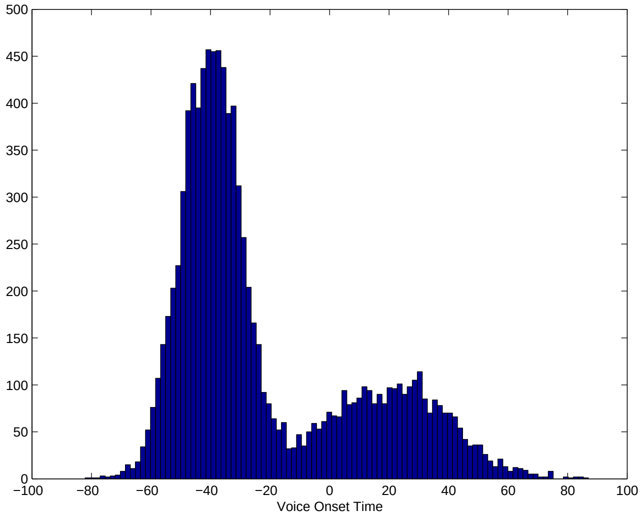

Figure 2.1: Histogram of the VOT stimuli similar to those used in Maye and Gerken (2000)

These VOT samples are generated from a 2-normal mixture that has the distribution function:

$$p(y|\theta) = \lambda f_1(y|\theta) + (1-\lambda) f_2(y|\theta) \tag{2.2}$$

λ is the proportion of samples from the sub-population modelled by f_1, $1-\lambda$ is the proportion of samples from the sub-population modelled by f_2. The mixture components f_1, f_2 are two normal distributions:

$$f_i(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y-\mu_i)^2}{2\sigma_i^2}\right), i = 1,2 \tag{2.3}$$

$\mu_1, \mu_2\, \sigma_1^2, \sigma_2^2$ are the parameters of the two normal distributions, and the unknown parameter $\theta = (\lambda, \mu_1, \mu_2\, \sigma_1^2, \sigma_2^2)$.

A mixture model is considered a *generative* model in statistics because one takes a parametric approach to the modeling problem – the data are assumed to be explained by some underlying causes – in this case, the mixture components. As an example, the generation of data from the above model, as illustrated by the histogram in Fig. 2.1, can be done by repeating the following two steps:

1. Flip a coin with heads-on probability λ;

2. If the outcome of the coin-toss is heads, then generate a random number from the distribution f_1, otherwise generate a random number from the distribution f_2.

In other words, there are two sources of uncertainty: one is the choice of probability distributions; the other is the probability distribution itself. Therefore going from the model to the data involves two rounds of random number generation.

2.2 The learning strategy for a finite mixture model

In statistical modeling, the word "learning" has the particular meaning of optimization or model fitting. A common criterion for fitting a probabilistic model, including the mixture model, is the maximum likelihood estimate (MLE). Given a set of samples $y = \{y_1, \cdots, y_N\}$, the estimate of θ is found by maximizing the log-likelihood (or equivalently, the likelihood) function over θ:

$$\sum_{i=1}^{N} \log p(y_i|\theta) = \sum_{i=1}^{N} \log \left[\sum_{m=1}^{M} \lambda_m(\theta) f_m(y_i|\theta) \right] \tag{2.4}$$

Here the summation sign follows from the standard assumption that each datum is generated from the same distribution independently. Hence the question of learning is posed as an optimization problem:

> Can we find values of θ, such that the value of the function (2.4) is maximized?

The standard constrained optimization technique, such as the method of *Lagrange Multipliers*, would require differentiating the objective function with regard to all the unknown parameters and solving a system of equations. However, this can be difficult for a function of the form (2.4): the summation sign inside the log is difficult to handle with conventional calculus. Intuitively, this difficulty arises precisely because of the nature of the learning problem: in order to learn each f_i, we would like to know which subsets of data belong to each f_i, or equivalently, the *membership* of each datum as belonging to each subset. Once the subset membership information is known, as given in supervised learning problems, then the problem of model-fitting is reduced to a standard type. However, as we start thinking about where to obtain such information, there is a dilemma of circularity: the only plausible way of inferring the membership information is by using a model – we compare the goodness-of-fit of each component (the likelihood), and then decide how likely a datum is to fall into each subset (the posterior). Clearly, some learning strategy is needed to guide us out of such a dilemma. One of these strategies is the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

The key idea behind the EM algorithm is to treat a subset of the variables as *missing data*. As an illustration of this idea, consider our example of learning 2 categories from the VOT data. The difficulty of maximizing (2.4) lies in the fact that only the VOT values $y = \{y_1, \cdots, y_N\}$ are observed: we do not see a straightforward way of doing direct maximization. But as we have commented above, it is helpful to imagine a somewhat idealized situation, one in which the category membership (e.g. voiced/voiceless) of each VOT sample is available. In other words, the data is presented in the form $\{(y_1, z_1), \cdots, (y_N, z_N)\}$, where y_i is the original data, and z_i is an indicator function that tells us which normal distribution y_i comes from. For purposes of computation, we would like to choose a particular coding of this indicator function $z_i = (z_i^1, z_i^2, \cdots, z_i^M)$, where exactly one among $z_i^m, m = 1, \cdots, M$ is 1, and the rest is 0. The advantage of adopting this coding is that the log-likelihood function of the complete data model has a simple linear form:

$$\sum_{i=1}^{N} \log p\left((y_i, z_i)|\theta\right) \quad = \quad \sum_{i=1}^{N} \sum_{m=1}^{M} z_i^m \log\left[\lambda_m(\theta) f_m(y_i|\theta)\right] \tag{2.5}$$

Here we have used the fact that z_i is an M-component indicator function, and $f_m(y_i|\theta)$ is the m-th component that generates y_i, according to the indicator z_i. Notice if $\{z_i\}$ are known, maximizing (2.5) over θ is a simpler matter, since there is no more summation inside the log, and only one term in $z_i^m, m = 1, \cdots, M$ is 1 and the rest are 0. In our VOT example, the maximum likelihood estimates of each normal distribution can be obtained in 2 simple steps:

1. Sample classification: divide all samples into 2 subsets according to the indicators $\{z_i\}$.

2. Subset fitting: fitting each mixture component to a subset of data according to maximum-likehood criteria.

The obvious difference between such an idealized situation and the original situation is the presence of $\{z_i\}$. Although the indicators z_i are not observable in reality, the important point is that they are random variables that observe a multi-nomial distribution specified by θ:

$$(z_i|\theta) \quad \sim \quad Multinomial(\lambda_1(\theta), \cdots, \lambda_M(\theta)) \tag{2.6}$$

One may think of z_i as outcomes of the coin-toss in the first step of generation described in the previous section. Once the outcome of z_i is known, the sample is generated from the component that the indicator z_i points to:

$$(y_i, z_i^m = 1|\theta) \quad \sim \quad f_m(y|\theta) \tag{2.7}$$

It can be verified that the density function derived from (2.6) and (2.7) is the same as (2.1) once we sum over all the z_i. Following the intuition outlined above, (2.1) is called the *observed data* model, (2.7) is called the *complete data* model, and (2.6) is referred to as *missing data.* The observed-data model is obtained from the complete-data model by integrating out/summing over the missing data. It is worth pointing out that the concept of missing data is proposed from the perspective used by model designers to facilitate computation. Therefore it should not be confused with data in the physical world from the perspective of language learners.

Although the concept of missing data helps clarify the structure of the problem, it is still not quite clear what strategy can take advantage of such a structure. In the above example, the "missing" $\{z_i\}$ clearly will help learning once they are known, yet they themselves depend on the learned model. Moreover, the original goal is fitting the observed-data model instead of the complete-data model.

Before a mathematically correct formulation appeared in Dempster, Laird and Rubin, (1977), the strategy many adopted was the following: if there is not enough information for us to choose a best hypothesis, we make a reasonable guess about the missing information based on the available information and our current hypothesis; once we have the observed information augmented by the inferred missing information, we can improve our hypothesis. When the hypothesis does not change as we repeat these two steps, it is the best hypothesis obtainable from this procedure.

Obviously, not all guesses will work, and one would expect a careful justification for what counts as a "reasonable guess". One way of making such guesses is based on a statistical principle – consider all different ways of filling in the missing data, and take the *average*. The proper meaning of "average", as made clear in Dempster, Laird and Rubin, (1977), is the *conditional expectation* of (2.5) given the observed data and the current estimate of the parameter. The two steps – one calculating the conditional expectation, the other updating the model parameters – were termed *Expectation* and *Maximization* respectively.

Put more formally, the EM algorithm tries to maximize the conditional expectation of (2.5) over θ in an iterative manner. First, it starts with an initial value of $\theta^{(0)}$. Then in the E step, it computes the conditional expectation of (2.5) using the observed data $\{y_i\}$ and the old parameter $\theta^{(t)}$; in the M step, it maximizes this expectation by finding a new parameter $\theta^{(t+1)}$. More succinctly:

$$z_i \leftarrow E(Z_i|y_i, \theta^t) \tag{2.8}$$

$$\theta \leftarrow \arg\max_{\theta} p\left(\{(y_i, z_i)\}|\theta\right) \tag{2.9}$$

To illustrate the details, an example using the mixture of two-Gaussians was included in Appendix 2.B. It was shown in Dempster, Laird, and Rubin, (1977) that each step of EM increases the original log-likelihood function (2.4) and converges to a local maximum. The argument from this paper is outlined in Appendix 2.A. This crucial result justifies the iterative strategy and supports the widespread use of EM in many different fields.

2.3 Connections with an exemplar-based model

The basic intuition underlying mixture modeling is that variation is generated by non-interacting independent sources, while observation is made from these sources one at a time. It is worth noting that a recent trend in linguistic thinking, the so-called *exemplar-based* models, which has its origin in psychology (Nosofsky, 1991), is also based on very similar intuition. In fact, the mixture model can be viewed as a formalization of the exemplar-based idea, and this connection can help us understand the learning algorithm for the mixture models from a non-statistical perspective. For example, an influential exemplar-based model (Johnson, 1997a) sets up the following learning scheme:

- All training tokens are represented in a low-dimensional space.
- The class membership of a new datum is determined jointly by all previously stored tokens as a weighted sum.
- The weights are adjusted when new tokens are added to the model.

The above implementation has a flavor of the EM algorithm, and this similarity may help us understand the rationale of the EM algorithm for mixture models: the E-step can be viewed as determining the membership of a new exemplar with regards to each class using the pre-stored exemplars, while the M-step can be viewed as shifting the centers of the exemplar ensemble by updating the contribution of each exemplar[3].

The main difference between our model and exemplar-based models has two analogs in other fields which are worth noting. The first analog is the difference

[3] Notice the use of intuitive concepts such as "center" reveals a connection with the parametric approach used in the current work.

between template-based and statistical speech recognition: one approach stores all exemplars and uses some template-matching techniques (for example, based on Dynamic Time-Warping) to determine similarity, the other approach uses models (for example, hidden Markov models, also see below) to generate the speech data, and uses likelihood to measure similarity. The second analog is the difference between non-parametric and parametric frameworks of statistics, in particular the estimation of probability distributions. More discussions of the mixture and exemplar models, as well their applications to the lexicon, will ensue in Chapters 5 and 6.

2.4 Choosing components of the mixture model

Although the conceptual framework of a mixture model is quite general, the choice of the components for the mixture model is very much dependent on the modeling domain. Since we are interested in applying the mixture model to acoustic speech signals, the component models must be flexible enough to deal with speech. Needless to say, finding good statistical models for speech sounds is a difficult matter. It has been a serious challenge for researchers from many disciplines. In the field of speech recognition, the core technology has more less remained the same for the last twenty years. Although there has not been significant interest in language acquisition in engineering research as a whole, the fact that both fields are dealing with the same kind of data – acoustic speech – suggests the possibility of adapting speech recognition tools for the purpose of modeling language acquisition.

Below, we will discuss a few fundamental characteristics of speech signals. The discussion is not intended to be comprehensive, but will rather focus on how these characteristics constrain the possible choices of the model, within the background of speech recognition.

2.4.1 High dimensionality of speech signals – front-end parameters

The first challenge is the dimensionality of a space that is appropriate for describing speech sounds. Recall that in the /t/-/d/ example used in this chapter, we have assumed that the data is taken from a 1-dimensional space – the Voice Onset Time. However, this is very much a simplifying assumption, as one may realize after inspecting the spectrogram of sounds from the continuum between /d/ and /t/. While low-dimensional phonetic representations have been identified for subsets of speech sounds, such as vowels, it is still an open problem whether a low-dimensional phonetic representation can be found in general. Although theoretically solid auditory representations for speech are yet to be established, many attempts have been made towards developing low-dimensional representations for applications, especially for speech recognition purposes. Such efforts mostly lie within the domain of *front-end* research – i.e. different ways of parameterizing acoustic signals for later processing. One such example is the *Mel-Frequency Cepstral Coefficients* (MFCCs), a popular choice for front-end parameters in many contemporary speech recognition systems. The procedure for computing MFCCs consists of the following steps:

1. (Time domain) Pre-emphasis, framing, windowing;
2. (Linear spectral domain) Compute the power-spectrum of the windowed short-time signal through a discrete Fourier transform;
3. (Mel spectral domain) Apply the Mel filter bank to weight the DFT coefficients, and obtain a non-linear down-sampling of the power spectrum;

4. (Cepstral domain) Take the log of the Mel spectrum, and apply a discrete cosine transform.

5. (Cepstral domain) Apply *liftering*[4] to the mel cepstrum.

After the last step, the first dozen Fourier coefficients obtained from the cosine transform are kept as MFCCs. Among these steps, the Mel filter bank is perceptually motivated by the non-linearity in the frequency resolution of the auditory system, while the transformation to the cepstral domain is more or less motivated by practical purposes[5]. As the result of signal processing and transformation, an MFCC vector is obtained for each sequence of windowed short-time speech signal, and the speech stream is represented as a matrix. The relatively low-dimensional, de-correlated MFCC vectors produce an approximate encoding of the spectral information in the original speech, and sacrifice the voice information. Roughly speaking, in terms of the /d/-/t/ distinction, while humans are sensitive to the voice onset time, MFCC places more emphasis on the differences in their power spectra than the differences in voicing.

Many other different choices of front-end processing have been proposed, including some that are designed carefully to match the results of physiological studies (Seneff, 1985; Seneff, 1986). However, when it comes to incorporating auditorily-motivated front-ends into the speech recognition system, factors that affect the system performance often dictate the choice of front-end processing, such as computational efficiency, performance under different conditions, etc. For our purposes, we are most interested in the kind of auditory representations for young children. Since such studies are difficult to conduct and the relevant data are scarce, MFCC will be applied throughout the current study as the front-end

[4] i.e. tapering the high-order coefficients

[5] For example, robustness under signal degradations, speaker independence, etc.

representation of speech, with the number of channels in the filter bank and the number of cepstral coefficients fixed to be 26 and 13, respectively[6]. As pointed out above, such a choice is not unproblematic, but since we are primarily interested in whether broad phonetic categories can be learned from the spectral information, and MFCC does encode such information, MFCC may serve as a reasonable starting point for studies like ours.

2.4.2 The dynamic nature of speech sounds – HMM and dynamic parameters

A second challenge is that speech sounds are naturally dynamic signals. This point has long been noticed by speech scientists (Stevens, 1971). Most obviously, the duration of a speech sound is highly variable and subject to many conditions, such as speech rate, speaking style, prosody, etc. Second, each speech sound may contain multiple points of acoustic change that all carry information for the sound (Stevens and Blumstein, 1978). Third, compared to the traditional view of speech perception as being based on static targets, more recent work has suggested the importance of dynamic cues (Kewley-Port, 1983; Nearey and Assman, 1986) which may have implications for the common phonological patterns (Warner, 1998).

The main tool for modeling the dynamics of speech that has dominated the *back-end* of speech recognition systems is a popular model for time-series data – the hidden Markov model (HMM). HMM received its name because it is a Markov model equipped with extra machinery to handle noise: the states are not directly observable, but only through a set of parameters specifying *output probabilities.* The duration of the time-series data is accounted for by a *state sequence*, and

[6] The number of cepstral coefficients does not include the delta coefficients, see below.

the possible state sequences are specified by the *transitions* in the Markov model. The output distribution on each state associates the state with the data at a given time. By assigning different output distributions to each state, points of acoustic change can potentially be captured by transitions between different states.

However, the basic HMM machinery still lacks the ability to model dynamic cues. As a remedy, people have proposed to augment the front-end with extra features that characterize the spectral change within a short time – the so-called *delta* features computed from local linear regression (Furui, 1986). In the current study, delta features are also used in front-end processing, thereby doubling the number of the front end parameters.

2.4.3 Inherent variability within acoustic speech – mixture output distributions

The variability in speech signals is yet another baffling difficulty. Although a rather different line of inquiry is directed towards discovering *invariance* in speech signals, the majority of the work in speech recognition has been concentrated on the modeling of variability with limited computational resources. A common strategy uses Gaussian mixture distributions as output distributions in each state of the HMM, with the mean and variance of each Gaussian being the parameters:

$$p(o_t|\theta_i) = \sum_j c_{i,j} N(o_t, \mu_{i,j}, \Sigma_{i,j})$$

where $\theta_i = \{(\mu_{i,j}, \Sigma_{i,j})_j\}$ is the parameter for the output distribution of state i, which includes a fixed number of Gaussian distributions.

In practice, combined with MFCC, this strategy has proved very effective, though the sources of variability in speech are still not well understood. In the current work, this same modeling strategy will also be adopted, primarily because the testing data for our study also contains multiple speakers and we do not plan to address the problem of speaker variability within the scope of this thesis. In our study, a mixture of two Gaussian distributions is used for each state, with the weights learned from data. Each Gaussian is parameterized with a mean vector of the same dimension as the front-end features, and a diagonal covariance matrix. It is a rather standard practice to adopt the combination of MFCC as front-end, and diagonal Gaussian mixtures as the output distributions in the backend.

2.5 Training Hidden Markov Models

Hidden Markov models are based on the assumption that real-world patterns are generated from Markov processes, but observed through a noisy channel. Despite this seemingly crude assumption, HMMs have been immensely popular in a number of fields, including speech recognition. In this section, we do not intend to provide a comprehensive introduction to HMMs. Instead, our goal is to highlight the connection between segmentation and model fitting, using notations borrowed from Rabiner (1989). Interested readers are referred to the standard tutorials [7] for the details that are omitted from our discussion.

The underlying Markov process for an HMM can be described by a finite-state machine, with states connected by probabilistic transitions. The word "hidden" derives from the fact that the state sequence is not directly observed from the data, but only through the output distributions on each state, which adds a great deal of flexibility to the application of HMM. On the state level, the segmentation problem for HMM is often formulated as Viterbi decoding – a dynamic programming problem that finds the single most likely state sequence given the

[7] e.g. (Rabiner, 1989; Bilmes, 1997; Jelinek, 1997).

model. When many models are composed into a finite-state network, this view also applies to segmentation into the individual models, since the result of Viterbi decoding in terms of a state sequence would also determine a unique model sequence.

However, the difficulty with using HMM lies in its training: to solve the problem of segmentation, one is required to provide a model for the sequence; but to train such a model, it would have been much easier if the state-level segmentation is known. Consider the following simple left-to-right HMM:

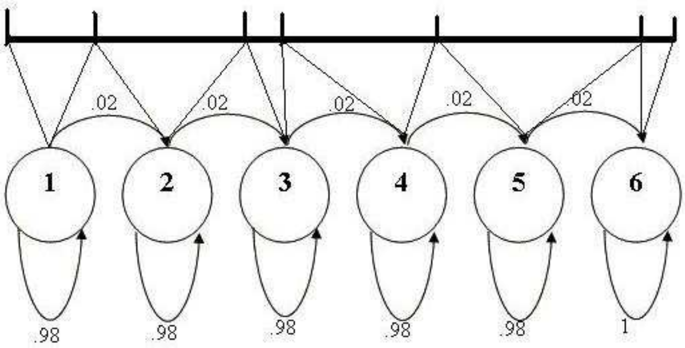

Figure 2.2: An example of a left-to-right HMM and an associated observation sequence

Let the state-level segmentation be represented by indicator functions $\psi_t(j)$, where $\psi_t(j) = 1$ if the observation vector at time t is emitted from state j, and 0 otherwise. As shown in Fig.(2.2), if we know in advance the values of $\psi_t(j)$ for all t and j, i.e. the state-level segmentation is given, then the estimation problem becomes easy. However the solution to the segmentation problem has to depend on the result of estimation.

One of the early ideas (Jelinek, 1976) looks quite ad-hoc: first guess a segmentation $\{\psi_t(j) : t = 1, \cdots, T, j = 1, \cdots, N\}$ to start with, and solve these two problems "in a cycle" by iterating:

1. Update the model using the segmented sequence;

2. Re-segment (using Viterbi decoding) the sequence with the new model.

It turns out that this "Viterbi training" algorithm[8] converges and provides a reasonable fit to the data in digit recognition applications. A similar idea also lies in the more general version developed by Baum (1970), where it is shown that the proper way of parameterizing segmentation for the E-step is to replace"hard" segmentation (i.e. indicator functions $\psi_t(j) \in \{0,1\}$) by "soft" ones (real-valued functions $\gamma_t(j) \in [0,1]$, where $\sum_j \gamma_t(j) = 1$). Computation of $\gamma_t(j)$ involves summing over all "soft" segmentations and will not be discussed here[9], but the structure of the algorithm remains the same and includes two steps: one step computing the "soft" segmentation, the other step updating the model.

As can be recalled from 2.2, this familiar case is yet another example of the missing data formulation: the "missing" part of the complete-data model is the "hidden" state sequence in the context of HMM. In parallel to the "soft" classifications w_i^m, the "soft" segmentations $\{\gamma_t(m)\}$ play the same role of missing data, where the index m points to the state to which observation vector t is classified. Like the EM algorithm introduced for mixture models, the Baum-Welsh algorithm iterates over computing the "soft" segmentation – the E-step – and updating model parameters – the M-step.

2.6 Mixture of Hidden Markov Models

The primary concern in choosing components of the mixture model is the nature of the data. Since acoustic speech segments may have different lengths and are

[8] Also called "segmental K-means" in Juang and Rabiner (1990).

[9] This is done through another dynamic programming technique called the *Forward-Backward* algorithm.

not stationary, the mixture model that we study must have the ability to handle time-series data. In principle, any probabilistic model that can produce a reasonable approximation to (2.14) for time series data can serve as components of the mixture. Therefore, the choice of models was not limited to HMM. For example, the mixture model in Kimball (1994) uses the model proposed in Ostendorf and Roukos (1989) as its components. Despite its limitations, we chose HMM based on a computational consideration. Many types of statistical inferences[10] can be done efficiently in HMMs because the natural ordering of the hidden variables allows the calculation to be done with dynamic programming. Since this dissertation depends heavily on iterative methods, choosing other models would incur a significantly higher computational cost.

On a historical note, the use of HMM for clustering originated in speech coding applications (Rabiner et al., 1989), and a mixture of HMMs was re-introduced by Smyth (1997) for clustering time series data. Since then it has been applied in the clustering of motion data (Alon et al., 2003) and ecological data (Li and Biswas, 2002).

The component HMMs in our mixture model are of a simple type. It has a left-to-right topology, with a continuous output probability distribution defined on each state. The output probabilities of our HMMs are chosen to be Gaussian mixtures also in order to facilitate computation. As we saw in 2.2, choosing the likelihood function from the exponential family allows us to represent the conditional expectation of the complete-data likelihood (2.16) as a linear combination of a set of sufficient statistics. Since we only need to compute the expectation of those sufficient statistics, the M-step has a rather standard solution.

The algorithm for fitting a mixture of HMMs extends the EM algorithm in

[10] Examples are likelihood calculation and segmentation.

Section 2.2 to handle the likelihood function of HMM. It follows the same "soft" classification-weighting scheme as described in Section 2.2. Two steps are iterated: one step assigns a "soft" label to each segment regarding their category membership, together with the necessary E-step for HMM; the other step uses the observed data as well as the labels to update the parameter of each HMM. Details of the algorithm, including equations and the explanation of notations, are included in Appendix 2.C.

For initial estimation of the EM algorithm, the segments were first clustered in the following procedure (Lee et al., 1989): the average spectrum of each speech segment is calculated using line spectral pairs, and a regular K-means clustering algorithm on the centroid vectors is based on the Itakura-Saito distance (Itakura, 1975). As a result, this procedure produces an initial clustering, which can be used to initialize the mixture weights as well as each HMM in the mixture.

2.7 Clustering experiments

2.7.1 Database

Most experiments in this dissertation will use data from the TIMIT database (Garofolo, 1988), a continuous spoken corpus designed to provide acoustic phonetic speech data for the development and evaluation of automatic speech recognition systems. It consists of read utterances recorded from 630 speakers from 8 major regions of American English, of which the data from the New England dialect region will be used. Because the goal of TIMIT was to train phone-based systems, the reading materials were chosen to be phonetically-balanced, and all data were manually transcribed by expert phoneticians. Since the manually transcribed segments provide a benchmark on the phone level, to date most phone recognition experiments have been evaluated on TIMIT.

2.7.2 Choice of parameters

In a given mixture model, all HMMs have the same architecture and the same number of parameters. For each experiment, the architecture and parameters of each HMM are shown in Table 2.1.

Experiment	Topology	Number of States	Mixture components per state
Phone	left-to-right	3	2
Diphone	left-to-right	6	2
Word	left-to-right	6	2

Table 2.1: Choice of model parameters in the clustering experiments

All experiments used MFCC+delta features as the front-end, with the number of cepstral coefficients set to 13.

2.7.3 Using HMM mixture in clustering speech sounds

As described in Section 2.6, a conventional K-means algorithm is used for the initialization of the subsequent HMM mixture learning. The difference between the two models lies mostly in their abilities to capture dynamics: while the K-means averages over each segment, HMM tries to capture the intra-segment dynamics with separate states with their respective output distributions. The following experiments are conducted in order to verify the effect of HMM mixture in clustering.

2.7.3.1 Phone clustering

The first two experiments use segments from a selected set of phone labels in the TIMIT transcription. The data are collected from 22 male speakers from the same area. The number of clusters is set to a fixed value (3 and 5, respectively) in each experiment, and the clusters are obtained first by running the K-means algorithm, then the EM algorithm for mixture of HMMs. Each segment is assigned to the cluster that corresponds to the maximal posterior probability for the given segment, and the cluster membership is calculated for all segments. The results are shown below in Table 2.2.

Since similar presentations of results are frequently used in this dissertation, we would like to explain how these tables should be read: the rows are instances of phone labels in TIMIT, and the columns are the cluster indices. Since the learning is unsupervised, the clusters are identified with integers, with an ordering chosen arbitrarily. Therefore each cell at (i, j) displays the number of instances with the i-th label that have the highest posterior probabilities for the j-th cluster. It may be seen that if the clustering entirely agrees with the phonetic labels, then with an appropriate ordering of the clusters, we should obtain a matrix with non-zero entries only on the diagonal. However, different clustering criteria can lead to different results, reflected by the distribution of non-zero entries in the matrix.

K-means:

	1	2	3
i	141	21	163
t	5	158	37
l	83	42	194

HMM mixture:

	1	2	3
i	240	1	84
t	0	200	0
l	106	6	207

K-means:

	1	2	3	4	5
i	5	21	119	88	92
s	127	140	18	97	1
tʃ	8	11	4	18	2
dʒ	25	24	4	25	1
n	37	79	103	96	30
l	9	46	144	85	35
k	18	53	78	82	17

HMM mixture:

	1	2	3	4	5
i	0	1	6	9	309
s	226	156	0	1	0
tʃ	17	7	0	19	0
dʒ	23	48	0	8	0
n	3	14	310	6	12
l	0	0	314	0	5
k	4	3	0	246	0

Table 2.2: Comparison of clustering methods in phone clustering experiments

2.7.3.2 Diphone clustering

Similar comparison was also conducted on larger units, such as diphones. Since the potential of HMM in describing dynamic patterns increases with the number of states, and diphones are complex units, 6 states are used for each HMM in the mixture. In order to collect enough instances of diphones, data from more than 300 speakers are used to provide the diphone sequences. The results are shown in Table 2.3.

K-means:

	1	2	3
ɪn	757	335	355
ɹi	243	421	303
ɑɹ	86	140	497

HMM mixture:

	1	2	3
ɪn	1437	5	5
ɹi	27	900	40
ɑɹ	13	41	669

K-means:

	1	2	3
ɪk	303	294	76
ɪs	363	116	11
ɹi	127	543	297
li	110	389	169
ɔɹ	40	258	235
ɔl	37	226	266

HMM mixture:

	1	2	3
ɪk	672	0	1
ɪs	490	0	0
ɹi	11	778	178
li	18	460	190
ɔɹ	0	29	504
ɔl	4	104	421

Table 2.3: Comparison of clustering methods in diphone clustering experiments

2.7.3.3 Word clustering

Lastly, we compare the clustering methods on a set of isolated words. The same set of data as used for diphone clustering are used in this experiment. The words are chosen from frequent words in TIMIT differ in several segments. The result is shown in Table 2.4.

K-means:

	1	2	3
water	276	2	58
she	8	230	170
ask	11	137	179

HMM Mixture:

	1	2	3
water	333	0	3
she	0	405	3
ask	0	11	316

Table 2.4: Comparison of clustering methods in word clustering experiments

As can be seen from Table 2.2 – Table 2.4, the more dynamics are contained within the sequences, the better the mixture model performs compared to the K-means method. This is expected, since K-means loses time resolution by averaging the spectra over the whole sequence, while HMM mixture captures the change within a sequence by different models. Moreover, as the number of clusters increase, not all clusters are distinct, possibly due to the fact that the EM algorithm only finds a local maximum within a complex likelihood space. Therefore, it may be an advantage if the search for clusters is carried out in a sequence of steps, each step adding a small number of clusters. This idea forms the basis of our conjecture on feature discovery. In the next chapter, we will be primarily interested in applying this method on the segmental level, with the goal of modeling the learning of phonetic categories.

2.A The convergence guarantee of the EM algorithm

This appendix summarizes the main arguments for the use of EM algorithm in Dempster, Laird and Rubin (1977). The original paper starts with the following identity based on the Bayes formula:

$$\log p(y|\theta) = \log p(y, z|\theta) - \log p(z|y, \theta) \tag{2.10}$$

Notice the first term is the likelihood of the complete data model, and the second term is a posterior predictive distribution. The main idea of EM lies in finding a way to increase the first term without an increase in the second term. It turns out that in order to make this work in general, the proper way is to take the expectation with regard to the predictive distribution in the second term. Since this distribution depends on θ in particular, we use $E_{\theta^{(t)}}[.]$ to indicate that $p(z|y, \theta^{(t)})$ is used to take the average. Hence:

$$\begin{aligned} E_{\theta^{(t)}}[\log p(y|\theta)] &= E_{\theta^{(t)}}[\log p(y, z|\theta)] - E_{\theta^{(t)}}[\log p(z|y, \theta)] \\ \log p(y|\theta) &= E_{\theta^{(t)}}[\log p(y, z|\theta)] - E_{\theta^{(t)}}[\log p(z|y, \theta)] \qquad (2.11) \end{aligned}$$

The left side remains the same since it has nothing to do with z at all. Now in order to find a $\theta^{(t+1)}$ that beats $\theta^{(t)}$, $\theta^{(t+1)}$ only needs to beat $\theta^{(t)}$ for the first term:

$$\text{Find } \theta^{(t+1)}, \text{ s.t. } E_{\theta^{(t)}}\left[\log p(y, z|\theta^{(t+1)})\right] > E_{\theta^{(t)}}\left[\log p(y, z|\theta^{(t)})\right] \tag{2.12}$$

The main reason why this suffices is because nothing can beat $\theta^{(t)}$ for the second term, because of Jensen's inequality:

$$\forall\theta,\ E_{\theta^{(t)}}\left[\log p(z|y,\theta)\right] = \int p(z|y,\theta^{(t)}) \log p(z|y,\theta)\, dz \le E_{\theta^{(t)}}\left[\log p(z|y,\theta^{(t)})\right] \tag{2.13}$$

Combining (2.11), (2.12) and (2.13), it can be seen that any $\theta^{(t+1)}$ that satisfies (2.12) is a better choice than $\theta^{(t)}$:

$$\log p(y|\theta^{(t+1)}) > \log p(y|\theta^{(t)})$$

EM actually finds the best possible $\theta^{(t+1)}$ by picking the one that maximizes $E_{\theta^{(t)}}\left[\log p(z|y,\theta)\right]$. Hence EM is the best solution following this strategy[11].

2.B An example of the EM algorithm as applied to a mixture of two Gaussian distributions

In this appendix, the E-step and M-step for fitting the mixture model (2.6) and (2.7) will be briefly described to illustrate the use of the EM algorithm.

> **E-step**: Since (2.5) is linear in z_i^m, to compute the conditional expectation of (2.5), we only need to take the conditional expectation

[11] Though a direct computation is not always possible.

of z_i^m:

$$\begin{aligned}
E_{\theta^{(t)},\{y_i\}}\left[z_i^m\right] &= E\left[z_i^m \mid \theta^{(t)}, \{y_i\}\right] \\
&= 1 \cdot Pr\left(z_i^m = 1|\theta^{(t)}, \{y_i\}\right), \text{ since } z_i^m \text{ is 0-1 valued} \\
&= \frac{Pr\left(y_i, z_i^m = 1|\theta^{(t)}\right)}{Pr\left(y_i|\theta^{(t)}\right)}, \text{by Bayes' formula} \\
&= \frac{Pr\left(y_i|z_i^m = 1, \theta^{(t)}\right) Pr\left(z_i^m = 1|\theta^{(t)}\right)}{Pr\left(y_i|\theta^{(t)}\right)} \\
&= \frac{Pr\left(y_i|z_i^m = 1, \theta^{(t)}\right) Pr\left(z_i^m = 1|\theta^{(t)}\right)}{\sum_{z_i^j=1} Pr\left(y_i, z_i^j = 1|\theta^{(t)}\right)}, \ z_i^j \text{ summed out} \\
&= \frac{\lambda_m(\theta^{(t)}) f_m(y_i|\theta^{(t)})}{\sum_{j=1}^{M} \lambda_j(\theta^{(t)}) f_j(y_i|\theta^{(t)})} \qquad (2.14)
\end{aligned}$$

For convenience, we also write $E_{\theta^{(t)},\{y_i\}}\left[z_i^m\right]$ as w_i^m. It is clear from (2.14) that $\{w_i^m\}$ satisfy $\sum_m w_i^m = 1$. In comparison to the indicator function z_i that corresponds to a "hard" classification of sample y_i into one of the mixture components, $w_i = (w_i^1, w_i^2, \cdots, w_i^M)$ is essentially a "soft" classification of y_i. The better a component f_m fits the data y_i, the closer w_i^m approaches 1. Intuitively, one may interpret the E-step as distributing a "fraction" of sample y_i into the m-th component according to w_i^m.

As an example, consider fitting the 2-Gaussian mixture in 2.1. The E-step computes the expected values of the membership indicators z_i as follows:

$$w_i = \frac{\lambda f_1(y_i|\theta^{(t)})}{\lambda f_1(y_i|\theta^{(t)}) + (1-\lambda) f_2(y_i|\theta^{(t)})}, i = 1, \cdots, N \qquad (2.15)$$

Where $f_1(y|\theta)$ and $f_2(y|\theta)$ are the likelihood of y_i as measured from the two normal distributions.

M-step: Since the likelihood function (2.5) has a linear form, the conditional expectation of (2.5) is obtained by simply substituting z_i^m with w_i^m. Moreover, it is convenient to write it as two terms:

$$\sum_{i=1}^{N}\sum_{m=1}^{M} w_i^m \log \lambda_m(\theta) + \sum_{i=1}^{N}\sum_{m=1}^{M} w_i^m \log\left[f_m(y_i|\theta)\right] \tag{2.16}$$

The second term of (2.16) uses w_i^m – the "fractions" to weight the contribution of y_i to each component. Since the E-step computes a conditional expectation, and w_i^m is a function of $\theta^{(t)}$, (2.16) can be viewed as a function of θ and $\theta^{(t)}$ (denoted as $Q(\theta, \theta^{(t)})$ in Dempster, Laird and Rubin (1977)). The next step is to maximize (2.16) over θ. This new value for θ will subsequently be used as the new parameter estimate $\theta^{(t+1)}$ in the next iteration.

In the example of 2-normal mixture, the M-step of model fitting has a solution $\theta^{(t+1)} = (\hat{\lambda}, \hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1^2}, \hat{\sigma_2^2})$, where the relevant equations for updating the model parameters are as follows:

$$\begin{aligned}
\hat{\mu_1} &= \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i} \\
\hat{\mu_2} &= \frac{\sum_{i=1}^{N} (1 - w_i) y_i}{\sum_{i=1}^{N} (1 - w_i)} \\
\hat{\sigma_1^2} &= \frac{\sum_{i=1}^{N} w_i (y_i - \hat{\mu_1})^2}{\sum_{i=1}^{N} w_i} \\
\hat{\sigma_2^2} &= \frac{\sum_{i=1}^{N} (1 - w_i)(y_i - \hat{\mu_1})^2}{\sum_{i=1}^{N} (1 - w_i)} \\
\hat{\lambda} &= \frac{\sum_{i=1}^{N} w_i}{N}
\end{aligned}$$

2.C The learning algorithm for a mixture of HMMs

To expand our notation for time-series data, we represent a sample as $O^{(s)} = o_1^{(s)}, o_2^{(s)}, \cdots, o_T^{(s)}$, where s is the sample index, T is the the length or the number of frames of the speech segment. The mixture model is parameterized by the following set of variable tuples:
$\{(\lambda_m, (a_{ij})^{(m)}, (\mu_{i,k})^{(m)}, (\Sigma_{i,k})^{(m)}, (c_{i,k})^{(m)}) : m = 1, \cdots, M\}$, A_{mn}.

E-step: The computation of (2.16) is reduced to the following terms.

$$\xi_t^{(s,m)}(i,j) = \frac{\alpha_t^{(s,m)}(i) \cdot a_{ij}^{(m)} b_j^{(m)}(o_{t+1}^{(s)}) \cdot \beta_{t+1}^{(s,m)}(i)}{p(O^{(s)}|\theta_m)} \tag{2.17}$$

$$\gamma_t^{(s,m)}(i) = \frac{\alpha_t^{(s,m)}(i)\beta_t^{(s,m)}(i)}{p(O^{(s)}|\theta_m)} \tag{2.18}$$

$$\gamma_t^{(s,m)}(i,k) = \gamma_t^{(s,m)}(i) \cdot \frac{c_{i,k}^{(m)} N(o_t^{(s)}, \mu_{i,k}, \Sigma_{i,k})}{\sum_j c_{i,j}^{(m)} N(o_t^{(s)}, \mu_{i,j}, \Sigma_{i,j})} \tag{2.19}$$

$$w_s^m = E_{\{O^{(s)}\},\theta}[z_s^m] = \frac{p(\theta_m)p(O^{(s)}|\theta_m)}{\sum_j p(\theta_j)p(O^{(s)}|\theta_j)} \tag{2.20}$$

Some explanations of the notations: $p(O^{(s)}|\theta_m)$ is the likelihood of sequence $O^{(s)}$ given mixture component θ_m. $\alpha_t^{(s,m)}(i), \beta_t^{(s,m)}(i)$ are the regular forward and backward probabilities computed for sequence $O^{(s)}$ using model parameter θ_m. Their role is to to facilitate the likelihood computation[12]. $a_{ij}^{(m)}$ are transition probabilities associated with model parameter θ_m. $b_j^{(m)}(.)$ are output probabilities associated with model parameter θ_m. $N(o_t^{(s)}, \mu_{i,k}, \Sigma_{i,k})$ are the likelihood functions of the normal components in the output mixture distribution $b_j^{(m)}(.)$.

[12] The reader is referred to standard tutorials, e.g. (Rabiner, 1989), for more details about the definitions of forward and backward probabilities.

In (2.17)(2.18)(2.19), the extra subscripts m and s indicate that there is a separate counter for each pair of HMM and observation sequence. (2.17) calculates the average number of transitions from state i to state j in model θ_m when sequence $O^{(s)}$ is observed. (2.18) calculates the "soft" segmentation of frame t to state i of model θ_m when sequence $O^{(s)}$ is observed. Finally, there are two steps that compute the "fractions": (2.19) calculates the fractions of individual frames in $O^{(s)}$ as assigned to each component in the normal mixture in state i of model θ_m; (2.20) calculates the fractions of sequences as assigned to each component in the HMM mixture. Notice that we have two different mixing mechanisms – one on the frame level, the other on the sequence level. It is the flexibility of the mixture model that allows them to be combined in one model.

> **M-step**: Due to the normality assumption, the solution to the M-step can be expressed as a series of weighted-sums.

$$a_{ij}^{(m)} = \frac{\sum_s w_s^m \sum_t \xi_t^{(s,m)}(i,j)}{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i)} \tag{2.21}$$

$$\mu_{i,k}^{(m)} = \frac{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,k) o_t^{(s)}}{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,k)} \tag{2.22}$$

$$\Sigma_{i,k}^{(m)} = \frac{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,k)(o_t^{(s)} - \mu_i)(o_t^{(s)} - \mu_i)^T}{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,k)} \tag{2.23}$$

$$c_{i,k}^{(m)} = \frac{\sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,k)}{\sum_j \sum_s w_s^m \sum_t \gamma_t^{(s,m)}(i,j)} \tag{2.24}$$

$$\lambda_m(\theta) = \frac{\sum_s w_s^m}{\sum_s \sum_j w_s^j} \tag{2.25}$$

The M-step further illustrates the idea outlined in 2.2: the "fractions" of individual frames are used to weight the sufficient statistics in (2.22) and (2.23);

In addition, the “fractions” of sequences are used to weight the sufficient statistics in (2.17)(2.18)(2.19). The parameters of a given model are then updated using the weighted sum of all counters associated with this model. Specifically, (2.21) updates the transition probabilities $a_{ij}^{(m)}$ between states; (2.22) updates the means $\mu_{i,k}^{(m)}$ in the normal mixtures; (2.23) updates the covariance matrices $\Sigma_{i,k}^{(m)}$ in the normal mixtures; and (2.24) updates the mixing priors of the normal mixture $c_{i,k}^{(m)}$. Finally, on the level of the HMM mixture, (2.25) updates the mixing priors of the mixture components.

CHAPTER 3

Learning features from waveforms

This chapter applies the methods introduced in Chapter 2 to the problem of clustering segments. Based on mixture models, our goal is to provide interpretations of two important concepts – phonetic categories and features. On one hand, the mixture model consists of components that each characterize a sound category, and learning categories is modeled as fitting the components of the mixture model. On the other hand, phonetic features will be obtained as a by-product of learning phonetic categories: starting from coarse-grained categories, the set of speech sounds is gradually refined to obtain the fine-grained sub-categories. Each step of refinement provides an interpretation of phonetic features based on acoustic criteria.

Since the word "feature" has been used with different meanings in different areas of linguistics, we should clarify what kind of features may conceivably be learned from data in the current statistical framework[1]. Within the scope of this dissertation, we take features to be *phonetic*, and avoid referring to "distinctive feature" as it is used in generative phonology. The main reason is that the word "distinctive" implies a rather different learning problem: finding which features are distinctive through access to all the lexical contrasts in the language. As explained in Section 1.3, this is not of primary interest to the current thesis.

[1] Outside the linguistics literature, there is another possible confusion with a completely different meaning in "feature selection" or "feature extraction" in the context of pattern classification.

Phonetic features are generally taken to be the distinctions between the individual sounds, and the qualification of features as phonetic distinctions dissociates the description of sounds from the mental lexicon (Ladefoged, 2001). With this narrower sense of feature, the heart of the learning problem is how features can be learned *before* lexical contrasts are established. Therefore the approach considered in this chapter can be characterized as *pre-lexical learning of phonetic features.*

3.1 Iterative refinement of the mixture model

The connection between clustering and phonetic features is motivated by the following thought experiment: suppose, for the sake of argument, that no features are known to a language learner *a priori.* Then how could the learner form any concept of features, such as [+voice], given only a distribution of speech sounds? The only plausible answer is that features must follow from a binary grouping of speech sounds. For example, a grouping of [p], [t] and [k] versus [b], [d] and [g]. Moreover, such a grouping can potentially be done in a hierarchical manner[2], since smaller sub-classes may be distinguished from within a given class of sounds.

As discussed in Chapter 2, the mixture model provides a method of grouping dynamic speech sounds into a fixed number of classes. In order to obtain a hierarchical grouping, we will need a simple extension of the clustering method introduced in Chapter 2, and perform clustering in a sequence of steps. In each step, a binary grouping is carried out within one of the existing classes, followed by a round of re-estimation, in which memberships of all the instances are adjusted according to the updated model. After a given class is partitioned into two

[2] A tree-like structure of features may remind the reader of the work on feature geometry (Clements and Hume, 1995), but it has a rather different meaning here.

sub-classes, the old class is subsumed by the new classes, thereby increasing the number of clusters by 1. Although the old class is not explicitly retained in the new mixture, the hierarchical structure is preserved through the order in which the new classes are added to the mixture: the initial values of the new classes are assigned by parameters of the old class. Therefore classes resulting from the same parent are generally more similar than the non-siblings in the model space. As an example, Figure 3.1 shows an example of such a hierarchy of classes.

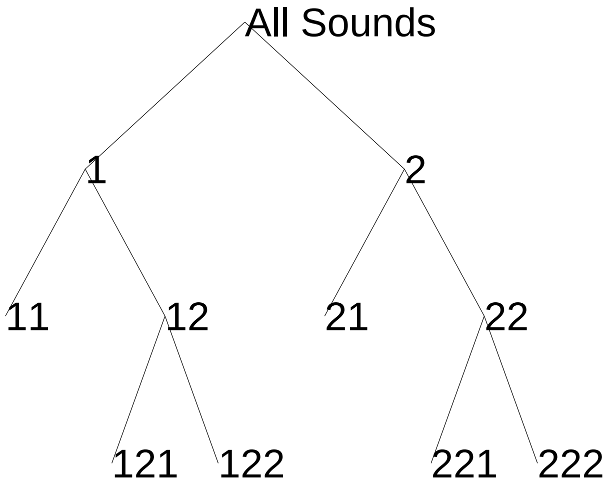

Figure 3.1: An example of of a hierarchy of 6 classes obtained through successive search (see Figures 3.2 – 3.6 for more details).

The above mentioned strategy of learning features is also motivated from the perspective of optimization. When the number of classes is large, due to the complex form of the likelihood function, finding the global maximum in the likelihood space can be very difficult. The strategy employed above can be seen as a search heuristic that starts with model with a small dimension, and gradually grows an extra dimension in order to achieve a better fit to the empirical distribution of data. The criterion for choosing which cluster to split is again based on likelihood.

Algorithm 1 Successive cluster splitting

```
Train a mixture of $k$ HMM's
repeat
    for each cluster $C_i$ do
        Split $C_i$ into $n$ clusters and obtain a new mixture model, record the gain in likelihood
    end for
    Choose the split that maximally increases the likelihood
    Retrain the new mixture model on all data
until stopping condition is satisfied
```

The intuition of Algorithm 1 is that new categories are identified from the largest or the most heterogeneous subset of data. The retraining step ensures that inaccuracies resulting from an earlier clustering can be potentially corrected as finer classes are added later[3]. Thus Algorithm 1 may be viewed as a strategy for inductively learning the sound categories as well as phonetic features from unlabeled data.

3.2 Experiment on learning features from the TIMIT data

3.2.1 The learned phonetic features

In the following experiment, we test Algorithm 1 on a set of TIMIT segments. The TIMIT segments are extracted from the phonetic transcriptions accompanying the corpus, but no segment labels are used in the clustering experiment. As a result, the learning data consist of 7166 instances of unlabeled TIMIT segments, derived from 2073 words by 22 male speakers. The use of hand-segmented phones allows us to focus on the problem of learning phonetic categories. It also enables easy comparisons between the clustering results and the knowledge-based phonetic labels.

[3] This is an advantage over clustering methods that use aggressive split/merge search.

Figures 3.2 – 3.6 illustrate how the phonetic segments are divided into two new clusters at each partitioning step. The phonetic labels use symbols from the TIMIT phonetic alphabet. The correspondence between TIMIT and IPA symbols is shown in Table 3.1. Six symbols: {bcl, dcl, gcl, pcl, tcl, kcl} represent the closure part of a stop or affricate and therefore do not have any IPA equivalents.

TIMIT	IPA	Example	TIMIT	IPA	Example
b	b	bee	hv	h̬	ahead
d	d	day	l	l	lay
g	ɡ	gay	r	r	ray
p	p	pea	w	w	way
t	t	tea	y	j	yacht
k	k	key	el	l̩	bottle
dx	ɾ	muddy	iy	i	beet
q	ʔ	bat	ih	ɪ	bit
jh	dʒ	joke	eh	ɛ	bet
ch	tʃ	choke	ey	eɪ	bait
s	s	sea	ae	æ	bat
sh	ʃ	she	aa	ɑ	bott
z	z	zone	aw	ɑʊ	bout
zh	ʒ	azure	ay	aɪ	bite
f	f	fin	ah	ʌ	but
th	θ	thin	ao	ɔ	bought
v	v	van	oy	ɔɪ	boy
dh	ð	then	ow	ʊ	boat
m	m	mom	uh	ʊ	book
n	n	noon	uw	u	boot
ng	ŋ	sing	ux	u̘	toot
em	m̩	bottom	er	r	bird
en	n̩	button	ax	ə	**a**bout
eng	ŋ̩	washington	ix	ɨ	deb**i**t
nx	˜ɾ	winner	axr	ɾ̩	butt**er**
hh	h	hay	ax-h	ə̥	s**u**spect

Table 3.1: TIMIT alphabet-IPA correspondence table

In the following figures, for each phonetic label, the position of the vertical bar indicates the percentages of the acoustic segments that were assigned to the left and right cluster. The two categories resulting from each partition are given names that are based on a subjective interpretation of the members. Therefore the names should be taken as mnemonics, rather than a kind of information available to the learner. For example in Figure 3.2, the bars corresponding to the voiced interdental fricative “dh” represent the result that 95% of acoustic segments labeled “dh” were assigned to cluster 1 (named “non-approximant”)[4] and 5% were assigned to cluster 2 (named “approximant”). The clusters were named using prefix coding. For example, a parent cluster named 12 was split into daughter clusters 121 and 122. To save space, each figure displays the subset of labels with more than half of the segments falling in the parent cluster. For example, labels included in Figure 3.4 (cluster 21 and 22) were those that have been mostly assigned to cluster 2 (“approximants”) in Figure 3.2. In particular, some labels are consolidated into one row in order to save space for better display.

[4] Notice another option for naming this feature is to call it [sonorant] and the two classes “obstruent” and “sonorant”. However since nasals are generally considered sonorants, we will try to avoid the confusion by using approximant instead.

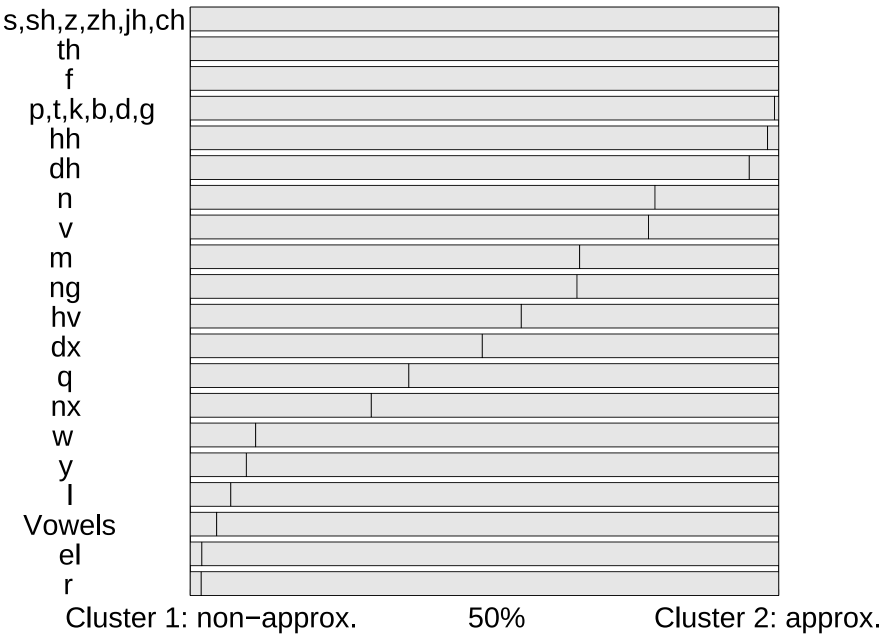

Figure 3.2: The first partition: [approximant]

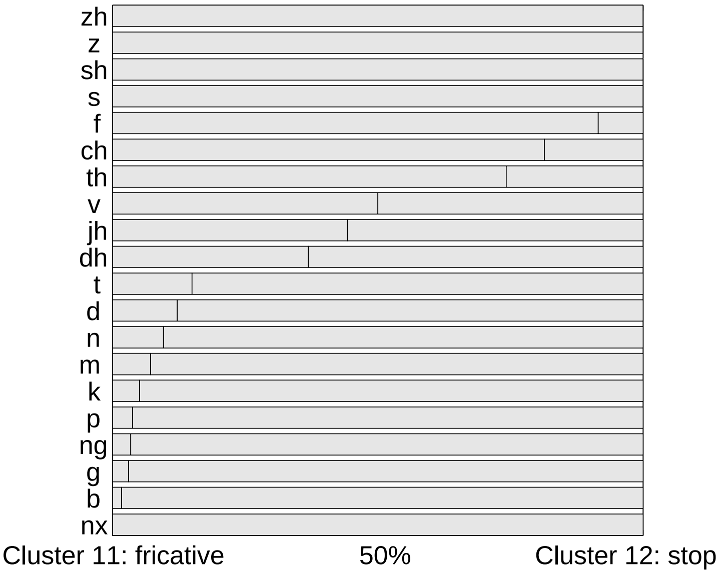

Figure 3.3: The second partition, of non-approximants: [fricative]

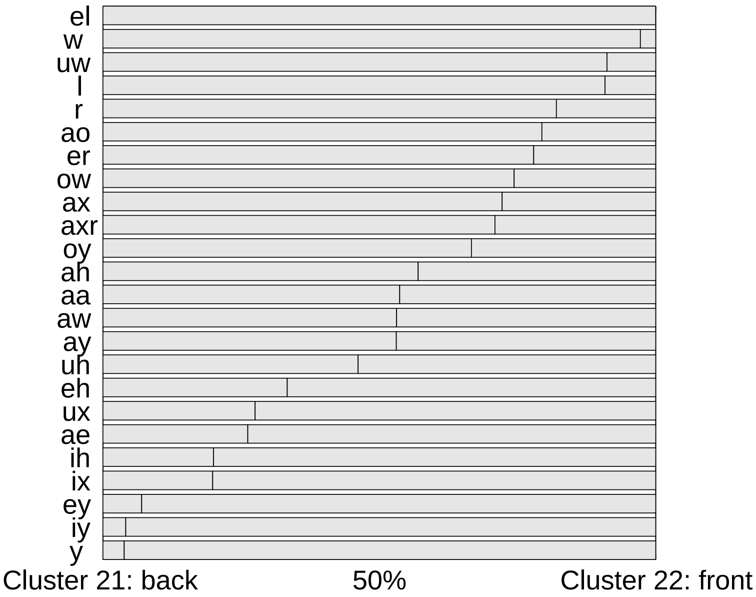

Figure 3.4: The third partition, of approximants: [back]

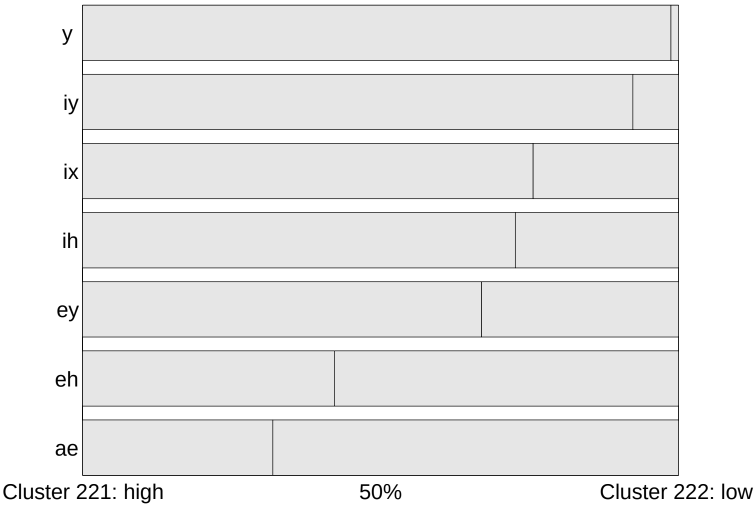

Figure 3.5: The fourth partition, of front approximants: [high]

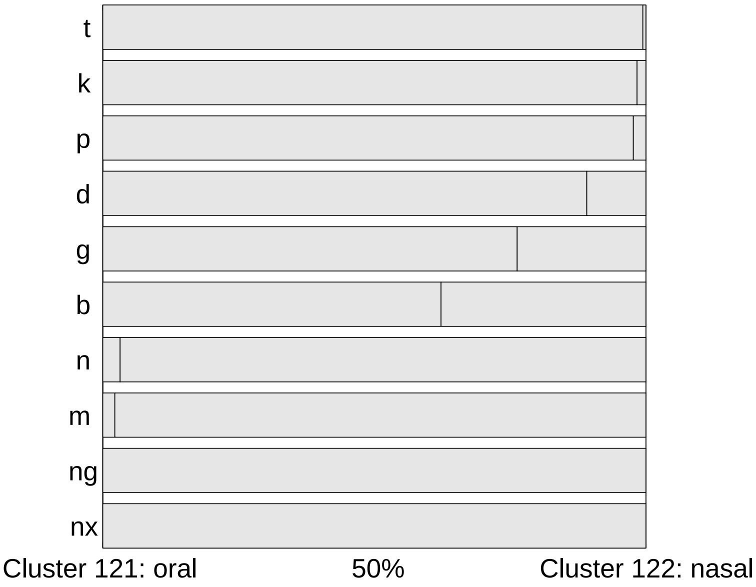

Figure 3.6: The fifth partition, of stops: [nasal]

The division of phonetic segments at each split suggests that the splits may be interpreted as gradient acoustic features that distinguish two classes of sounds by the general shapes of their spectral envelopes. For convenience, these features were named using linguistic terms. The percentages may depend on the distribution of sounds in the training data set, but they reflect some general patterns of contextual variation in phonetic segments.

3.2.2 Features values and phonetic contexts

The bottom-up approach described in 3.2.1 treats the broad classes as context-independent units. Therefore, the gradient feature values for individual sounds are to a large extent due to the specific phonetic contexts in which these sounds occur. As a representative example, let us consider the distribution of the voiced labiodental fricative [v]. The fact that in continuous speech, [v] is often produced as an approximant without significant frication noise is reflected in the ambiguous status of [v] in Figure 3.2 and Figure 3.3. To examine the effect of phonetic context on the distribution of [v] among the clusters, three different contexts are distinguished – word initial, word medial and word final positions. As a baseline, the distribution of [v] in these three positions is shown below:

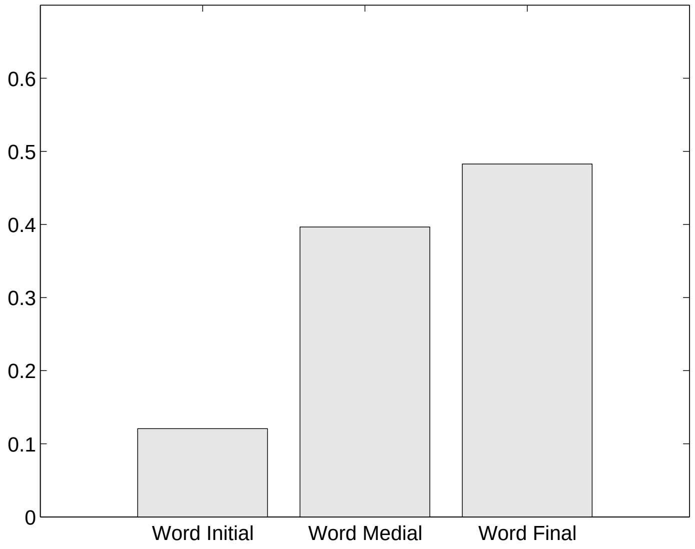

Figure 3.7: The distribution of [v] in the whole TIMIT database

For Cluster 1 and 2, the percentages of [v] that fall within each of the three categories are shown in Figure 3.8.

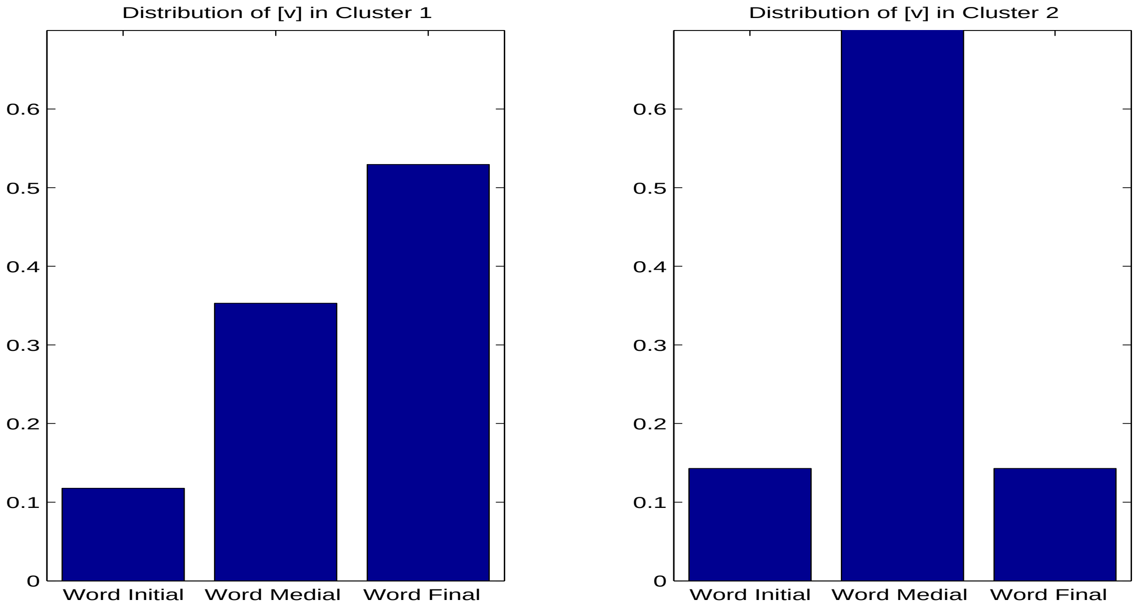

Figure 3.8: The distribution of [v] in obstruent versus sonorant classes

Recall that Cluster 1 groups together sounds that have primarily the characteristic of obstruents, while Cluster 2 corresponds to sonorants (as shown in Figure 3.2). Therefore, if we compare the above results with the baseline distribution, Figure 3.8 is in accord with the phenomenon that [v] behaves much more like an approximant in word-medial position than in word-final position.

We also consider the distribution of [v] within the two smaller sub-classes: Cluster 11 and Cluster 12. The result is shown in the next figure.

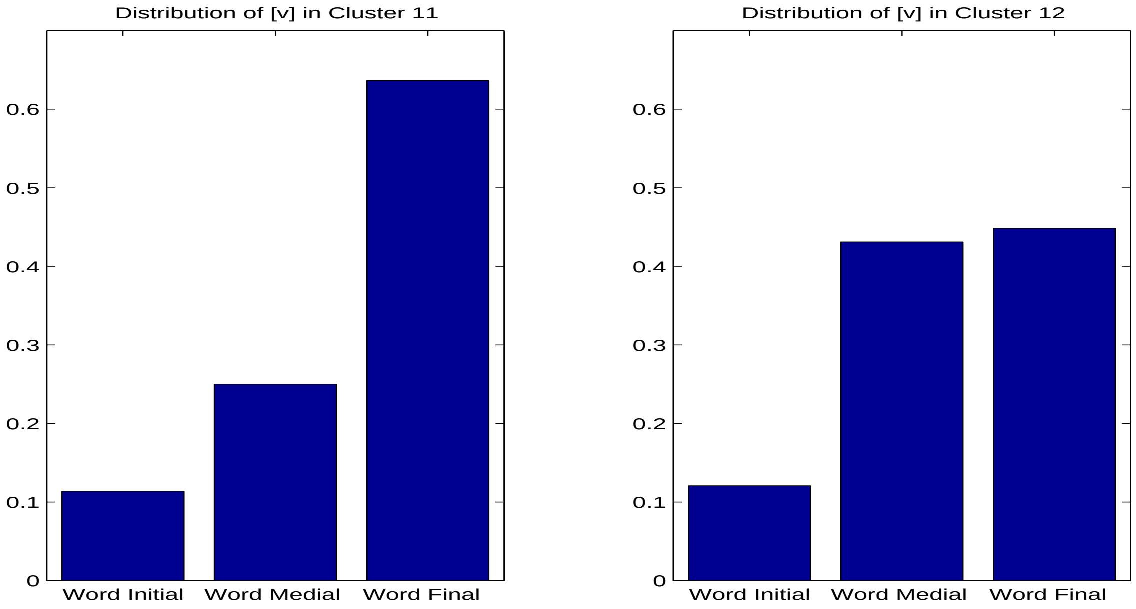

Figure 3.9: The distribution of [v] in fricative versus non-fricative obstruent classes

Compared with Figure 3.3, which characterizes the distinction between Cluster 11 and Cluster 12 as the feature [fricative], Figure 3.9 suggests that the fricative realization of [v] is overwhelmingly word-final, while the stop realization is slightly more frequent word-medially than word-finally, as compared to the baseline distribution of [v] in different word positions. Again, this agrees with the behavior of [v] as an ambivalent fricative in continuous speech, as documented for the TIMIT database (Keating et al., 1994).

A few other observations can also be made from the results shown in Figure 3.2 to Figure 3.6:

- ʔ, h, ɾ are also ambiguous between sonorants and obstruents.
- ɹ, l, w and l̩ all belong to the same class as the back vowels, and j stays with front high vowels.

- Affricates do not have their own category. They go with either fricatives or plosives.
- Among the stops, p, t, k show less variation across different classes than b, d, g.
- Canonical members of the classes have near-categorical feature values. For example: s, k, n, ə, w, i.

To a greater or lesser degree, these observations are in accordance with the familiar contextual variations of English segments.

3.2.3 Assessing the phonetic categories obtained via unsupervised learning

The examination of individual features provides a subjective evaluation of the classes obtained via unsupervised learning, and we have depended on our phonetic knowledge in making the assessments. However, a quantitative measure of the classes may be desirable. However, this is not as straightforward as the supervised tasks because no "answer" was given during learning. In order to create the answer, the reference label for each segment is created to reflect the subjective interpretations of those classes. For simplicity, these labels are based only on a rough division of the TIMIT phones and do not take into account any realizations of the segment. In evaluation, the reference labels are matched with the results output by the mixture model trained by the 22 speaker data set. As expected, the broad phonetic classes learned from data only roughly correspond to the broad classes determined by knowledge.

	(11)	(121)	(122)	(21)	(221)	(222)	% Match
Fricative	**958**	175	87	10	7	0	77.4
Plosive	48	**835**	90	1	2	0	85.6
Nasal	43	47	**734**	68	66	40	73.5
Back	21	33	147	**1088**	108	255	65.9
High	60	1	72	48	**817**	217	67.2
Central	36	48	59	152	268	**525**	48.3

Table 3.2: Distribution of broad classes in the training set

	(11)	(121)	(122)	(21)	(221)	(222)	% Match
Fricative	251	54	28	11	1	1	72.5
Plosive	16	218	27	0	0	0	83.5
Nasal	3	10	216	18	32	15	73.5
Back	1	10	33	329	29	86	67.4
High	3	1	22	21	244	77	66.3
Central	9	15	14	50	88	151	46.2

Table 3.3: Distribution of broad classes in the test set

Data set	Speakers	Phones	% Match
Train	22	7166	69.17
Test	7	2084	67.61

Table 3.4: Summary of the matching results on training and test sets

The similarity of performance on training and test sets suggests that the mixture model characterizes the general acoustic properties of the broad classes, as shown through Figure 3.2 – Figure 3.6 as well as Table 3.4. These results are comparable to previous work on classification of broad classes[5] (Leung and Zue, 1988; Meng and Zue, 1991; Cole and Hou, 1988; Juneja and Espy-Wilson, 2003), which also showed that the difficulty of classification depends on the confusability between different classes. Interestingly, similar confusions have also been observed

[5] Although the use of classifiers based on non-linear transformations of data has improved the performance of such works.

from observations of child speech production (Smith, 1973), thereby suggesting the possible difference between a child's internal phonological system and that of an adult's. Thus, categories derived from the mixture model offer an alternative view of the basic units of child phonology, leading to interesting predictions about the status of certain speech sounds in the development of speech perception.

CHAPTER 4

Learning segments from waveforms

The discussion in Chapter 3 assumed that instances of speech sounds are given for learning phonetic categories and features. However, this is a simplifying assumption at the first stage of building a model. In the situation of learning directly from word-level acoustic signals, the sequence of speech sounds in a speech signal poses another kind of "hidden structure" problem for learning. Segmentation, the procedure of associating each word signal with a sequence of category models, is the computation that infers such hidden structures. Thus learning not only involves identifying segment-sized phonetic units from word-level continuous speech signals, but also implies assigning segmental representations to these signals as well. The key to solving this problem is to extend the same learning strategy, applied in the previous chapter, to the situation where segmentation is the unknown information. Since the sub-problem of category learning has been addressed separately, the incremental approach allows us to add one component of unknown information at a time: first we consider adding segmentation, then adding phonotactics to the model. At each stage, the sub-model is augmented with an extra set of parameters to obtain a larger model; learned parameters of the sub-model are also used to set initial values of the larger model in the next stage of iterative learning.

4.1 The role of segmentation

As mentioned in Chapter 1, the main difficulty of learning phonetic units from word-level signals lies in the lack of cues of segment boundaries. If segment boundaries are given for each word, then phonological acquisition could in principle proceed along a line similar to the model that we presented in Chapter 3: starting with a small number of categories, the phonetic inventory will increase in size as smaller categories are discovered within the bigger ones. However, it is clear that what we have considered is, at the very best, an idealized situation: no children have been reported to benefit from hearing isolated speech sounds in their development. In order to justify such an idealization, we should be able to let go of the segment boundary assumption and directly learn from word-level signals. Since segmentation is no longer assigned to a fixed value, we will again rely on an iterative learning strategy; the EM algorithm introduces in Chapter 2 gives such an example. Based on the observation that identifying hidden variables can simplify the structure of the learning problem, the concept of "missing data" is introduced for the category membership of each segment. The strategy consists of two parts: one part is to first make guesses about the missing data based on the current model and observed data; the other part is to update the model using the observed and missing data together.

The current situation of learning from word-level signals bears a strong similarity to the previous one, especially with regard to the role that segmentation plays in phonological acquisition. Although there are some debates about the time order with regard to the knowledge of phonetic categories and the ability to identify these units from words (Werker, 2003), it is clear that a computational model would need to improve on both at the same time. Better knowledge of units leads to more reliable identification of these units from the speech stream, and improved segmentation would definitely help improve the knowledge of units.

These observations motivate the strategy that will be used in this chapter: first, each word is segmented based on what we know about sound categories; once the segmentation is known for each word, we can improve the sound categories based on the new segmentations. In what follows, we will present an algorithm based on this idea, and try to provide formal justifications in the appendix.

Another interesting aspect of phonological acquisition is the phonotactics of a language, i.e. the kind of knowledge that tells an English speaker *blick* is a better word than *bnick*. Taking the simplifying stance that phonotactics can be modeled as finite-state transitions between different units, phonotactics naturally becomes part of the iterative learning framework: results of segmentation can inform phonotactics, and phonotactics also constrains segmentation. Combined with the learning of units, the augmented model places phonological acquisition at a pre-lexcal stage: units and phonotactics are learned only from acoustic signals, without any concept of words.

4.2 Breaking words into segments

In Chapter 2, the problem of learning phonetic categories is characterized as the following optimization problem:

Finding *units*, so that the function $p(segments|units)$ is optimized.

The probability function $p(segments|units)$ tells us how well the segments are captured by the the unit mixture models. The situation changes somewhat when we turn to the problem of learning phonetic categories from holistic words. Intuitively, we would like to formulate a similar optimization problem, one that allows us to assign probabilities to words from the knowledge of units. However, an obvious difficulty is we do not yet know how to calculate such a function: we need a theory about how phonetic categories can be used to express words. The key observation, as was discussed in the introduction, is that once we know how the word w_i is segmented, then the problem should reduce to a familiar kind – since we already know how to learn units from segments. The missing piece that needs to be added is the role of segmentation in optimization. Care needs to be taken when one refers to segmentation as a variable – in this case, when segmentation is instantiated with a value, it should not only specify the boundary of each segment, but also the unit label for each segment in the given word.

A simple way of adding segmentation into the picture is introducing it as a parameter of the model. In other words, in addition to units, segmentation also appears at the right side of the conditional probability, and learning is set equivalent to the following problem:

> Find units and segmentation, so that $p(words|units, segmentation) = \prod_i p(word_i|units, segmentation_i)$ is optimized

Here *units* refer to the unit model, which itself is a mixture of HMMs and therefore include the parameters of each HMM as well as the weights in the mixture. The equality is based on the assumption that the probability of each word is independent of the others, and is only relevant to the unit model and its segmentation, written as $segmentation_i$. The explicit formulation of the hidden structure as a parameter allows us to do the probability calculation explicitly, based on a crucial assumption: given the unit model, the probability of words is a product of its component segments. Writing $word_i$ for a given word, $segmentation_i$ for the segmentation of $word_i$ and $segment_i^1, \cdots, segment_i^{k(i)}$ for the sequence of segments as the result of assigning a value s_0 to $segmentation_i$, this yields:

$$\begin{aligned} p(word_i|units, segmentation_i = s_0) &= p(segment_i^1, \cdots, segment_i^{k(i)}|units) \\ &= \prod_{j=1}^{k(i)} p(segment_i^j|units) \end{aligned}$$

This *conditional independence* assumption, taken literally, does not fit our understanding of coarticulation: probabilities of segments are actually correlated within a word. But for the current purpose, which seeks to identify relatively context-independent units at the level of broad classes, the conditional independence assumption allows us to break words down into segments:

$$\begin{aligned} p(words|units, segmentation) &= \prod_i p(word_i|units, segmentation_i) \\ &= \prod_i \prod_{j=1}^{k(i)} p(segment_i^j|units) \end{aligned} \tag{4.1}$$

Since the form of (4.1) is essentially the same as the mixture learning problem discussed in Chapter 2, the simple intuition outlined above suggests the following strategy for optimizing $p(words|units, segmentation)$: in order to do optimization over two sets of parameters, we may fix the values of one set, and try to optimize over the other set of parameters. Intuitively, this strategy resembles hill climbing in two coordinates, one along the coordinate of units, the other along the coordinate of segmentation[1]. As shown briefly in Appendix 4.A, this strategy would bring the function to a local maximum. Each step of optimization can be summarized as follows (the boldface indicates the variable to optimize, and the plain letters indicate fixed values):

[1] Since both the units and segmentation will each consist of many parameters, the coordinates are taken in an abstract sense.

1. Update units to optimize $p\,(words|\mathbf{units}, segmentation_i)$

2. Update segmentation of each word, to optimize $p\,(word_i|units, \mathbf{segmentation_i})$

Step 1 brings us back to a familiar problem, which itself was solved using iterative methods – the EM algorithm in Chapter 2. On the other hand, Step 2 involves a statistical inference that finds the best segmentation of each word given the unit models. This step is carried out by the Viterbi algorithm. Given an HMM θ and a sequence x, the Viterbi algorithm finds an optimal state sequence q, in the sense that q optimizes $p(x|\theta, Q = q)$. A short note on the computation of Viterbi is included in Appendix 4.B.

The use of Viterbi in finding an optimal state sequence can be extended to the case of finding the optimal unit sequence in a word. By specifying the search space for the possible sequences of unit models, unit HMMs can be composed into a finite state machine. When the Viterbi algorithm is run on the state space of the larger model, segmentation into units can be inferred from the optimal state sequence, since the state spaces of the unit models are disjoint. For the current purpose, the search space is chosen to include all unit sequences no longer than M_0; since it is assumed that all such unit sequences are equally likely[2], the result of Viterbi decoding would only depend on the signal and the status of units.

The role of M_0, a free parameter, is to prevent segments that have very short duration from occurring. In practice, when the models are poorly estimated in the beginning, unconstrained segmentation tends to result in many very short segments, and the algorithm is often trapped in a local maximum. Similar problems have also been reported in previous work on the design of data-driven speech

[2] In Section 4.4, this assumption is replaced by explicit modeling of phonotactics.

recognition (Bacchiani, 1999), where a threshold was used on the number of frames per segment. Since we do not yet have an assessment of segmental duration in our model, it is difficult to decide an appropriate threshold. Instead, a constraint is imposed on the total number of units allowed for each word. This constraint reflects the assumption that lexical representations are not coded by arbitrarily long sequences of categories, and may be interpreted as a kind of memory constraint on the learner. In practice, the value of M_0 is chosen to be between 8 and 10.

Based on the high-level view presented above, we introduce an algorithm for learning units from holistic words. The equations used therein are explained in Appendix 4.C.

Algorithm 2 Learning with segmentation as parameter

```
Require: an initial estimate $units^0$ for the unit models, $\epsilon > 0$
construct the unit sequence space $H$
$t \leftarrow 0$
repeat
    set of segments $Y \leftarrow \phi$
    for each $word_i$ do
        $segmentation_i^{(t)} \leftarrow Viterbi(word_i, units, H)$
        $\{segment_i^1, \cdots, segment_i^{k(i)}\} \leftarrow (word_i, segmentation_i^{(t)})$
        $Y \leftarrow Y \cup \{segment_i^1, \cdots, segment_i^{k(i)}\}$
    end for
    $k \leftarrow 0;$
    repeat
        for each $y_i \in Y$ do
            $z_i \leftarrow \text{E-step}(y_i, units^{(k)})$
        end for
        $units^{(k+1)} \leftarrow \text{M-step}(\{z_i\}, Y)$
        $k \leftarrow k + 1$
    until $\log p(Y|units^{(k+1)}) - \log p(Y|units^{(k)}) < \epsilon$
    $units^{(t+1)} \leftarrow units^k$
    $t \leftarrow t + 1$
until $\log p(words|units^{(t+1)}, segmentation^{(t+1)})$
          $- \log p(words|units^{(t)}, segmentation^{(t)}) < \epsilon$
```

4.3 Initial acoustic segmentation

In general, iterative methods require a reasonable initial estimate of the parameters. As an instance of iterative methods, Algorithm 2 requires initial unit models as part of the initial conditions. Since holistic words are the only available data in the beginning, an initial estimate of the unit models must be obtained through a set of initial segments. The role of producing initial segments is played by *acoustic segmentation* – a segmentation of the waveform without any unit models. For this purpose, we adapt an acoustic segmentation algorithm that has been used in previous work (Svendson and Soong, 1987; Bacchiani, 1999).

Roughly speaking, the goal of acoustic segmentation can be seen as finding discontinuities, or landmarks[3], in the speech signal. This task has been approached in many different ways, for example, in a supervised classification framework (Cole and Hou, 1988; Juneja and Espy-Wilson, 2003). For the current study, acoustic segmentation is posed as the problem of dividing waveforms into relatively stationary regions. When the number of segments M is fixed, acoustic segmentation can be formalized as another optimization problem:

> Given $1, \cdots, N$ frames of speech and $M > 1$, find $s_0 < s_1 < \cdots < s_M$ subject to constraints $s_0 = 1$, $s_M = N$, such that $\sum_{i=0}^{M-1} d(X_{s_i, s_{i+1}})$ is minimized, where $d(X_{s,t})$ is a function that measures the cost of the segment $X_{s,t} = (x_s, x_{s+1}, \cdots, x_t)$.

M_1 – the number of segments in the word – serves as a free parameter. When the value of M_1 is fixed, a solution to this problem can be obtained by applying a dynamic programming strategy: let $D(m, n)$ be the minimal cost for n frames to be divided into m segments, then:

$$\begin{aligned} D(1,n) &= d(X_{1,n}), n = 1, \cdots, N \\ D(m,n) &= \min_{m<t\leq n} \left(D(m-1,t) + d(X_{t,n})\right), m > 1, n = 1, \cdots, N \qquad (4.2) \end{aligned}$$

The segment boundaries are found by backtracking from the last boundary that minimizes $D(M, N)$. In Svendson and Soong (1987), d is chosen to be the Itakura-Saito distortion (Itakura, 1975), which has connections with an all-pole model of speech production. However, since it is expensive to compute the

[3] The original definition of landmarks (Stevens, 1998) refers to locations of change in feature values. But here we are using landmarks to loosely refer to possible segmental boundaries in the signal.

Itakura-Saito distortion, we followed a simpler approach and let d be the following function (Bacchiani, 1999):

$$d(X_{s,t}) = \max_{\mu} \left(\log(p(x_s, \cdots, x_t) | \mu, \Sigma_0) \right), \text{ where } \Sigma_0 = cov(X_{1,N}); \tag{4.3}$$

where Σ_0 is the diagonal covariance matrix calculated using the whole sequence of data $X_{1,N}$. (μ, Σ_0) are the parameters of each Gaussian distribution over the MFCC features, where μ is unknown. $\max_{\mu} \left(\log(p(x_s, \cdots, x_t) | \mu, \Sigma_0) \right)$ can be calculated using $\mu = \hat{\mu}$, the maximum likelihood estimate of μ from the sequence $X_{s,t}$. In practice this produces similar results to the metric using the Itakura-Saito distortion. The following figure illustrates the result of applying the algorithm to the same signal of "yumyum", preceded and followed by silence. The segment number is increased from 3 to 8. When $M_1 = 8$, the resulting segmentation roughly locates the expected segmental boundaries in the word. However, as can be seen from the figure, it is difficult to choose an *a priori* criterion for deciding the number of segments in the word.

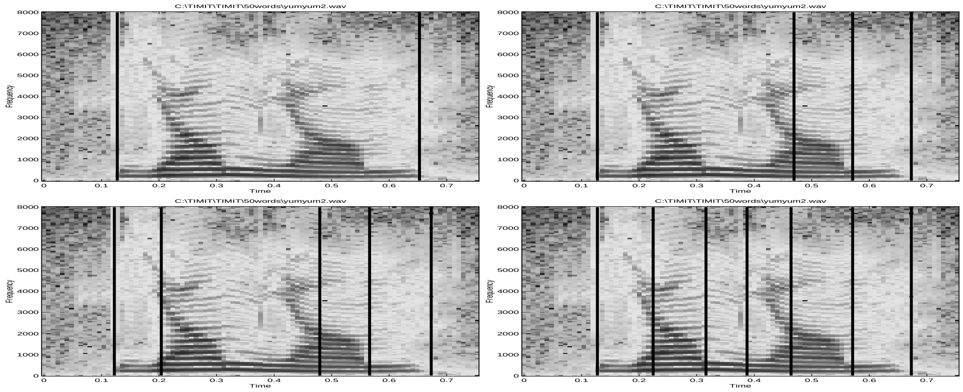

Figure 4.1: Maximum-likelihood segmentation of word "yumyum", M=3,5,6,8

In the experiment reported below, the total number of segments M_1 in the initial acoustic segmentation is set to be equal to the largest allowed number of segments M_0, so that the two constraints are made consistent with each other. As the result of initial segmentation, the set of initial segments is used to produce an initial estimate of the category models, using the procedure introduced in Chapter 3.

4.4 Incorporating phonotactics into learning

In both algorithms discussed above, all sequences of units are treated as equally likely, as reflected in the construction of the search space – it is assumed that each unit can be followed by any other unit, with equal probability. However, it has been shown that knowledge of phonotactics is acquired quite early (Jusczyk and Aslin, 1995), and the sensitivity to which sound sequences are more likely than others provides a basis for the acquisition of phonotactics (Saffran, Johnson, and Aslin, 1996).

The integration of phonotactics in the current learning model is straightforward, thanks to the flexibility of the iterative learning framework. Compared to the previous approach of treating all unit sequences as equally probable, the current approach treats phonotactics as a Markov chain, for which the state space consists of all the units, plus an initial and a termination state that do not have any output observation. The Markov chain starts from the initial state, and ends at the final state with probability 1. An example of such a phonotactic model, defined over a system of two units, is shown in Table 4.1.

	start	Sonorant	Obstruent	**end**
start	0	0.3	0.7	0
Sonorant	0	0.2	0.4	0.4
Obstruent	0	0.7	0.2	0.1

Table 4.1: An example of a phonotactic model based on Markov chain with starting and ending states

Thus learning phonotactics implies estimating the transition probabilities between units. Again, it may be seen that phonotactics and segmentation stand in an interdependent relationship:

- Knowledge of phonotactics affects segmentation. In particular, segmentation is determined jointly from the acoustic score calculated from the unit models and the phonotactic score.
- Phonotactics needs to be re-estimated from segmentation. When segmentation of each word is known, the resulting unit sequences should be used to derive the maximum likelihood estimate of phonotactics.

One way of taking advantage of such a relationship is extending the previous model to include phonotactics as part of the parameters. For example, extending the formulation in Section 4.2, the model is set up as:

$$p(words|units, phontactics, segmentation) \tag{4.4}$$

and the optimization is carried out in the following steps:

1. segmentation $\leftarrow \arg\max_s p(words|units, phontactics, segmentation = s)$
2. phonotactics $\leftarrow \arg\max_p p(words|units, phontactics = p, segmentation)$
3. units $\leftarrow \arg\max_u p(words|units = u, phontactics, segmentation)$

Step 1 amounts to finding the most likely sequence given the unit models and between-unit transitions. Since the new search space can be represented as a larger HMM itself, this step can be done with the Viterbi algorithm. In Step 2, since the segmentation is known for each word, and phonotactics governs the transitions between symbolic units, it suffices to update each entry $a_{i,j}$ of the phonotactics transition matrix to maximize the probability of all the unit sequences:

$$\hat{a}_{i,j} = \frac{\text{Number of transitions } unit_i \rightarrow unit_j}{\text{Number of occurrences of } unit_i} \tag{4.5}$$

Step 3 brings us back to the familiar problem discussed in Section 4.2: clustering the segments with the mixture-based unit model. After Step 3, the iteration then continues from Step 1. Based on the justification given in Appendix 4.A, this algorithm will also converge to a local maximum for (4.4).

As an extension of Algorithm 2, the algorithm for learning phonotactics together with units is presented below. Compared to Algorithm 2, it contains only a few extra steps of using phonotactics in segmentation and updating phonotactics.

Algorithm 3 Learning phonotactics and units

```
Require: an initial estimate $units^{(0)}$ for the unit models: $unit_1, \cdots, unit_M$; $\epsilon > 0$
initialize phonotactics $\{a_{i,j}\}$: $\forall i = 1, \cdots, M, a_{i,j} \leftarrow \frac{1}{M}$
$t \leftarrow 0$
repeat
    set of segments $Y \leftarrow \phi$
    for each $word_i$ do
        $segmentation_i^{(t)} \leftarrow Viterbi(word_i, units^{(t)}, phonotactics^{(t)})$
        $\{segment_i^1, \cdots, segment_i^{k(i)}\} \leftarrow (word_i, segmentation_i^{(t)})$
        $sequence_i \leftarrow segmentation_i^{(t)}$
        $Y \leftarrow Y \cup \{segment_i^1, \cdots, segment_i^{k(i)}\}$
    end for
    $phonotactics^{(t)} \leftarrow \text{Transition}(\bigcup_i \{sequence_i\})$
    $k \leftarrow 0$;
    repeat
        for each $y_i \in Y$ do
            $z_i \leftarrow \text{E-step}(y_i, units^{(k)})$
        end for
        $units^{(k+1)} \leftarrow \text{M-step}(\{z_i\}, Y)$
        $k \leftarrow k + 1$
    until $\log p(Y|units^{(k+1)}) - \log p(Y|units^{(k)}) < \epsilon$
    $units^{(t+1)} \leftarrow units^{(k)}$
    $t \leftarrow t + 1$
until $\log p(words|units^{(t+1)}, phonotactics^{(t+1)}, segmentation^{(t+1)}) - \log p(words|units^{(t)}, phonotactics^{(t)}, segmentation^{(t)}) < \epsilon$
```

4.5 Experiments on learning broad classes from holistic words

4.5.1 Data and procedure

The first experiment reported here is based on the same set of data from the TIMIT database as used in the previous chapter. However, an important difference lies in the use of segment boundaries. The segment boundaries provided in TIMIT are ignored in the learning stage of this experiment, and are only used in later comparisons with the results found by the learner. Consequently, the learning algorithm only has 2068 word-level acoustic signals[4] as input, and proceeds as follows:

- Use the acoustic segmentation algorithm (described in 4.2) to obtain an initial set of segments for each word;
- Run the K-means algorithm and the successive splitting algorithm (Algorithm 1 in 3.1) on the set of initial segments, and obtain an initial set of category models;
- Use Algorithm 3 to include the updating of unit models, segmentation and phonotactics, until the convergence criterion is met.

As before, 6 broad classes are derived as the result of successive splitting; and then the unit models are passed as initial values to the joint learning of units and phonotactics. Algorithm 3 converges after a certain number of iterations. As shown in Figure 4.2, as the data likelihood increases, the total number of segments decreases in each iteration.

[4] A few article words are excluded because they were assigned very short duration (< 10ms) in TIMIT transcription.

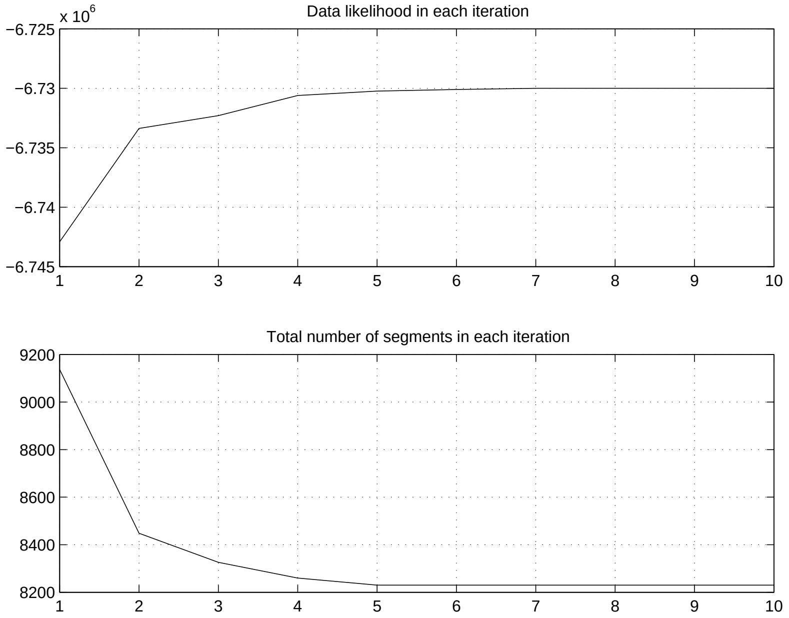

Figure 4.2: Data likelihood and total number of segments in 10 iterations

4.5.2 Broad class representations of the holistic words

As a result, a total of 8230 segments were returned for the 2068 words used as learning data. Some examples of the resulting broad class representations are shown in Figure 4.3 – Figure 4.5. The top panels show the spectrograms of the example words, overlaid with representations obtained through unsupervised learning; while the bottom panels indicate the TIMIT transcriptions for each word. The interpretation of these results will be included in the next few sections, where the issue of evaluation will be addressed.

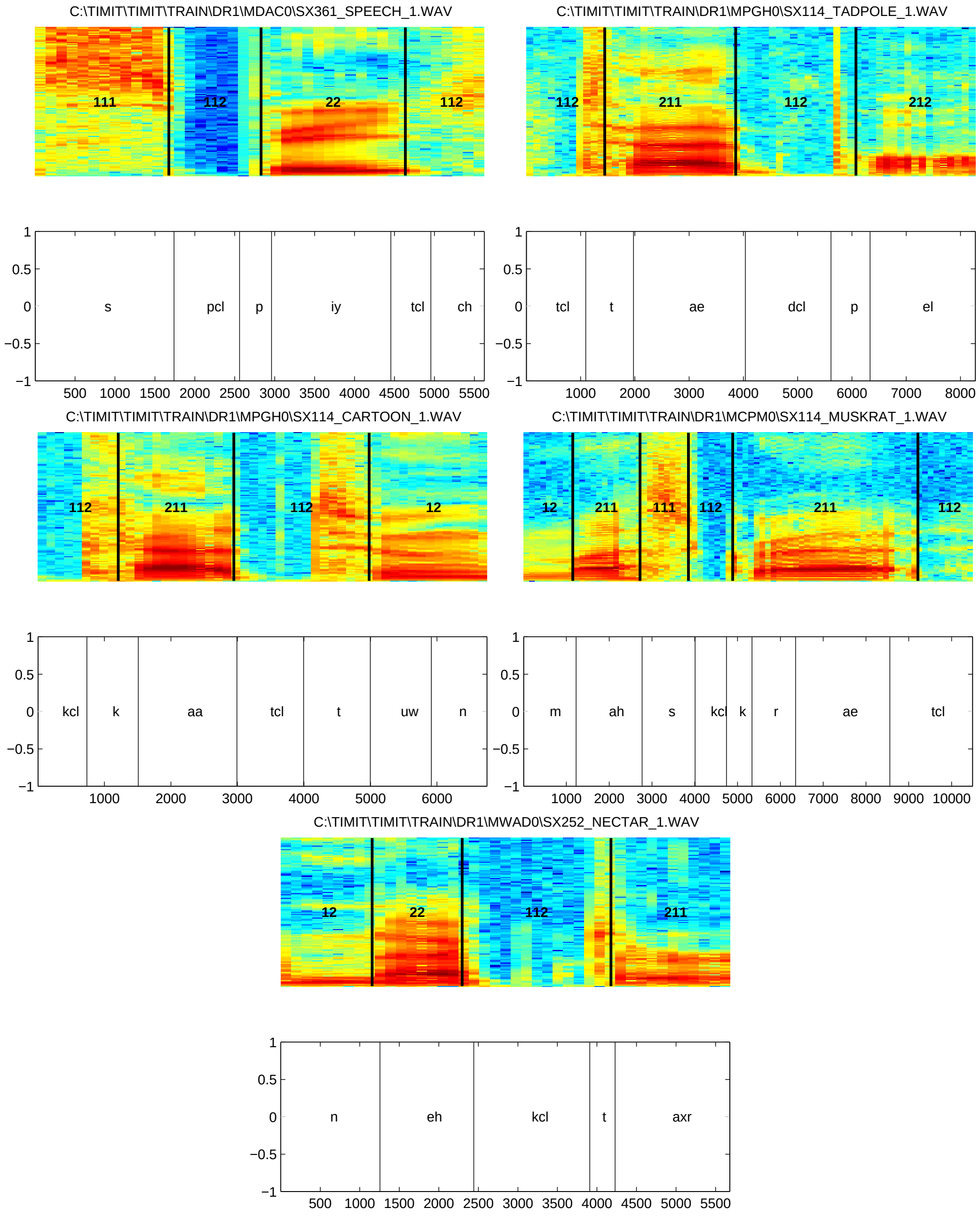

Figure 4.3: The learned and TIMIT segmental representation for the word "speech", "tadpole", "cartoon", "muskrat" and "nectar" in the last iteration.

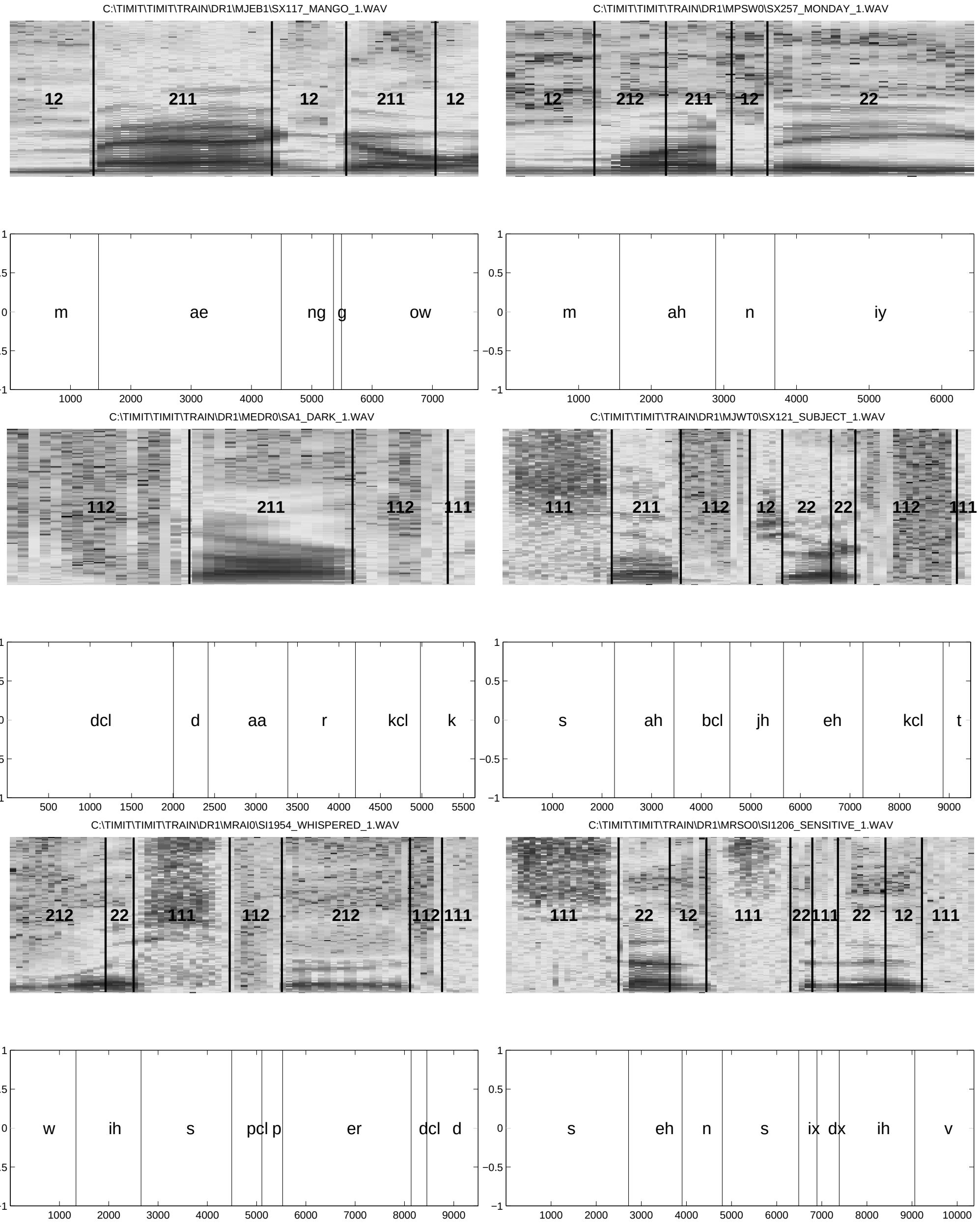

Figure 4.4: The learned and TIMIT segmental representation for the words "mango", "money", "dark", "subject", "whispered" and "sensitive" in the last iteration.

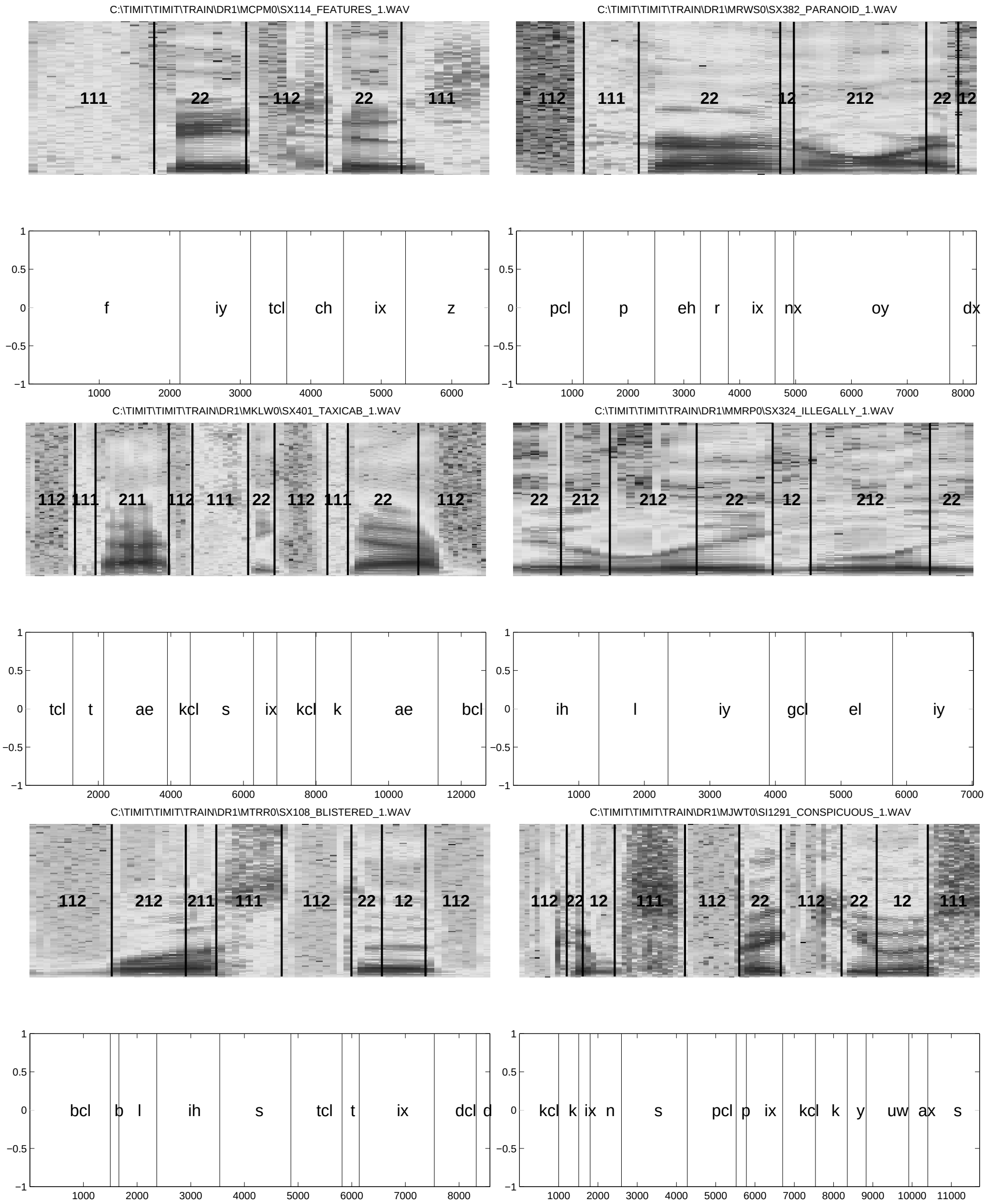

Figure 4.5: The learned and TIMIT segmental representation for the words "features", "paranoid", "taxicab", "illegally", "blistered" and "conspicuous" in the last iteration.

4.5.3 The problem of evaluation

Examination of the individual representations found for each word gives us a subjective evaluation of the model. However, an objective summary of the model's performance is often helpful, especially since we are dealing with a large amount of learning data. However, for unsupervised learning models, choosing appropriate benchmarks can be a problem. If the goal of the model is to capture certain aspects of human learning, then the ideal benchmark would be based on results from experiments in which human learners are presented the same set of data as the model. Obviously, this type of behavioral data is difficult to obtain, and we must look for alternative ways of evaluating the model.

Instead of committing ourselves to some "gold standard" evaluation metric, the philosophy adopted here is to stay open-minded, and approach the issue from three different perspectives. First, subjective interpretations of the learning results are made explicit by creating maps to available knowledge sources that can be assessed quantitatively; second, classes at each level of hierarchy are examined in the same way as in Chapter 3, thereby once again relating the hierarchy to the phonetic features; third, model parameters are also examined directly by waveform synthesis.

4.5.4 Evaluation of the broad class boundaries

From the figures shown in 4.5.2, it can be seen that although the broad classes cover a number of different phonemes, the time resolutions of the broad classes generally coincide with the segments identified by expert transcribers. To give a rough estimate of how the broad class boundaries are aligned with the TIMIT transcription, the following measures of accuracy are also calculated based on the best segmentation of each word. *Precision* is defined as the percentage of learned segmental boundaries that coincide with the TIMIT boundaries; while *recall* is defined as the percentage of TIMIT boundaries that correspond to learned representations. In order to accommodate the error caused by short-time processing, an error threshold of 3 frames is allowed in aligning the broad class boundaries with the expert transcription. The precision and recall values evaluated from two data sets, one training and one test, are included in Table 4.2. These numbers suggest that the learned representations are roughly made of segment-sized units.

	Precision	Recall
Training	75.6%	81.0%
Test	78.0%	77.1%

Table 4.2: Precision and recall on the segment boundary detection task

4.5.5 Evaluation of the broad classes

As a benchmark, the TIMIT phonetic transcriptions are not without problems for our purpose. Yet they provide the only available evaluation criterion for an unsupervised approach such as ours. Based on subjective interpretation of the broad classes, a correspondence between these classes and segment labels used in TIMIT is set up manually, as shown in Table 4.3[5]. It should be noted that such correspondences are made without regard to the specific phonetic contexts where these segments occur. As will be seen later, this may account for some of the errors in evaluation.

[5] These labels are different from the previous experiment in Chapter 3 since a different set of segments is used in the initial estimation.

111	fricative (f v th dh s sh z zh hh)
112	plosive (p t k b d g jh ch q epi)
12	nasal (em en nx eng ng m n)
211	central vowels (ae ay ux uh ah ax)
212	back sonorant (aa aw ao el r l w ow uw oy er axr)
22	front high sonorant (iy ih eh ey y ix)

Table 4.3: Correspondence between the broad classes and TIMIT segments based on subjective interpretation

Such correspondences are taken as the basis for constructing a benchmark for evaluating the broad classes. Since the broad classes and the TIMIT segments do not always agree in number, the broad class evaluation is done in a manner similar to continuous speech recognition: for each word, the learned broad class sequence is aligned with the benchmark sequence using dynamic programming, with the standard weights assigned on deletion, insertion and substitution[6]. The result on the training set, represented in a format similar to a confusion matrix, is included in Table 4.4:

	111	112	12	211	212	22	Deletion
Fric	776	96	119	24	17	30	146
Stop	52	1174	94	47	19	61	154
Nasal	4	9	558	28	22	16	72
Cent	3	10	43	493	64	91	71
Back	31	28	65	215	989	71	228
Front	11	8	93	114	13	1024	138
Insertion	232	152	401	280	78	264	

Table 4.4: Confusion matrix-type evaluation of the broad classes

The above table can be also summarized with the following statistics, which

[6] The same set of weights are the same as the ones used in the standard NIST scoring package: 3, 3, and 4 for insertion, deletion and substitution, repectively.

are also conventionally done in continuous speech recognition.

$$
\begin{aligned}
\text{correct\%} &= \frac{\text{Correct} - \text{Substitution} - \text{Deletion}}{\text{Total}} \\
\text{accurate\%} &= \frac{\text{Correct} - \text{Substitution} - \text{Deletion} - \text{Insertion}}{\text{Total}}
\end{aligned}
$$

We also report these summary statistics in Table 4.5. The results include both the training set and the test set. Since the alignment of learned and reference label sequences is sensitive to weights used in the dynamic programming procedure, these numbers should be taken as a qualitative, rather than quantitative assessment of the model.

	Correct%	Accurate%
Training	68.3%	49.3%
Test	62.6%	42.1%

Table 4.5: Summary statistics of broad class evaluation

4.5.6 Comparison of unsupervised and supervised approaches

A successful unsupervised learning model is expected to discover the regularities in the signal when it does not have access to such information in the learning stage. Therefore it may be informative to compare this scenario with a *supervised* one, where the same type of model and same number of parameters are used, but all the answers are explicitly given to the model. This type of comparison is useful for us to assess the power of the learning machine in the presence of all the necessary information that it is designed to discover, as well as the best generalization that can be achieved with the machinery.

In the experiments reported below, the fully supervised model has not only the segmental boundary information, but also the broad class labels that the unsupervised approach intends to learn. During the learning stage, each model is directly trained with the TIMIT segments (according to the broad class definition given in Table 4.3) with the standard Baum-Welsh algorithm, and then tested under the same condition as in the previous section. The results are shown below. In Table 4.6, the unsupervised results are the same as the training set in Table 4.2; in Table 4.7, the unsupervised results are the same as Table 4.5.

	Precision	Recall
Unsupervised	75.6%	81.0%
Supervised	79.4%	84.6%

Table 4.6: Unsupervised versus supervised learning on the segment boundary detection task (training set)

		Correct%	Accurate%
Training	Unsupervised	68.3%	49.3%
	Supervised	75.6%	61.7%
Test	Unsupervised	62.6%	42.1%
	Supervised	73.0%	60.8%

Table 4.7: Unsupervised versus supervised learning on the broad class identification task

Results from the supervised learning tasks are comparable to some of the previous benchmark results reported from supervised broad class classification (Juneja and Espy-Wilson, 2003)[7]. Although all the information was made available during supervised learning, Table 4.6 and Table 4.7 demonstrate that unsupervised learning performs reasonably well. In fact, the lack of superior performance indicates a fundamental limitation of the context-free broad classes. Conceivably, the overlap between these knowledge-based classes accounts for most of the error for both the supervised and the unsupervised learning results.

[7] Note that the definition of the broad classes used here differ slightly from those works.

4.5.7 Learning features without segmental boundaries

In Chapter 3, we have observed that the successive partitioning of sound categories can be seen as the discovery of acoustic phonetic features. Such observation is verified through the experiment conducted on the TIMIT database, with the assumption that the segmental boundaries are given for learning phonetic categories. Since our goal in this chapter is exploring the possibility of jointly learning phonetic categories and segmental representations at the same time, it is interesting to explore whether the partitioning of sound categories can still be interpreted as feature discovery, since the discovery of broad classes follows more or less the same procedure as before. Figure 4.6 shows a hierarchical structure of the 6 broad classes that have been derived using Algorithm 3. As in category learning experiments of Chapter 3, the leaf nodes are the actual classes, and the binary tree structure shows the history of partitioning.

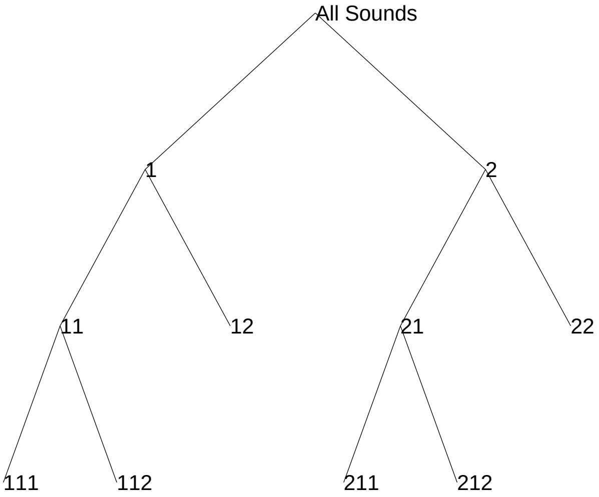

Figure 4.6: The 6 classes used in Algorithm 3.

It is not possible to conduct exactly the same assessment as was done in the previous clustering experiment, since the learned segments no longer stand in one-one correspondence with the phonetically-transcribed segments in TIMIT. In order to give a rough interpretation of the partitions, the learned representation is aligned with the expert transcriptions, and each TIMIT phone is set to correspond to the segment with the maximal overlap in time. For example, according to the representation shown for the word "nectar" in Figure **??**, the correspondence between the TIMIT phones and the learned categories is calculated as follows: [n]–12; [eh]–22; [kcl]–112; [t]–112; [axr]–211.

Using the procedure as described above, the distributions of TIMIT phones among the 6 categories are shown in the following figures. Similar to the order in Chapter 3, we start with the first partition of all sounds into approximants and non-approximants. The order of partitioning is not exactly the same as Chapter 3, due to a different initial condition. Therefore the partitions also differ somewhat from the previous ones. In places where they are comparable, the old results are also included below for the sake of comparison.

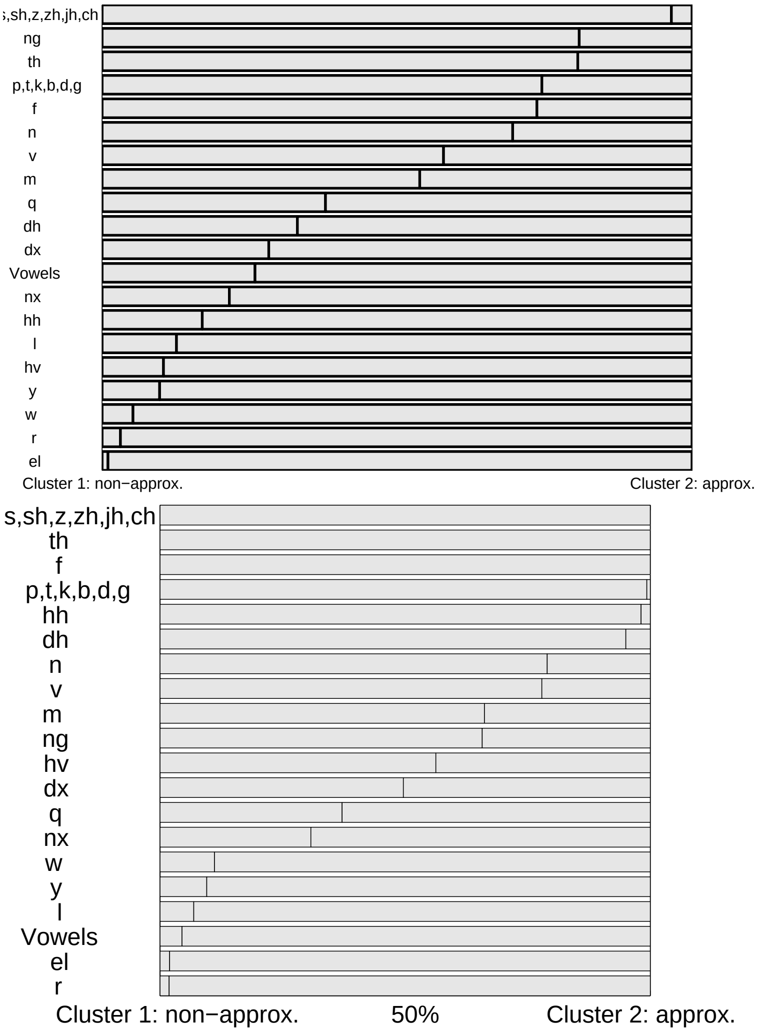

Figure 4.7: The partition of all sounds, into approximants and non-approximants. Top: when segmental boundaries are learned from data. Bottom: when segmental boundaries are given by TIMIT.

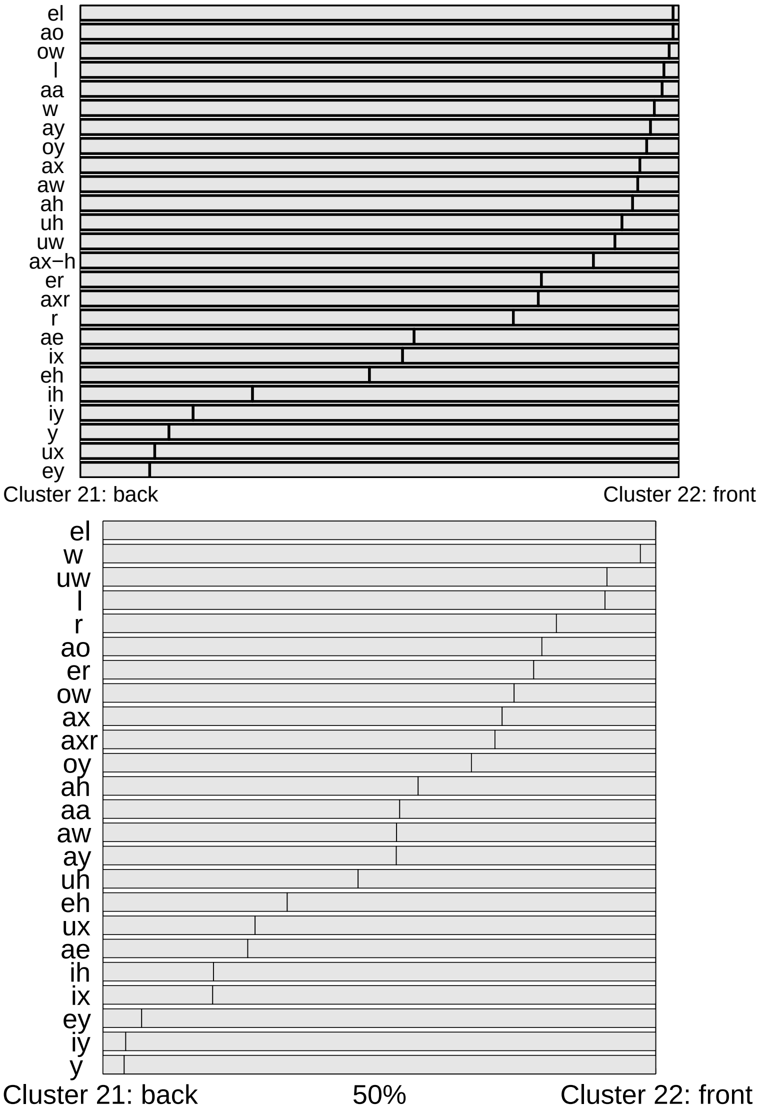

Figure 4.8: The partition of approximants. This partition mainly separates front high sonorants from the rest of the sonorants. Top: segmental boundaries learned from data. Bottom: segmental boundaries given by TIMIT.

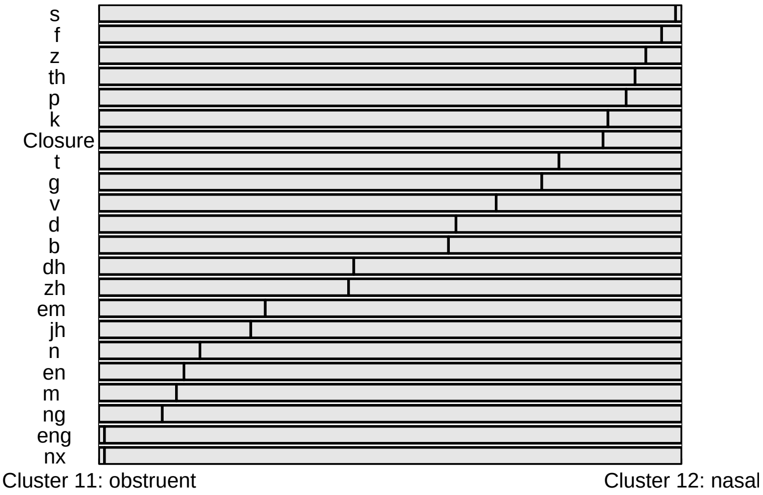

Figure 4.9: The partition of Cluster 1, into nasals and obstruents

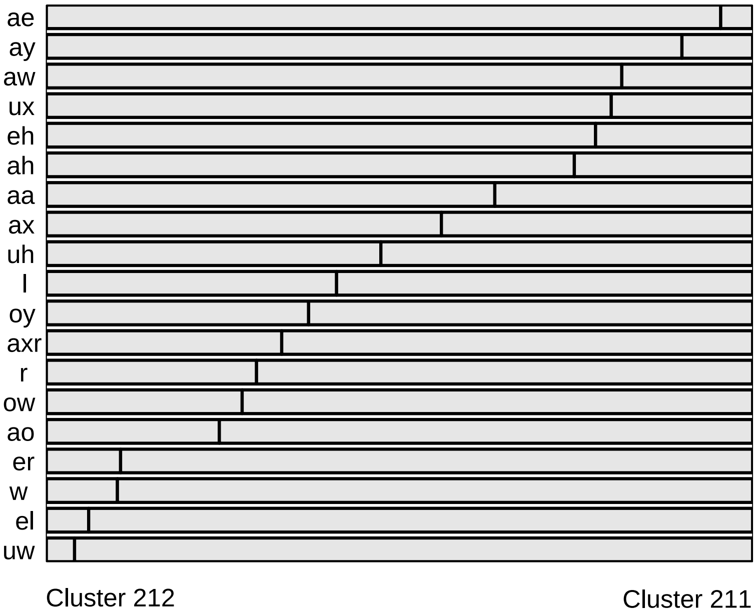

Figure 4.10: The partition of Cluster 21, the non-front high sonorants. This partition may again be interpreted as a feature that distinguishes backness of the sonorant.

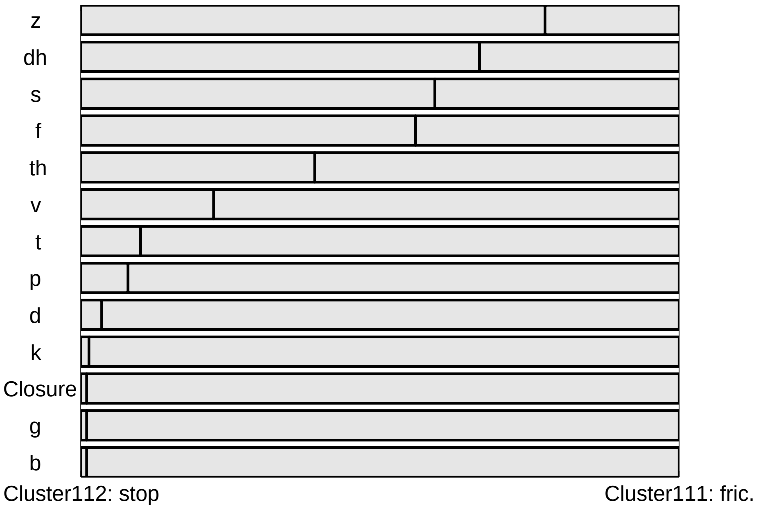

Figure 4.11: The partition of Cluster 11, into plosives and fricatives.

These results further support the view that features can be learned by inductively refining the phonetic categories, since learning only takes word-level acoustic signals as input. It is noteworthy that the classes that result from unsupervised learning without segmental boundary information are very similar to the previous ones obtained using such information. An interesting implication is that these classes may be the most robust categories that can be easily identified from the waveforms.

4.5.8 Learned phonotactics

The phonotactics, defined as the transition probabilities between the six broad classes, are displayed in Table 4.8. Since the database does cover contain a substantial number of English words, no attempt is made to evaluate phonotactics with a quantitative metric. For the interest of the reader, the learned transition probabilities and the ones derived from the TIMIT bechmark are shown in Table 4.8 and Table 4.9.

	Fric	Stop	Nasal	Centr	Back	Front
Fric	0.023	0.103	0.042	0.112	0.116	0.276
Stop	0.215	0.056	0.095	0.091	0.141	0.136
Nasal	0.074	0.149	0.023	0.070	0.093	0.224
Centr	0.128	0.228	0.206	0.058	0.132	0.097
Back	0.042	0.131	0.116	0.264	0.095	0.136
Front	0.099	0.147	0.232	0.157	0.071	0.060

Table 4.8: Phonotactics defined over the 6-class inventory learned from data

	Fric	Stop	Nasal	Centr	Back	Front
Fric	0.007	0.109	0.015	0.151	0.144	0.239
Stop	0.062	0.037	0.024	0.102	0.280	0.184
Nasal	0.073	0.169	0.001	0.079	0.106	0.159
Centr	0.213	0.323	0.159	0.010	0.079	0.014
Back	0.085	0.112	0.067	0.110	0.222	0.201
Front	0.177	0.180	0.189	0.022	0.107	0.046

Table 4.9: Phonotactics calculated from the benchmark

4.5.9 Synthesis from the learned models

The last evaluation that we would like to explore is trying to synthesize speech from these models, since it may give us an idea what acoustic properties have been captured by the models. However, two problems need to be solved before any waveform can be synthesized from the model. One is related to the front-end, while the other is related to the back-end. First, in the signal representation used in the model, only spectral information is preserved, and it is rather difficult to reconstruct speaker/voice information from the spectra (MFCC in particular) alone. Second, HMMs are not expected to perfectly characterize the distributions that they are fitted to, and directly generating random sequences from an HMM is unlikely to result in intelligible speech, especially with regard to segmental duration[8].

Therefore, we pursue a synthesis procedure that draws information from two different sources: the spectral information comes from the trained model, and the pitch and duration information come from the real speech. The motivation for carrying out such a test is to see whether our model captures the spectral information of the phonetic categories. The synthesis procedure from the learned models is described by the following algorithm:

Algorithm 4 Synthesis from the learned word models

```
Require: original waveform of a word $w$, the unit models and segmental representation of $w$ as the result of learning
Extract pitch from the waveform
Generate power spectra from the output distributions of the unit models
for each short time window do
    Calculate the amplitude, frequency and phase of a set of sinusoidal from the power spectra and pitch
    Perform short-time synthesis using sinusoidal harmonics
end for
Synthesize the output signal using overlap-add method
```

There are a variety of pitch extraction algorithms available. In our experiment, we have used the one by Sun (2002); while the smoothed power spectra reconstruction uses the inversion technique reported in Milner and Shao (2002; 2000). The idea of sinusoidal synthesis comes from McAuley and Quatiery, (1986), and a simplified version is used in the current evaluation. In order to overcome the between-frame discontinuities caused by the conditional independence assumption of HMM, the dynamic features used in the front-end are incorporated in reconstructing the optimal output sequence of spectra, based on ideas from recent work on HMM-based speech synthesis (Tokuda, Kobayashi, and

[8] Depending on the model structure, the distribution of segment duration for an HMM-based model is generally some convolution of geometric distributions, which state-wise duration observes.

Imai, 1995; Tokuda et al., 2000). More details of these methods, including the equation used to generate power spectra, are included in Appendix 4.E.

The experimental data set consists of 283 isolated words, selected from a list of 50 most frequent lexical entries reported from parent surveys (Fenson et al., 1991). The words are recorded from a female speaker, who imitates a child-directed speech style in a sound booth. These data are used as the input to the learning procedure described in Section 4.5.1. As the output, a set of unit models is obtained as well as the representations for each word signal. Figure 4.12 shows three power spectra synthesized from the 3 different representations learned from the data, each associated with a different phonetic inventory:

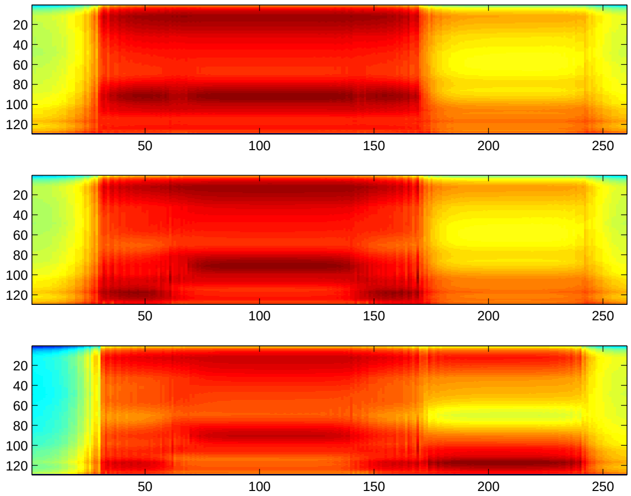

Figure 4.12: Synthesized power spectra from the learned models of "mommy" using static and dynamic features. Number of units = 2, 4, 6

As can be seen from the figure, as the number of units increases, the spectral/temporal resolution of the learned representation also improves. With a synthetic pitch signal, these power spectra can also be used to generate waveforms that increasingly resemble real speech[9]. Two more examples, synthesized

[9] Although the contribution of pitch to the quality of the synthesis should be carefully excluded from the evaluation.

from the models for “thank you” and “grandpa”, are included in Figures 4.13 and 4.14. Each of these models is composed by a concatenation of smaller unit models. The spectral change between the units can be seen from the changing intensity of the pixels.

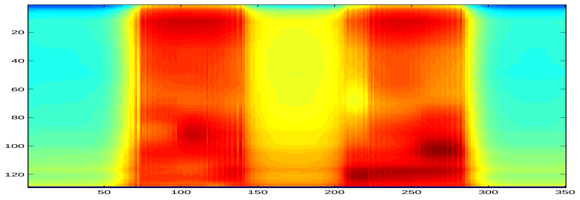

Figure 4.13: Synthesized power spectra for “thank you” using static and dynamic features. Number of units = 6

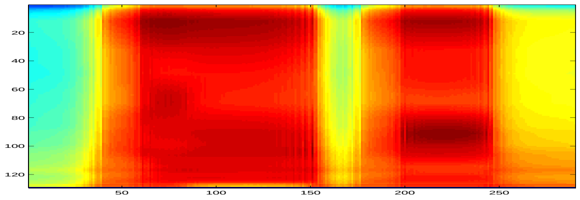

Figure 4.14: Synthesized power spectra for “grandpa” using static and dynamic features. Number of units = 6

4.6 Summary and discussion

The basic finding of this chapter is that segment-sized units that roughly correspond to broad phonetic classes can be discovered from waveforms, using an iterative learning strategy. Without the segmental boundary information, the learned features are comparable to those obtained from clustering the TIMIT segments, and they also form a hierarchy that supports the proposed mechanism of feature discovery. Moreover, synthesis experiments further confirm that the unit models individually characterize spectral properties of the target broad classes.

Due to the difficulty of evaluating the results of unsupervised learning, much of the effort in this chapter was spent on finding a proper evaluation procedure for the unsupervised learning. Although it may be possible to set a gold-standard for evaluation based on some type of expert transcription, it should be noted that we do not know whether a child would represent a spoken word the same as adults do. Instead, an appropriate evaluation metric must be based on understanding of the development of lexical representations, and this is another scientific question that requires extensive research by itself. An examination of the literature reveals two contrasting views: one sees the early lexical representations as completely "holistic", thereby leading to the inability of children to distinguish words (Jusczyk, 1986); while the other posits a lexical representation that is detailed enough to distinguish two words that differ by one feature (Gerken, Murphy, and Aslin, 1995). However, neither has made predictions about what kind of units a child might use for her lexical representations. The learning strategy outlined in our model suggests a line between those two views: the representations that have been learned by the algorithm are segmental, but do not capture as many details as the adult phonology would contain. Whether such an intermediate position is in accord with the "protracted process of phonological acquisition" (Gerken, Murphy, and Aslin, 1995) remains to be tested.

One may expect that a more detailed phonetic inventory might result if we proceed with the partitioning process, rather than stopping at an arbitrary inventory size (say 6). This is not the case. Empirically, it is observed through experiment that partitioning the broad classes even further often results in clusters that do not correspond to finer natural classes, and an adult-like phoneme set seems far out of reach. Therefore our model is inevitably faced with the obvious question: "when to stop"? In the rest of this thesis, we would like relate this question to the lexicon, and the next chapter will explore models of the lexicon in more detail.

Finally, we would like to note a subtle point related to the formulation of the learning problem. In Algorithms 2 and 3, segmentation is treated as a parameter. However, the meaning of segmentation is different from units and phonotactics, since the latter two are clearly established through experimental research, and can potentially be quantified through certain kinds of measurements. Hence it is somewhat unsatisfactory that segmentation remains as a parameter[10]. In principle, this could be avoided by treating segmentation as "missing data" and trying to use the EM algorithm for learning. However, since it is most natural to view the computation of segmentation as an operation of finding a maximum (or mode of a distribution), segmentation is a quite different operation from one that finds an average over a set of possibilities, as an E-step would generally imply. Moreover, averaging over all possible segmentations is computationally expensive when a dynamic programming scheme is not available. After all, given all other information in the model, if the uncertainty with regard to segmentation is small, then the maximization strategy is unlikely to differ greatly from the expectation-type ones, since the two operations should yield similar results (see Appendix 4.D for a formal discussion).

[10] In statistics literature, such parameters are called *nuisance* parameters.

4.A The use of alternating maximization in learning with segmentation as a parameter

We would like to show that the method used in Section 4.2 always increases the likelihood of the function (4.1) and is thus guaranteed to converge. In the notation used below, $\{w_i\}$, U, $\{S_i\}$ will be used for words, units, and segmentation, respectively. The superscript $^{(t)}$ denotes the values of the parameters at iteration t, and $u^{(t)}$ stands for the (vector-valued) assignment of the unit models at t, $s_i^{(t)}$ for the segmentation of word i at t.

Suppose at $t+1$, the segmentation of each word is set to $s_i^{(t+1)}$, i.e. :

$$s_i^{(t+1)} = \arg\max_{s_i} p(w_i|U = u^{(t)}, S_i = s_i)$$

By (4.A), $p(w_i|U = u^{(t)}, S_i = s_i^{(t+1)}) \geq p(w_i|U = u^{(t)}, S_i = s_i^{(t)})$, with equality holds iff. $s_i^{(t+1)} = s_i^{(t)}$. On the other hand:

$$p\left(\{w_i\}|U = u^{(t)}, \{S_i = s_i^{(t)}\}\right) = \prod_i p(w_i|U = u^{(t)}, S_i = s_i^{(t)})$$

Thus we have:

$$p\left(\{w_i\}|U = u^{(t)}, \{S_i = s_i^{(t+1)}\}\right) \geq p\left(\{w_i\}|U = u^{(t)}, \{S_i = s_i^{(t)}\}\right)$$

Similarly, we can also show if the unit models are updated at a given iteration $t+1$, then:

$$p\left(\{w_i\}|U = u^{(t+1)}, \{S_i = s_i^{(t)}\}\right) \geq p\left(\{w_i\}|U = u^{(t)}, \{S_i = s_i^{(t)}\}\right)$$

In other words, each step in the algorithm always increases the likelihood. The argument applies to any finite number of parameter subsets, since each subset can be regarded as a separate coordinate along which optimization is carried out.

4.B Viterbi algorithm

The Viterbi algorithm intends to solve the following problem: given a sequence of observations $o_1, \cdots, o_T$, and an HMM θ, find the state sequence $q_1, \cdots, q_T$, such that:

$$(q_1, \cdots, q_T) = \arg \max_{q_1, \cdots, q_T} p(o_1, \cdots, o_T | q_1, \cdots, q_T, \theta)$$

The key to solving this equation efficiently is to use the dynamic programming method based on the recursive relation:

$$p(o_1, \cdots, o_{t+1} | q_1, \cdots, q_{t+1}, \theta) = \max_q p(o_1, \cdots, o_t | q_1, \cdots, q_t, \theta) \cdot a_{q_t, q} b_q(o_{t+1}) \quad (4.6)$$

In other words, the best state sequence of length $t+1$ is obtained after considering all different ways of extending the best sequence of length t one step further. Hence the above recursive definition lets us work "backwards", and reduce the size of the problem as follows:

$$\begin{aligned} \delta(1, m) &= b_1(m) \\ \delta(t+1, m) &= \max_j \left(\delta(t, j) \cdot a_{j,m} \cdot b_{t+1}(m)\right) \end{aligned}$$

When $t = T$, the dynamic programming programming terminates, and the best sequence can be recovered by tracing back each state q that maximizes (4.6).

4.C Equations used in Algorithm 2 and Algorithm 3

Using the superscript $^{(t)}$ for the values of a parameters at iteration t, the main update equations for Algorithm 2 are the following:

$$\begin{aligned} units^{(t+1)} &= \arg\max_u p\left(words|units = u, segmentation_i^{(t)}\right) \\ segmentation_i^{(t+1)} &= \arg\max_{s_i} p\left(words|units^{(t+1)}, segmentation_i = s_i\right) \end{aligned}$$

Here arg max is understood to be a local maximum. Hence the objective (4.1) in Section 4.2 is equivalent to maximize:

$$\log p\left(words|\mathbf{units}, segmentation_i^{(t)}\right) = \sum_i \sum_{j=1}^{k(i)} \log p(segment_i^j|\mathbf{units}) \tag{4.7}$$

Renumber these segments as $segment_1, \cdots, segment_K$, and note $unit = \{\lambda_j, \theta_j\}$, i.e. a mixture model, then (4.7) can be maximized by the EM algorithm, which again includes two equations:

$$\begin{aligned} Z_i^{(k+1)} &= E\left[Z_i \middle| unit^{(k)}, \{segment_i\}\right] \\ unit^{(k+1)} &= \arg\max_u p\left(\{segment_i, Z_i = Z_i^{(k+1)}\} \middle| unit = u\right) \end{aligned}$$

Details of these update equation can be found in the appendix of Chapter 2.

4.D Hidden variables as missing data versus as parameter

Suppose we are interested in predicting data x with a model θ, yet need the help of a hidden variable z in the calculation. There are two choice of incorporating z:

1. Include z as missing data, and do EM:

$$\begin{aligned} z &\leftarrow E[z|x,\theta] \\ \theta &\leftarrow \text{maximize}(x,z|\theta) \end{aligned}$$

2. Include z as parameter, and do two maximizing steps:

$$\begin{aligned} z &\leftarrow \text{maximize}(x|\theta,z) \\ \theta &\leftarrow \text{maximize}(x|\theta,z) \end{aligned}$$

The θ update steps are often the same, since z usually clarifies the structure of x in a way that makes the calculation of $p(x, z = z_0|\theta)$ and $p(x|\theta, z = z_0)$ identical. Now suppose $p(z|x, \theta)$ has very small uncertainty, i.e., approximately a point mass at $\delta(x, \theta)$ (assuming some appropriate parametrization of z), then:

$$\begin{aligned} E[z|x,\theta] &= \int_{\chi} p(z|x,\theta) \cdot z \, dz \\ &\approx 1 \cdot \delta(x,\theta) \end{aligned} \quad (4.8)$$

On the other hand, since:

$$p(z|x,\theta) = \frac{p(x|z,\theta)p(z|\theta)}{\sum_z p(x|z,\theta)p(z|\theta)}$$

If $p(z|x, \theta)$ tends to a point mass at $\delta(x, \theta)$, then assume $p(z|\theta)$ is uniform, then $\frac{p(x|z,\theta)p(z|\theta)}{\sum_z p(x|z,\theta)p(z|\theta)} \rightarrow 1$ implies $\delta(x, \theta) = \arg\max_z p(x|z, \theta)$, i.e. the E-step for the missing data z and the direct maximization over parameter z are expected to return similar results.

4.E Synthesis techniques used in evaluation

4.E.1 HMM-based synthesis using dynamic features

The main idea in Tokuda et al. (1995) is to consider synthesis not as sampling from a distribution, but as the solution of the following optimization problem:

$$\arg\max_{O} P(O|\lambda, Q) \tag{4.9}$$

where λ is the model parameter, Q is the state sequence of length T, and O is a sequence of observation with dimension, each coming from the output distribution on the state of Q. Hence O can be thought of as a matrix of dimension $M \times T$. The key to solving (4.9) is by incorporating the dynamic features indirectly as constraints on the static features. Specifically, rearranging terms of O as a column vector, this constraint can be written as a matrix multiplication:

$$O = W \cdot C \tag{4.10}$$

O is the whole observation, C is the static features, and W is a differential operator that works on each dimension of C at different times to produce W. Notice O and C must be rearranged as vectors in order to make this work. Let we let T be the duration in terms of frame number, and M be the cepstrum order, then O is $2TM \times 1$, C is $TM \times 1$, and W is $2TM \times TM$.

In synthesis, the term directly related to power spectra is C. So the problem becomes:

$$\arg\max_{C} P(W \cdot C|\lambda, Q) \tag{4.11}$$

Here W, λ and Q are assumed to be known. Assuming terms have been arranged appropriately, this reduces to solving:

$$\frac{\partial}{\partial C}(W \cdot C - M)^T \Sigma^{-1}(W \cdot C - M) = 0 \tag{4.12}$$

Expanding and using the matrix calculus, this produces a linear term and a constant term:

$$2 \cdot W^T \Sigma^{-1} W \cdot C - 2W^T \Sigma^{-1} M = 0 \tag{4.13}$$

The problem reduces to solving a linear system of a large order, once appropriate representations of W, Σ and M are given. Hence the key step is arranging the term appropriately so the relation (4.10) goes through. The following is a possible arrangement of the terms that differs slightly from the original paper:

Let $c_t(m)$ be the m-th order cepstrum at time t. So in theory, there are $T \times M$ such terms in C. A way to write down C in terms of its components is first going by time: $1, \cdots, T$, then going by cepstrum order $1, \cdots, M$.

$$\begin{aligned} C &= [c_1(1), c_2(1), \cdots, c_T(1), \\ &\quad c_1(2), c_2(2), \cdots, c_T(2), \\ &\quad \cdots \\ &\quad c_1(M), c_2(M), \cdots, c_T(M)]^T \end{aligned}$$

The advantage of using this notation is that W is relatively simple to write down, since W is an operator that works on each dimension of C. Following the same enumeration order as C, O can be written as a $2TM \times 1$ vector (assuming Δ features are used. Δ^2 can be treated similarly):

$$\begin{aligned} O \;=\; & [c_1(1), c_2(1), \cdots, c_T(1), \Delta c_1(1), \Delta c_2(1), \cdots, \Delta c_T(1), \\ & c_1(2), c_2(2), \cdots, c_T(2), \Delta c_1(2), \Delta c_2(2), \cdots, \Delta c_T(2), \\ & \cdots \\ & c_1(M), c_2(M), \cdots, c_T(M), \Delta c_1(M), \Delta c_2(M), \cdots, \Delta c_T(M)]^T \end{aligned}$$

So every $2T$ dimensions of O can be obtained from T dimensions of C, and W can be constructed as a block diagonal matrix with the following block:

$$W^0_{2T\times T} = \begin{bmatrix} I_{T\times T} \\ D_{T\times T} \end{bmatrix} \tag{4.14}$$

Here $I_{T\times T}$ is the identity matrix, and $D_{T\times T}$ is a differential operator that applies to each $(c_1(i), c_2(i), \cdots, c_T(i))$, $i = 1, \cdots, M$. So W has the following structure:

$$W = \begin{pmatrix} W^0_{2T\times T} & 0 & \cdots & 0_{2T\times T} \\ 0_{2T\times T} & W^0_{2T\times T} & & \cdots \\ 0 & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & W^0_{2T\times T} \end{pmatrix} \tag{4.15}$$

For fast solution of (4.13), one needs to observe that $W^T\Sigma^{-1}W$ is positive semi-definite and symmetric, therefore having a Cholesky decomposition:

$$W^T \Sigma^{-1} W = D^T \cdot D \tag{4.16}$$

D is upper-triangular, and also a sparse matrix. Therefore the answer to (4.13) comes from solving two systems, from left to right – one is $D^T \cdot (D \cdot C) = W^T \Sigma^{-1} M$, the other is $D \cdot C = D^{-T} \cdot W^T \Sigma^{-1} M$.

4.E.2 Sinusoidal synthesis

HMM-based synthesis produces time-smoothed MFCC sequences. From these, power spectra can be reconstructed by simply inverting all the steps in computing MFCCs. The reconstructed power spectrum is a smoothed version of the original, due to the information loss in the cepstral domain (recall the result of discrete Cosine transform is truncated to obtain the MFCCs used in the front-end). Moreover, since phase information is not preserved, perfect reconstruction is not possible.

Sinusoidal synthesis is a method developed in speech coding that tries to reconstruct speech with a summation of sine tones. A simplified version similar to the one used in Chazan (2000) is used in the current work. Within each short-time window, the reconstructed signal is expressed as:

$$x(t) = \sum_{i=1}^{N} A_i cos(\omega_i t + \phi_t) \tag{4.17}$$

where ω_i is the frequency of each component. For voiced frames, $\omega_i = i \cdot f_0$, i.e. frequency of the harmonics; for voiceless frames, ω_i is set to random. A_i can be estimated from the reconstructed power spectrum. ϕ_t is the phase term to ensure smooth transition between frames, and depends on the frequency and the frame index.

CHAPTER 5

Learning a lexicon from waveforms

In Chapter 3 and 4, we focused on the problem of pre-lexical learning, i.e. the problem of identifying units and phonotactics without a lexicon. In other words, the information about the specific lexical entry for each instance of a spoken word is not available to the model. This type of learning can be thought of as bottom-up, since the concept of a "word" did not exist in the model. Even though the model did not know what words are, the results from previous chapters showed that the model still identifies segment-sized units and features during the course of model refinement.

However, by forcing phonological acquisition to be pre-lexical, we may be making the problem harder than the one the child is solving. For example, when a baby hears "toy", there is a good chance that she can perceive the presence of an object the word is referring to. Therefore waveforms can very well be related to concepts represented by the spoken words, and such information should also be made available to the model.

The problem to be addressed in this chapter is how a lexicon can constructed from waveforms. Building upon the work in the previous chapters, our approach is again incremental: we consider adding a lexical component to the bottom-up learning model, and extending the learning algorithm to incorporate the updating of lexical entries. However, unlike the previous chapters, the discussions in this chapter is more speculative and consists of a few disjoint, yet related, topics.

First, we present a view of the lexicon based on mixture models. We then proceed to discuss a few key issues related to the lexicon: namely, how a lexical entry can be formed using the segmental representations of words; how lexical access can proceed with such a lexicon; how statistical theories can inform us about the nature of lexical neighborhoods; and how the refinement of the model can be driven by lexicon-level optimization.

5.1 Lexical items as mixture models

Up until now, the input data were assumed to be "holistic words", and the goal of learning segments was stated as:

> Find units and phonotactics, such that the function
> $p(words|units, phonotactics, segmentation)$ is optimized.

In order to calculate this function, the probability of each word is assumed to be independent of others:

$$p(words|units, phonotactics, segmentation) =$$
$$\prod_i p(word_i|units, phonotactics, segmentation)$$

Therefore, even if two holistic words were instances of the same lexical entry, the model did not make use of such information. Metaphorically, it may be thought of as a passive learner that has no other cognitive functions except for hearing. Given a few examples of "cheese", the learner pays attention to the similarity between the individual sounds within each word, yet does not take advantage of the global similarity among the sound sequences underlying the words because the concept of a lexical item does not exist. If any attempt is made to relate this model to language development, it thus seems most appropriate to refer to this type of learning as *pre-lexical*, in other words, a stage of phonological acquisition without the effect of a lexicon.

On the other hand, since each word is treated as independent of others, there is in general no guarantee that different instances of the same lexical item will have the same segmental representation. By following a completely bottom-up approach, the segmental representations for each holistic word are determined solely by properties of the signal and the current state of the units and phonotactics. To give some examples of the lexical variation that exists in the learning data, let's consider some of the segmental representations that have been assigned to two words from the TIMIT database (Figure 5.1 and Figure 5.2):

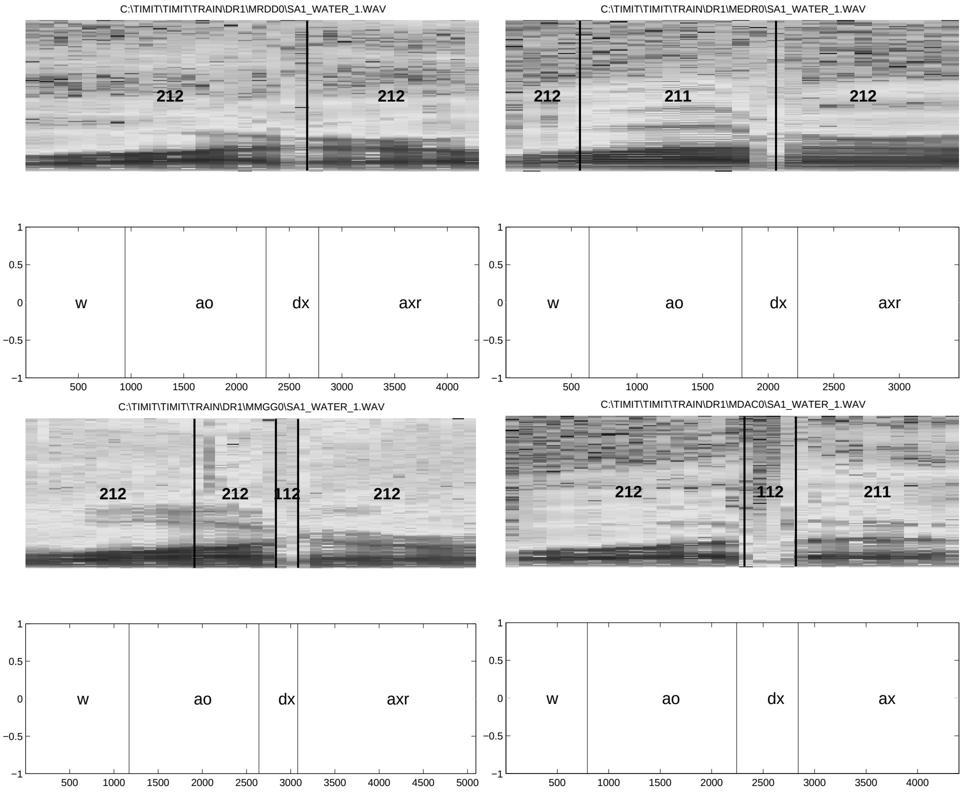

Figure 5.1: Four different tokens of "water" in TIMIT

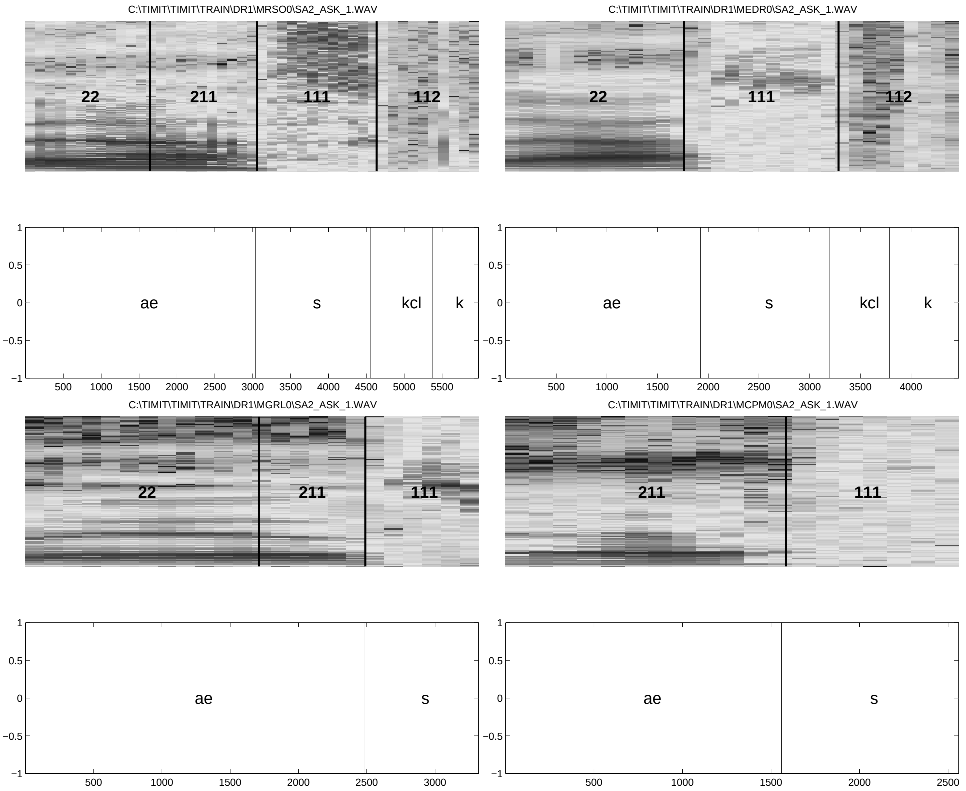

Figure 5.2: Four tokens of "ask" in TIMIT

As seen from the figures, representations resulting from bottom-up learning are evidently sensitive to the variation in the waveforms. Depending on whether an amplitude drop is present between the two sonorants, different representations are assigned to the instance of "water". The different representations for "ask", on the other hand, are mostly due to different degrees of coda deletion. Beyond the case of flapping and coda deletion, the existence of lexical variation seems widespread (Keating, 1998). For example, given a vocabulary of 838 words (with varying frequencies in the TIMIT training data), a total of 1452 different segmental representations were found. Considering the fact that those TIMIT words are extracted from continuous speech, the larger number of learned representations is partly due to context-dependent allophones. However, compared to the 1032 representations derived from TIMIT's allophonic transcriptions, much of that variation occurs at a level of phonetic details that are finer than allophones, henceforth referred to as the sub-allophonic level. To illustrate, consider the transcriptions of "water" in TIMIT versus the results from bottom-up learning:

```
1 WATER  [|w|aa|dx|axr|]
1 WATER  [|w|aa|dx|er|]
1 WATER  [|w|ao|dx|ah|]
1 WATER  [|w|ao|dx|ax|]
18 WATER [|w|ao|dx|axr|]
3 WATER  [|w|ao|dx|er|]
```

Table 5.1: Lexical representations for "water" in TIMIT transcriptions.

```
5 WATER [|212|]
1 WATER [|212|112|211|]
1 WATER [|212|112|211|212|]
2 WATER [|212|112|212|]
1 WATER [|212|211|]
1 WATER [|212|211|211|]
4 WATER [|212|211|212|]
1 WATER [|212|211|212|211|]
1 WATER [|212|211|212|212|]
6 WATER [|212|212|]
1 WATER [|212|212|112|212|]
1 WATER [|212|212|211|211|]
```

Table 5.2: Lexical representations for "water" in the result of bottom-up learning.

Implications of acoustic variation must be taken into account when we consider the problem of building a lexicon. Based on the more traditional view of the lexicon, each lexical entry or underlying form is a string of units, and phonology is responsible for modifying these strings with various kinds of rules or constraints. However, such a view of the lexicon relies heavily on top-down knowledge when we consider the problem of learning – the discovery of lexical representations from acoustic data. As is well documented in the literature (e.g. Bell and Jurafsky (2003)), spoken words are often pronounced differently from their citation forms in continuous speech, therefore significantly deviating from what is suggested by the underlying representations. From a learner's point of view, discovering the (presumably) invariant underlying form involves a critical *induction* step, and top-down knowledge needs to override bottom-up evidence. Since children's lexicons seem to differ from adults' in significant ways (c.f. the discussion in Chapter 1), it is rather unclear what kind of top-down knowledge would be appropriate in order to achieve such robust generalization.

Perhaps motivated by the desire to pay more attention to lexical variation, some current researchers have adopted a view of the lexicon as "episodic" (Goldinger, 1997), or based on exemplars (Pierrehumbert, 2001). Typically, the postulate of lexical exemplars in memory is meant to explain various types of short-term memory effects; an important aspect of learning in such a model is the storage of large amount of exemplars. In abandoning the rigid, invariant underlying forms, exemplar-based ideas seem to swing to the opposite end of the spectrum, and hence leave open the problem of *generalization* (see Section 5.4.4 for more discussions).

The stance taken in the current work with regard to lexical variation can be considered intermediate between the above two approaches. Our approach is again based on the mixture model – the same framework used for learning phonetic categories. As can be recalled from Chapter 2, under the mixture model assumption, the lexical variation of each entry is seen as generated by discrete sources. In this case, each component[1] in the lexical mixture is a concatena-

[1] For convenience, we will refer to the mixture components as "exemplars". Therefore, for us exemplars are always models (HMMs) constructed by composition, therefore they should not be confused with other uses in the literature, such as raw acoustic trace in memory, etc.

tion of atomic unit models as specified by the segmental representation, and the weight on each component can be derived from the frequency of each segmental representation in the data. Compared to the two views above, the mixture model entertains lexical variation, yet not at the cost of missing generalization: the multiple components within the mixture describe the variation of each entry, and generalization is achieved by composition – the reuse of unit models in the higher level. Compared to the "massive storage" proposal, the model captures variation by summarizing important statistics from the exemplars rather than storing them all.

For example, given the set of representations learned for "water", the mixture model for the lexical entry **water** will have the following form:

$$p(word|\mathbf{water}) = \sum_i \lambda_i p(word|\text{water-exemplar}_i)$$

where $exemplar_i$ stands for the i-th model composed from the unit sequence, and λ_i stands for its relative frequency ($\sum_i \lambda_i = 1$). As previously shown, both the compositional model and the frequency can be obtained from the result of bottom-up learning. Here the higher-level lexical knowledge is embodied as the prior probability over the lexical exemplars. As a consequence, frequent exemplars will have more influence on the lexical entry than less frequent ones.

5.2 Learning the lexicon together with units and phonotactics

Once the bottom-up learning is complete (the algorithm converges), a lexical model based on mixtures of exemplars can be constructed from the individual representations of the holistic words: we simply collect all the possible representations for each lexical entry, and assign the mixing weights to be the relative frequency of each exemplar. However, we have not answered the question:

What good can a lexicon do for learning?

Although lexical mixtures are formed by summarizing the results of bottom-up learning, the lexicon still plays no role in learning. The learner does not yet have the concept of a "word" when listening to the waveforms, and he simply has no way of relating two instances of "water" in learning units and phonotactics.

However, the situation can be changed if the weights of the lexical mixture are treated as part of the model parameter. Once this is done, the effects of the lexicon on other components would occur within the optimization process, since the lexicon needs to be jointly optimized together with the other parameters. The specific mechanism for integrating the lexicon lies behind the familiar scenario: on the one hand, the frequency of the lexical exemplars should have a say in deciding the representation of each word, since the prior knowledge would prefer more frequent exemplars; on the other hand, frequency also needs to be updated when a new set of exemplars is available. As has happened a number of times with regard to the learning of units and phonotactics, this type of interdependence is a cue for designing a strategy to integrate the lexicon into the model.

In particular, taking the bottom-up model $p(words|units, phonotactics, segmentation)$ as a starting point, what the above suggests is augmenting the model with the mixture-based lexicon, i.e. constructing a model with one more set of parameters. The structure of the augmented model is illustrated below. As shown in Figure 5.3, each lexical entry is a mixture of lexical exemplars, whereas each exemplar is composed of atomic unit models. Therefore the two levels have different meanings: the first level represents a hierarchical structure, while the second indicates a compositional structure of the lexical exemplars.

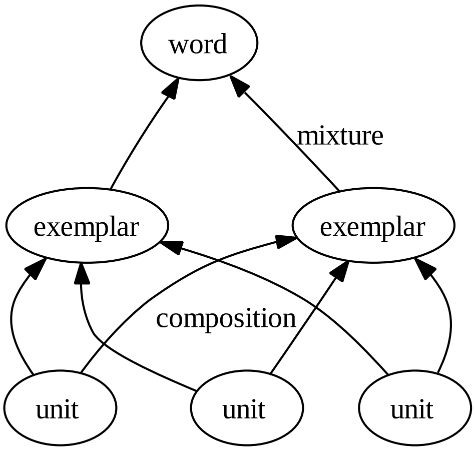

Figure 5.3: The structure of a lexical model.

With the additional lexical component, the optimization procedure can be extended as follows:

1. Segment each word using the waveform, the phonotactics, and the lexical mixture to which the word belongs;
2. Update phonotactics by counting the unit sequences;
3. Update the unit models with the result of segmentation;
4. Update the mixture weights in the lexicon with the exemplar counts.

In Step 1, the search space for segmentation consists of all the exemplars for the given lexical entry, and the mixture weights are used together with phonotactic probabilities in constructing the finite state network. An illustration of the composition of the lexicon, phonotactics and units in searching for the segmentation is shown in Figure 5.4.

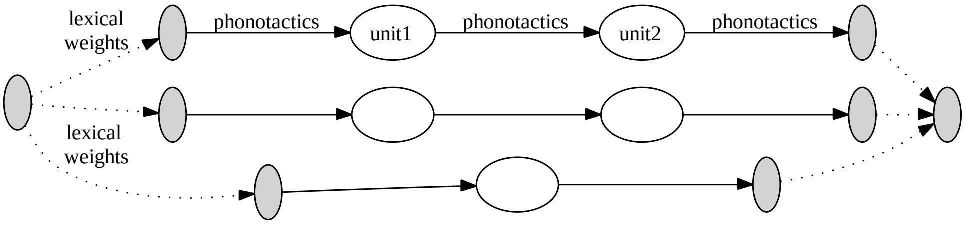

Figure 5.4: Composition of the lexicon, phonotactics and units in constructing the search space. The parallel paths depend on the number of exemplars for each lexical entry.

The filled nodes do not have any output distribution associated with them, and the transparent nodes are the unit models, which themselves are HMMs. The dotted lines represent transitions with probabilities assigned to the lexical knowledge – the weights of the lexical mixture. The solid lines are the transitions whose probabilities are inherited from the phonotactic model.

Since the search is limited within the set of lexical exemplars for each entry, this method can also be understood as *re-clustering* of words using the lexical mixture, and Step 1 may be seen as assigning each word to an exemplar – the classification step in a clustering scheme. Notice the phonotactic model is the same Markov chain with a fully connected state space as in Chapter 4. Without the use of the lexicon, the search space would have been the same Markov chain as the one used for learning phonotactics. Hence what Figure 5.4 illustrates is the use of lexical information in reducing the search for a segmentation, instead of a graphical model describing the joint distribution[2].

After the initial classification/re-segmentation of words, Steps 2 and 3 perform the same computations as in the previous bottom-up learning procedures. Step 4 re-calculates the weights in the lexical mixture. Taken together, Steps 2, 3 and

[2] Otherwise the model would not be consistent, since the transitions leaving a node do not sum up to 1.

4 correspond to the model update step of a clustering scheme. Repeating these steps again brings the likelihood function to a local maximum. For a more formal presentation, the reader is referred to Appendix 5.A, where the update equations and the objective function are explained in greater detail.

Algorithm 5 provides an implementation of Steps 1 – 4. Compared to Algorithm 3, it adds the steps for building a lexicon and using the lexicon to guide a search, as described above. As part of the requirement, bottom-up learning is used to provide the initial conditions for Algorithm 5. When bottom-up learning reaches convergence, the initial values for unit models and phonotactics will be set to the ones obtained from bottom-up learning, and the initial lexical mixtures can be calculated as described in Section 5.1.

Algorithm 5 Learning phonotactics, units and lexicon

```
Require: an initial estimate $units^{(0)}$ for the unit models: $unit_1, \cdots, unit_M$;
Require: initial phonotactics $phonotactics^{(0)}$;
Require: lexical entry for each $word_i$: $entry_i = \{S_{i,1}, \cdots, S_{i,M(i)}\}, i = 1, \cdots, K$,
    where each $S_{i,j}$ is a composition of units
Require: initial weights of lexical exemplars $lexicon^{(0)} = \{\lambda_{i,j} : i = 1, \cdots, K; j = 1, \cdots, M(i)\}$, $K$ = number of entries, $M(i)$ = number of exemplars for entry $i$;
Require: $\epsilon > 0$
$t \leftarrow 0$
repeat
    set of segments $Y \leftarrow \phi$
    for each $word_i$ do
        $(Z_i, segmentation_i^{(t)})$
        $\leftarrow Viterbi(word_i, entry_i, units^{(t)}, phonotactics^{(t)}, lexicon^{(t)})$
        $\{segment_i^1, \cdots, segment_i^{k(i)}\} \leftarrow (word_i, segmentation_i^{(t)})$
        $sequence_i \leftarrow segmentation_i^{(t)}$
        $Y \leftarrow Y \cup \{segment_i^1, \cdots, segment_i^{k(i)}\}$
    end for
    $phonotactics^{(t+1)} \leftarrow \text{Transition}(\bigcup_i \{sequence_i\})$
    for each $entry_s$ do
        for each lexical exemplar $j = 1, \cdots, M(s)$ do
            $\lambda_{s,j} = \left( \sum_{word_i \sim entry_s} Z_i^j \right) \Big/ \left( \sum_{word_i \sim entry_s} \sum_j Z_i^j \right)$
        end for
    end for
    $lexicon^{(t+1)} \leftarrow \{\lambda_{s,j} : s = 1, \cdots, K; j = 1, \cdots, M(s)\}$
    $k \leftarrow 0$;
    repeat
        for each $y_i \in Y$ do
            $z_i \leftarrow \text{E-step}(y_i, units^{(k)})$
        end for
        $units^{(k+1)} \leftarrow \text{M-step}(\{z_i\}, Y)$
        $k \leftarrow k + 1$
    until $\log p(Y|units^{(k+1)}) - \log p(Y|units^{(k)}) < \epsilon$
    $units^{(t+1)} \leftarrow units^{(k)}$
    $t \leftarrow t + 1$
until $\log(likelihood^{(t+1)}) - \log(likelihood^{(t)}) < \epsilon$
```

In order to see the effects of introducing a lexicon, the result of running Algorithm 5 is compared with that of bottom-up learning. The database is the same TIMIT subset that is used in the previous chapter, and the bottom-up learning has also been reported in Chapter 4. Table 5.3 shows the results on the segment boundary detection task. Table 5.4 shows the results on the broad class identification task.

	Precision	Recall
Bottom-up	75.6%	81.0%
Lexicon	79.4%	84.6%

Table 5.3: Effects of lexical modeling in the segment boundary detection task

	Correct%	Accuracy%
Bottom-up	68.3%	49.3%
Lexicon	70.8%	51.4%

Table 5.4: Effects of lexical modeling in the broad class identification task

These results suggest that adding a lexical component to the model improves the results for both the boundary detection and the broad class identification task. Since introducing a lexical model is equivalent to re-clustering words within the lexical entry, such improvements can be attributed to the integration of top-down knowledge (exemplar weights) with the bottom-up evidence.

5.3 Lexical access based on a constructed lexicon

Lexical access is often used as a general term for the retrieval of lexical entries from certain types of sensory input. Within a probabilistic framework, there are several options to formalize this notion. One possibility is to use the posterior probability of a lexical entry given an acoustic presentation of a word.

$$p(entry_j|word_i) = \frac{p(word_i|entry_j)p(entry_j)}{\sum\limits_j p(word_i|entry_j)p(entry_j)} \tag{5.1}$$

Here $p(entry_j)$ is the relative frequency of the lexical entry, and $p(word_i|entry_j)$ is the likelihood of the word given the entry. Since each lexical entry itself is modeled as a mixture of exemplars, the posterior probability is calculated as:

$$p(entry_j|word_i) = \frac{\sum\limits_k \lambda_{j,k} \cdot p(word_i|exemplar_{j,k}) \cdot p(entry_j)}{\sum\limits_j \sum\limits_k \lambda_{j,k} \cdot p(word_i|exemplar_{j,k})p(entry_j)} \tag{5.2}$$

Following the commonly used terminology, $p(entry_j|word_i)$, taking a value between 0 and 1, can be thought of as the "activation" of a lexical item for an acoustic input, and the calculation for each item could be thought of as parallel (Marslen-Wilson, 1987). An overall accuracy measure of lexical access can then be obtained as the sum of the posterior probabilities of the correct lexical items. As an alternative to posterior probability, another way of measuring lexical access is to simply identify the maximum of $p(word_i|entry_j)p(entry_j)$, and let the retrieved item be the one that maximizes this joint probability. The overall evaluation would thus be much like the one used in isolated word speech recognition. Because the calculation of the posterior probability involves evaluating the likelihood on all word models, the latter evaluation metric is adopted because of its simplicity[3].

To verify the consequences that inventory size has on lexical access, a sequence of experiments was conducted on an isolated word database. The vocabulary of 50 words was selected from a frequent word list reported from a survey of children's receptive lexicons (Fenson et al., 1991). One female speaker produced all 283 tokens in the sound booth. In each of the experiments, the model described

[3] Since the maximum can be found with Viterbi decoding, a much faster procedure.

in Section 5.2 was tested on this data set in several runs, each with a fixed number of phonetic classes. The size of the inventory increases from 2 to 6 in five steps. At each step, the phonetic inventory was obtained from the results of an initial segmentation, using Algorithm 1. As a result, broad classes that are similar to those derived from the TIMIT experiments were obtained. Since no transcription is available for this data set, and there is little overlap between the two vocabularies, similar evaluations to those in the previous chapters are not performed. However, the lexical access accuracy score for each number of classes is calculated and shown in the following figure.

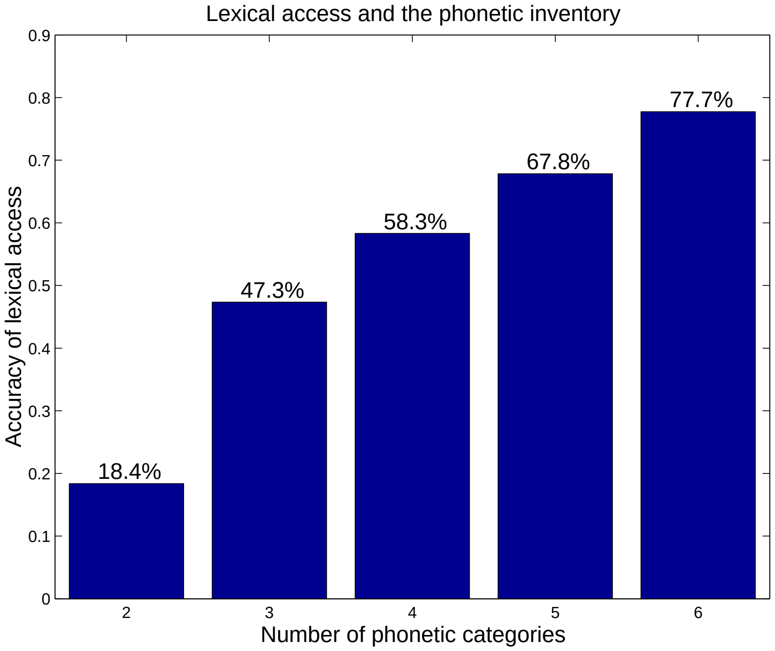

Figure 5.5: Lexical access accuracy versus the number of phonetic classes

It may be seen from Figure 5.5 that accuracy of lexical access improves as more units are added in the phonetic inventory, reaching a reasonable level when 6 units are used in the inventory. Although acoustic variation must be taken into account, such a result reminds us of the lexicon-based studies of Charles-

Luce and Luce (1990), where it is shown that there are few close neighbors in a vocabulary of 50 words. With a constructive approach to the lexicon and a stochastic view of the lexical neighbors, the result in Figure 5.5 may be seen as an indirect argument for the possible use of a broad class-based inventory in the small vocabulary period of language acquisition.

5.4 Probabilistic formulations of lexical distance and lexical neighborhood

Lexical distance is a frequently used notion in the literature related to lexical access and representation. Since it is an intuitive idea to think of the lexicon as a set of words located in some space, a proper measure of distance becomes important. Among the various distance measures that have been considered, the most commonly used metric of lexical distance is the string edit-distance, perhaps dating from as early as Greenberg and Jenkins, (1964): given two words, represented by two sequences of segments, the edit distance is the minimum number of deletion/insertion/substitutions between them. Since edit-distance does not consider the difference in segments, and insertion/substitution/deletion are treated as equal, the edit-distance-based measure is commonly considered a crude approximation to the psycholinguistic distance between words, though competitors that correlate well with behavioral data have been scarce (Bailey and Hahn, 2001).

Among all the infant-oriented lexicon studies, the most relevant is Charles-Luce and Luce (1990), where it was demonstrated that early words tend to be far away from each other by the edit-distance measure. However, since their calculation of distance is based on phonemic representations used by adults, it may be interesting to consider how lexical neighborhood can be informed by the lexical models in this dissertation. In particular, the current effort in formalizing lexical distance will occur in two directions. First, we would like to generalize the edit-distance-based measure to the mixture model. Second, we also explore how the acoustic component can be taken into account when calculating the lexical distance.

5.4.1 Extending the edit-distance to mixture of exemplars

Recall that a mixture-based lexical model assumes the following scenario: given a lexical item $entry_k$, the chance of seeing $exemplar_{k,j}$ is $\lambda_{k,j}$, as given by the prior distribution. Therefore, although each exemplar has a symbolic representation that can be used to calculate edit-distance, the prior distribution must be taken into account when we consider the distance between two lexical entries. In order to arrive at a distance measure between lexical entries, the most natural way is to take the average, or the *expectation* of distances between exemplars:

$$\begin{aligned} dist(entry_m, entry_n) &= E_{p_m,p_n}\left[dist(exemplar_{m,i}, exemplar_{n,j})\right] && (5.3) \\ &= \sum_{i,j} \lambda_{m,i}\lambda_{n,j} dist(exemplar_{m,i}, exemplar_{n,j}) && (5.4) \end{aligned}$$

Like edit-distance, the resulting metric is still symmetric, non-negative and satisfies the triangular inequality. As an example, the distance matrix calculated from a lexicon constructed from 844 words in TIMIT is displayed in Figure 5.6 as a gray-level 2D image. The dark color of the diagonal line corresponds to the zero distance from a lexical entry to itself.

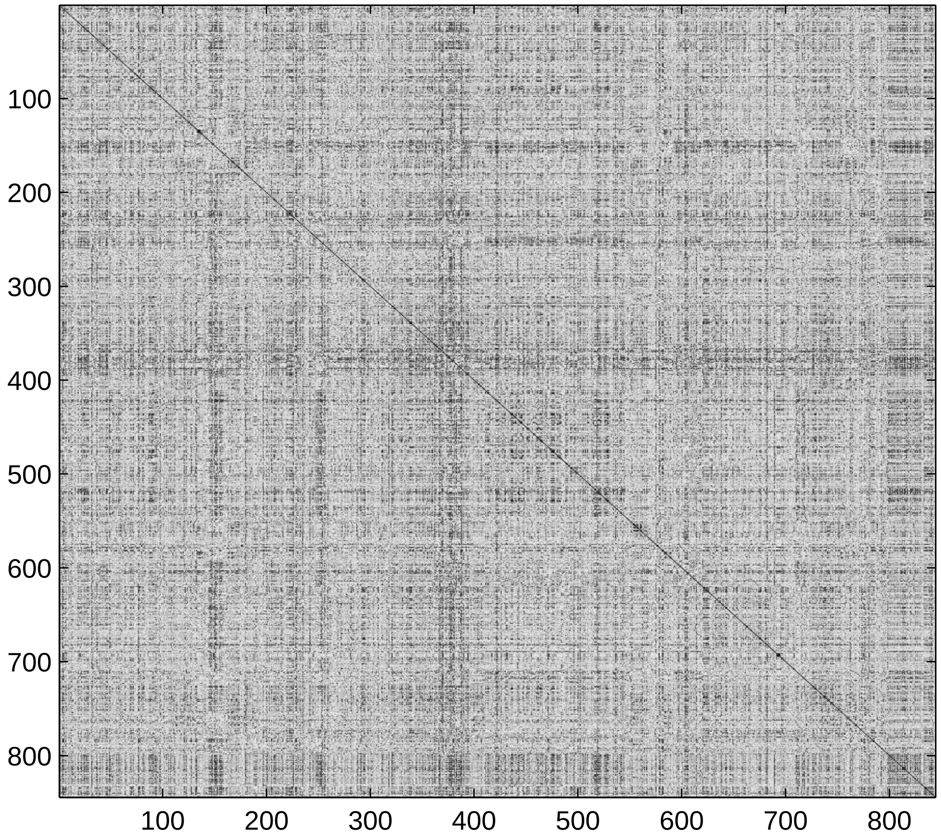

Figure 5.6: Distance matrix calculated from the expected edit-distance between lexical mixtures. The lexical entries are ordered alphabetically. The darker colors are associated with larger distances. The distances vary from 0 to 8, with a mean of 4.45.

5.4.2 An Information-Theoretic approach to lexical distance based on acoustic models

Although edit-distance ignores the differences between different sounds, alternative models that take into account such phonetic differences are few. The main difficulty in building such models seems to be the following issue: how can the differences between different sounds be quantified in a principled manner? Such efforts are occasionally seen in phonology (Frisch, Broe, and Pierrehumbert, 1995), yet their dependence on pre-conceived feature systems makes them unfavorable as we focus our attention to a constructive approach on phonological acquisition.

The view suggested here is that a coherent way of quantifying distances should be derived from sound distributions[4]. Information Theory provides an appropriate framework in which the notion of distance between distributions can be formalized. A commonly used measure is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), which sometimes is also referred to as relative entropy. Given two distributions, $p(.)$ and $q(x)$, the KL-divergence from $p(.)$ to $q(.)$, represented by $D\left(p||q\right)$, is defined as:

$$\begin{aligned} D\left(p||q\right) &= E_p\left[\log\frac{p(x)}{q(x)}\right] \\ &= \int \log\frac{p(x)}{q(x)}p(x)dx \end{aligned} \tag{5.5}$$

For complex generative models, the KL-divergence can be approximated with a Monte-Carlo procedure, which involves generating a large number of data from the model with a computer program. Details of the simulation method are included in Appendix 5.B. As an example, the KL-divergence matrix between 50 words in the infant-directed database is shown in Figure 5.7.

[4] An example is the distance calculation based on exemplars in Johnson (1997b).

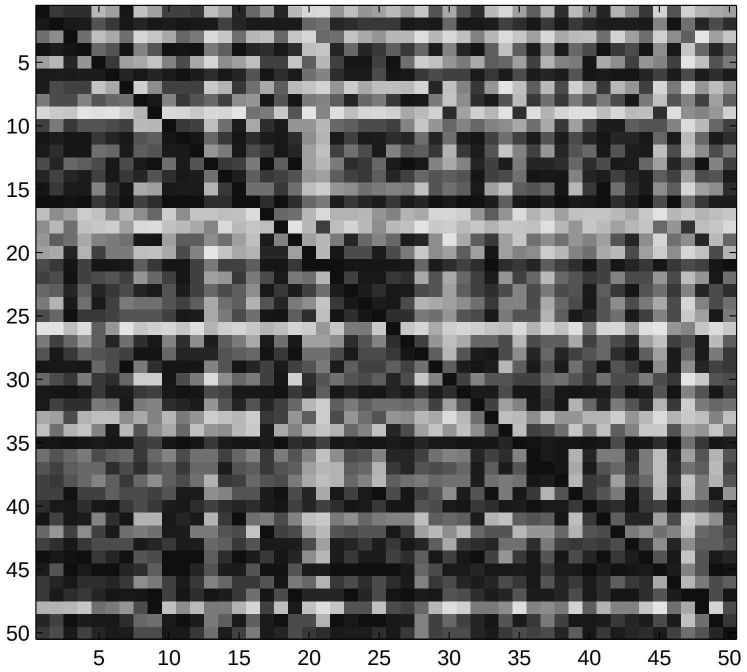

Figure 5.7: The KL-divergence matrix between 50 words derived from the infant-directed database

Although there is a correlation between the edit-distance between lexical mixtures and the one based on KL-divergence, important differences should be noted. First, edit distance does not take the acoustic model into account, and is strictly local – any two strings that differ in one unit will have edit distance 1. KL-divergence, on the other hand, is based on the overall compositional model of acoustic signals, and there is generally no guarantee that any two models that differ in one unit will have the same distance. On the other hand, edit-distance is symmetric; while KL-divergence is not[5]. To compare the generalized edit-distance and KL-divergence, a set of 8 frequent words are selected from TIMIT, and two distance matrices based on the learned lexicon are presented in Table 5.5 and Table 5.6. Potentially, these predictions can be compared with the behavioral data on confusability of word pairs.

[5] It is near-symmetric when the two distributions are very close.

	water	year	wash	suit	rag	oily	greasy	dark
water	0.00	2.49	2.35	3.77	3.15	2.95	5.20	3.77
year	2.49	0.00	3.26	2.90	3.44	3.32	4.65	3.65
wash	2.35	3.26	0.00	3.66	2.54	3.07	4.58	3.29
suit	3.77	2.90	3.66	0.00	3.59	4.15	4.36	3.52
rag	3.15	3.44	2.54	3.59	0.00	3.06	4.63	3.01
oily	2.95	3.32	3.07	4.15	3.06	0.00	4.17	3.94
greasy	5.20	4.65	4.58	4.36	4.63	4.17	0.00	4.51
dark	3.77	3.65	3.29	3.52	3.01	3.94	4.51	0.00

Table 5.5: Expected edit-distance between 8 lexical items

	water	year	wash	suit	rag	oily	greasy	dark
water	0.0	116.0	76.0	162.7	102.0	37.5	102.4	137.9
year	159.8	0.0	35.4	65.7	41.1	31.7	113.0	71.8
wash	97.0	242.3	0.0	202.5	65.1	74.6	113.3	164.7
suit	421.6	312.8	254.7	0.0	204.5	240.3	155.2	135.5
rag	231.6	285.8	127.7	224.6	0.0	123.7	159.3	186.4
oily	195.1	171.8	260.6	182.4	114.7	0.0	93.3	264.5
greasy	657.6	485.1	455.6	264.3	448.9	344.9	0.0	334.7
dark	361.8	399.5	257.7	104.9	199.1	238.0	168.6	0.0

Table 5.6: KL-divergence between 8 lexical items. The numbers have not been re-scaled.

5.4.3 The geometry of lexical neighborhoods

In addition to lexical distance, the introduction of the term *lexical neighborhood* (Luce, 1986) brings a strong geometric flavor to the view of the lexicon. Intuitively, a lexical neighborhood consists of words that "sound similar" (Luce, 1986), where similarity is again based on the lexical distance as discussed in the previous section. Since much of the effort in the current chapter can be seen as a probabilistic characterization of the lexicon, it is interesting to contemplate whether the intuitive notion of lexical neighborhood can be characterized formally in probabilistic terms.

A framework that brings in a mathematical perspective on lexical neighborhood is Amari(1986)'s work on *Information Geometry*, the study of probability distributions as geometrical objects. Intuitively, in order to set up a space for the distributions of words, two things need to be specified: one is how to locate a distribution; the other is how to calculate the distance. Both of them are linked to statistical theory in Information Geometry: first, when a distribution is described by a parametric model, the parameters can be thought of as the *coordinates* for the *manifold* of probability distributions. Second, when such a manifold is equipped with the *Fisher information metric*[6], it becomes a *Riemannian structure.*

Once a coordinate system is found, it is natural to ask whether we can find the shortest (or "geodesic") distance between two distributions along the manifold. Unfortunately, this is difficult in general, especially for complex models like HMM. However, we would like to point out a connection between the length of line segments in Information Geometry and KL-divergence, in a local sense. Let $p(x|\theta)$ and $p(x|\theta + \Delta\theta)$ be two probability distributions, differing in coordinates by $\Delta\theta$. Let Δs be the length of the line segment between $p(x|\theta)$ and $p(x|\theta + \Delta\theta)$. Then:

$$D\left(p(x|\theta)||p(x|\theta + \Delta\theta)\right) = \frac{1}{2}\Delta\theta^T I(\theta)\Delta\theta = \frac{1}{2}\Delta s^2 \tag{5.6}$$

The approximation depends on the assumption that $\Delta\theta$ is small as compared to θ. A short proof of (5.6), accompanied with a brief introduction to Information Geometry, is included in Appendix 5.C. Hence the KL-divergence, such as the one assessed directly from the empirical data in Section 5.4.2, may serve as a rough estimate of the true information distance and provide a quantitative assessment of the distance between lexical entries.

[6] Fisher information is an important concept in the theory of statistical estimation.

5.4.4 Parametric versus nonparametric approaches to lexicon modeling

The view of lexical neighborhoods through information geometry, as shown above, follows a parametric approach. In other words, the parameters of the model provide coordinates for the space of probability distributions, where the distance is based on the Fisher information matrix. In practice, we have first fitted parametric models (mixture of HMMs, in particular) to the data, then computed the relative entropy using the estimated models.

However, starting from a different perspective, one could take a non-parametric approach to the notion of relative entropy, and try to derive the Kullback-Leibler divergence based on non-parametric estimates of the distributions. Similar kinds of ideas, phrased somewhat differently, often appear in the increasingly popular exemplar-based views of lexicon: typically, each lexical entry is seen as a large collection of lexical exemplars, while the notion of distance is based on the lexical exemplars as a whole. However, when one considers formalizing such a proposal, a (possibly high dimensional) space will need to be set up for the lexical exemplars, and in order to obtain a non-parametric estimate of the distribution, histograms will have to be calculated from the examples. However, as the dimension of the space increases, obtaining reliable estimates of the histogram will require an astronomical amount of data[7], and one must consider reducing the dimension of the data for the purpose of learning. Since the parametric approach amounts to finding appropriate coordinates that reduce the dimension of data, it is favorable in terms of both computation and variance in estimation.

It is true that by taking a parametric approach, we are committed to certain assumptions about the kinds of distributions among data. These assumptions

[7] This is also known as "the curse of dimensionality" (Duda, Hart, and Stork, 2000).

may not be adequate, and they can lead to inappropriate modeling of the true distribution. However, such bias in building a model may not be easy to escape. As future progress is made towards better parametric models of speech, the route that we have outlined can potentially lead to a more accurate modeling of lexical neighborhood effects.

From here, we will leave the discussion of lexical distance and neighborhoods. The rest of the chapter turns to the question of what principles can be used to increase the size of the phonetic inventory.

5.5 Towards lexicon-driven discovery of sub-lexical units

5.5.1 Model refinement based on lexicon-level optimization

A major claim made by some researchers in phonological acquisition is that the phonetic inventory correlates with the size of the lexicon. As briefly discussed in Chapter 1, this involves choosing the number of parameters in the model, and is best identified as a model selection problem. For a complex model with many parameters such as ours, two crucial steps need to be taken: the first is to design a search strategy in the space of all possible models that would gradually increase the dimension of the model; the other is to choose a function that evaluates each model on the set of data.

In Chapter 3, model refinement was discussed within the context of segment-level optimization. Given N clusters, there are N choices of refining the mixture model and moving to the subspace of $N+1$-cluster models. In that case, each of the N choices is attempted, and the one with the highest likelihood is chosen as the optimal solution.

Although the current model has incorporated phonotactics and lexicon, the issue of model refinement is still very similar to our very first model $p(segments|units)$. Since the criterion of optimization is still based on the likelihood score, it is straightforward to re-apply the likelihood-based search strategy. Crucially, although the current model contains more sets of parameters, the size of the phonetic inventory is still the sole determinant of the overall model complexity, since the state space of phonotactics and the lexical exemplars both depend on the inventory size. Therefore, model refinement would still proceed from the mixture-based unit model.

A main distinction between the current situation and the previous one is the objective of optimization: the current approach optimizes $p(words|units, phonotactics, lexicon, segmentation)$, while previously the objective was $p(segments|units)$. In implementation, after each step of unit splitting, the whole model is updated, including phonotactics and lexicon. The likelihood score is calculated after the learning algorithm reaches convergence. Details of this refinement procedure are described by the following algorithm.

Algorithm 6 Model refinement with a lexical component

```
Require: an optimized model including $N$ units, phonotactics and lexicon
Ensure: a locally optimal model with $N + N_0$ units and updated phonotactics
    and lexicon
$k \leftarrow 0$
while $k < N_0$ do
    for each unit model $unit_i$ in the inventory do
        identify all segments associated with $unit_i$, and split the cluster into two;
            Replace $unit_i$ with these two models in the mixture
        run Algorithm 5 (including phonotactics and lexicon update) until it
            converges, and calculate the likelihood of the model with increased dimension
    end for
    select the split with the highest likelihood, and increase the inventory size
        from $N + k$ to $N + k + 1$
    $k \leftarrow k + 1$
end while
```

In the following experiment, we run Algorithm 6 on the same TIMIT data set that has been used a number of times already. The inventory size was increased from 6 to 10, and the inventories for two of the models are shown in Figure 5.8. The initial values are set by the result of the first lexicon building step described in Section 5.2. After the first iteration, results from the previous iteration are used as initial values for the next iteration.

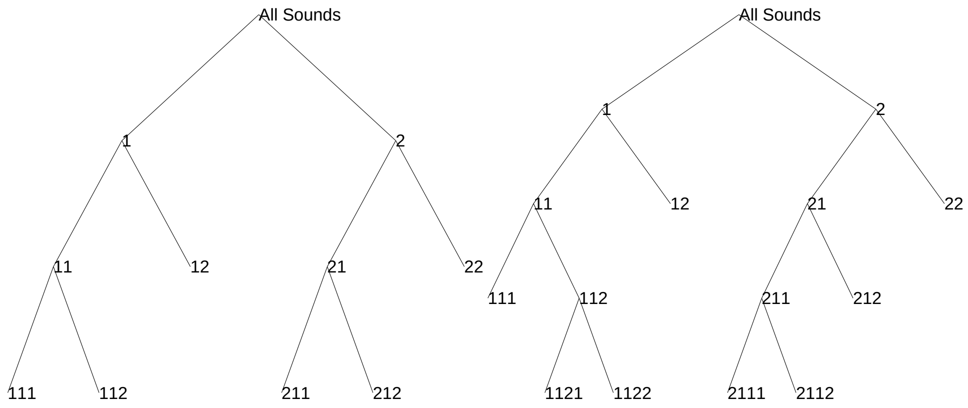

Figure 5.8: Model refinement driven by lexicon-level optimization. Left: the initial 6-class inventory; Right: the 8-class inventory after 2 steps of refinement, including lexicon building.

In the right panel of Figure 5.8, although the refinement of the cluster "211" in the left tree does not produce new classes that are phonetically significant, the refinement of "112" does have the consequence of identifying [s] and [z] as a new class. Table 5.7 shows the interpretations of the 7 classes, obtained after the splitting of "112" on the left side of Figure 5.8:

111	[s] and [z]
1121	plosive (p t k b d g jh ch q epi)
1122	nasal (em en nx eng ng m n)
12	other fricatives (f v th dh sh zh hh)
211	central vowels (ae ay ux uh ah ax)
212	back sonorant (aa aw ao el r l w ow uw oy er axr)
22	front high sonorant (iy ih eh ey y ix)

Table 5.7: Interpretation of the 7 classes after one step of refinement

Notice the order of the classes, especially the obstruents, is not the same as the 6-class case, due to the effect of retraining after the splitting of each cluster[8]. Using these interpretations as a benchmark, an evaluation based on a confusion

[8] Noticeably, the nasals have migrated from class 12 to class 1122, a subclass with a different parent class.

matrix is also conducted, as shown in Table 5.8. The same measure as used previously produces a correctness score of 70.6% and accuracy score of 47.1%.

	111	1121	1122	12	211	212	22	Deletion
[s],[z]	485	29	24	52	7	0	5	25
Stop	28	1077	106	149	47	24	33	137
Nasal	2	4	562	13	29	10	10	79
Other fricative	17	18	50	441	10	6	1	38
Central vowels	1	9	27	22	748	36	65	63
Back sonorant	2	16	44	112	287	974	35	157
High sonorant	0	5	24	18	175	6	883	94
Insertion	68	109	258	661	354	78	195	

Table 5.8: Confusion matrix of the 7 classes as compared to the benchmark

5.5.2 Stopping of model refinement

As a consequence of an expanding inventory, the likelihood score increases with the model complexity. Figure 5.9 shows the increasing likelihood score from an inventory of 6 units to 11 units. In each step, the split that leads to the maximal gain in likelihood is chosen to increase the size of the inventory.

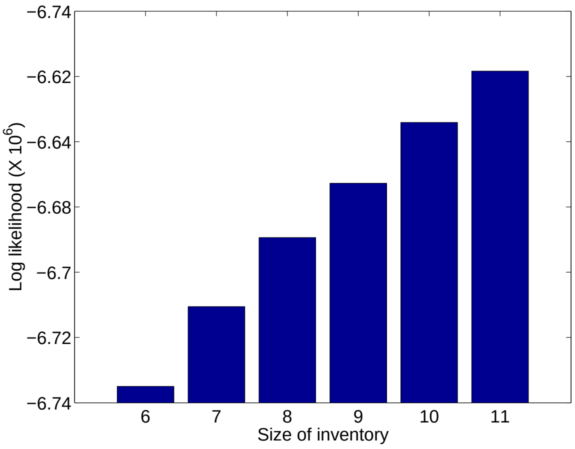

Figure 5.9: Likelihood increase from a unit inventory of 6 classes to 11 classes

As can be seen from the figure, increasing the size of the inventory always results in a gain in likelihood, and over-fitting will eventually occur as the inventory size grows too large. Unfortunately, this problem does not have a general solution, and we are faced with many options for a stopping criterion, each with its own problems.

- Following the discussion in 5.4, the most natural stopping criterion would be based on the distances between lexical entries, either using a generalized edit-distance 5.4.1, or the KL-divergence 5.4.2. For example, using the generalized edit-distance, the lexical distance matrix for a randomly selected set of 100 words in the TIMIT lexicon is calculated for 6 different lexicons, each learned from a phonetic inventory of a different size. The sum of these distance matrices is shown in Figure 5.10.

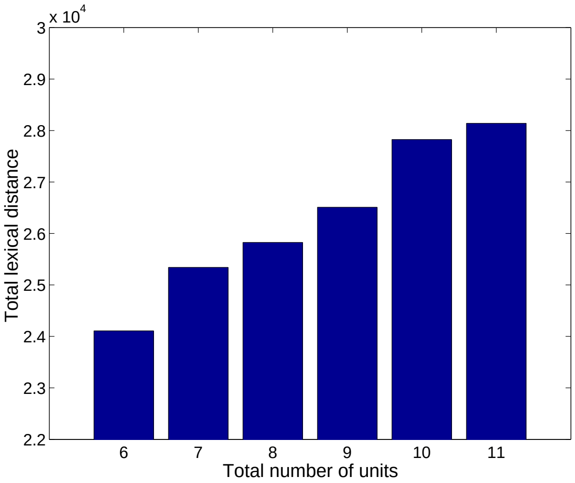

Figure 5.10: The increase in lexical distance from a unit inventory of 6 classes to 11 classes.

As can be seen from the figure, the larger the phonetic inventory, the more distant words are from each other. Therefore in order to address the stopping problem, a proper threshold needs to be chosen so that the lexical distance will not increase indefinitely. One can also use the KL-divergence for similar calculations, but the large computational cost in computing the KL-divergence can be an issue.

- A computationally simple way is to find a tradeoff between likelihood score and number of parameters[9], such as Akaike Information Criterion, Bayesian Information Criterion, Minimal Description Length, etc. However, the theoretical guarantees of these model selection criteria may not apply when their assumptions are not valid. For example, the current model derives likelihood scores not directly from the time-domain, but from front-end representations derived via transform of the original signals. Thus the score is not a true likelihood.

- If one is willing to pay the computational cost, then one could even consider introducing a prior distribution over all possible models differing in their number of parameters and doing some inference from the posterior distribution. Although theoretically appealing, this leaves the burden of choosing an appropriate prior.

- A widely used method in practice is *cross-validation*: a separate set of data not seen in training is used to evaluate the lexical access of each of the refined models. When the model dimension grows so that *overfitting* on the training data occurs, its performance on the test set will be hurt[10]. However, the use of cross validation needs a measure of error for the predictor. As mentioned previously, determining a meaningful measure of error is also a problem.

[9] A more technical term is "smoothing" or "regularization".

[10] This strategy has been used in acoustic sub-word unit systems for speech recognition (Singh, Raj, and Stern, 2002).

Among these options, the first one explicitly keep words as distinct as possible, while for the other ones, their connections with lexical distance are less direct. Due to the limitations of time, these options for controlling model refinements will not be discussed further in this dissertation. Although the discussion of the "lexically-driven" model is coming to an abrupt end, the author's hope is that enough preliminary steps have been outlined so that further investigations may be carried out in future work.

5.5.3 Other approaches to the model selection problem

All of the model selection methods suggested above are predominantly *statistical* – in other words, the likelihood score decides optimality. Although the simplicity of likelihood-based optimization is attractive, it also obscures one of the goals of this model – testing correlation of lexicon size and the unit inventory. Currently, the effect that the lexicon has on the inventory can be seen as rather indirect: a larger set of words are generally better modeled with more unit models. However, since models of higher complexity already bring in a better fit to the data, it is unclear how the effect of vocabulary size can be separated from merely the amount of data being fed into the model, especially if the optimization criterion is simply finding some best combination of likelihood score and model complexity.

The main reason why the lexicon-unit relationship remains obscure is that the concept of *contrast*, or *functional load*, does not hold a place in the optimization criterion we have adopted. As previous studies of lexicon (Shipman and Zue, 1982) have shown, the values of different contrasts in establishing lexical distinctions differ, and such differences can be quantified in different ways (Carter, 1987). On the other hand, in the current work, words are treated as independent of each other, which allows their individual probabilities to be multiplied together. But the calculation of probabilities certainly does not account for how *different* words are, let alone how a given contrast can distinguish similar words. Along this line of thinking, the discussion of lexical distance and neighborhood (Section 5.4) can be potentially enlightening, since these formulations do intend to quantify the distinction between lexical items.

Some major gaps need to be filled before such a direct link to the lexicon can be forged. The first issue is the treatment of binary distinctions. Traditionally, contrast is often thought of as a relationship between a pair of phonemes, while the functional load is thought of as the contribution of a contrast to the lexical distinctions. In the examples included in this dissertation, binary distinctions are followed in all the searches for new units. However, from an optimization/likelihood perspective, consistently pursuing a binary search may not be optimal. As an example, suppose the data consists of three well-separated groups. In this case, a better strategy may be splitting the data into 3 clusters, rather than doing a binary split in two steps. Moreover, if components of the mixture model poorly captures the underlying structure that generates the data, then splitting may not work at all (see Figure 5.11). Thus it remains to be seen what types of formulations of contrast need to be sought after in order to be compatible with the optimality criterion.

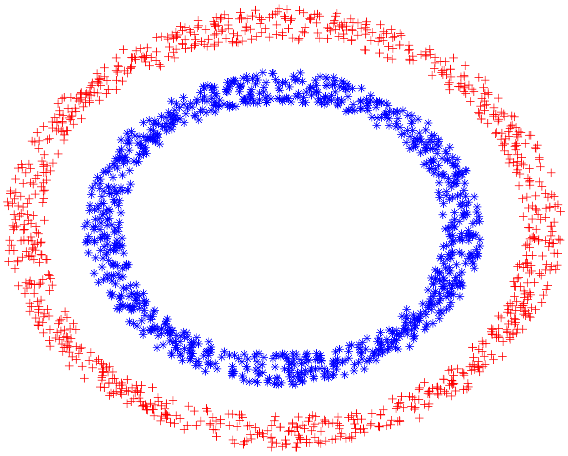

Figure 5.11: A data set that cannot be separated with a mixture-of-Gaussian model

The second problem is probably not merely a technical point, but related to some deeper issues: how can traditional views of lexicon and the one taken in this dissertation be reconciled? The notion of contrast, as was developed in the history of linguistics, is more or less based on the assumption of lexical representations as rigid, symbolic entities. The current view of the lexicon is noticeably different: each lexical item is seen as a collection of lexical exemplars, each a product of unit composition. Lexical distance, as seen in Section 5.4, has a gradient nature, since the basis for deriving these distances is the acoustic model, and variation is embedded within the lexical entry. Does the symbolic nature of contrast indicate a need for another higher-level model of the lexicon? Or is there a unified view of the symbolic contrast and the numerical distance? These interesting questions will be left for future explorations.

5.A The objective function optimized by Algorithm 5

The purpose of this section is to state an objective function that is optimized by Algorithm 5 and clarify the notation used therein.

The objective function is the following:

$$\sum_i \log p(w_i|U, P, (Z_i, S_i), \{\lambda_{i,j}\}) \tag{5.7}$$

where the notations are:

- w_i = holistic word
- U = mixture-based unit model (including M units)
- P = between-unit transition matrix $(a_{ij}) \sim$ phonotactics
- Z_i = the exemplar membership indicator for w_i. $Z_i^j = 1$ iff. w_i is assigned to the j-th exemplar corresponding to w_i's lexical entry; otherwise $Z_i^j = 0$.
- S_i = segmentation of w_i according to the exemplar indicated by Z_i
- $\{\lambda_{k,j} : k = 1, \cdots, K; j = 1, \cdots, M(k)\}$ = the weights of the lexical exemplars. For each lexical entry k, the lexical mixture weights satisfy $\sum_{j=1}^{M(k)} \lambda_{k,j} = 1$. The collection of these weights form the lexical knowledge, denoted as L.

Algorithm 5 include the following equations. The superscript $^{(t)}$ indicates the values of the parameter at each time t. For notational convenience, we use $f : \{w_i\} \to \{1, \cdots, K\}$ for the function that looks up the lexical entry for the word w_i, so that the index for words and that for lexical entries can be distinguished. Also note that $L^{(t)} \equiv_{def} \{\lambda_{i,j}^{(t)}\}$, $P^{(t)} \equiv_{def} \left(a_{i,j}^{(t)}\right)$.

$$(Z_i^{(t+1)}, S_i^{(t+1)}) = \arg\max_{(Z_i, S_i)} \log p\left(w_i | U^{(t)}, P^{(t)}, (Z_i, S_i), L^{(t)}\right) \tag{5.8}$$

$$a_{i,j}^{(t+1)} = \frac{\sum\limits_k count(i, j, S_k^{(t+1)})}{\sum\limits_k count(i, S_k^{(t+1)})}, i, j = 1, \cdots, M \tag{5.9}$$

$$\lambda_{k,j}^{(t+1)} = \frac{\sum\limits_{f(w_i)=k} Z_i^j}{\sum\limits_{j=1}^{M(k)} \sum\limits_{f(w_i)=k} Z_i^j}, j = 1, \cdots, M(k); k = 1, \cdots, K \tag{5.10}$$

$$U^{(t+1)} = \arg\max_U \sum_i \log p\left(w_i | U, P^{(t+1)}, (Z_i^{(t+1)}, S_i^{(t+1)}), L^{(t+1)}\right) \tag{5.11}$$

(5.8) finds the most likely segmentation of w_i, within the set of exemplars for the lexical entry $f(w_i)$. Z_i and S_i are computed in one step by the Viterbi algorithm. (5.9) updates the phonotactic probabilities (hence also $P^{(t)}$) by recounting the unit sequences. (5.10) renormalizes the counts of the exemplars within each lexical entry to get the weight on each exemplar. At last, (5.11) re-estimates the unit models so as to increase the likelihood function. In Algorithm 5, (5.11) corresponds to the lines that perform the EM algorithm.

By iterating the equations (5.8) – (5.11), each step of Algorithm 5 performs a maximization step over a different set of parameters. As noted in the appendix of Chapter 3, this strategy is guaranteed to increase the likelihood and converge.

The use of Z_i as parameters can be avoided by adopting a view of Z_i as missing data. Similar to the formulation in Chapter 3, this approach would replace (5.8) with an E-step, and assign "soft" exemplar labels to words rather than Z_i. However, this is associated with a higher computational cost and may not differ significantly from the approach taken here.

5.B The approximation of KL-divergence using Monte-Carlo simulation

Since a parametric approach assumes that the observed distribution is generated by a model, the way we calculate (5.5) is first fitting two parametric models to the target distribution of interest, and take $p(.)$ and $q(.)$ to be the distributions generated by the models.

Nevertheless, computing (5.5) directly can be a problem in practice, because our models – HMMs – use hidden variables that must be integrated out in calculating a quantity like (5.5). To solve the problem of computation, an Monte-Carlo methods is used to simulate (5.5). The use of Monte-Carlo is motivated by the following result in probability theory:

$$E_p\left[\log\frac{p(x)}{q(x)}\right] = \lim_{N\to\infty}\frac{1}{N}\left(\log\frac{p(x_1)}{q(x_1)} + \cdots + \log\frac{p(x_N)}{q(x_N)}\right) \tag{5.12}$$

where $x_1, \cdots, x_N$ are samples independently drawn from the same distribution $p(.)$. The advantage of this method rests on the relative ease of generating sequences from HMMs: in order to generate a sequence x from $p(.)$, we actually generate two sequences from a joint distribution $P(X, Q)$, such that $p(X)$ is a marginal distribution of $P(X, Q)$:

1. Generate a random sequence q from the Markov chain of HMM
2. Generate a vector sequence x from the output distributions from each state in q

If we use $A(Q)$ for the Markov chain, and $B(X|Q)$ for the output distribution of the HMM, then the above procedure is equivalent to generating (q, x) from the distribution $P(X, Q) = A(Q)B(X|Q)$. By using the fact that the HMM distribution $p(X) = \int A(Q) \cdot B(X|Q)dQ$, the x sequences by itself form a sample of $p(.)$, and (5.12) can be approximated to any precision without dealing with integration.

In order to calculate the KL-divergence between lexical entries, the approximation method of (5.12) is particularly useful. In this case, each lexical entry is a mixture of lexical exemplars, which themselves are composed of unit models. The analytic form of $p(.)$ is more complex, since $p(.)$ now has a mixture form. But generating sequences from these lexical mixtures would only add one more step to the generation process outlined above:

1. Generate a random number r and decide which exemplar to use
2. Generate a random state sequence q from the exemplar
3. Generate an vector sequence x from the output distributions on each state in q

5.C A brief introduction to Information Geometry and its relation to KL-divergence

In this section, we will briefly introduce the use of Fisher information in Information Geometry, and prove (5.6). For a comprehensive overview of Information Geometry, the reader is referred to Amari (1986) and Amari and Nagaoka (2000).

The key to the basic setup of Information Geometry is to recognize the parameters of a probabilistic model $\theta = (\theta^1, \cdots, \theta^n)$ as *coordinates* for the *manifold* of probability distributions. Such a manifold is generally non-Euclidean. When it is equipped with a *metric tensor* that defines how distance is calculated using coordinates, it becomes a *Riemannian structure.*

Abstractly, the metric tensor at a given point p is an inner product between two tangent vectors:

$$g_p(\vec{v}, \vec{w}) = \langle \vec{v}, \vec{w} \rangle_p$$

When the two tangent vectors both live in a finite-dimensional tangent plane with a basis $e_1, \cdots, e_n$, let $\vec{v} = \sum_i v_i \cdot e_i$, $\vec{w} = \sum_j w_i \cdot e_j$ be the coordinate representations, then the metric tensor has a matrix representation:

$$\begin{aligned} g_p(\vec{v}, \vec{w}) &= \langle \vec{v}, \vec{w} \rangle_p \\ &= \langle \sum_i v_i \cdot e_i, \sum_j w_j \cdot e_j \rangle_p \\ &= \sum_{i,j} v_i \cdot w_j \cdot \langle e_i, e_j \rangle_p \end{aligned}$$

Therefore once we fix the basis $e_1, \cdots, e_n$, the inner product of two vectors $\vec{v}$ and $\vec{w}$ can be simply written as matrix multiplication:

$$g_p(\vec{v}, \vec{w}) = v^T G(p) w$$

where v, w are the coordinates of $\vec{v}$ and $\vec{w}$ under the basis $e_1, \cdots, e_n$, and the (i, j)-entry of $G(p)$ is $g_{ij}(p) = \langle e_i, e_j \rangle_p$.

Through Amari's work, it becomes clear that a statistically meaningful metric tensor is provided by *Fisher information*, an important concept in statistical theory[11]. The (i, j)-th entry in the Fisher information matrix of a probabilistic

[11] E.g. the connection with the *Cramer-Rao* theorem.

distribution $p(x|\theta)$ at a certain point θ_0 is given by:

$$\begin{aligned} I_{ij}(\theta_0) &= -E_p\left[\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}\log p(x|\theta)|_{\theta_0}\right] \\ &= E_p\left[\frac{\partial}{\partial\theta^i}\log p(x|\theta)|_{\theta_0}\cdot\frac{\partial}{\partial\theta^j}\log p(x|\theta)|_{\theta_0}\right] \end{aligned} \tag{5.13}$$

where $\frac{\partial}{\partial\theta^i}$ means the partial derivative with regard to the i-th coordinate of the parameter θ.

To see how Fisher information can be used as a Riemannian metric for probability distributions, note that the set of random variables $\left\{\frac{\partial}{\partial\theta^i}\log p(x|\theta)\right\}$ are linearly independent. Moreover, they form a basis for the tangent plane to $p(x|\theta)$ at θ. Since Fisher information can be thought of as an inner product of two basis vectors: $\frac{\partial}{\partial\theta^i}\log p(x|\theta)$ and $\frac{\partial}{\partial\theta^j}\log p(x|\theta)$, we can set the metric tensor $G(\theta)$ to the Fisher information matrix:

$$g_{ij}(\theta) = I_{ij}(\theta) = E_p\left[\frac{\partial}{\partial\theta^i}\log p(x|\theta)\cdot\frac{\partial}{\partial\theta^j}\log p(x|\theta)\right]$$

Therefore to compute local distances around a point θ, we can use the Fisher information metric as follows: consider a model $p(x|\theta)$ and another model in its local[12] neighborhood $p(x|\theta+\Delta\theta)$. Then the length Δs of the line segment from $p(x|\theta)$ to $p(x|\theta+\Delta\theta)$ is calculated by the metric tensor:

$$\begin{aligned} \Delta s^2 &= \sum_{i,j}\Delta\theta^i\Delta\theta^j I_{ij}(\theta) \\ &= \Delta\theta^T I(\theta)\Delta\theta \end{aligned} \tag{5.14}$$

Once we obtain the length of the line segment, the distance between any

[12] Here "local" means $\Delta\theta$ is sufficiently small such that their distance can be approximated on the tangent plane.

two points can be obtained through line integrals along the geodesic between the two points. However, to explicitly carry out the calculation, one still needs to obtain the closed form of Fisher information, which is difficult to do in our case. However, at least in a local sense, there is a connection between (5.14) and the KL-divergence, a quantify that can be assessed empirically. In order to see such a connection, let's consider the KL-divergence from the model $p(x|\theta)$ to $p(x|\theta+\Delta\theta)$, written as $D(p_\theta||p_{\theta+\Delta\theta})$:

$$\begin{aligned} D(p_\theta||p_{\theta+\Delta\theta}) &= E_p\left[\log \frac{p(x|\theta)}{p(x|\theta+\Delta\theta)}\right] \\ &= \int p(x|\theta) \log \frac{p(x|\theta)}{p(x|\theta+\Delta\theta)} dx \end{aligned}$$

If we write $l(\theta) = \log p(x|\theta)$, then by carrying out multi-variate Taylor expansion on $l(\theta)$, we thus obtain:

$$\begin{aligned} D(p_\theta||p_{\theta+\Delta\theta}) &= \int p(x|\theta)(l(\theta) - l(\theta+\Delta\theta))dx \\ &= \int p(x|\theta)\left(-\Delta\theta^T \left(\frac{\partial}{\partial\theta^i} l(\theta)\right)_i - \frac{1}{2}\Delta\theta^T \left(\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j} l(\theta)\right)_{i,j} \Delta\theta\right) dx \end{aligned}$$

Here we have ignored the higher-order terms in the Taylor expansion since $\Delta\theta$ is considered to be small. The subscripts i and i, j indicate the vector and matrix form of the functions, respectively. The first term equals zero because of the following:

$$\begin{aligned}
E_p(\frac{\partial}{\partial\theta^i}l(\theta) &= \int p(x|\theta)\frac{\partial}{\partial\theta^i}l(\theta)dx \\
&= \int p(x|\theta)\frac{\partial}{\partial\theta^i}\log p(\theta)dx \\
&= \int p(x|\theta)\frac{1}{p(x|\theta)}\frac{\partial}{\partial\theta^i}p(\theta)dx \\
&= \frac{\partial}{\partial\theta^i}\int p(x|\theta)dx \\
&= \frac{\partial}{\partial\theta^i}\cdot 1 = 0
\end{aligned}$$

Hence the KL-divergence from the model $p(x|\theta)$ to $p(x|\theta+\Delta\theta)$ can be written as:

$$\begin{aligned}
D(p_\theta||p_{\theta+\Delta\theta}) &= \int p(x|\theta)\left(-\frac{1}{2}\Delta\theta^T\left(\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}l(\theta)\right)_{i,j}\Delta\theta\right)dx \\
&= -\frac{1}{2}\Delta\theta^T\left(\int p(x|\theta)\frac{\partial}{\partial\theta^i}\frac{\partial}{\partial\theta^j}l(\theta)dx\right)_{i,j}\Delta\theta
\end{aligned}$$

Compare this with the definition of Fisher information in (5.13); the expectation of the mixed partial derivatives is exactly the corresponding entry $I_ij(\theta)$ in the Fisher information matrix. We thus obtain:

$$\begin{aligned}
D(p_\theta||p_{\theta+\Delta\theta}) &= \frac{1}{2}\Delta\theta^T I(\theta)\Delta\theta \\
&= \frac{1}{2}\Delta s^2
\end{aligned}$$

CHAPTER 6

Summary and discussion

6.1 Mechanisms for phonological acquisition

The approach taken in this dissertation can be seen as *constructive*: phonological acquisition is seen as the discovery of recurring patterns in acoustic signals. Using the "code-breaking" metaphor in Kuhl (2004), the present model tries to "break" the code by constructing its own representation of the code. Intuitively, such a discovery is guided by two kinds of processes. One is grouping similar sounds into classes, the other is counting these sound classes from a signal varying with time. Since a fundamental characteristic of speech signals is the lack of cues that separate different speech sounds, grouping and counting must be done together, with the help of processing that unveils the underlying structure of the signals as inferred from the current state of knowledge.

The probability calculus applied throughout this dissertation provides a coherent framework in which learning from uncertainty can be characterized as computation, and is becoming the dominant paradigm in a number of other fields. Within such a framework, phonological acquisition is cast as an optimization problem, and iterative techniques are used to find possible solutions to the problem. To recapitulate, Chapters 2 – 5 present a sequence of models that represents a progression of the optimization-based approach. In the pre-lexical stage, the first optimization problem is learning units from segments, i.e. the problem of grouping (Chapters 2 and 3). The transition from segments to words is made by adding segmentation and phonotactics to the model, therefore also taking the counting problem into consideration (Chapter 4). In the post-lexical stage, a lexicon is constructed and optimized together with units and phonotactics, thereby adding a higher level of representation to the model (Chapter 5). Through experiments on acoustic speech data, with expert transcriptions as an evaluation metric, it is demonstrated that these models perform reasonably well in the designated task of discovering sequential patterns in the word-level waveforms. Thus, these models support the standpoint that early phonological acquisition could benefit from sensitivity to those patterns in the input data.

It is rather tempting to connect the sequence of models with the actual stages of phonological development. However, such a connection is not grounded in empirical evidence, since we are not aware of any studies showing that knowledge of phonotactics and lexicon is developed later than features and segments. Instead, the order in which these models are presented is motivated by a computational goal: the model at each stage addresses a sub-problem of the next stage, and its solution can be used to set the initial values for a larger model containing more parameters. For complex models with many local maxima, solving smaller problems before bigger ones is one of the few alternative strategies for model fitting. Whether a similar strategy is also employed in higher-level cognition remains a curious question.

6.2 Linking intuition with model and computation

The technical tools used in the current work are motivated by various assumptions we have made with regard to phonological acquisition. Most of these tools are not new from an engineering perspective, but the novel aspect of the current thesis is applying these tools to address questions of linguistic interest. A summary of those assumptions about acquisition, the corresponding modeling assumption, and computation is presented in Table 6.1:

Phonological acquisition	Probabilistic model	Computation
sensitivity to acoustic discontinuities	initial segmentation of all words	segmentation based on dynamic programming (4.3)
similar sounds are grouped into units	clustering segments with unit models	mixture model and EM algorithm (2.6)
representations of words are composed of units	word representation is assigned to the optimal unit sequence	iterative segmentation and clustering (4.2, 4.4)
the lexical variation of each item is contained in the lexicon	a set of lexical exemplars is assigned to each lexical item	lexical mixture and joint optimization (5.2)

Table 6.1: A review of the modeling assumptions and the computational tools used in the current project

One characteristic of the current model is the reduced emphasis placed on the role of initial knowledge in phonological acquisition. However, the use of statistics is not intended to be a challenge to the idea of an initial bias. Instead, the type of probabilistic modeling used in this dissertation follows much traditional thinking, and is better seen as a way of quantitatively specifying the bias. To illustrate this point, we reproduce Figure 5.3 below:

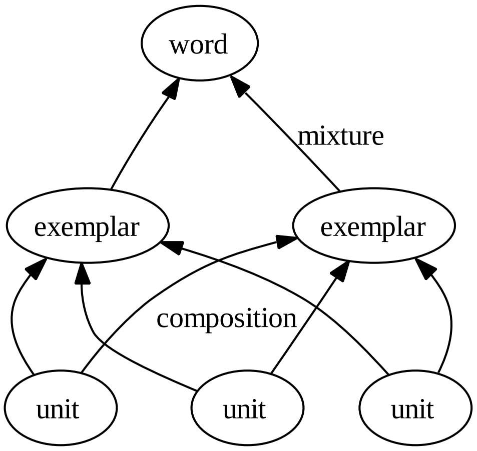

Figure 6.1: The structure of a lexical model. (reproduced)

The use of a mixture model and the associated EM algorithm captures the idea that the building blocks of phonology are a set of discrete units, and that a lexical entry may be realized in different forms. Moreover, the principle of composition – arguably the most important insight of linguistics – is rather central to the effectiveness of modeling, since it allows us to characterize a large number of words with a small number of units[1]. On the other hand, rules of composition in the model[2] remain in the same finite-state class as most of traditional phonology, but are made more flexible and sensitive to the input data with the use of phonotactics and lexical exemplars. Therefore, such an approach represents an enrichment of, rather than a departure from, traditional views of phonology.

The way those ideas are formalized is still quite preliminary, and from an engineering perspective, the data sets used in this study are probably not "large enough". Moreover, because of our unusual standpoint, it is difficult to find other work on unsupervised learning of phonology to use as a comparison. But our formulation is precise enough for quantitative assessment to be carried out

[1] A modest version of the more dramatic statement "infinite use of finite means".

[2] In particular, the Markov processes that underly the unit models and phonotactics.

(although the issue of assessment is also non-trivial and should be given equal attention). It is the author's hope that the work presented here can serve as a starting point for more sophisticated models.

6.3 Further clarifications

Before closing, we would like to address some objections that may arise:

- *A model that does not take into account the role of articulation is not interesting.*

It is widely accepted that phonological structures do not develop from speech perception alone (Locke, 1993). In fact, dimensions chosen for acoustic signals may not capture all the important characteristics of speech data (Stevens, 1998). But as we argued in Section 1.4, the fact that perception precedes production suggests that focusing on perception is a necessary step for abstracting the complex process of phonological acquisition. It is true that the current model does not have the ability to distinguish certain phonetic classes by place of articulation, a basic kind of distinction that is believed to be acquired early in acquisition. However, it is worth noting that the approach taken in this thesis can also be applied to articulatory data and be further used to explore the relation between perception and production. Such extensions may be left for future research.

- *Speech is organized by syllables, not by segments.*

It should be pointed out that the current work does not deny syllables as units of speech. Rather, it shows the possibility of reaching the segment level without learning syllables as an intermediate stage. Each holistic word is presented as a waveform, and segment-sized broad classes are hypothesized as the only level of structure that generates the waveform. These broad classes are presumed to be relatively homogeneous in time, and syllables do not fit into this profile since they are temporally complex units. However, with extra hidden variables representing syllable-sized units, the same framework in this thesis could be applied to learning syllable-sized units[3]. Since the identification of syllables may rely on other perceptual cues (e.g. visual), such questions are also not within the scope of the present thesis.

- *Models of human language should avoid the designer's bias since such bias does not exist in humans.*

This objection may arise from the connectionist approach to phonological acquisition, such as Plaut and Kello (1999). In fact, attempts to discredit the initial bias continue to arise from studies based on neural networks (Elman, 1993; Elman et al., 1996). Since there have been plenty of arguments in psycholinguistics, I present a few arguments from statistics and machine learning, where neural networks fall in the category of *non-parametric models.*

First, neural networks are not truly general learning machines that can learn anything. Results in statistical learning theory show that no machine can learn an arbitrary set of functions (Vapnik and Chernovenkis, 1971; Blumer et al., 1989). A neural network has its own bias and one can prove when it reaches its limits (Baum and Haussler, 1989). In addition, neural networks also face the problem of other non-parametric approaches: models with little bias need a prohibitively large number of examples[4] to control the variance (Geman, Bienenstock, and Doursat, 1991), or to reliably identify the right solution. To address

[3] Assuming there is some way of defining syllables on the waveform.

[4] Informally, the problem can be described as "new data never stops coming in".

such problems, recent approaches in non-parametric learning have focused on *dimension reduction*. Again, this assumes a certain kind of structure underlies the data (Tenenbaum, de Silva, and Langford, 2000; Belkin and Niyogi, 2003) and is clearly not a universal solution.

In the author's opinion, if neural networks are to be the main framework for modeling phonological acquisition, then the fundamental challenge is constructing a compositional representation[5] rather than learning *per se*. Although much effort has been made in this direction (Smolensky, 1990; Hinton, 1990), it has only achieved limited success and has received no mention in current neural network models of phonological acquisition.

- *Segment is an emergent property of language.*

The point of divergence here lies in our use of the word "discovery" as our modeling objective instead of "emergence". We emphasize that statistical learning is responsible for fitting a model to a set of data, but not for coming up with the model. It is our job to design the structure for each model and the structure of the hypothesis space, and "discovery" is defined as selecting a hypothesis based on certain criteria. In fact, many emergentist ideas[6], such as Lindblom (2000), have a strong flavor of discovery, in the sense defined above. As an example, consider the following metaphor used in Tomasello (2003): each utterance is compared to a curve on a transparency along a time axis, and infants notice the structure within the utterances by "overlaying the transparencies". His metaphor actually summarizes what is done in this thesis: if we regard each holistic word

[5] For example, trying to add constraints to architectures of the network (Waibel et al., 1989; Koizumi et al., 1996).

[6] One exception is the discussion of emergence within the context of language evolution or "self organization". As mentioned in 1.2, this type of idea has a rather different focus from the current thesis.

as a “transparency” with a variable length, and think of the higher dimensional acoustic signals as the patterns, then the statistical learning of segmental structure is in fact analogous to “overlaying” of transparencies.

- *HMMs are highly unconstrained and cannot be used to model language acquisition.*

The flexibility of the HMM in modeling time series data is well-known to many fields (Bilmes, 2002). To address the concern that an HMM may be too unconstrained a learning device, we need to make a few clarifications about our use of HMM. First, as was made clear in Chapter 2, the mixture model as a whole is our formal proposal for an exemplar-based approach. The components of the mixture model can be chosen from other models in speech recognition (Ostendorf, Digalakis, and Kimball, 1996; Deng, 1999) and HMM is not the only choice. With other types of components, the iterative learning procedure will remain more or less the same. Second, in our model, phonological representation is constrained in the sense that the acoustic signals are mapped to sequences of category models. The restricted, rather than arbitrary, set of possible representations crucially distinguishes the current work from the neural network-based approach mentioned above.

6.4 Implications and future work

The implications of the current work are relevant to several fields. Most directly related is the empirical work on phonological acquisition. As discussed in Chapter 1, infant experts have revealed the following puzzle: while 14-month old infants can distinguish between minimal pairs in a discrimination task, they fail to do so in a word learning task, which is designed to test whether they really represent the pairs differently (Stager and Werker, 1997). One possible explanation is to differentiate a re-organization of the peripheral auditory perception, as shown in the discrimination tasks, from a word-learning task that requires a linguistic representation[7]. If the coarse-to-fine refinement strategy outlined in the current thesis is plausible, then it may be worthwhile considering whether a similar developmental process also occurs in children's phonological systems. These hypothesized systems are combinatorial, but differ from the adults' systems in certain aspects, such as having a smaller inventory of units. To the author's knowledge, the possibility of a "sketchy phonology" in development has not been investigated, and we expect new interpretations of the existing results, for example of the acquisition of phonotactics (Jusczyk, Luce, and Charles-Luce, 1994), to arise from this perspective.

The difficulties encountered in learning segments from waveforms also lead us to reflect on the interface between phonetics and phonology. Transcribers of a speech corpus like TIMIT received significant phonetic training on a specified phonetic alphabet before doing their job, which involves projecting their segmental analysis onto the waveform. In practice, much compromise needs to be made between the rigid, sequential phonological representation and the prevalent ambiguities in the waveform (Keating et al., 1994). Infants, as envisioned in the current thesis, works in the opposite direction: they not only need to discover which units are present in what they heard, but also have to decide the sequence of units in each word[8]. Traditional discussions of the phonetics-phonology interface often encapsulate phonology as a separate module, and the basis of such a module is the phonological units. Such a top-down perspective is analogous to the view of the phonetic transcribers. The problem of learning forces us to change

[7] LouAnn Gerken, personal communication.

[8] In fact, we assume they will eventually discover the entire phonetic alphabet.

the view from transcribers to infants, and reconsider our understanding of the form-substance relationship[9]. Tools used in the current thesis can potentially be employed to initiate such investigations.

A change of view for the phonetics-phonology interface may also benefit the neighboring field of speech recognition. Although it is widely acknowledged that speech recognition should incorporate phonetic knowledge, knowledge-based systems have only demonstrated limited success (Hasegawa-Johnson et al., 2005)[10] Curiously, most of these efforts have also taken the perspective of a phonetic transcriber: phonological structures[11] based on some type of dictionary representations are either specified as the goal of classification (Cole and Hou, 1988; Juneja and Espy-Wilson, 2003), or built into the statistical model (Deng, Ramsay, and Sun, 1997). No matter whether they are based on linear or multiple-tiered representations of phonology, those structures or rules were proposed from the transcriber's view and may not be supported by a learning procedure. Since the unsupervised approach focuses on the learner, we hope it will help reduce the transcriber's bias and build parsimonious models that better characterize the data.

There are a number of directions in which the work in this dissertation can be extended.

First, more sophisticated models of speech will add to the strength of the statistical approach. Due to its inappropriate assumptions, HMMs do not approximate the true distribution of the data very well. In optimization, the mix-

[9] Also see Pierrehumbert (2002) for a similar discussion.

[10] In the experiments reported in Hasegawa-Johnson et al. (2005) and previous work, the knowledge-based system is taken as a discriminative model that is used to rescore the N-best hypotheses output by a baseline system. It should also be noted that an engineering goal (e.g. word error rate) always guides the measure of success.

[11] For example, broad phonetic classes defined by manner of articulation, or feature bundles obtained through the application of some spreading rules.

ture of HMMs is faced with many spurious local maxima, thereby affecting the modeling of phonological units. Future models are expected to overcome these shortcomings by adopting better assumptions and optimization algorithms.

Second, it may be worthwhile to consider other types of signal representations, for the benefits of learning distinctions between classes that are not well characterized by MFCCs. All the phonetic classes, no matter how they are defined, are unlikely to completely fall within a binary tree-like structure. Therefore in order to provide a more complete characterization of a feature system, multiple trees need to be constructed from different signal representations. This can be not only critical for separating classes that will not be distinct under MFCC (e.g. voiced vs. voiceless), but also give rise to the question of how different perceptual domains may jointly define the natural classes. Hence new architectures and search algorithms are also in demand.

Third, it may be worthwhile creating a database that better characterizes infant-directed speech and allows us to focus on the issues of interest. Ideally, the database would control speaker variability, and be representative of the inputs actually directed to children (MacWhinney and Snow, 1985; Brent and Siskind, 2001). Moreover, a better understanding of child phonology would produce an annotation scheme that can be used to conduct more appropriate evaluation.

When the model's characterization of the phonetic inventory becomes sufficiently fine-grained, it will be interesting to pursue the route of classical phonological analysis, but from a constructive perspective: allophones, alternations and perhaps some fragment of morpho-phonology may eventually find their place in the model. These goals are ambitious, but will hopefully appear within our reach as necessary progress is made towards the difficult problem of modeling language acquisition.

References

Albright, Adam and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90:119–161.

Allen, J. 1994. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2:567–577.

Alon, J., S. Sclaroff, G. Kollios, and V. Pavlovic. 2003. Discovering clusters in motion time-series data. In *Proc. IEEE Computer Vision and Pattern Recognition Conference*.

Amari, Shun-ichi. 1986. *Differential-geometrical methods in statistics*. Springer-Verlag, Berlin.

Amari, Shun-ichi and Hiroshi Nagaoka. 2000. *Methods of Information Geometry*. American Mathematical Society.

Anisfeld, M. 1984. *Language Development from Birth to Three*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bacchiani, M. 1999. *Speech recognition system design based on automatically derived units*. Ph.D. thesis, Boston University.

Bacchiani, M. and M. Ostendorf. 1999. Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, 29:99–114.

Bailey, Todd M. and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44:568–591.

Bates, E. and J. Goodman. 1997. On the inseparability of grammar and lexicon: Evidence from acquisition, aphasia and real time processing. *Language and Cognitive Processes*, 12:507–587.

Baum, E. and D. Haussler. 1989. What size net gives valid generalization? *Neural Computation*, 1(1):151–160.

Baum, Leonard E. 1972. An inequality and associated maximization technique in statistical estimation for probabalistic functions of markov processes. *Inequalities*, 3:1–8.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occuring in the statistical analysis of probabalistic functions in Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Beckman, M. E. and J. Edwards. 2000. The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*, 71:240–249.

Belkin, M. and P. Niyogi. 2003. Using manifold structure for partially labelled classification. In *Proceedings of Advances in Neural Information Processing Systems.*

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, 113(2):1001–1024.

Benedict, H. 1979. Early lexical development: Comprehension and production. *Journal of Child Language*, 6:183–200.

Bilmes, J. 2002. What can HMMs do? Technical Report 0003, Dept. of Electrical Engineering, University of Washington.

Bilmes, Jeff A. 1997. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models. Technical Report TR-97-021, International Computer Science Institute.

Blumer, Anselm, Andrzei Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36:929–965.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulary and perceptual drives.* Ph.D. thesis, University of Amsterdam.

Brent, Michael R. and Timothy A. Cartwright. 1996. Lexical categorization: Fitting template grammars by incremental MDL optimization. In Laurent Micla and Colin de la Higuera, editors, *Grammatical Inference: Learning Syntax from Sentences.* Springer, NY, pages 84–94.

Brent, Michael R. and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:B33–B34.

Browman, Catherine and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology*, 6:201–251.

Carter, David M. 1987. An information-theoretic analysis of phonetic dictionary access. *Computer, Speech and Language*, 2:1–11.

Cartwright, Timothy Andrew and Michael R. Brent. 1994. Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition. In *Proc. of the 16th Annual Meeting of the Cognitive Science Society*, Hillsdale, New Jersey.

Charles-Luce, J. and P. A. Luce. 1990. Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, 17:205–515.

Chazan, Dan. 2000. Speech reconstrution from Mel-cepstral coefficients and pitch. In *Proceedings of ICASSP.*

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English.* MIT Press, Cambridge, Massachusetts.

Clements, G. N. and E. Hume. 1995. The internal organization of speech sounds. In J. Goldsmith, editor, *Handbook of Phonological Theory*. Blackwell, pages 245–306.

Cole, R. and L. Hou. 1988. Segmentation and broad classification of continuous speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Cole, R. A. and C. A. Perfetti. 1980. Listening for mispronunciations in a children's story. *Journal of Verbal Learning and Verbal Behavior*, 19:297–315.

Cover, T. M. and J. A. Thomas. 1991. *Elements of Information Theory*. Willey & Sons.

de Marcken, Carl. 1995. The unsupervised acquisition of a lexicon from continuous speech. Massachusetts Institute of Technology, Technical Report, A.I. Memo 1558.

de Marcken, Carl. 1996. *Unsupervised language acquisition*. Ph.D. thesis, Massachusetts Institute of Technology.

de Saussure, Ferdinand. 1907. *Premier Cours de Linguistique Générale*. Pergamon, NY. 1996 French-English edition, with English translation by George Wolf, edited by Eisuke Komatsu.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B(39):1–38.

Deng, L., G. Ramsay, and D. Sun. 1997. Production models as a structural basis for automatic speech recognition. *Speech Communication*, 22:93–111.

Deng, Li. 1999. Computational models for speech production. In Keith Ponting, editor, *Computational Models of Speech Pattern Processing*, NATO ASI series. Springer, Berlin, pages 199–213.

Duda, R. O., P. E. Hart, and D. G. Stork. 2000. *Pattern Classification.* Wiley-Interscience, 2 edition.

Edwards, J., M. E. Beckman, and B. Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47:421–436.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. 1971. Speech perception in infants. *Science*, pages 303–306.

Elman, Jeffrey L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development.* Bradford Books/MIT Press.

Fenson, Larry, Philip S. Dale, J. Steven Reznick, and Donna Thal, 1991. *Technical Manual for the MacArthur Communicative Development Inventories.* San Diego State University, November.

Ferguson, Charles and C. Farwell. 1975. Words and sound in early language acquisition: English initial consonants in the first fifty words. *Language*, 51.

Fernald, Anne and Patricia Kuhl. 1987. Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 8:181–195.

Firth, J. R. 1948. Sounds and prosodies. *Transactions of the Philological Society.*

Flemming, Edward. 1995. *Auditory Representations in Phonology.* Ph.D. thesis, UCLA, Los Angeles.

Fletcher, H. 1953. *Speech and Hearing in Communication.* Kreiger, New York.

Fowler, A. E. 1991. How early phonological development might set the stage for phonological awareness. In S. Brady and D. Shankweiler, editors, *Phonological Processes in Literacy: A Tribute to Isabelle Y. Liberman*. Lawrence Erlbaum, Hillsdale.

Fowler, Carol. 1986. An event approach to the study of speech perception from a direct-realistic perspective. *Journal of Phonetics*, 14:3–28.

Frisch, S. A., M. B. Broe, and J. B. Pierrehumbert. 1995. The role of similarity in phonology: Explaining OCP-place. In K. Elenius and P. Branderud, editors, *Proceedings of the 13th International Congress of the Phonetic Sciences*, pages 544–547.

Furui, Sadaoki. 1986. On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America*, 80:1016–1025.

Garlock, V. M., A. C. Walley, and J. L. Metsala. 2001. Age-of-acquisition, word frequency and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language*, 45:468–492.

Garnica, O. K. 1973. The development of phonemic perception. In T. E. Moore, editor, *Cognitive Development and the Acquisition of Language*. Academic Press, New York.

Garofolo, J. S. 1988. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST).

Geman, S., E. Bienenstock, and R. Doursat. 1991. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Gerken, L. A. 1994. Child phonology: Past research, present questions, future directions. In M. A. Gernsbacher, editor, *Handbook of Psycholinguistics*. Academic Press, New York.

Gerken, L. A., W. D. Murphy, and R. N. Aslin. 1995. Three- and four-year-olds' perceptual confusions for spoken words. *Perception and Psychophysics*, 57:475–486.

Gildea, Daniel and Daniel Jurafsky. 1996. Learning bias and phonological rule induction. *Computational Linguistics*, 22:497–530.

Glass, James R. 1988. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Glass, James R. and Victor W. Zue. 1988. Multi-level acoustic segmentation of continuous speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 429–432.

Goldinger, S. D. 1997. Words and voices: Perception and production in an episodic lexicon. In Keith Johnson and John W. Mullennix, editors, *Talker Variability in Speech Processing.* Academic Press, San Diego, CA, pages 9–32.

Greenberg, J. H. and J. J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word*, 20:157–177.

Grenander, Ulf. 1996. *Elements of Pattern Theory.* Johns Hopkins University Press.

Guenther, F. H. 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102:594–621.

Harm, Michael W. and Mark S. Seidenberg. 1999. Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, 106(3):491–528.

Hasegawa-Johnson, Mark, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchho, Karen Livescu, Srividya Mohan, Jennifer Muller, Kemal Sonmez, and Tianyu Wang. 2005. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. In *Proceedings of ICASSP*.

Hastie, T., R. Tibshirani, and J. Friedman. 2002. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Hayes, Bruce P. 2004. Phonological acquisition in Optimality Theory: the early stages. In Rene Kager, Joe Pater, and W. Zonneveld, editors, *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, pages 158–203.

Hillenbrand, J., L. Getty, M. Clark, and K. Wheeler. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97:3099–3111.

Hinton, G. E. 1990. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46:47–75.

Hockett, Charles F. 1960. The origin of speech. *Scientific American*, 203:88–96.

Ingram, D. 1979. Phonological patterns in the speech of young children. In P. Fletcher and M. Garman, editors, *Language Acquisition*.

Itakura, Fumitada. 1975. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. ASSP*, 23:67–72.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jakobson, Roman. 1941. *Child Language, Aphasia and Phonological Universals*. Mouton, the Hague, Paris.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1952. *Preliminaries to Speech Analysis*. MIT Press, Cambridge, MA.

Jelinek, F. 1976. Continuous speech recognition by statisical methods. *IEEE Proceedings*, 64(4):532–556.

Jelinek, Fred. 1997. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA.

Johnson, Elizabeth K. and Peter W. Jusczyk. 2001. Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567.

Johnson, Keith. 1997a. The auditory/perceptual basis for speech segmentation. In *OSU Working Papers in Linguistics*, volume 50.

Johnson, Keith. 1997b. Speech perception without speaker normalization: An exemplar model. In Keith Johnson and John W. Mullennix, editors, *Talker Variability in Speech Processing.* Academic Press, San Diego, CA, pages 145–166.

Juang, Biing-Hwang and Lawrence Rabiner. 1990. The segmental K-means algorithm for estimating parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1639–1641.

Juneja, A. and C. Espy-Wilson. 2003. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In *Proceedings of International Joint Conference on Neural Networks*, Portland, Oregon.

Jusczyk, P. W. 1993. From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21:3–28.

Jusczyk, P. W. and R. N. Aslin. 1995. Infants' detection of sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1–23.

Jusczyk, P. W., J. Bertoncini, R. Bijeljac-Babic, L. J. Kennedy, and J. Mehler. 1990. The role of attention in speech perception by infants. *Cognitive Development*, 5:265–286.

Jusczyk, P. W., A. Cutler, and N. Redanz. 1993. Preference for the predominant stress patterns of English words. *Child Development*, 64:675–687.

Jusczyk, P. W. and E. A. Hohne. 1997. Infants' memory for spoken words. *Science*, (277):1984–1986.

Jusczyk, Peter. 1986. Toward a model of the development of speech perception. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*. Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, New Jersey, pages 1–18.

Jusczyk, Peter. 1992. Developing phonological categories from the speech signal. In Charles Ferguson, Lise Menn, and Carol Stoel-Gammon, editors, *Phonological Development: Models, Research, Implications*. York Press, Timonium, Maryland.

Jusczyk, Peter. 1997. *The Discovery of Spoken Language*. MIT Press, Cambridge, MA.

Jusczyk, Peter W. 1994. Infants' speech perception and the development of the mental lexicon. In Judith C. Goodman and Howard C. Nusbaum, editors, *The Development of Speech Perception*. MIT Press, Cambridge, MA.

Jusczyk, Peter W., Angela D. Friederici, Jeanine M. I. Wessels, Vigdis Y. Svenkerud, and Ann Marie Jusczyk. 1993. Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, (32):402–420.

Jusczyk, Peter W., Paul A. Luce, and Jan Charles-Luce. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, (33):630–645.

Kaisse, Ellen M. 2000. Laterals are *[-continuant]*. manuscript, University of Washington.

Keating, P. A. 1998. Word-level phonetic variation in large speech corpora. In A. Alexiadou, editor, *ZAS Papers in Linguistics: The Word as a Phonetic Unit*, volume 11, pages 35–50.

Keating, P. A., Peggy MacEachern, Aaron Shryock, and Sylvia Dominguez. 1994. A manual for phonetic transcription: Segmentation and labeling of words in spontaneous speech. In *UCLA Working Papers in Phonetics*, volume 88, pages 91–120, Los Angeles.

Kewley-Port, Diane. 1983. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73:1779–1793.

Kimball, Owen. 1994. *Segment Modeling alternatives for continuous speech recognition.* Ph.D. thesis, Boston University.

Koizumi, Takuya, Mikio Mori, Shuji Taniguchi, and Mitsutoshi Maruya. 1996. Recurrent neural networks for phoneme recognition. In *Proceedings of the 4th International Conference on Spoken Language Processing*, volume 1.

Kuhl, P. K. 1993. Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, 21:125–139.

Kuhl, Patricia K. 2004. Early language acquisition: Cracking the speech code. *Nature Reviews: Neuroscience*, 5:831–843, Nov.

Kullback, S. and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March.

Lacerda, Francisco. 1998. An exemplar-based account of emergent phonetic categories. *The Journal of the Acoustical Society of America*, 103(5):2980–2981, May.

Lacerda, Francisco and Ulla Sundberg. 2004. An echological theory of language learning. In *148th Annual Meeting of the Acoustical Society of America*, San Diego.

Ladefoged, P. 2001. *A Course in Phonetics.* Harcourt Brace, 4th edition.

Lasky, R. E., A. Syldal-Lasky, and R. E. Klein. 1975. VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 20:215–225.

Lecanuet, J. and C. Granier-Deferre. 1993. Speech stimuli in the fetal environment. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, and J. Morton, editors, *Developmental Neurocognition: Speech and face processing in the first year of life*. Kluwer Academic.

Lee, C. H., B. H. Juang, F. K. Soong, and L. R. Rabiner. 1989. Word recognition using whole-word units and sub-word units. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 683–686.

Leopold, W. F. 1947. *Speech development of a bilingual child: Sound learning in the first two years*, volume 2. Northwestern University, Evanston.

Leung, Hong C. and Victor W. Zue. 1988. Some phonetic recognition experiments using artificial neural nets. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Li, C. and G. Biswas. 2002. Applying the hidden Markov model methodology for unsupervised learning of temporal data. *International Journal of Knowledge-based Intelligent Engineering Systems*, 6(3):152–160.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and Studdert-Kennedy M. 1967. Perception of the speech code. *Psychological Review*, 74:431–461.

Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer. 1977. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber and D. L. Scarborough, editors, *Toward a Psychology of Reading*. Lawrence Erlbaum.

Lindblom, Björn. 1992. Phonological units as adaptive emergents of lexical development. In Charles Ferguson, Lise Menn, and Carol Stoel-Gammon, editors, *Phono-*

logical Development: Models, Research, Implications. York Press, Timonium, Maryland.

Lindblom, Björn. 2000. Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, 57:297–314.

Linde, Y., A. Buzo, and R. M. Gray. 1980. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28:84–95.

Locke, John L. 1983. *Phonological acquisition and change.* Academic Press, New York.

Locke, John L. 1993. *The child's path to spoken language.* Harvard University Press.

Luce, P. A. 1986. Neighborhoods of words in the mental lexicon. Technical Report 6, Department of Psychology, Indiana University, Bloomington, IN.

MacNeilage, P. F. 1998. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21:499–511.

MacWhinney, B. and C. Snow. 1985. The child language data exchange system. *Journal of Child Language*, 12:271–296.

Marr, David. 1982. *Vision.* Freeman, San Francisco.

Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word-recognition. In U. H. Frauenfelder and L. K. Tyler, editors, *Spoken Word Recognition.* MIT Press.

Mattys, S. L. and P. W. Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78:91–121.

Maye, J. and L. Gerken. 2000. Learning phoneme categories without minimal pairs. In *Proceedings of the 24th Annual Boston University Conference on Language Development*, pages 522–533, Somerville, MA. Cascadilla Press.

Maye, J., J. F. Werker, and L. Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:101–111.

McAuley, R. J. and T. F. Quatiery. 1986. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Tran. ASSP*, 34:744–754.

McLachlan, G.J. and K.E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Mehler, J., P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, (29):143–178.

Meng, Helen M. and Victor W. Zue. 1991. Signal representation comparison for phonetic classification. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*

Menn, Lise. 1983. Development of articulatory, phonetic and phonological capabilities. In B. Butterworth, editor, *Language Production*, volume 2. Academic Press.

Menn, Lise. 1992. The two-lexicon model of child phonology: Looking back, looking ahead. In Charles Ferguson, Lise Menn, and Carol Stoel-Gammon, editors, *Phonological Development: Models, Research, Implications.* York Press, Timonium, Maryland.

Menyuk, P., L. Menn, and R. Silber. 1986. Early strategies for the perception and production of words and sounds. In P. Fletcher and M. Garman, editors, *Language Acquisition.* Cambridge University Press, pages 198–222.

Metsala, J. L. and A. C. Walley. 1998. Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala and L. C. Ehri, editors, *Word Recognition in Beginning Literacy.* Lawrence Erlbaum, New York, pages 89–120.

Mielke, Jeff. 2004. *The Emergence of Distinctive Features.* Ph.D. thesis, The Ohio State University, Columbus, OH.

Milner, Ben and Xu Shao. 2002. Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In *Proceedings of ICSLP.*

Mitchell, T. 1997. *Machine Learning.* McGraw Hill.

Mumford, David. 2002. Pattern theory: The mathematics of perception. In *Proceedings of the International Congress of Mathematics.*

Nearey, Terrance M. 1997. Speech perception as pattern recognition. *Journal of the Acoustic Society of America*, 101(6):3241–3254.

Nearey, Terrance M. 2001. On the factorability of phonological units. In *LabPhon 6: The Sixth Conference on Labotory Phonology.*

Nearey, Terrance M. and Peter F. Assman. 1986. Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80:1297–1308.

Ninio, A. 1993. On the fringes of the system: children's acquisition of syntactically isolated forms at the onset of speech. *First Language*, (13):291–314.

Nittrouer, S., M. Studdert-Kennedy, and R.S. McGowan. 1989. The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, 32:120–132.

Nosofsky, Robert M. 1991. Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, (17):3–27.

Ostendorf, M., V. Digalakis, and O. Kimball. 1996. From HMM's to segment models: A unified view of stochastic modelling for speech recognition. *IEEE Transactions on Speech and Audio processing*, 4(5).

Ostendorf, Mari and S. Roukos. 1989. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. ASSP*, 37(12):1857–1869.

Paliwal, K. K. 1990. Lexicon-building methods for an acoustic sub-word based speech recognizer. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Pater, J., C. Stager, and J. Werker. 2004. The lexical acquisition of phonological contrasts. *Language*, 80.

Peters, A. 1983. *The Units of Language Acquisition*. Cambridge University Press, Cambridge.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper, editors, *Frequency Effects and the Emergence of Linguistic Structure*. John Benjamins, Amsterdam, pages 137–157.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In C. Gussenhoven and N. Warner, editors, *Laboratory Phonology VII*, pages 101–140. Mouton de Gruyter.

Pierrehumbert, Janet. 2003. Probabilistic phonology: Discrimation and robustness. In R. Bod, J. Hay, and S. Jannedy, editors, *Probabilistic Linguistics*. MIT Press.

Pike, K. L. 1943. *Phonetics*. University of Michigan Press.

Plaut, David C. and Christopher T. Kello. 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In Brian MacWhinney, editor, *The Emergence of Language*. Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, New Jersey, pages 3–50.

Rabiner, L. and B. H. Juang. 1993. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ.

Rabiner, L. R., C. H. Lee, B. H. Juang, and J. G. Wilpon. 1989. HMM clustering for connected word recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 405–408.

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Ratner, N. B. 1996. From "signal to syntax": But what is the nature of the signal? In James Morgan and Katherine Demuth, editors, *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* Lawrence Erlbaum Associates, chapter 9, pages 135–150.

Riccardi, Giuseppe. 2000. On-line learning of acoustic and lexical units for domain-independent ASR. In *Proceedings of the 4th International Conference on Spoken Language Processing.*

Saffran, J., R. Aslin, and E. Newport. 1996. Statistical learning by 8-month old infants. *Science*, (274):1926.

Saffran, J. R., E. K. Johnson, and R. N. Aslin. 1996. Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35:606–621.

Seneff, S. 1986. A computational model for the peripheral auditory system: Application to speech recognition research. In *Proceedings of ICASSP*, pages 1983–1986.

Seneff, Stephanie. 1985. *Pitch and Spectral analysis of speech based on an auditory synchrony model.* Ph.D. thesis, MIT, Cambridge, Mass.

Shipman, David W. and Victor W. Zue. 1982. Properties of large lexicons: Implication for advanced isolated word recognition systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 546–549.

Singh, Rita, Bhiksha Raj, and Richard M. Stern. 2000. Structured redefinition of sound units for improved speech recognition. In *Proceedings of the 6th International Conference on Speech and Language Processing*, Beijing, China.

Singh, Rita, Bhiksha Raj, and Richard M. Stern. 2002. Structured redefinition of sound units for improved speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(2), February.

Smith, Neilson. 1973. *The Acquisition of Phonology.* Cambridge University Press.

Smolensky, Paul. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence*, 46.

Smyth, P. 1997. Clustering sequences with hidden markov models. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing*, volume 9. MIT Press, Cambridge, MA.

Stager, C. L. and J. F. Werker. 1997. Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature*, 388:381–382.

Stevens, K. N. and S. E. Blumstein. 1978. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64:1358–1368.

Stevens, Kenneth. 1998. *Acoustic Phonetics.* MIT Press, Cambridge, MA.

Stevens, Kenneth N. 1971. The role of rapid spectrum changes in the production and perception of speech. In L. L. Hammerich, Roman Jakobson, and Eberhard Zwirner, editors, *Form and substance: Phonetic and Linguistic papers.* Akademisk Forlag, Copenhagen.

Storkel, H. L. 2001. Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, 44:1321–1337.

Streeter, L. A. 1976. Language perception of 2-month old infants shows effects of both innate mechanisms and experience. *Nature*, 259:39–41.

Sun, X. 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of ICASSP*.

Svendsen, T., K. K. Paliwal, E. Harborg, and P. O. Husøy. 1989. An improved subword based speech recognizer. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Glasgow.

Svendson, T. and F. K. Soong. 1987. On the automatic segmentation of speech signals. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 77–80.

Tanner, Martin and Wing Hung Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398).

Tenenbaum, J., V. de Silva, and J. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290.

Tesar, Bruce and Paul Smolensky. 1996. Learnability in optimality theory (long version). Technical Report JHU-CogSci-96-3, Department of Cognitive Science, Johns Hopkins University, Baltimore, Maryland.

Tokuda, K., T. Kobayashi, and S. Imai. 1995. Speech parameter generation from HMM using dynamic features. In *Proceedings of ICASSP*, pages 660–663.

Tokuda, Keiichi, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of ICASSP*.

Tomasello, Michael. 2003. *Constructing a Language.* Harvard University Press.

Tu, Z. W. and S. C. Zhu. 2002. Image segmentation by data-driven Markov Chain Monte Carlo. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 24(5):657–673, May.

Vapnik, V. N. 2000. *The Nature of Statistical Learning Theory.* Springer, NY.

Vapnik, V. N. and A. Y. Chernovenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.

Vernooij, G.J., G. Bloothooft, and Y. van Holsteijn. 1989. A simulation study on the usefulness of broad phonetic classification in automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 85–88.

Vihman, M. M. 1996. *Phonological Development: The Origins of Language in the Child*. Blackwell.

Waibel, Alexander, Toshiyuki Hazanawa, Geoffrey Hinton, and Kiyohiro Shikano. 1989. Phoneme recognition using time-delay neural networks. *IEEE Trans. ASSP*, 37(3):328–339.

Walley, A. C. 1993. The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review*, 13:286–350.

Walley, A. C. and C. Ancimer. 1990. Spoken word recognition by young children in individual and successive presentation formats of the gating paradigm. In *Proceedings of the Conference on Human Development*, Richmond, VA.

Walley, A. C. and J. L. Metsala. 1990. The growth of lexical constraints on spoken word recognition. *Perception and Psychophysics*, (47):267–280.

Warner, Natasha. 1998. *The role of dynamic cues on speech perception, spoken word recognition and phonological universals*. Ph.D. thesis, University of California, Berkeley.

Weijer, J. van de. 2001. The importance of single-word utterances for early word recognition. In *Proceedings of ELA 2001*, Lyon, France.

Weijer, J. van de. 2002. How much does an infant hear in a day? In J. Costa and M. Joao Freitas, editors, *Proceedings of the GALA 2001 Conference on Language Acquisition*, pages 279–282.

Werker, J. F. 2003. The acquisition of language specific phonetic categories in infancy. In *Proceedings of the 15th International Conference of Phonetics Sciences*, Barcelona.

Werker, J. F., J. E. Pegg, and P. McLeod. 1994. A cross-language comparison of infant preference for infant-directed speech: English and Cantonese. *Infant Behavior and Development*, (17):321–331.

Werker, J. F. and R. C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, (7):49–63.

Werker, J. F. and R. C. Tees. 1992. The organization and reorganization of human speech perception. In M. Cowan, editor, *Annual Review of Neuroscience*, volume 15. pages 377–402.

Wu, Yingnian. 1996. *Modeling general mixture components, with application to schizophrenic eye-tracking.* Ph.D. thesis, Harvard University, Cambridge, MA.