Chapter 3

A Lexical Contrast Model of Phonological Acquisition

If we take phonological units as a set of symbols that can be used combinatorially in lexical representation, models of phonological acquisition should aim to satisfactorily explain how such symbolic units emerge. This chapter presents a model of phonological acquisition that accounts for the simultaneous learning of abstract phonological categories, their mapping onto the relevant acoustic features, and symbolic lexical representations using the acquired phonological units. This learning model introduces a mechanism of phonological category creation and refinement without the assumption of innately available phonological features. Central to this model is the idea that the need to represent lexical contrast is the driving force behind the creation and adjustment of phonological categories. The model, like the infant learner, begins with no phonological knowledge. As the model acquires words with distinct meanings, the need for abstract representation arises, and the model creates phonologically meaningful contrasts within the acoustic space to allow appropriate representations of the words in the learner's lexicon.

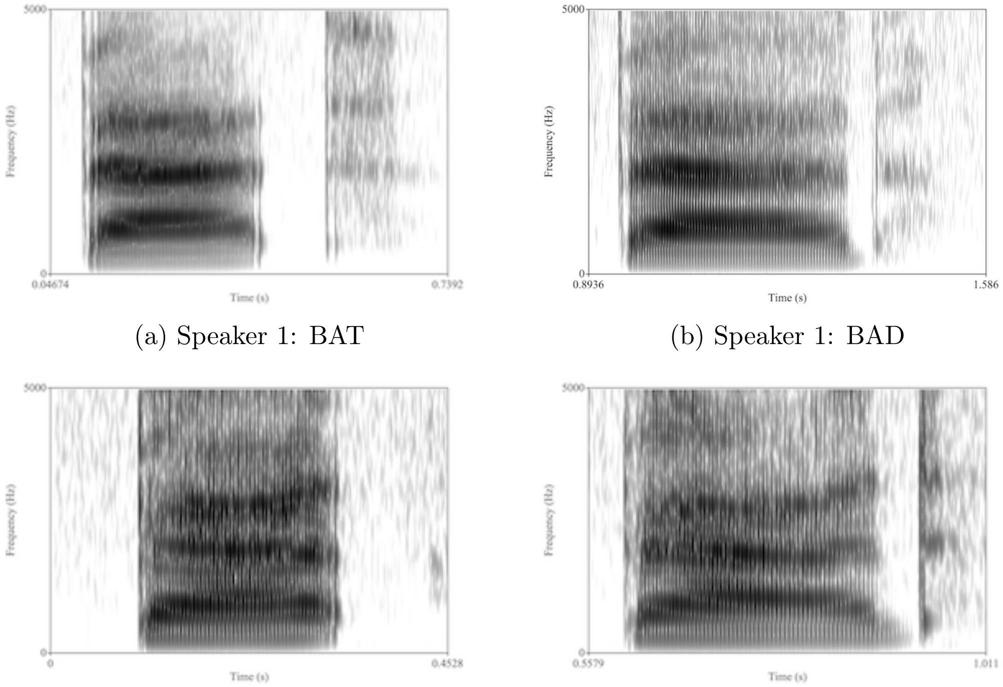
3.1 Lexical contrast and phonological acquisition

The notion of lexical contrast has a long history in phonology and was especially important in early approaches in phonology although it has received less attention in recent years (see Dresher, 2016, for a review). In phonological analysis, phonological distinctions are diagnosed via lexical contrast through the minimal pair test. More recently, researchers in language acquisition have given word learning a more central role in the acquisition of phonological knowledge (Jusczyk, 1997; Werker and Curtin, 2005). This section reviews and discusses the importance of lexical contrast in phonological representation and offers motivation for a path of acquisition through the continuous restructuring of the phonological space to accommodate lexical distinctions.

3.1.1 Minimal pairs and lexical contrast

Phonological analysis operates on the symbolic level, which rests on the identification of abstract units of representation. Minimal pairs are a very efficient way of doing so. A minimal pair is two words that have distinct meanings and differ by only one unit. The unit is often assumed to be a segment. For English, “bin” and “pin” can be used to establish that /b/ and /p/ are distinct segments, i.e., phonemes. In commonly used feature theories, /b/ and /p/ are also minimal in the sense that they differ by only one phonological feature [voice]. Words such as “shin” and “bin” are a minimal pair and differ by one phoneme, but /ʃ/ and /b/ differ by more than one phonological or articulatory feature. While /ʃ/ is a voiceless alveolar fricative, /b/ is a voiced bilabial stop. As such, [ʃ] and [b] would also be more acoustically distinct than [b] and [p]. Additionally, for languages with suprasegmental features, minimal pairs can be found with words that share the same segments but differ in other aspects of articulation, such as pitch or phonation.

What role do minimal pairs play in phonological acquisition? Approaches that emphasize phonetic learning view minimal pairs as unnecessary (Maye and Gerken, 2000) and favor statistical learning. This approach often draws heavily from the perceptual discrimination results. However, as discussed extensively in Chapter 2, although perceptual discrimination provides compelling evidence for early phonetic development on the perceptual level, these results do not necessarily map directly to the development of abstract phonological categories. In addition to understanding the developmental trajectory of the discriminatory abilities themselves, it is equally important to carefully consider whether and how phonetic discrimination is used by the learner to parse linguistic input.



(a) Speaker 1: BAT

(b) Speaker 1: BAD

(c) Speaker 2: BAT

(d) Speaker 2: BAD

Figure 3.1: Spectrograms of the minimal pair “bat” vs. “bad” by two speakers.

The picture becomes more complicated when the details acoustic realizations are taken into consideration. Take the minimal pair “bat” and “bad” in English. When transcribed phonemically, they are respectively /bæt/ and /bæd/. Based on the phonemic analysis of the adult grammar, one might expect a minimal pair-based learner to identify the last segment as distinct phonemes. However, the actual acoustics of the two words suggests that this process is far more involved. Figure 3.1 illustrates the complications from the acoustic signal. Figure 3.1a and Figure 3.1b are the minimal pair produced by speaker 1. As can be seen, the acoustic distinctions between these two words are far from minimal. First, the vowel of “bad” is longer (each pair is plotted on the same time scale). The closure for the /t/ in “bat” is longer than the /d/ in “bad”, and “bat” has a stronger release than “bad”. There is a small amount of voicing for the /d/ in “bad”. Since multiple acoustic cues differ between these two words, how does the learner figure out which ones are relevant? It would not be unreasonable to hypothesize that vowel length is the distinction between these two words, rather than the final consonant. Tokens from a second speaker further illustrate the challenge of learning from the acoustic signal. In Figure 3.1c, the final /t/ is unreleased. Figure 3.1d has more prevoicing and a fairly strong burst and release. Similar to speaker 1, the vowel in “bad” is longer than the vowel in “bat”. Clearly, a minimal phonological contrast does not correspond to a minimal phonetic contrast both within each speaker and across different speakers.

What, then, can the learner abstract away from knowing that the signal for “bat” and the signal for “bad” have different meanings and sounds? From two words that are acoustically different and referentially different, there is enough evidence that some contrast between them needs to be represented. This information is not sufficient to pinpoint the exact nature of this contrast, but learner can make an initial hypothesis about what to represent from the signal. Perhaps vowel length would be identified as the contrastive feature between “bat” and “bad”, if the learner happens to perceive duration as the most salient difference between these two words. Then, as the learner acquires from words with /æ/ or encounter /t/ and /d/ in other contexts, the learner can use the additional lexical knowledge to evaluate the hypothesis that vowel length is the distinctive feature between “bat” and “bad”. The important takeaway from these observations is that while the phonologist knows that “bat” and “bad” are a minimal pair, the learner does not. All the information the learner has is that these two words sound different and mean different things.

If a difference in signal and a difference in meaning are the only cues necessary for learning contrasts, the learner does not require phonological minimal pairs to start acquiring phonological contrasts. It is really the notion of lexical contrast that is important here. The words “fish” and “dog” differ by all three segments in adult English phonology. However, if these are the only two words a learner knows, the learner only needs two abstract symbols to represent them and can assign some acoustic salient cues to each symbol. In this initial state of phonology of the learner, “fish” and “dog” would actually be a minimal pair since they differ in sound and differ by one phonological unit of representation. Indeed, the phonological abstraction of what is contrastive is only *as detailed as the learner’s lexicon needs it to be.* Minimal pairs in adult phonology may not correspond to minimal pairs in a developing phonology because these phonologies can be very different. The minimal pairs in adult phonology are the end result of generalizing lexical contrasts over the acoustic space. Although the learner does not require minimal pairs to begin phonological acquisition, minimal pairs are nevertheless essential to the eventual refinement of phonological categories. Minimal pairs in the input grammar are words of high phonological signal, and they can help the learner to better pinpoint the relationship between abstract phonological units and their surface phonetic distinctions.

3.1.2 Phonological representation and lexical access

The phonological representations of words are accessed in word recognition. In mature adult phonology, homophones should have the same underlying phonological units, and experimental evidence suggests that this is in fact the case. Lexical decision tasks with homophones and non-word homophones show that words are phonologically encoded in the lexicon and that phonological processing occurs in the word recognition process. Some of this evidence comes from visual word recognition. Early work by Rubenstein et al. (1971) suggests that phonological processing does occur in lexical recognition. When subjects are presented with a homophonous non-word (e.g., brane), the reaction time is slower than phonotactically legal non-words without homophones. The longer latency for homophonous non-words is interpreted as longer search time as a result of phonemic matching. A separate experiment with all real words show that there is also a word frequency effect; low frequency homophones have higher latency and lower accuracy. Additionally, homophones facilitate the access of semantically related items (e.g., rows for flower, chare for table) (Van Orden, 1987; Lukatela and Turvey, 1991). Even though these experiments used orthography, the results indicate that orthography is parsed into some abstract phonemic representation, resulting in the observed effects from phonological homophones.

In the acoustic domain, word recognition is clearly not solely based on acoustics but rather combines acoustic and contextual cues. Because of the close association between phonology and phonetics, it would be easy to assume that phonology provides the mapping between acoustics and abstract forms. This is partially correct. Phonology is a function that combines all levels of information (phonetic, phonological, morphological, syntactic, semantic, and pragmatic) to produce an abstract representation. When listening to prose, subjects sometimes fail to identify words with a phoneme mispronounced, especially in word initial positions (Cole, 1973; Cole et al., 1978). The retrieval of words is highly dependent on context. Syntactic and semantic context play a role in lexical parsing (Marslen-Wilson, 1975; Marslen-Wilson and Welsh, 1978), and listeners struggle to identify words when they are removed from their conversational context (Pollack and Pickett, 1963). On the segmental level, phoneme identification is also associated with contextual predictability of the words they occur in (Morton and Long, 1976).

3.1.3 Early lexical representation and underspecification

Research in lexical acquisition shows that word learning begins early (Borden et al., 1983; Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012), and that infants are aware of phonetic details in familiar words (e.g., Jusczyk and Aslin, 1995; Swingley, 2005, 2009; Mani and Plunkett, 2010). However, not all phonetic details may be encoded as phonologically relevant by the learner (Van der Feest and Fikkert, 2015). When the nuances of perceptual identification are investigated, it appears that certain aspects of words are remembered better than others. For example, the stressed portion of the word is better represented. For bisyllabic words, 11-month-old French infants failed to recognize familiar words when the medial consonant was modified, but still recognized the words when the initial consonant was changed in manner or voicing (Hallé and de Boysson-Bardies, 1996). The stress pattern in English is different, and early perception reflects this difference. At 11 months, English-learning infants did not recognize familiar words when the initial consonant was modified, but tolerated modifications to the medial consonant (Vihman et al., 2004).

Another line of research suggests that early representation is more holistic than segmental. In production especially, word forms appear be represented more holistically early on, and often only salient details are retained (Ferguson and Farwell, 1975; Walley, 1993). A number of studies suggest that early lexical representation may be phonologically underspecified (Hallé and de Boysson-Bardies, 1996). Moreover, young children process phonetic similarity on the syllabic level rather than phonemic level, and they are better at identifying items that share multiple phonemes than a single phoneme (Treiman et al., 1981; Walley et al., 1986). Also, children are more influenced by coarticulatory cues. For example, they rely more on vowel formant transitions in identifying fricatives than adults (Nittrouer and Studdert-Kennedy, 1987; Nittrouer et al., 1989).

3.1.4 Word learning and referent resolution

How young children learn the meaning of words is an important research question. Much like acoustic data, the signal for word-referent mappings is extremely noisy. Even nouns referring to concrete objects can be difficult to identify since many interpretations can fit the scene in which they are uttered. However, even at a very early stage of word learning, infants are able to identify the intended referents to their acoustic forms (Bergelson and Swingley, 2012; Mani and Plunkett, 2010; Tincoff and Jusczyk, 2012). Different mechanisms have been proposed to account for the acquisition of word-referent mapping. Mutual exclusivity (i.e., no two words can have identical meaning) can help constrain the learning of new words (Markman and Wachtel, 1988; Markman et al., 2003). Cross-situational statistics, through which the learner keeps track of common signal and objects across multiple scenes, offers one account for the learning of word-referent mappings (Smith and Yu, 2008).

There is a lot of active research in this area, but it is beyond the scope of this dissertation to address how referents are identified. The model described in the next section incorporates a random element in the acquisition of words, but it does not propose a mechanism through which the correct identification of the referent is achieved.

3.2 A model of phonological emergence

This section introduces a concrete mechanism whereby the learner acquires discrete phonological representations from continuous, variable acoustic signal. Given a set of words in a lexicon and their corresponding acoustic realizations, the model arrives at the relevant phonological features that best represent the contrasts in the lexicon. The two components of the model are the lexicon and its associated phonology. The lexicon stores each word's phonetic representation including exemplars, frequency, and its abstract representation according to the current state of the learner's phonology. The learner's phonological knowledge describes the relationship between acoustic cues and abstract phonological categories. For each phonologically contrastive dimension, the phonological knowledge enables the learner to transform the acoustic signal into abstract representations by paying attention to the cues that are informative for each contrast. At the end of learning, the model acquires 1) the appropriate number of phonological contrasts that are best suited to represent the lexicon, 2) which acoustic cues matter for each contrast, and 3) the abstract symbolic representation for each word in the lexicon.

This section describes the components and operations of the model and discusses the emergent properties of the model. To fully validate the model, the results from a computational experiment using acoustic data extracted from the Philadelphia Neighborhood Corpus is presented in the following section.

3.2.1 Lexical learning

Lexical learning begins early and forms the foundation of phonological learning (cf. Section 3.1.3). In this model, the lexicon module stores information about words that the learner has been exposed to. The learner keeps track of three pieces of information for each referent: its average (i.e., prototypical) acoustic signal, phonological representation, and frequency. The structure of the lexicon is illustrated in Figure 3.2.

The learner begins with no words in the lexicon. At each learning iteration, the learner is presented with the referent of a word and its acoustic signal. The model assumes that the learner is always able to correctly identify an acoustical signal with its referent, as in a perfect lab learning situation. The mapping between the signal and its referent is by no means a simple problem in language acquisition, but it is not a problem that this model

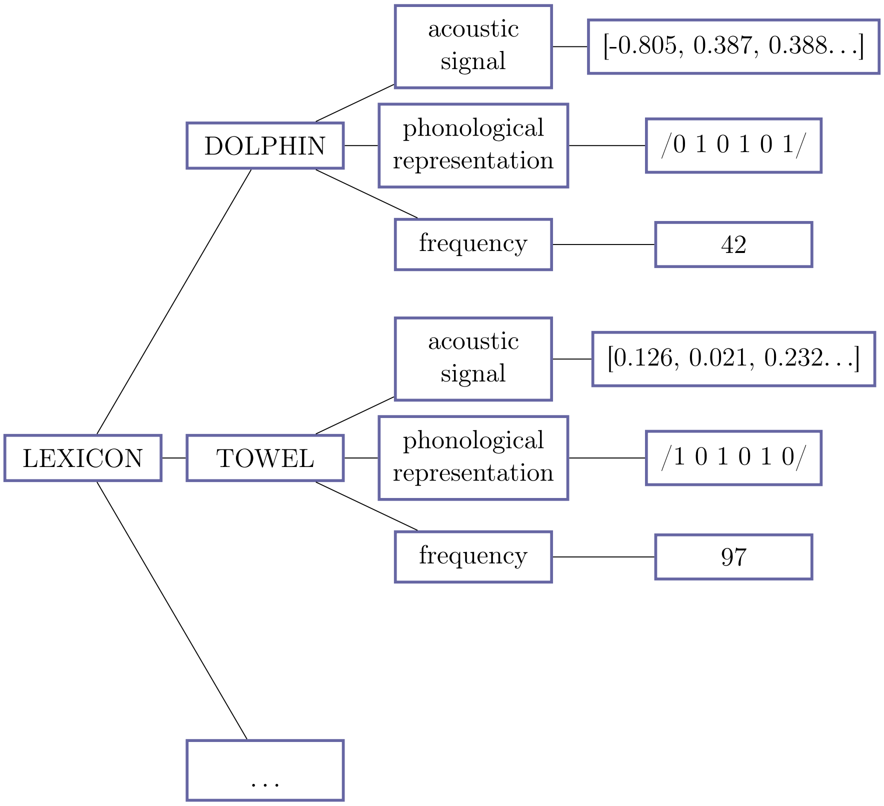


Figure 3.2: The structure of the lexicon.

aims to solve. As acoustic tokens for each referent are presented, the learner begins building up their knowledge of the phonetic forms that are associated with each referent. Since this model is primarily concerned with phonological acquisition, I make simplifying assumptions about the representation of a word's syntax, semantics, and pragmatics. The phonological learning part of this model only requires the learner to identify words as distinct in meaning along any of the dimensions of linguistic contrast.

The phonetic knowledge part of the lexicon reflect the learner's overall experience with phonetic forms of a word, and it includes any acoustic cue that the learner perceives from the input, both phonologically relevant cues and cues that do not contribute to any phonological contrast in the language. This phonetic knowledge is represented as the average of all the acoustic realizations corresponding to a referent, and it is updated each time an acoustic token for a referent is heard. As a result, after hearing a number of acoustic realizations identifying a referent, the learner knows what a typical realization sounds like for this referent, and this process effectively creates an acoustic prototype for the phonetic realization of a word. After each iteration, the acoustic knowledge according to Equation 3.1, and frequency is updated according to Equation 3.2.

$$s = \frac{s \times f + s_i}{f + 1} \tag{3.1}$$

$$f = f + 1 \tag{3.2}$$

where:

f = word frequency; the number times a word has been heard

s = the existing prototypical (average) signal of a word

s_i = a specific acoustic token of the word

Before a word can make an impact on phonological learning, the learner needs enough familiarity with the word to be able to recognize it consistently. To simulate the increasing familiarity with a word with exposure, a simple frequency-based memory system is used to model the acquisition of words. The more frequently a word has been heard, the more likely that it is acquired by the learner and used in phonological learning. Before a word is acquired, the learner only updates their knowledge of the word on the phonetic level, and its phonological form is determined at the point of word acquisition. The acquisition of phonological contrasts and representations will be discussed in the following section.

The acquisition of a word is implemented as a probabilistic process with the likelihood increasing as the frequency of the word increases. After each token is heard, a random acquisition threshold t is generated from a uniform distribution between 0 and 1 (Equation 3.3). A random threshold is used to implement some noise in the learning process. The familiarity of a word is modeled as a logistic function (cf. Anderson et al., 1998) in Equation 3.4 (illustrated in Figure 3.3 for $k = 20$). If the familiarity r of the word is greater than the threshold t, the word is marked as acquired and pass onto the phonology module to be assigned a phonological representation.

$$t = \text{unif}(0, 1) \tag{3.3}$$

$$r = \frac{1.0}{1.0 + e^{-(f-k)}} \tag{3.4}$$

where:

t = threshold at which a word is considered acquired

r = familiarity to the word

k = the word frequency at which $r = 0.5$

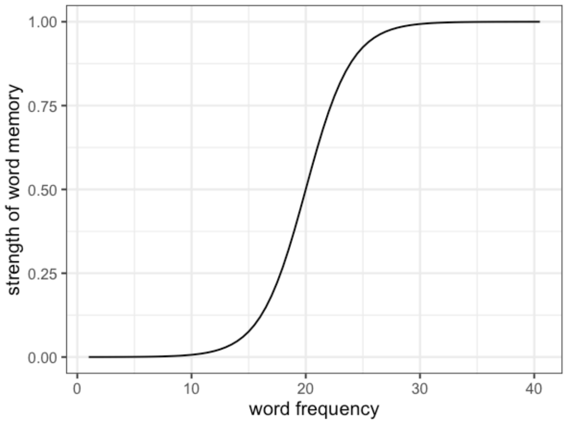
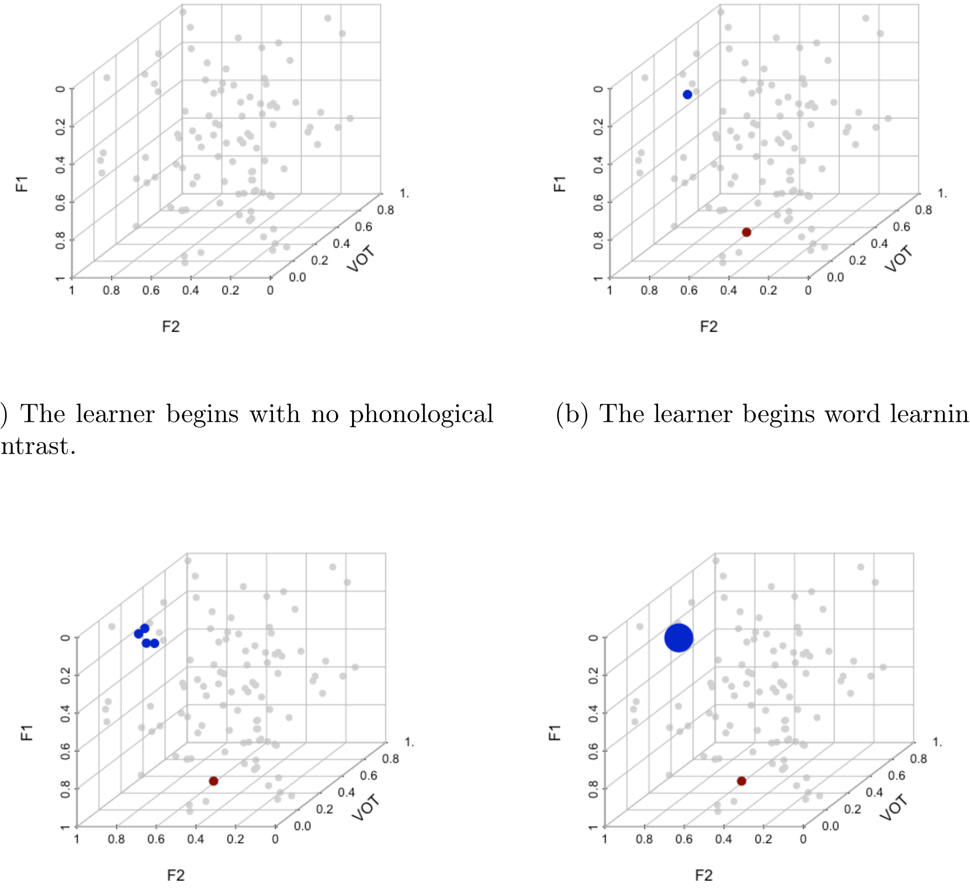


Figure 3.3: The probability of word familiarity as a function of word frequency.

Figure 3.4 illustrates this process of word learning. These illustrations assume a toy language with only three acoustic dimensions (VOT, F1, F2) on the phonetic level and an unknown number of words. Figure 3.4a represents the stage prior to any lexical learning, and each grey dot represents some acoustic token of the words in this language. In Figure 3.4b, the learner begins paying attention to certain words, as represented by the BLUE and RED dots. Dots of the same color represent acoustic tokens that have the same referent. In Figure 3.4c, the learner is exposed to more tokens of BLUE. After some amount of exposure, the learner acquires BLUE, as represented by the big BLUE dot in Figure 3.4d). Further lexical



(a) The learner begins with no phonological contrast.

(b) The learner begins word learning.

(c) The learner hears many tokens of a BLUE.

(d) The learner acquires BLUE.

Figure 3.4: An illustration of lexical acquisition.

acquisition occurs the same way. After the learner hears tokens of the same word multiple times, the learner acquires this word and can use this word in phonological acquisition.

3.2.2 Phonological learning

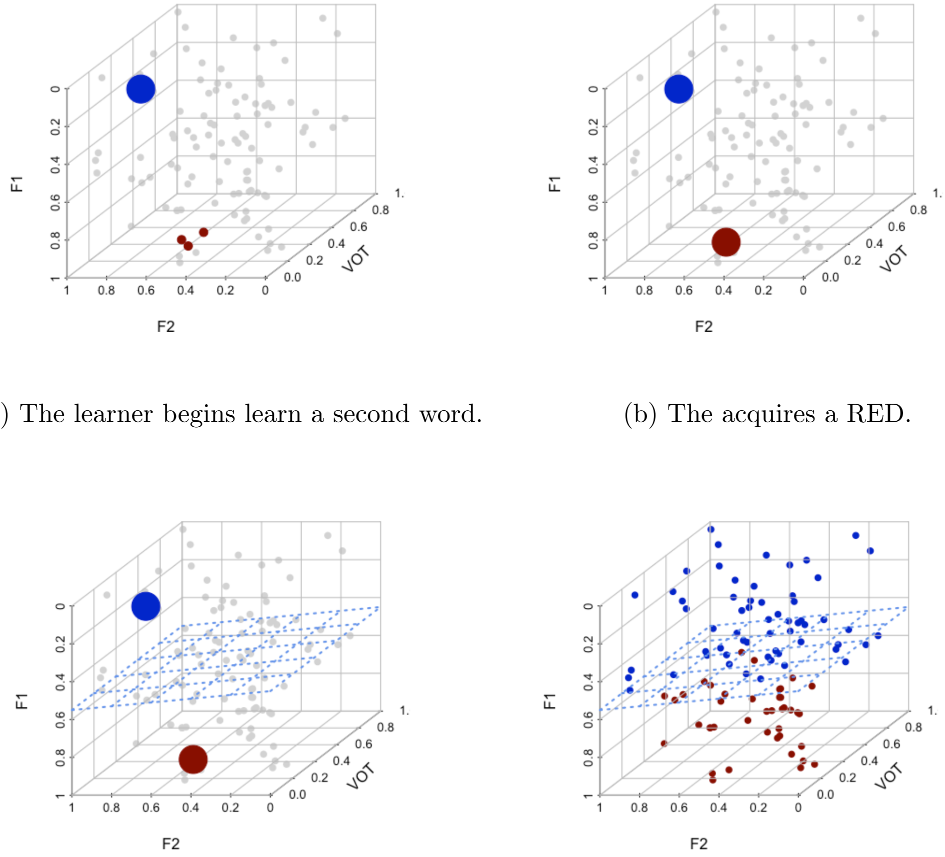
Phonological learning occurs as the learner continuously makes hypotheses about how to transform the phonetic signal into abstract phonological categories that best represent the current lexical distinctions in the learner's lexicon. The learning is unsupervised and non-parametric; the learner does not know which phonological distinctions exist in the input and is not given target representations. The learner's representations of words are updated dynamically as the learner acquires words and phonological contrasts.

The phonological module of the model consists of three processes: contrast creation, contrast adjustment, and contrast consolidation. In contrast creation, the learner adds a phonological contrast when the current number of contrasts is insufficient for representing the lexicon. After its initial creation, each contrast is updated as more words are learned and assigned to either side of the phonological boundary. Finally, should two contrasts become functionally the same after updates, they are consolidated into one contrast.

3.2.2.1 Contrast creation

After a period of lexical learning, the learner will begin to recognize familiar words. When the learner acquires two words that are distinct in meaning, the learner needs to create the first phonological contrast that allows them to represent these two words distinctly. This is illustrated in Figure 3.5b, where the learner has acquired both BLUE and RED. To create the first contrast, the learner creates a division in the phonetic space that separates these two words based on the salience of the acoustic cues that distinguish these two words. The light blue plane in Figure 3.5c represents phonological CONTRAST #1, created after the learner has acquired BLUE and RED. Since these two words appear to be most distinct in F1, the plane cuts through the acoustic space mostly along the F1 dimension, with some tilt along the F2 dimension. The learner will be able to represent any subsequent acoustic tokens along this contrastive plane (Figure 3.5d). If the learner identifies another pair of words as distinct in meaning but current phonology represents them in the same way (BLUE and PURPLE in Figure 3.6a), the learner can create an additional contrast (the mostly vertical plane CONTRAST #2) to accommodate this need for distinct representation (Figure 3.6b). The number of phonological contrasts grows as the learner gains more vocabulary.

In the computational implementation, the learner's phonological knowledge is represented as a matrix W, where each column corresponds to an acquired phonological plane that divides the multidimensional acoustic space (Equation 3.5). At the beginning of learn-



(a) The learner begins learn a second word.

(b) The acquires a RED.

(c) The learner creates a phonological contrast in the acoustic space.

(d) The learner can use this acquired contrast to classify any token in this acoustic space.

Figure 3.5: An illustration of phonological contrast creation.

ing W is empty. Upon acquiring the first two words, the first phonological contrast is created. To create this contrast, the model compares the acoustic signals of the two words and determines the most acoustically salient cues between the two words. The relative salience of cues is calculated as the absolute value of the differences between each cue of the two words. Then, a phonological contrast is constructed as the plane equidistant from the most distinctive acoustic cues in the two words (Equation 3.6). Subsequent phonological contrasts are created in the same fashion, and phonological representations are assigned to each word using sigmoidal activation (Equation 3.7).

$$W = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{pmatrix} \tag{3.5}$$

$$\begin{aligned} W_{2:m,j} &= a_1 - \frac{a_1 + a_2}{2} \\ W_{1,j} &= -W_{2:m,j} \cdot \frac{a_1 + a_2}{2} \end{aligned} \tag{3.6}$$

$$p = \frac{1.0}{1.0 + e^{-Ws_i}} \tag{3.7}$$

where:

W = a matrix where each column is a phonological division in the acoustic space

$W_{1:m,j}$ = weights for the jth phonological contrast

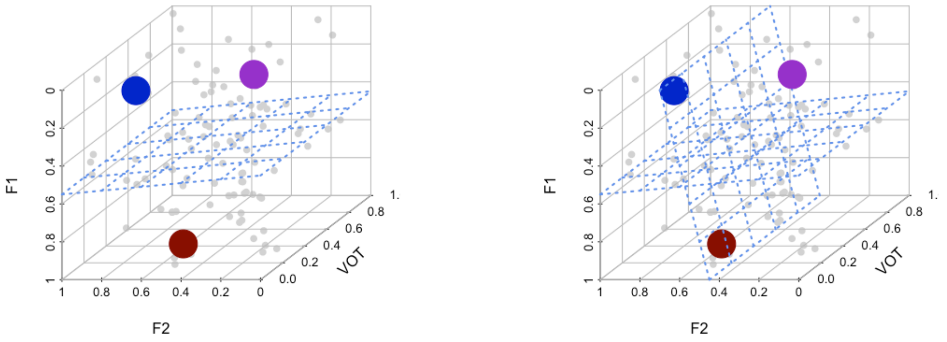
a_1, a_2 = the acoustically salient part of the signals of two distinct words

s_i = the acoustic signal from some word

p = the phonological representation

3.2.2.2 Contrast update and adjustment

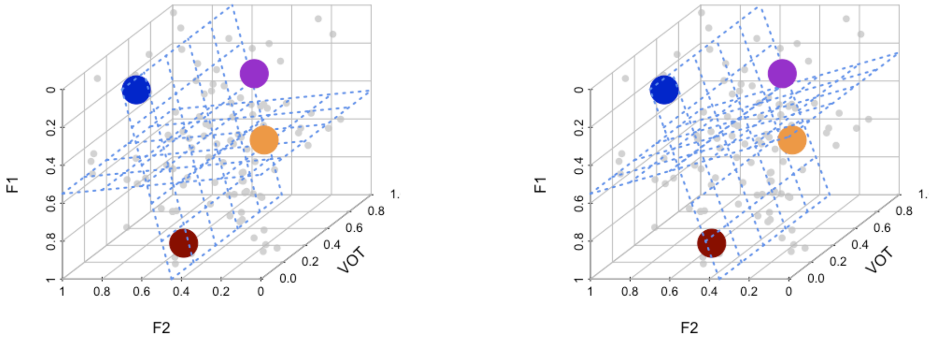
In addition to creating more phonological distinctions to represent the growing vocabulary, the phonological planes can also shift to to distinguish newly acquired word distinctions. This operation can be observed in Figure 3.7. In 3.7a, a new word, ORANGE has been acquired, and it falls in the same phonologically delineated space as PURPLE. In 3.7b, the



(a) The learner begins learn a third word PURPLE.

(b) The learner creates a second contrast.

Figure 3.6: The number of contrasts increases to accommodate the bigger vocabulary size.



(a) The learner acquires a new word ORANGE.

(b) The learner adjusts a phonological contrast to accommodate the lexicon.

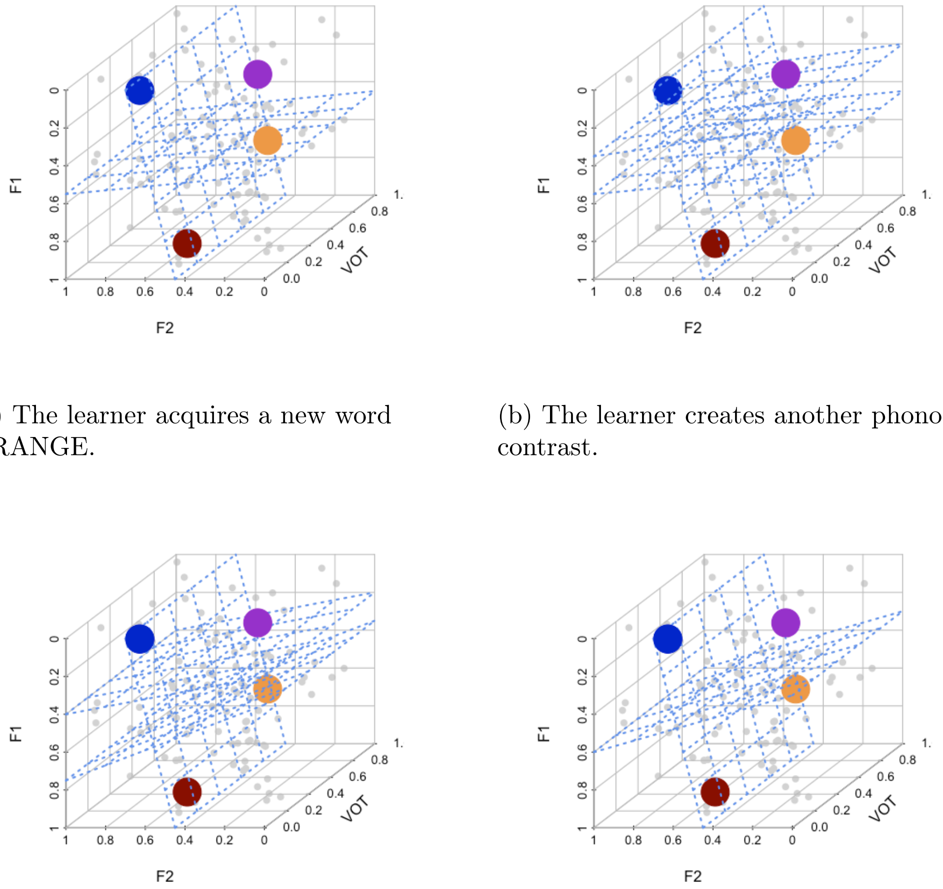
Figure 3.7: The number of contrasts increases to accommodate the increased vocabulary size.

existing horizontal CONTRAST 1 tilts upward to phonologically separate PURPLE and ORANGE in the acoustic space.

As new tokens of existing words are heard and as new words are acquired and assigned phonological representations, all contrastive planes shift to best reflect the acoustic distinctions of the words assigned to either side of each boundary. For example, in 3.7b, there is also a slight shift in the vertical CONTRAST 2. The shift is the result of adjusting to the opposition of RED+BLUE vs. **PURPLE+ORANGE**, rather than just RED+BLUE vs. **PURPLE** (cf. Figure 3.6b). The plane is updated using Equation 3.6, where a_1 =mean(RED, BLUE) and a_2 = mean(PURPLE, ORANGE).

3.2.2.3 Contrast consolidation

Because phonological contrasts are created based on prominent acoustic features of specific words, these contrasts can be word-specific initially. As more words are learned and contrasts become generalized across more lexical items, it is possible for two contrasts to become more and more phonologically similar. This scenario is depicted in Figure 3.8. Upon learning ORANGE 3.8a, rather than adjusting the boundary as in Figure 3.7, another possibility is that the learner creates an additional contrast as in Figure 3.8b. After learning more words (not represented in the plots to avoid visual clutter) and updating the boundaries, it is possible for two categories to become functionally equivalent. Illustrated in Figure 3.8c, both horizontal planes that create divisions mostly along F1 separate RED+ORANGE from BLUE+PURPLE. Because these two contrasts are functionally the same in this lexicon, they consolidate into one contrast (Figure 3.8d). In this case, consolidating the categories does not affect the system of contrast within the lexicon: BLUE remains distinct from RED, and PURPLE remains distinct from ORANGE. The developmental interpretation for this consolidation of categories is that learners tend to learn word-specific contrasts initially. The learner might acquire a contrast /b/ vs. /d/ from "ball" and "doll", then acquire a similar contrast /b'/ vs. /d'/ from "boo" and "do" because the phonetic realizations of /b/ and /d/ might be different as the result of coarticulation with the following vowel. As the learner acquires more words and adjust the phonological boundaries, word-specific phonetics will be attenuated, and /b/ vs /d/ and /b'/ vs /d'/ will become more similar and eventually consolidated as the same categories.



(a) The learner acquires a new word ORANGE.

(b) The learner creates another phonological contrast.

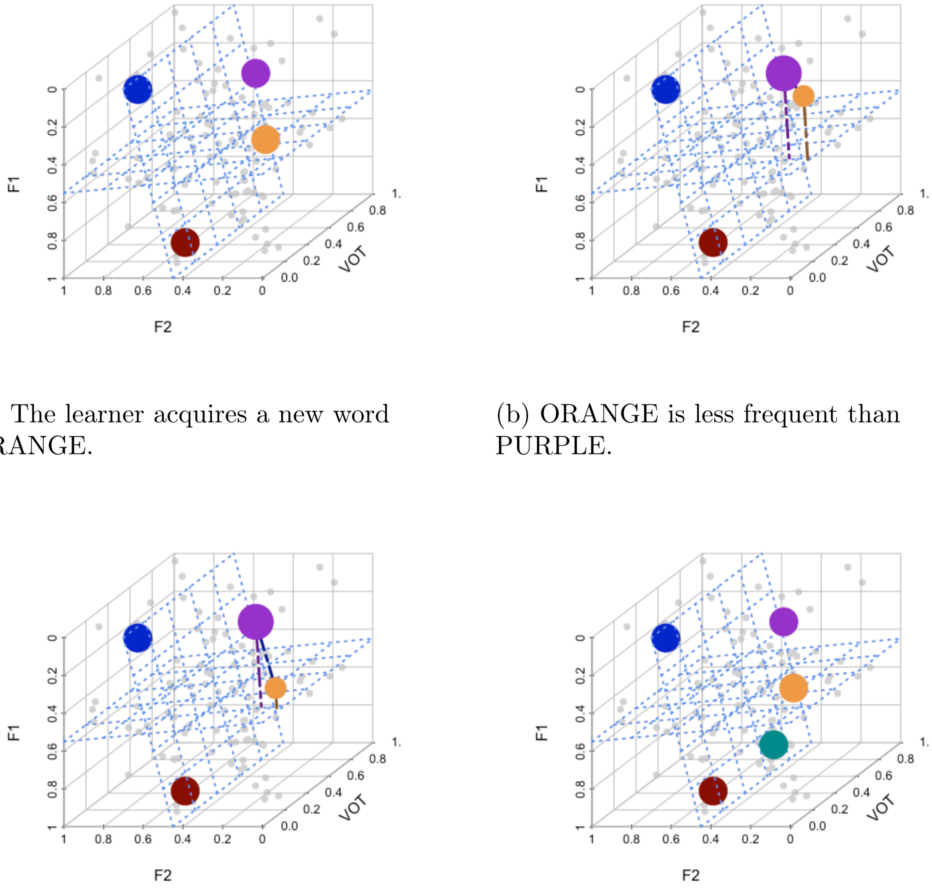
(c) The two contrasts become functionally the same.

(d) The two contrasts consolidate.

Figure 3.8: An illustration of phonological contrast consolidation.

3.2.2.4 Contrast determination

The above presents two mechanisms that two words can be represented as distinct. The model can create a new phonological contrast or adjust an existing contrast to accommodate the increasing lexical distinctions that need to be represented. However, homophones exist in language, and mergers as a sound change are very common. A model of phonological acquisition should be able to account for the existence of true homophones. How does the model choose between 1) creating a new contrast, 2) adjusting an existing contrast, and 3)



(a) The learner acquires a new word ORANGE.

(b) ORANGE is less frequent than PURPLE.

(c) ORANGE is acoustically similar to PURPLE.

(d) The learner acquires a new word ORANGE.

Figure 3.9: An illustration of phonological contrast generalization and merger.

representing two words as homophones?

How does the learner conclude which items in their lexicon are better represented homophones? The choice depends on the acoustic distance between the two words in question, the existing phonological contrasts, and the relative frequencies of the two words. The motivation for this decision comes from psycholinguistic findings about lexical access. When two words have the same phonological form, the more frequent of a homophonic pair is accessed first regardless of syntactic and semantic context (Boland and Blodgett, 2001; Caramazza et al., 2001; Bonin and Fayol, 2002). The processing cost of representing two words as homophonous can thus be quantified as the relative frequencies of the two words.

Returning to the learning stage where ORANGE has just been acquired (Figure 3.9a). There are possible scenarios it might be advantageous for the learner to represent PURPLE and ORANGE as homophones rather than representing them distinctly. For example, if ORANGE is less frequent than PURPLE and the two words are acoustically close (Figure 3.9b), a learner that assigns the same representation to PURPLE and ORANGE would still correctly identify PURPLE as the intended referent most of the time. If the intended referent is ORANGE, the learner would access PURPLE first and need additional processing to access the less frequent form ORANGE. This delay in processing can be quantified using the frequencies of the two words. With homophonous representations, the delay in processing from representing the two words as homophones can be quantified as follows:

$$C_{homophone} = \frac{\text{freq(ORANGE)}}{\text{freq(ORANGE)+freq(PURPLE)}} \tag{3.8}$$

On the other hand, if ORANGE and PURPLE are more acoustically distinct (Figure 3.9d), it might make sense to represent them distinctly even if PURPLE is far more frequent. Two factors need to be considered in making this determination. First, are PURPLE and ORANGE sufficiently acoustically distinct to warrant the creation of a new contrast? Second, is ORANGE frequent enough to warrant a distinct lexical representation? The first factor can be quantified using a measure of acoustic confusability between the two words:

$$\text{confusability} = \frac{\text{d(PURPLE, boundary)}}{\text{d(PURPLE, boundary)+d(PURPLE, ORANGE)}}$$

where:

$$d(a_1, a_2) = \sqrt{\sum_{i=1}^{m}(a_{1i} - a_{2i})^2}$$

The closer ORANGE is to PURPLE, the more acoustically similar they are. If they are too acoustically similar, creating a contrast between them will likely result in confusion in perception. This confusability measure is calculated based purely on acoustics, and it is still necessary to take into account the relative frequencies of the two words. If both words have the same frequency, they would be confused with each other by this measure. However, since the two words are not equally frequent, a weighted confusability measure can be used to quantify the processing cost of having contrastive representations:

$$C_{contrastive} = \frac{\text{freq(PURPLE)}}{\text{freq(ORANGE)+freq(PURPLE)}} \times \text{confusability}$$

If the processing cost of homophonic representation is greater than contrastive representation ($C_{\text{homophone}} > C_{\text{contrastive}}$), the learner either adjusts existing contrasts or create a new contrast to be able to represent these two words distinctly. Otherwise, homophonic representations are tolerated.

Lastly, one more scenario is illustrated in 3.9d, where shifting the existing phonological plane to distinguish PURPLE and ORANGE would make ORANGE homophonic with GREEN. Therefore, creating a new contrast would be the only option here if the learner determines that PURPLE and ORANGE need to be represented distinctly.

3.2.3 Emergent representations and properties of the model

This learning mechanism outlined in this section has several emergent properties, which are discussed below.

Phonological features. Some prominent treatment of phonology assume innate, universal phonological features (Jakobson, 1968; Chomsky and Halle, 1968; Reiss, 2018). This learning model illustrates a concrete path by which phonological features can be acquired using only acoustic and lexical cues. There is no need to assume innate phonological features. Some abstract category formation mechanism would be sufficient, either domain-general or guided by UG.

Acoustic cues for phonological features. While learning phonological categories, the model simultaneously learns the mapping between these categories and the relevant acoustic cues. By comparing the acoustics of lexical items, the model identifies which acoustic cues are meaningful to a phonological contrast and their relative contribution to the identification of the phonological contrast.

Discrete lexical representations. Discrete lexical representation are assigned to each word as soon as it is acquired. The creation of phonological boundaries enables to learner to transform the acoustics of each word into phonological distinctions.

Increasing specificity of lexical representation. The learning mechanism naturally address early underspecification that has been reported by many studies (Hallé and de Boysson-Bardies, 1996; Vihman et al., 2004; Fikkert and Levelt, 2008). The lexical representations themselves become more specified when more words are learned, and the differences between infant and child language can be largely explained in terms of the size of the vocabulary. With few words, the apparent underspecification can come from two sources. First, the learner does not need as many symbols to represent fewer words, leading to the generalization of more phonetic information over fewer symbols. Second, with a smaller vocabulary, the learner may be inaccurate in determining which specific acoustic cues matter for a phonological contrast or fail to compensate for coarticulatory effects.

Minimal pairs. Because phonological representations are built on lexical contrast, minimal pairs arise naturally as the result of the learning process.

Feature economy. Feature economy refers to the idea that languages tend to maximize the use of contrastive dimensions (Clements, 2003). Because phonological contrasts are only created as needed from lexical and acoustic cues, the resulting system is naturally economical. As more words are acquired, more dimensions of contrasts are learned, but the growth of contrasts is much slower than the growth in vocabulary.

3.2.4 Advantages of the model

A general approach to phonological acquisition. This is a general and integrated model for phonological learning and aims to learn any phonological contrasts. While many computational models focus on specific contrasts and only use cues for the contrast in question, this model makes use of the acoustic information over an entire word to learn cue weighting, and abstract lexical representations simultaneously.

Minimal theoretical assumptions. The applicability of this model is not dependent on existing phonological frameworks. The abstract representations in learned through the model can be used for further phonological analysis.

Minimal memory requirement. Because the learning is online, this model does not require calculations over a large number of input items. This model only requires the learner to remember the general acoustic shape of each word, their phonological representations, and the cue weights for each learned phonological contrast.

Non-parametric learning. This model is completely unsupervised and nonparametric. The learner does not know what contrasts exist and which cues matter for particular contrasts, both of which are discovered in the learning process. Also, the learning result is consistent and not dependent on the initialization of parameters. Third, this model can makes use of dynamic and overlapping acoustic information in word learning.

3.3 Experiment

The learning mechanism described in Section 3.2 is implemented computationally to test its validity. Acoustic measurements are extracted from the Philadelphia Neighborhood Corpus as input to the model, and the learning outcomes for phonological contrasts, acoustic cue weights, and lexical representations are presented.

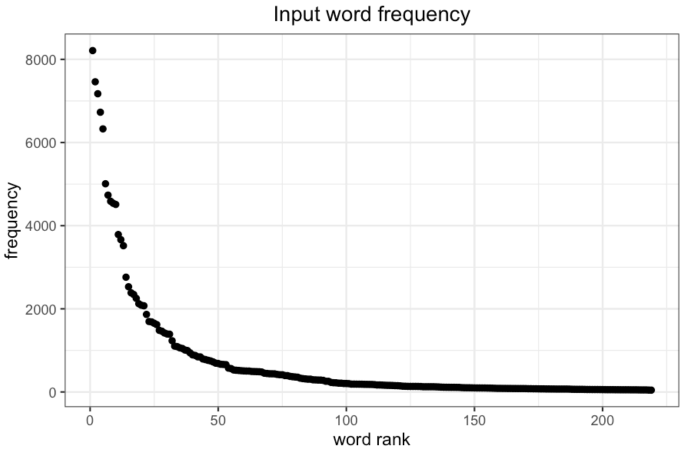


Figure 3.10: Input word frequencies.

3.3.1 Input preparation

Most of the previous work in the computational modeling of phonological/phonetic acquisition use simulated data as input (e.g., Vallabha et al., 2007; Feldman et al., 2013a). In order to better represent the noisy data that the child learner is faced with, this study uses real acoustic measurements from the Philadelphia Neighborhood Corpus (Labov and Rosenfelder, 2011). The input is limited to monosyllabic words with the syllable structures V, CV, VC, and CVC. Words containing nasal segments were excluded because of difficulty with automatically tracking measures of nasality across a large number of speakers. Words with frequencies 20 or fewer in the entire corpus are omitted.

3.3.1.1 Measurement extraction

A Praat script was written to automatically extract measurements from the corpus. For each segment, measurements were taken at 25%, 50%, and 75% of the duration of the segment. For all consonants, duration, center of gravity, jitter, shimmer, HNR (harmonics-to-noise-ratio), and autocorrelation were extracted. For sonorant consonants, f0, F1, F2, F3, B1, B2, and B3 were also extracted. Most vowel measurements, including F1, F2, F3, B1, B2, and B3 are available with the PNC. An additional measurement f0 is extracted for vowels.

3.3.1.2 Measurement normalization

Because the measurements were extracted automatically, normalization was carried out to replace potential tracking errors. The formant values were transformed onto the bark scale, and the f0 values were transformed onto semitones for each speaker. Measurements below 10% and above the 90% percentiles on the group level were changed to the group mean, and all the measurements were z-scored.

3.3.1.3 Descriptive statistics of the input

There are measurements from a total of 383 subjects from the PNC. Overall, there are 219 word types. Out of the word types, there are 162 CVC words, 30 CV words, 24 VC words, and 3 V words. There are 153,438 total word tokens, and 62909 CVC, 59934 CV, 28166 VC, and 2429 V word tokens. There are 16 onset phonemes (including null onset), 11 nucleus vowels, and 14 coda phonemes represented in the input data (including null coda). In total, 42 phonological oppositions are present among the phonemes in each position (Table 3.1).

Onset	anterior, approximant, back, consonantal, continuant, coronal, delayed release, distributed, dorsal, front, labial, labiodental, lateral, round, sonorant, strident, voice
Nucleus	back, diphthong, front, front.diphthong, high, labial, long, low, round, stress, tense
Coda	anterior, approximant, consonantal, continuant, coronal, delayed release, distributed, dorsal, labial, labiodental, lateral, sonorant, strident, voice

Table 3.1: Actual phonological contrasts in the input words for each position.

3.3.1.4 Representation of the input

Each segment of a word is represented as a 14-element vector with the measurements in the follow order: phoneme duration, f0, F1, F2, F3, B1, B2, B3, center of gravity, voicing, jitter, shimmer, autocorrelation, HNR. If a segment is null (for instance, for VC words the onset is null), a vector containing 14 0's is used. Each instance of a word is represented as a 42-element vector (14 cues × 3 segments).

3.3.1.5 Learning trials

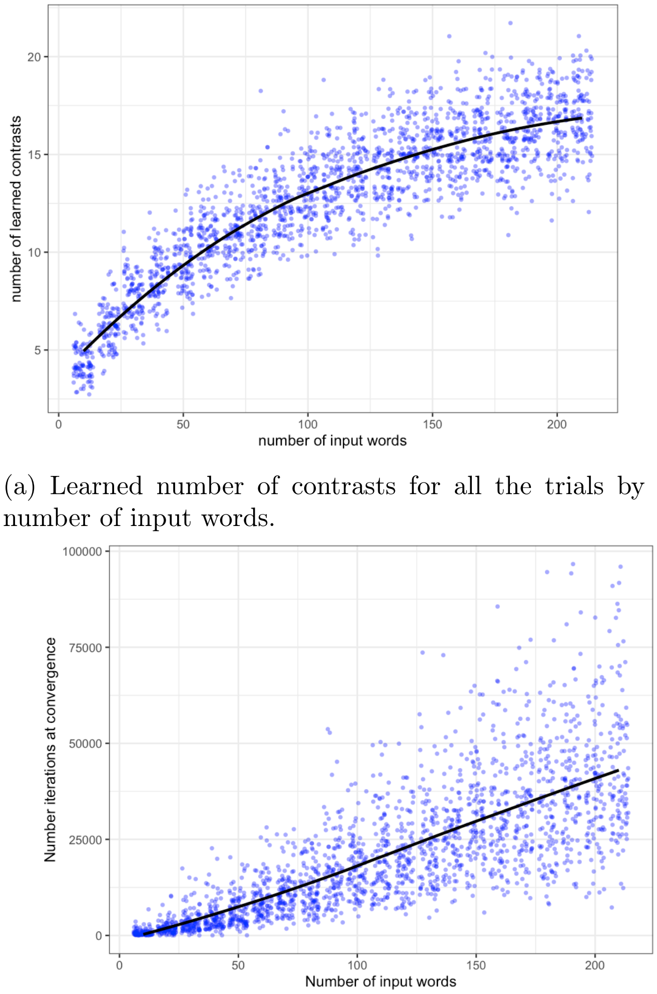
In total, there were 2100 trials with 21 input vocabulary sizes in increments of 10 (10 words, 20 words, 30 words, etc.) and 100 learning trials for each input vocabulary size. That is, for 100 trials, the model randomly picks 10 out of the 219 word types and uses the acoustic tokens of these 10 word types as input for learning. For each trial, 10 different words are randomly sampled. After 100 trials with 10 input words are terminated, 100 trials with 20 random input words are run, and so on. Learning is terminated after the number of phonological contrasts has stayed stable for 20,000 iterations. To evaluate the learning process and outcome of the learning model, the learned phonological weights, lexical representations, and word frequencies are logged every time there is a phonological change (i.e., addition or consolidation of phonological contrasts) and also every 1000 iterations.

3.3.2 Results

Overall, the model learns reasonable numbers of categories for the number of input words, and phonological contrasts converged for all trials. Case studies of specific learning trials show that the learned representations and the acoustic cues approximate phonological features commonly used for phonological analysis.

3.3.2.1 The effect of input vocabulary size

Across 2100 learning trials with varying input sizes, the model learned more contrasts for larger numbers of words. Figure 3.11a plots the number of categories the model learned for all learning trials, and a numerical summary of the results is presented in Table 3.2. The effect of vocabulary size on the number of contrasts learned is apparent. As the number of input words increased, the model learned more phonological categories to represent the words that have been acquired in the lexicon. The growth of phonological contrasts appears to flatten



(a) Learned number of contrasts for all the trials by the number of input words.

(b) Number of iterations needed for phonological convergence by the number of input words.

Figure 3.11: Learning outcome as the number of input words increases.

out with more number of words. This behavior of the model is expected, since the theoretical minimum number of binary contrasts required to represent N words is $\log_2(N)$. For 210 words, the minimum number of contrasts needed is eight ($\log_2(210) = 7.71$). The model learns on average twice the number of the theoretical minimum for 210 words. This could be partially the result the actual number of contrasts that exists in the input words. When all input words are considered, there are 42 distinct features using a feature system proposed for phonological analysis (Table 3.1). Compared to the 42 actual distinctive features that can be identified from these words, 16-17 learned features is reasonable for the given vocabulary.

# input words	contrasts	sd
10	4.09	1.13
20	5.93	1.43
30	7.54	1.40
40	8.50	1.61
50	9.47	1.54
60	10.65	1.31
70	11.03	1.34
80	11.63	1.63
90	12.39	1.48
100	13.03	1.48
110	13.46	1.59
120	14.08	1.64
130	14.25	1.73
140	14.95	1.60
150	15.41	1.46
160	15.46	1.77
170	15.99	1.51
180	15.74	1.92
190	16.64	1.48
200	16.56	1.70
210	17.06	1.77

Table 3.2: Average number of phonological contrasts learned over 100 learning trials for increasing numbers of input words.

Figure 3.11b displays the number of trials needed before the model converges on a set of phonological contrasts. As defined in Section 3.3.1.5, phonological convergence is achieved when there have been no changes to phonological contrasts for 20,000 iterations. The average number of iterations needed for convergence increases as the number of input word increases, but the variance also becomes greater as the number of input word increases. With more words, the model needs to account for a wider range of phonetic variation. Because word learning is probabilistic, in some cases, the model might acquire more generalizable contrasts earlier, resulting in the lower number of iterations needed for convergence. It is also possible that the model will need to re-tune the phonological contrasts many more times before achieving a stable state, thus resulting in a greater number of trials needed before convergence.

3.3.2.2 An example of learned representations

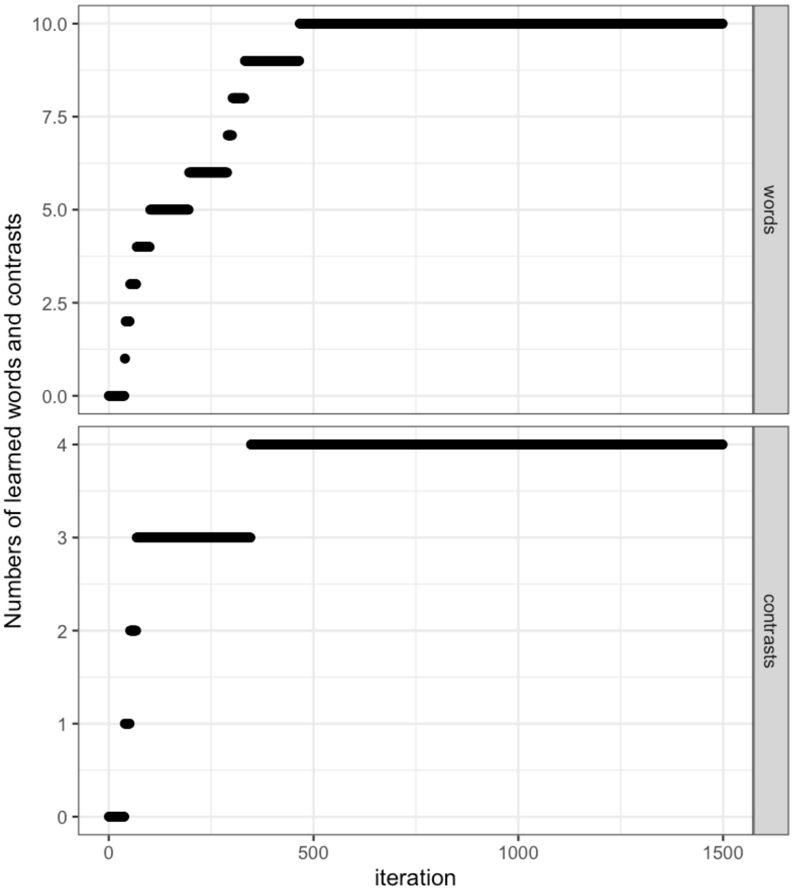


Figure 3.12: Word and contrast learning trajectories for a 10-word trial.

Each learning trial produces three results: the number of contrasts, cue weights for each contrast, and lexical representations based on these learned contrasts. This section presents a typical learned outcome from a learning trial with 10 words. The small number of words makes the results more easily interpretable. This particular instance of the learning outcome produced 4 phonological contrasts. The acquisition trajectory on the word level and the phonological level is illustrated in Figure 3.12. In this particular trial, a stable phonological state is reached on iteration 347, before all ten words have been acquired on iteration 466. The rest of this section will present the learning outcome of this trial by referencing the learned contrasts, learned representations, and comparisons to actual contrasts from a phonological analysis of these words.

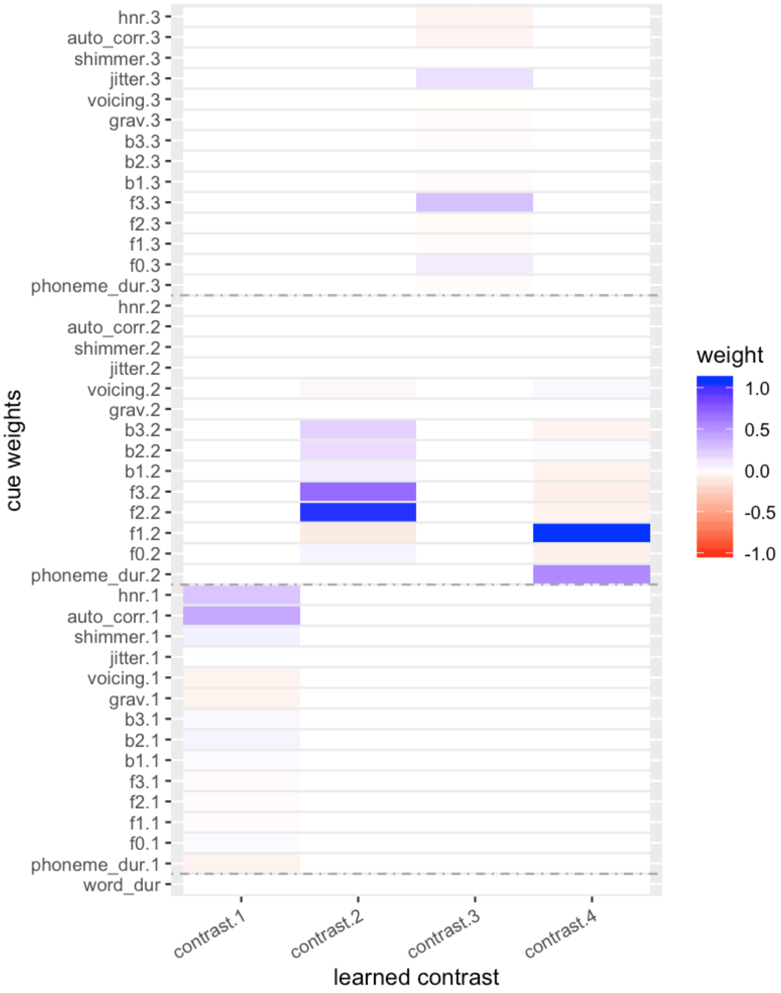


Figure 3.13: Learned weights for each of the four contrasts for a 10-word learning trial.

Figure 3.13 illustrates the learned cue weights in the form of a heatmap. Each column in the plot corresponds to one phonological contrast. Darker colors (either more blue or more red) indicate that the acoustic cue is more important for the contrast. Table 3.3 presents the learned lexical representations according to the four phonological contrasts. The use of “0” and “1” are purely symbolic and they merely indicate distinction along a phonological dimension. All the words that have the representation “0” fall on one side of the phonological division in acoustic space, while all the words with the representation “1” fall on the other. Which words are assigned “0” and which are assigned “1” is arbitrary. Figure

3.14 shows how the learned representations correspond to phonological features typically used in phonological analysis.

referent	contrast 1	contrast 2	contrast 3	contrast 4
FAR	0	0	0	1
ARE	1	0	0	1
OR	1	0	0	0
DEAL	1	1	0	0
WE'VE	1	1	1	0
FEEL	0	1	1	0
TOOK	0	0	1	0
COP	0	0	1	1
PAID	0	1	1	1
CAT	0	1	0	1

Table 3.3: Learned lexical representations with 10 words in the input.

According to Figure 3.13, the first contrast learned in this trial is an onset contrast, and the relevant cues are autocorrelation and HNR. These acoustic cues are typically associated with the manner or voicing of consonants. In the learned representations (Table 3.3), words with voiceless onsets (FAR, FEEL, TOOK, COP, PAID, CAT) are separated from words with voiced onsets or no onsets (ARE, OR, DEAL, WE'VE). Indeed, when comparing the learned contrast with phonologically analyzed contrasts, this dimension correlates highly with voicing and manner features (Figure 3.14). Moving on to the second learned contrast, the heavily weighted phonetic cues are in the nucleus, and the most important cue is F2 (Figure 3.13), which usually indicates differences in the frontness or backness of the vowel. Indeed, in the learned representations, this contrast marks the distinction between back vowels (FAR, ARE, OR, TOOK, COP), and front vowels (DEAL, WE'VE, FEEL, CAT), and the learned representations correspond to [front] and [back] features in traditional phonological analysis (Figure 3.14).

As for contrast 3, the phonetic weighting indicates that this is a coda contrast based on F3 differences (Figure 3.13). Unlike contrast 1 and contrast 2, this contrast does not correspond neatly to any phonologically analyzed contrasts. For the most part, words with sonorant codas (FAR, ARE, OR, DEAL) are separated from stop and fricative co-

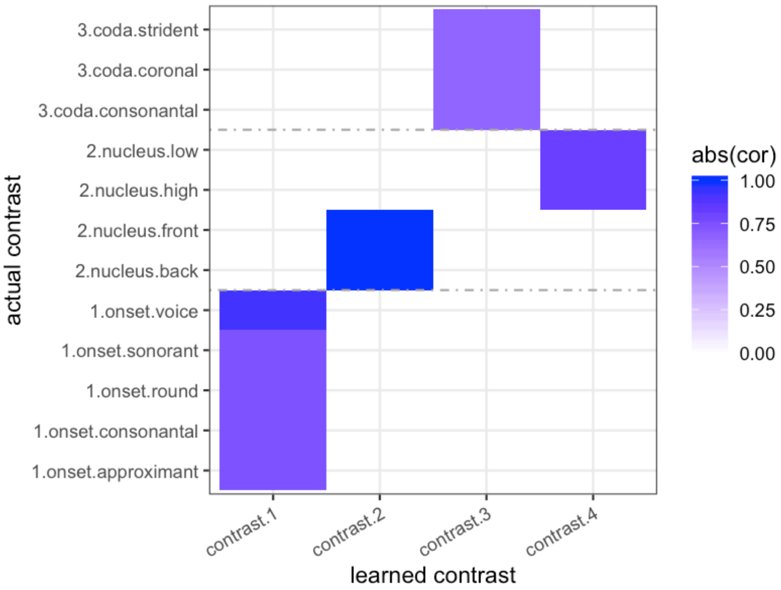


Figure 3.14: Correlation of learned representations to to actual phonological features for a 10-word trial.

das (WE'VE, TOOK, COP, PAID). However, CAT and FEEL do not fit this pattern. The learned representation for CAT is /0 1 0 1/. If CAT is represented as /0 1 1 1/, it would be homophonous with PAID. Perhaps this is the reason that the model adjusts this contrast to accommodate CAT vs. PAID in the existing phonological space rather than creating a new contrast. The assignment of FEEL to /1/ for contrast 3 is anomalous and may be the result of the specific acoustic measurements of FEEL. A more general voicing contrast may be acquired with more words. Finally, the last learned contrast distinguishes vowel height, with F1 as the most prominent acoustic feature. Table 3.3 shows that words with low vowels (FAR, ARE, COP, PAID, CAT) are separated from words with high vowels (OR, DEAL, WE'VE, FEEL, TOOK), and this is confirmed by the high correlations to manner features in Figure 3.14. There are minimal pairs in the learned phonological representations, but these minimal pairs are defined within the phonological contrasts learned from these 10 input words. FAR /0 0 0 1/ and ARE /1 0 0 1/ differ by the onset Contrast 1. This corresponds to the actual phonological contrast that FAR has an onset /f/ and ARE has null onset. The rest of these two words have the same representations. Similarly, ARE /1 0 0 1/ and OR /1 0 0 0/ form a minimal pair and differ only in Contrast 4, a vowel height contrast.

However, within this phonology, TOOK /0 0 1 0/ and COP /0 0 1 1/ are also a minimal pair even though in actual English phonology they differ by all three segments. These two words are fairly acoustically similar: They both have a voiceless stop in the onset and a voiceless stop in the coda. With a small vocabulary of 10 words, representing TOOK and COP as a minimal pair is entirely reasonable. The difference in the vowel – the distinctive part between these two words – is enough for the learner to identify the contrast between these two words within this small lexicon. The learner is being efficient (or economical) in this kind of use of their phonological space. With a larger vocabulary, the learner will need to create more fine-grained contrasts between the different stops, but this is not necessary given the acoustics and the lexical contrasts in the input of this trial.

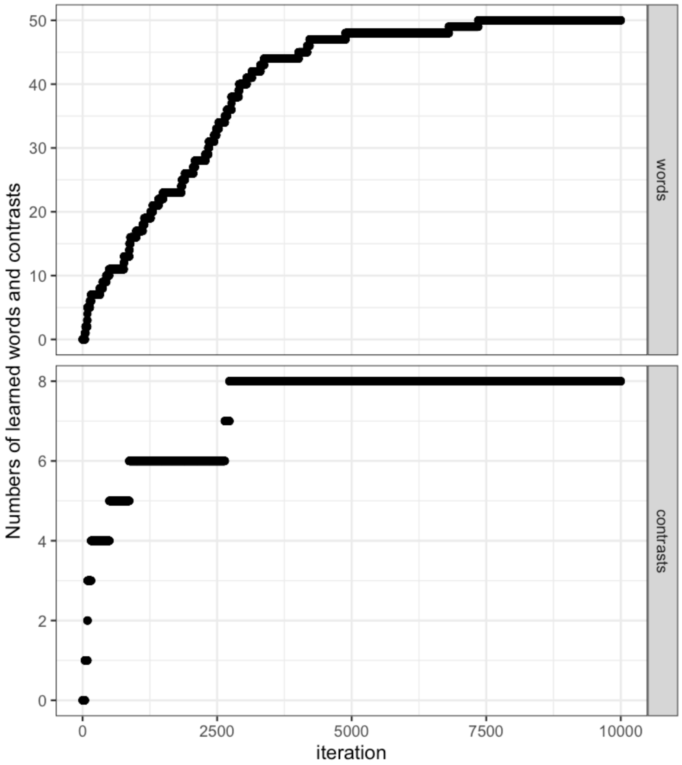


Figure 3.15: Word and contrast learning trajectories for the 50 word trial.

3.3.2.3 Learning outcome with 50 words

The model is successful at discovering meaningful contrasts for the 10-word trial presented above. Does this result generalize to the learning of more words? In this section, the results from a 50-word learning trial are presented. The learning trajectory for words and phonological contrast is shown in Figure 3.15. For this case, the number of phonological contrast stabilizes at iteration 2728, when 36 words have been acquired. These learned representations are sufficient to accommodate the words that have not yet been learned. All 50 words are acquired at iteration 7352.

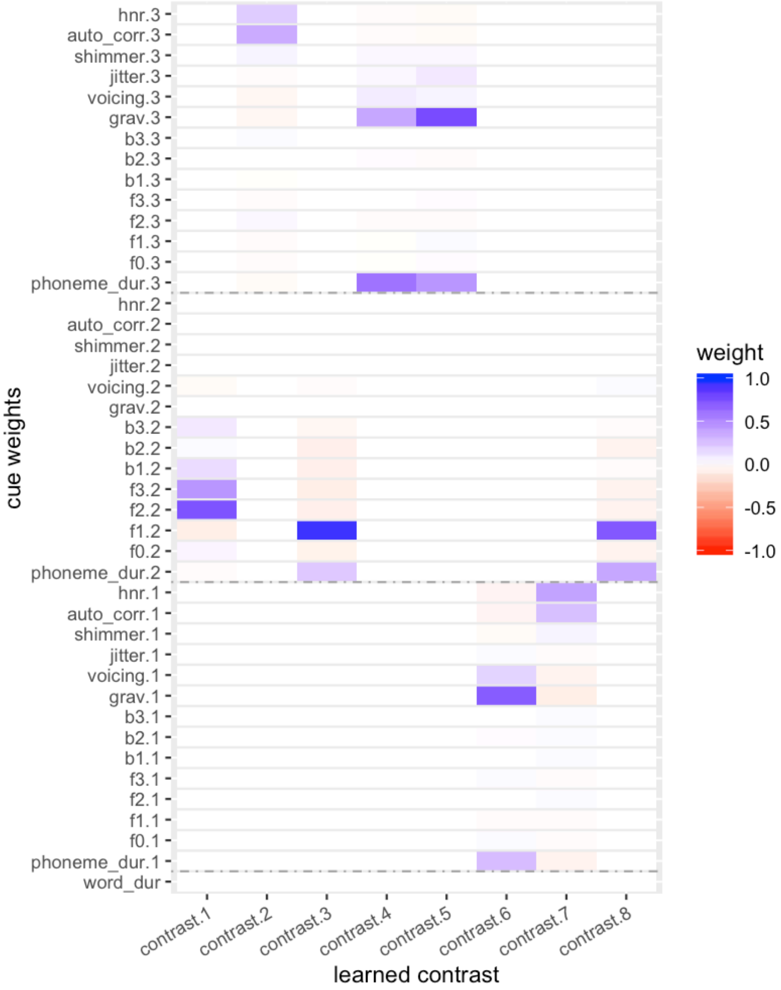


Figure 3.16: Learned contrasts for 50 words.

Figure 3.16 shows the learned cue weights. A total of 8 contrasts were learned, among which there are two onset contrasts (#6 and #7), three vowel contrasts (#1, #3, #8), and three coda contrasts (#2, #4, #5). Since listing all the learned representations is not as easily interpretable as the 10-word trial, I will discuss the learned representation results on the segmental level.

	Contrast 6		Contrast 7	
	Fricatives + t		Voice	
	0	1	0	1
∅	100.00	0.00	0.00	100.00
b	100.00	0.00	0.00	100.00
p	100.00	0.00	100.00	0.00
d	100.00	0.00	50.00	50.00
t	0.00	100.00	100.00	0.00
g	100.00	0.00	25.00	75.00
k	50.00	50.00	100.00	0.00
f	33.33	66.67	100.00	0.00
s	0.00	100.00	100.00	0.00
ʃ	33.33	66.67	100.00	0.00
l	100.00	0.00	50.00	50.00
ɹ	100.00	0.00	0.00	100.00
w	100.00	0.00	0.00	100.00
j	100.00	0.00	0.00	100.00

Table 3.4: Percentages of each onset phoneme assigned to each side of a learned phonological contrast.

The learned onset distinctions are compared to actual phonemic representations in Table 3.4. For each learned phonological contrast, this table presents the percentages of the learned representations for each phoneme. For instance, /p/ is assigned /0/ for the learned Contrast 6 in 100% of the word types it occurs in, while /t/ is assigned /1/ for Contrast 6 in 100% of the word types it occurs in. According to Figure 3.16, Contrast 6 separates onset phonemes by the acoustic cue center of gravity. Comparing this to the assigned representations, it appears that Contrast 6 separates fricatives from the rest of the phonemes. The phoneme /t/ is grouped with the fricatives, possibly because its coronal place of articulation results in similar average frequencies as /s/ and /ʃ/. Contrast 7 is more straightforward; it creates a boundary between voiced and voiceless onsets by dividing the acoustic space mostly along

HNR and autocorrelation, both are measures of periodicity in the signal.

	Contrast 1		Contrast 3		Contrast 8	
	Front/Back		High/Low		High/Low	
	0	1	0	1	0	1
i	0.00	100.00	100.00	0.00	100.00	0.00
ɪ	100.00	0.00	100.00	0.00	100.00	0.00
e	0.00	100.00	0.00	100.00	40.00	60.00
ɛ	16.67	83.33	0.00	100.00	83.33	16.67
æ	66.67	33.33	0.00	100.00	0.00	100.00
ɑ	100.00	0.00	0.00	100.00	20.00	80.00
ʌ	100.00	0.00	0.00	100.00	50.00	50.00
ɔ	100.00	0.00	0.00	100.00	0.00	100.00
o	100.00	0.00	0.00	100.00	100.00	0.00
ʊ	100.00	0.00	100.00	0.00	100.00	0.00
u	85.71	14.29	100.00	0.00	100.00	0.00

Table 3.5: Percentages of each vowel phoneme assigned to each side of a learned phonological contrast.

The learned representations for each vowel is presented in Table 3.5. Contrast 1 separates the phonemes mostly along F2, which corresponds to the frontness or backness of the vowel. The acoustic boundary separates /i e ɛ/ from the rest of the vowels. This learned boundary appears to be very "front": The vowel /ɪ/ and /æ/ are typically described as front in phonological analysis, but they are grouped with the back vowels in this learned contrast. Contrast 3 clearly distinguishes high vowels from non-high vowels. Contrast 8 is a second high-low contrast, but the boundary appears to be "lower" than Contrast 3. Contrast 8 separates the mid vowels /o/ and /ɛ/ from the low vowels, but /e/ is ambiguously represented by this contrast.

There are three contrasts learned for the coda (Table 3.6). Contrast 2 corresponds to voicing and separates the voiceless codas /p t k s ʃ θ/ from the voiced ones /d v z ɹ/. Both Contrast 4 and Contrast 5 weigh mostly heavily the cues center of gravity and phoneme duration. Contrast 4 distinguishes fricatives and the phoneme /k/ from non-fricatives, but it is ambiguous for the phonemes /g/ and /l/. Contrast 5 groups /g/ and /l/ with non-fricatives. All the fricatives have the same representation except for /v/, but this might be because the only word type with /v/ in the coda is "of."

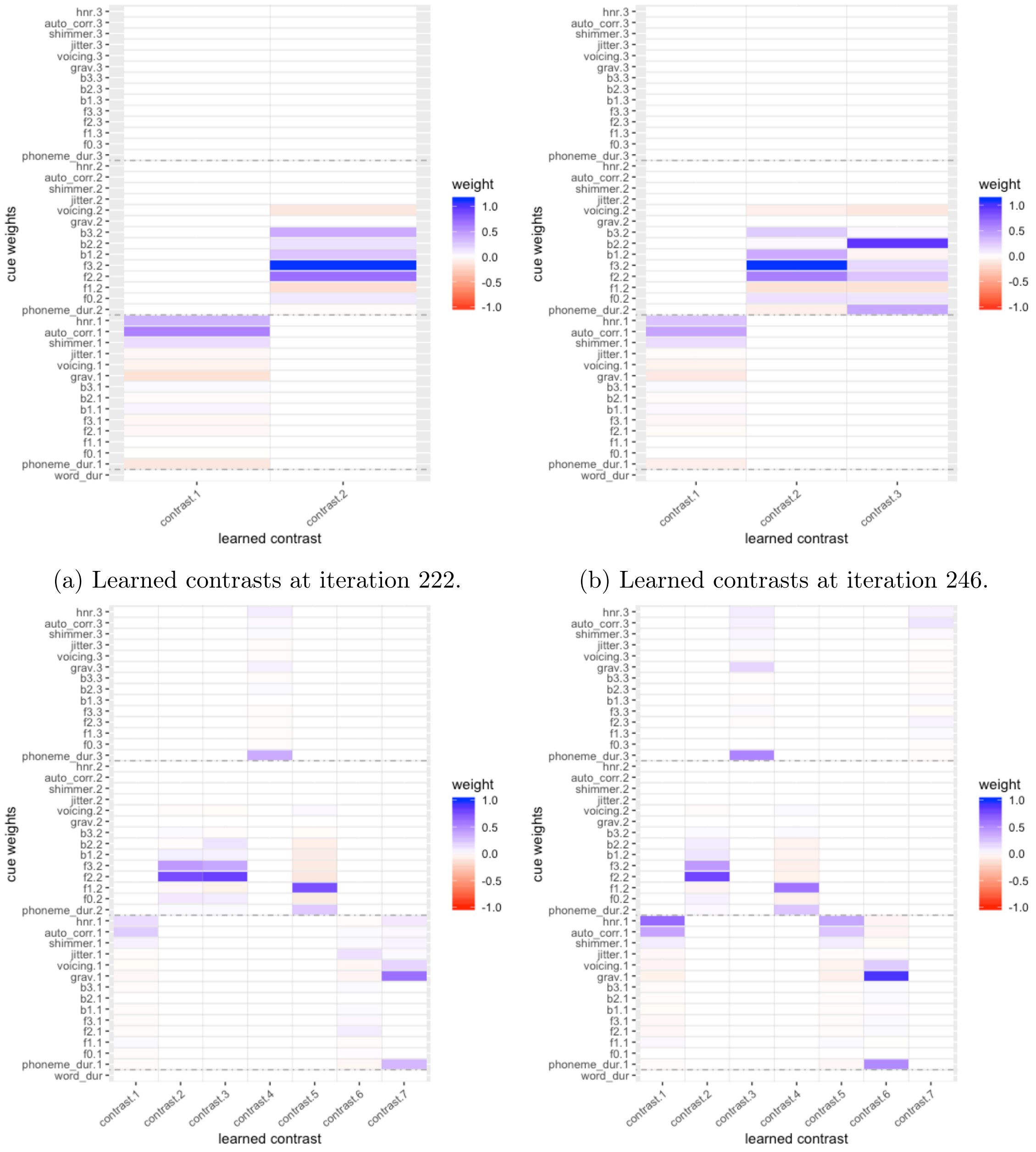
	Contrast 2		Contrast 4		Contrast 5	
	Voicing		Fricatives + k		Fricatives - v	
	0	1	0	1	0	1
∅	14.29	85.71	71.43	28.57	100.00	0.00
p	100.00	0.00	83.33	16.67	83.33	16.67
d	28.57	71.43	100.00	0.00	85.71	14.29
t	80.00	20.00	80.00	20.00	100.00	0.00
ɡ	50.00	50.00	50.00	50.00	100.00	0.00
k	100.00	0.00	0.00	100.00	100.00	0.00
f	100.00	0.00	0.00	100.00	0.00	100.00
v	0.00	100.00	100.00	0.00	100.00	0.00
s	100.00	0.00	0.00	100.00	0.00	100.00
ʃ	100.00	0.00	0.00	100.00	0.00	100.00
θ	100.00	0.00	0.00	100.00	33.33	66.67
z	20.00	80.00	0.00	100.00	0.00	100.00
l	50.00	50.00	50.00	50.00	100.00	0.00
ɹ	0.00	100.00	66.67	33.33	100.00	0.00

Table 3.6: Percentages of each consonant phoneme assigned to each side of a learned phonological contrast.

3.3.2.4 An example of category consolidation

The learning mechanism outlined in Section 3.2.2.3 describes a scenario where two contrasts can become similar and consolidate without any changes to the system of lexical contrast. This section shows a specific example of how this process is played out during the course of learning by the model.

Figure 3.17 and Table 3.17 show the four snapshots of the learning process in a particular trial. On iteration 222 (3.17a), the model learns a vowel contrast (Contrast 2) from WE and BUT since Contrast 1 does not distinguish between them. On iteration 246, the model learns another vowel contrast (Contrast 3) from DO and BUT. By iteration 974, Contrast 2 and Contrast 3 have become fairly similar. On iteration 1400, CONTRAST 2 and Contrast 3 are consolidated into one category. When the contrasts are initially learned, the acoustics cues that were weighted the heaviest do not make much sense based on what we know about English phonetics. F3 for Contrast 2 and B2 (bandwidth of F2) for Contrast 3 are not the most important acoustic cues when it comes to vowel distinctions (Figure 3.17b). As more words are learned and classified, both contrasts update with the phonetics of newly



(a) Learned contrasts at iteration 222.

(b) Learned contrasts at iteration 246.

(c) Learned contrasts at iteration 974.

(d) Learned contrasts at iteration 1400.

Figure 3.17: An illustration of contrast generalization.

acquired words and begin to make more sense phonetically. F2, which corresponds to the frontness/backness of the vowel, becomes more heavily weighted for both contrasts and eventually the two become functionally the same and are consolidated.

	contrast							
referent	1	2	3	4	5	6	7	8
iteration 222								
WE	1	1						
BUT	1	0						
iteration 246								
WE	1	1	1					
BUT	1	0	0					
DO	1	0	1					
iteration 974								
WE	1	1	1	0	0	0	0	
BUT	0	0	0	0	1	0	0	
DO	1	0	1	0	0	1	0	
iteration 1400								
WE	1	×	1	0	0	1	0	0
BUT	0	×	0	0	1	0	0	0
DO	0	×	1	0	0	1	0	0

Table 3.7: Evolution of learned lexical representations.

3.4 Discussion

The model presented in this chapter makes several important contributions to the understanding of first language acquisition and phonological representation. The model succeeds in learning phonological contrasts appropriate for a given lexicon by identifying meaningful boundaries in the multidimensional acoustic space. These results demonstrate the efficacy of a nonparametric and unsupervised approach to modeling phonological acquisition and that phonological features are an emergent property from structuring the acoustic space to accommodate lexical contrast.

3.4.1 Computational approach

The model advances the computational study of phonological acquisition in a number of ways. First, this model is a general model of phonological acquisition. Many previous computational models of speech category acquisition focus on specific contrasts and use cues relevant to those contrasts as input for learning, such as vowels (Vallabha et al., 2007; Feldman et al., 2013a; Dillon et al., 2013) and voicing (Toscano and McMurray, 2010). The model presented in this chapter is not limited to specific contrasts but aims to learn any contrast in a given lexicon.

Second, it is common practice to use artificially generated data as input (e.g., Vallabha et al., 2007; Toscano and McMurray, 2010; Feldman et al., 2013a). This model achieved reasonable results using natural acoustic measurements taken from a speech corpus. Moreover, the input consists of acoustic measurements from entire words. The acoustic representations used in this study better approximate the multidimensional and continuous nature of the speech signal a learner receives. Although the approach used in this model is not a perfect representation of continuous speech signal, it nevertheless is an important step forward towards more realistic input representation in acquisition modeling.

Third, the model is set up to more closely simulate the actual learning process of a child. This model also has the advantage of being nonparametric. In contrast, models that rely on statistical learning, such as Bayesian models (e.g., Feldman et al., 2013a), need parameter tuning to achieve the best results. Additionally, the learning is completely online. The learner hears the input one at a time and updates their phonological knowledge as needed at each iteration of learning, just as a child might as they are exposed to more and more linguistic input. In contrast, many existing models rely on batch learning. While these algorithms can be adapted to be online (e.g., Vallabha et al., 2007), their implementation are often parametric. Moreover, the learning in this model is unsupervised. It does not learn from target representations, but rather discovers both contrastive dimensions and appropriate phonological representations through learning. Acoustics and lexical contrast are sufficient for the learner to form appropriate abstract representations. All of these properties closely approximate the actual challenge faced by the learner.

Finally, the learning outcome from the experiment validates the learning mechanism described in the model. The model learns the appropriate numbers of phonological contrasts given the size of the input lexicon, and it also learns the appropriate phonetics for each phonological contrast. Because phonological contrasts and lexical representations interact and update dynamically, this model can offer some explanations for the developmental trajectory of phonology. At the beginning of learning, the model had limited numbers of contrastive dimensions because only a few words need to be assigned abstract representations. However, with more input and sufficient word frequency, the model learns more distinct representations for different lexical items. This can in part explain why early lexical representation appears to be underspecified. With a small vocabulary, the learner does not need phonologically detailed representations because there are fewer word distinctions that need to be represented. The success of the model so far indicates representational pressures indeed play a role in phonological acquisition.

3.4.2 Theoretical implications

Phonological features are a useful tool of phonological analysis, but as reviewed in Section 2.2.2, assuming a universal set of innate features has a number of issues. The model presented in this chapter operationalizes the acquisition of emergent phonological features, and the experiment results indicate that the learning mechanism proposed in this chapter is computationally viable. One important theoretical advance from this model is that it outlines a concrete path from multidimensional acoustic input to abstract representation. Although many conceptual models of phonological acquisition incorporate lexical learning (e.g., Jusczyk, 1997; Dresher, 2004; Werker and Curtin, 2005), most of these models have not been implemented computationally and tested.

The learning is both phonetically and linguistically motivated, and the acoustic input and learned contrasts reflect the multidimensional nature of phonetic cues in production and perception. The hypothetical binary contrastive dimensions can offer insights into why phonological systems tend to be symmetrical. For example, if a contrastive dimension is created to distinguish vowel height for /i/ and /æ/, it is easy to extend the same contrast to vowels like /u/ and /ɑ/ since there are shared acoustic cues. Lastly, this model can capture the role of language experience. Depending on the input, the specific order of acquisition of contrasts can differ, but the end result will converge to distinct phonological representation of all the lexical items when the critical number of lexical items has been acquired.

3.4.3 Future directions

There are a few aspects of the work that needs further development. First, it would be ideal if the model learns contrastive dimensions and cue weights that more consistently align with results from linguistic analysis. Although the results presented above are fairly close to linguist contrasts, the learning results vary from trial to trial. Part of this variation is expected, since there is a random element in word acquisition. However, the learning results might be more consistent with additional acoustic measurements. Second, at the maximum, only 210 lexical types were used as input to the model. It would be interesting to see how further input would alter the learning outcome of the model. Third, as this model is intended to be a general model of acquisition, the learning mechanism in the model should validated with results from additional languages. Lastly, this model only learns position-specific contrasts. Generalization across different positions is an important part of phonological learning and should be incorporated into a model phonological acquisition.

3.5 Conclusion

The learning model presented in this chapter makes several important contributions. First, it demonstrates that innate features are not necessary for the acquisition of discrete phonological representation. Second, it contributes to the research on emergent phonological features by proposing a clear mechanism whereby phonological contrasts can be learned from the input in a nonparametric and unsupervised fashion. Third, the model provides explanations for the trajectory of phonological acquisition observed in developmental studies. Overall, the results in the chapter suggest that phonological representations can emerge from the interaction of acoustics and lexical contrast without innate features or statistical learning.