Modeling Phonetic Category Learning from Natural Acoustic Data

Stephanie Antetomaso, Kouki Miyazawa, Naomi Feldman, Micha Elsner, Kasia Hitczenko, and Reiko Mazuka

1 Introduction

Languages make different distinctions regarding what constitutes a meaningful difference in sound, yet children quickly learn to distinguish between these sounds in their native language. A major question in the field of language acquisition regards what strategies children use to acquire knowledge of their language from linguistic input. Empirical research shows that children are able to track some statistical properties of their input (Saffran et al., 1996; Maye et al., 2002). One way to convert these statistical observations into linguistic knowledge is by tracking distributional regularities of the input, such as regularities in the frequency or duration of sounds in the child's environment, a process known as distributional learning. This distribution of sounds is hypothesized to be important for phonetic category learning (Maye et al., 2008).

Computational models help us explore how well linguistic structures can be learned from input given a set of statistical tools. Perhaps we wish to test how completely the phonetic categories of some language can be learned given information about the distribution of sounds in natural speech. Provided with this input, a model could be programmed with a learning algorithm taking advantage of these data, and its output (the categories recovered by the model) could be compared to the categories existing in the language.

In practice, however, the input for such models is often simplified compared to the input available to children: acoustic vowel information is sampled from unrealistically parametric distributions whose parameters are taken from carefully enunciated words produced in laboratory settings, and lexical information providing context for these sounds is transcribed phonemically, rather than phonetically. These simplifications occur both due to technical limitations of the models and

* Corresponding author: Stephanie Antetomaso, The Ohio State University, antetomaso.2@osu.edu. This research was supported by NSF grants IIS-1422987, IIS-1421695, and DGE-1343012. We thank Laura Wagner and the OSU Language Acquisition discussion group for helpful comments and discussion.

because realistic annotated data are expensive. Because the purpose of modeling language acquisition is to test how linguistic structures can be learned from input, it is important to ensure that these simplifications are not impacting model results in unexpected ways.

In this paper, we explore the effect of input simplification on models of phonetic category learning, specifically on the learning of vowel categories. We run a lexical-distributional phonetic category learning model on more realistic input data, containing vowel variability caused by reduction or changes in duration. We examine two different input languages, English and Japanese, to see whether model performance is affected similarly by input simplifications in different languages. We show that as variation in the input increases, model performance – unlike the performance of a child – decreases. We also show that although model performance decreases on both English and Japanese, it does so in different ways. By characterizing the ways in which realistic input changes the phonetic category learning problem, our goal is to aid in extending computational models of phonetic category acquisition to use natural acoustic data rather than artificially constructed regular data, while still maintaining linguistically relevant output.

2 Phonetic category acquisition

2.1 Phonetic variability

Infants are initially able to discriminate among a universal set of acoustically perceptible sounds, even if a particular contrast is not functional in their native language. However, as children become more experienced with their language, their discrimination of non-native phonetic contrasts declines. This occurs in their first year, specifically around 10-12 months for consonants (Werker and Tees, 1984) and beginning around 6 months for vowels (Polka and Werker, 1994). During this learning process, children must deal with a large amount of variation. Two speakers saying the same word in the same language may have differences in pronunciation based on dialect, age, or gender. Even the sounds produced by a single individual are affected by speaking rate, pitch, co-articulation with nearby segments, and prosodic effects such as vowel lengthening and reduction. Empirical results have shown that children are able to generalize at a relatively early age across affect (Singh et al., 2004), amplitude and pitch (Singh et al., 2008; Quam and Swingley, 2010), accents with vowel shift (White and Aslin, 2011), gender (Houston and Jusczyk, 2000), and speaking rate (Weismer and Hesketh, 1996). Given the apparent 'inconsistency' of the child's input, learning the phonetic categories of their language quickly seems like a difficult prospect. However, there is mounting evidence that variability actually facilitates language learning. This has been shown in grammatical domains (Gómez, 2002), word recognition (Rost and McMurray, 2009; Singh, 2008; Quam et al., 2017), and phonetic categorization

(Scarborough, 2003). This variability causes learners to shift their focus to sources of more stable structure, better defining their phonetic categories.

Models of phonetic category learning have only rarely been tested on input data with the high level of variability characteristic of children's input, and our goal is to characterize how increases in variability impact performance. Empirical evidence shows that the types of variability encountered by children, for example speaking rate (Fourakis, 1991; Moon and Lindblom, 1994; Lindblom, 1983; Stack et al., 2006) and gender and dialect (Byrd, 1994), affect vowel acoustics. This work will focus primarily on two types of acoustic variability due to prosody. The first, vowel reduction, is a change in the acoustic quality of the vowel – generally a weakening and centralization within vowel space – caused by the vowel being unstressed, short in duration, quickly articulated, etc. The second, phrase-final lengthening, causes lengthening of vowels around the boundary of a phrase. English and Japanese, the languages used in our simulations, are interesting with respect to these prosodic effects. English has extensive vowel reduction in conversational speech (Johnson, 2004) but, although phrase-final lengthening occurs, we might not expect it to strongly affect category learning of English vowels as vowel duration is not phonemic in English. In contrast, Japanese does have phonemic vowel duration, indicating that phrase-final lengthening could affect categorization of long versus short vowels; however, it is a mora- rather than stress-timed language and therefore less likely to have substantial vowel reduction (Grabe and Low, 2002).

2.2 Models of phonetic learning

A leading hypothesis in phonetic category acquisition is *distributional learning*, the idea that learners attend to distributions of sounds in their input as a cue to category membership (Maye et al., 2002). Several computational models have implemented distributional learning algorithms and shown their effectiveness at learning phonetic categories from corpus data. An early model gives evidence of the learnability of distributional information by using Expectation Maximization of a mixture of Gaussians to show that cardinal vowels /i/, /a/, and /u/ can be learned when the system is given a target number of vowel categories (De Boer and Kuhl, 2003). Later models extended this to discovering categories without knowing the number of vowels in advance and without having access to all the data at once, simultaneously estimating the number and parameters of vowel categories. These models extend the mixture of Gaussians algorithm to include competition, allowing the model to account for developmental changes in over-generalization (Vallabha et al., 2007; McMurray et al., 2009). Vallabha et al. (2007) also extend their model to look at both English and Japanese, as we do here. Bion et al. (2013) show that when using natural adult- and child-directed Japanese speech, distributions of short and long vowel categories are overlapping and unlikely to be captured well by a

purely distributional model.

The data used in these studies are often from laboratory elicitation settings where participants are recorded producing stressed, clearly enunciated words that are often single-syllable and/or consist of a particular consonant frame. Phonetic categories are assumed to approximate Gaussian (normal) distributions and the data from these laboratory productions are used to parametrize these distributions, giving each vowel a mean and variance characteristic of clear laboratory speech. To produce sufficient data for the model to analyze, additional data points are sampled from these distributions. This results in input which is artificially Gaussian in nature (a criticism also made by Dillon et al. (2013), who use a non-parametric resampling method instead) and, due to how the original laboratory vowels were produced, does not take vowel reduction or phrasal lengthening effects into account. In cases where sounds are elicited in a consistent consonant frame, it also produces data with no variation due to co-articulation.

Purely distributional models of category learning are only concerned with the categorization of individual sounds; they do not assume any more complex hierarchical structure of language. However, such models struggle to identify categories when faced with overlapping distributions of sound. Feldman et al. (2013) showed that results on overlapping data can be improved by taking into account a structure wherein the child's word learning influences their learning of distinct sounds and vice versa. For models using such lexical information, this generally comes from a phonemically transcribed context, in which each token of a word is transcribed the same way. However, for naturally produced acoustic data we would expect to see vowel sounds changing based on their position in a phrase, how carefully they were enunciated, and what segments they were produced near (Harmegnies and Poch-Olivé, 1992; Picheny et al., 1986; Ferguson, 2004). Thus, although these models show great success in categorizing phonetic input with low variability, it is unclear whether their results would be similarly successful on input including the effects of variability present in conversational speech.

Speech recognizers using Hidden Markov Model representations of speech tend to use more realistic data but to have fewer linguistic constraints on the type and number of categories that are produced (Lee and Glass, 2012; Varadarajan et al., 2008; Jansen and Church, 2011). These models tend to over-produce categories due to co-articulation, as they have no way to explain the sound change in the vowel which occurs as a consequence of its phonetic context. In this work, we are specifically interested in how children arrive at a more linguistically relevant output. We would like a model to produce generalizable categories, rather than assuming a new vowel category for every possible co-articulated pair.

For our simulations, we adopt the model from Feldman et al. (2013). This model has two main advantages. First, it performs well on data with low variability: for overlapping distributions of vowel formants, Feldman et al. (2013)

Table 1: The three types of input corpora differed in their degree of simplification, yielding different patterns of variability at the lexical and acoustic levels.

	Transcription	Vowel Acoustics
Simulation 1	phonemic	re-sampled from lab productions
Simulation 2	phonetic	re-sampled from lab productions
Simulation 3	phonetic	measured from the corpus

showed that including lexical information to disambiguate between overlapping sounds improves performance over a purely distributional model, raising their phonetic categorization results from an F-Score of 0.45 (out of 1.0) for a purely distributional model, up to a F-Score of 0.76 for a model including both lexical and distributional information. Second, the inclusion of additional structure for the input is developmentally motivated, as children begin segmenting speech and learning forms and meaning for simple words around the same time that they are acquiring phonetic categories (Jusczyk and Aslin, 1995; Swingley, 2009). These attributes make it a good candidate for testing the effects of realistic input and for testing model results on multiple languages.

3 Simulations

To test how variation in language affects model performance, we ran a series of simulations with increasing levels of lexical and acoustic variability (Table 1). The first simulation replicates previous work by using input with simplification of both lexical and acoustic information, testing whether there are language specific effects of this simplification by extending a model previously only run on English to Japanese. The second simulation replaces the phonemic lexical transcription with a phonetic one, increasing lexical variability while maintaining artificial acoustics. In this simulation, words may be pronounced differently at different times but the vowel acoustics themselves will always be normally distributed. Finally the last simulation is most faithful to input found in the real world; we continue to use phonetic transcriptions of lexical items while also using acoustic information directly from the audio recordings.

3.1 Model

The model from Feldman et al. (2013) takes as input a set of acoustic values and lexical contexts. For our simulations, these acoustic values consist of first and second formant pairs for English vowels and formant pairs plus duration for Japanese vowels. Our lexical contexts consist of categorical consonant frames from either phonetic or phonemic transcriptions of conversational English and

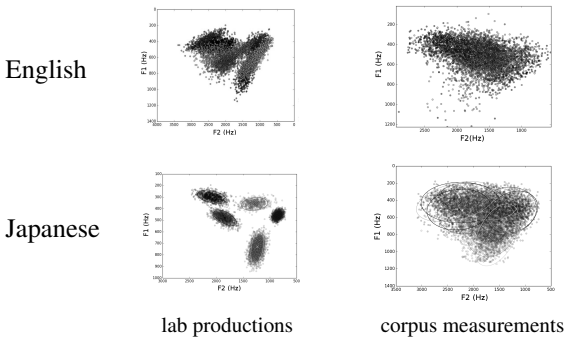

Figure 1: Variability in formant values that are resampled from lab productions (left) and measured from corpus recordings (right).

Japanese speech. Parameter values for the model are the same as in Feldman et al. (2013), with lexical and phonetic concentration parameters set to 10 (see their paper also for details of the mathematical model). The model produces vowel categories, defined as Gaussian distributions of sounds, as well as lexical categories which consist of sequences of phones with vowel values selected from the model-determined vowel categories. The model itself does not give labels to the categories it produces, but for example, one lexical category might consist of all word tokens of the form /k Cluster1 n/, where Cluster 1 consists of /æ/ vowels; this lexical category would represent what we would think of as the word *can*. Although the model is allowed to propose a potentially infinite number of categories, making it language independent, it has a bias toward fewer categories, resulting in a more realistic output. The lexical-distributional format of the model means that there is an organization of sounds into words rather than each sound standing independently. Not only are learners optimizing their set of sound categories to best describe the acoustic data they receive, but they are doing so in light of the lexical categories they think generated the words of the corpus. This gives the output of the learning process a higher-level structure.

3.2 Input data

Following previous work, vowel formants were measured in Hertz. Vowel duration was measured in log(ms), and log vowel durations were then scaled by 100 so that their magnitude was comparable to that of formant values. First and second formant values in Simulations 1 and 2 were sampled from a Gaussian distribution generated by vowels produced and recorded in laboratory settings (see Figure 1 for a comparison of acoustic values between these simulations and Simulation 3). Lab production vowels from English were sampled from categories based on the

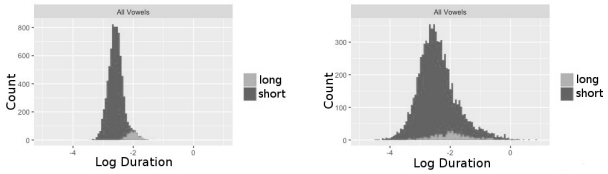

Figure 2: Short and long Japanese vowels as resampled from adult-directed productions (left) and measured from child-directed recordings (right).

recordings of Hillenbrand et al. (1995), which consists of men and women from the Upper Midwest reading a series of words consisting of a vowel inserted into a h_d frame. Formants for each of these vowels were measured once, at the maximally steady point of the vowel, resulting in a clean set of acoustic values, containing minimal co-articulation and reduction. Lab production vowels for Japanese were sampled from categories based on Mokhtari and Tanaka (2000), which includes vowels spoken by men from the ETL-WD-I and II balanced word dataset. Words from the dataset were chosen if they contained vowels with the longest steady-state nuclei and least co-articulatory influences. Adult productions from the R-JMICC corpus (Mazuka et al., 2006) were used to estimate the Gaussians for Japanese durations (Figure 2).

The phonemic and phonetic lexical contexts for English, as well as the natural acoustic data for Simulation 3, were taken from the Buckeye Speech Corpus (Pitt et al., 2007), a hand-aligned corpus of transcribed interviews with men and women from the Columbus, Ohio area. Data consist of transcripts that have been hand-aligned to corresponding audio files so particular acoustic values (including the first and second formant) of each vowel can be identified within their actual context. F1 and F2 values were extracted automatically using Praat (Boersma, 2001) and averaged over the middle third of the vowel to avoid formant artifacts and accidental sampling from neighboring segments. Only monophthongs were used.

The phonemic and phonetic lexical contexts for Japanese, as well as the natural acoustic data for Simulation 3, were taken from the R-JMICC corpus (Mazuka et al., 2006). This corpus consists of recordings of 22 mothers speaking to their children, ages 18-24 months, and to an adult experimenter. The B and T sections of the corpus were used as child-directed speech and the A section of the corpus was used as adult-directed speech. Words that included singing, laughing, coughing, onomatopoeias, and fragments the transcriber could not understand were excluded. In addition, words were excluded if they included more than one vowel for which the annotators could not judge onset or offset time; this automatically excluded any word that included a sequence of two vowels. Formants were extracted automatically using Praat (Boersma, 2001). To identify tracking errors, the means

Table 2: Pairwise F-scores for models trained on adult-directed (AD) and child-directed (CD) speech.

	Phonetic F-score				Lexical F-score			
	English		Japanese		English		Japanese	
	AD	CD	AD	CD	AD	CD	AD	CD
Simulation 1	0.78	0.80	0.96	0.98	0.96	0.94	0.98	0.98
Simulation 2	0.46	-	0.95	0.95	0.63	-	0.97	0.99
Simulation 3	0.13	-	0.24	0.22	0.41	-	0.59	0.61

and variances of the categories in child-directed speech and adult-directed speech were computed, and word tokens were excluded if either of the first two formant values (as measured by Praat) were more than two standard deviations away from the mean of the category they belonged to.

We ran simulations on both adult- and child-directed speech in Japanese. In English, we ran simulations on adult-directed speech, but did not have access to a phonetically annotated corpus of child-directed speech. English child-directed results are given for Simulation 1, with phonemic transcription of words taken from the CHILDES parental frequency count (Li and Shirai, 2000; MacWhinney, 2000); these are comparable with previous results on similar input reported by Feldman et al. (2013). The fact that Japanese CDS results are similar to Japanese ADS results for all three simulations might lead us to predict similar behavior for English ADS and CDS, but this is still an open empirical question.

3.3 Simulation 1

Our first simulation uses a simplified lexicon and acoustic values to replicate previous work and extend this work to Japanese. For example, in this corpus, every time the word *can* is present, it is represented as /k æ n/, with the /æ/ replaced by first and second formant values sampled from the Gaussian estimated by all /æ/ laboratory productions. Input data for both languages consisted of 5,000 word frame tokens from a phonemic transcript with consonants represented categorically and vowels replaced with data points sampled from distributions over lab productions. For English, these 5,000 words are made up of 1099 word types, while for Japanese, they are made up of 751 types. Both languages perform very well on this simplified input corpora (Table 2).

We see similar results for the English data as in previous work; a phonetic category recovery with a F-Score of 0.78 compared to the 0.76 from Feldman et al. (2013). When we extend the model to Japanese, it performs even better (F-Score of 0.98). This high performance in Japanese may be partially due to the quality of distributional information in the model's input. A comparison of the re-sampled vowels for English and Japanese (Figure 1) shows that even for the Gaussian distributions over laboratory produced vowels, the Japanese vowels show

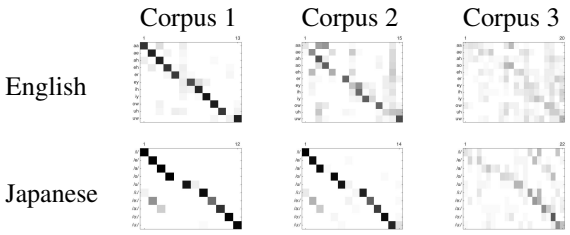

Figure 3: Vowel confusion matrices. Rows correspond to true categories; columns correspond to model categories. The color of a square indicates the proportion of vowel tokens from the true category that are assigned to a model category, with darker squares corresponding to a higher proportion of tokens.

greater separability in F1 and F2.

Using confusion matrices (Figure 3), we can see which categories of sounds the model is most likely to confuse. The model correctly categorizes most of the sounds – that is, tokens belonging to the same vowel category are categorized as being the same sound by the model. We can see, however, that in Japanese the model has a slight tendency to create categories containing both short and long sounds (becoming more pronounced in Simulation 3, below) whereas in English the model has a tendency to create some categories containing tokens of many vowel types (a 'catch all' category).

These results go beyond previous findings to show that the high performance of the model extends to another language, Japanese, and that learning performance on adult- and child-directed corpora is comparable in this model. They provide a baseline for Simulations 2 and 3, which investigate the effect of increased variability on learning performance.

3.4 Simulation 2

Simulation 2 uses more realistic lexical information in the form of phonetic transcripts for lexical context, but retains the simplified acoustic information from Simulation 1. In this corpus, the word *can* may vary each time it is present depending on how it was produced in the original audio. For example, it may be represented as /k ɪ n/, /k ɛ n/, or /k ə n/ as well as /k æ n/. In each of these cases, the vowel is replaced by first and second formant values sampled from the Gaussian estimated by laboratory productions of the corresponding vowel. The most notable contrast between English phonemic and phonetic lexical transcripts involved phonetic reduction. Frequent words in English are often reduced in natural speech, resulting in words with a single phonemic transcription for Simulation 1 having numerous phonetic variations in Simulation 2. This vowel reduction

generated numerous 'minimal pairs' in English which appeared to contain different vowels but were in reality reduced versions of a single word type. This type of prosodic variation affected English lexical information considerably more than Japanese. For English, the phonetic transcription of the same 5,000 words that were used in the Simulation 1 contained 1813 types, in comparison to the 1099 types from the phonemic transcription. In contrast, for Japanese, the phonetic transcription of the 5,000 words only increased to 791 word types (from 751 in the phonemic transcription). This is mirrored in the results, as we see a sharp drop in performance for English once the lexicon is represented more realistically, but no corresponding drop in performance for Japanese (Table 2).

These results indicate that variation in the input does affect model performance, but that certain types of simplification – such as representing lexical items phonemically rather than phonetically – may hide more variation in some languages than in others. For languages with a large amount of phonetic reduction, such as English, a single word type can surface with many different pronunciations depending on how fast or carefully it is produced. These different pronunciations cause the model to create more categories containing multiple vowel types as it attempts to categorize all these tokens as types of the same word (Figure 3), lowering model performance. In the example given above, if the multiple pronunciations of *can* are correctly identified as tokens of a single lexical item, this encourages the model to create a single vowel category varying over [ɪ/ɛ/ə/æ].

3.5 Simulation 3

The final simulation was run on the most realistic input, with lexical items represented by phonetic transcription and vowels replaced by acoustic values taken from the spoken corpus. For example, every presentation of the word *can* was represented as /k (F1,F2) n/ with formant values taken directly from the audio recording of that particular word instance. The increased variation present in this input caused a large drop in performance for both English and Japanese (Table 2). From the confusion matrix for Japanese, we see that additional categories were created for the short vowels (evidenced by the faint diagonal line in the upper right corner of the matrix; Figure 3). This appears to be due to prosodic lengthening affecting vowel durations, in some cases encouraging the model to create extra durational categories for vowels in words frequently found immediately preceding prosodic boundaries, which are known to lower speaking rate and thus to increase vowel duration (Bion et al., 2013). Although these confusions are present in Japanese but not English, they do not account for the majority of confusions in Japanese and, when normalized out (by normalizing speaking rate for each word in the corpus), performance in Japanese does not improve.

These results show that model performance does not only decrease when run

on more realistic acoustic input; it plummets. Although there appear to be some language specific effects on the increased variability due to lengthening in Japanese, they do not account for much of the confusion of model vowel categories. The drop in performance for both languages here is primarily due to the non-normal and highly overlapping distribution of actual acoustic values seen in Figure 1.

3.6 Summary of Results

We find a decline in model performance on both English and Japanese as the input corpora become more faithful to the original acoustics and thus contain more variability. Although increasing vowel variability impacts phonetic learning in both languages, the consequences of that impact differ. Both languages perform very well when phonemic transcriptions are used for lexical items and acoustic values are sampled from lab productions for vowels, but English suffers a sharp drop in performance as soon as the lexicon is represented phonetically, while scores stay high for Japanese until the sampled lab vowels are replaced with acoustics taken directly from the spoken corpus.

4 Discussion

In this paper, we investigated the impact of simplified input to computational models of phonetic acquisition. Three simulations were run on corpora of decreasing simplification. As input to the model becomes more natural and contains more variability, model performance declines considerably. Model performance is much better on certain types of input, specifically input where words are represented as having a consistent pronunciation and vowel values make up clear, relatively non-overlapping, normally distributed categories. Unfortunately, this is not the same type of input that children receive. The model's poor performance on realistic data indicates that it may be missing cues utilized by children in learning vowel categories, particularly given the existing experimental results showing that children learn better from variable data.

The magnitude of the decline in model performance indicates a serious mismatch between previous models and real speech data. If previous modeling results were taken as evidence that distributional and lexical information can help children acquire phonetic category systems from data, the present results should serve to qualify that claim: distributional and lexical information can help children acquire phonetic categories from realistic data only when equipped with some way to compensate for other sources of variation. This steep decline in performance is not likely to be a failing of the particular model we tested; rather, we would expect the impact of acoustic variation to extend to more traditional distributional models as well, which would still have the added difficulty in disambiguating overlapping

categories. Given that performance seems to decline due to irregularities or variability in the data received by the model, there seem to be two possible solutions: either the selection of a low-variability subset of tokens which the model uses to learn vowel categories, ignoring messier examples of vowel data (Adriaans and Swingley, 2012) or a simultaneous normalization process by which the model uses some information about prosody, context, etc. to normalize variable tokens so they can be utilized for learning (Dillon et al., 2013).

These results show that simplifying input to computational models of phonetic category learning can drastically impact model performance. At least part of this difference arises from the lack of prosodic variability, such as vowel reduction and phrase-final lengthening, in simplified input. In future work it will be important to consider how prosody is taken into account during phonetic learning. More generally, our results underscore the importance of ecologically valid datasets. For computational modeling to be useful in exploring child language acquisition, we must account for the variation in the input children actually receive.

References

Frans Adriaans and Daniel Swingley. Distributional learning of vowel categories is supported by prosody in infant-directed speech. In *Cognitive Science*, pages 72–77, 2012.

Ricardo A. H. Bion, Kouki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka. Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE*, 8(2):e51594, 2013.

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10): 341–345, 2001.

Dani Byrd. Relations of sex and dialect to reduction. *Speech Communication*, 15(1-2): 39–54, 1994.

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

Brian Dillon, Ewan Dunbar, and William Idsardi. A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, 37(2):344–377, 2013.

Naomi H Feldman, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4): 751, 2013.

Sarah Hargus Ferguson. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 116(4):2365–2373, 2004.

Marios Fourakis. Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical Society of America*, 90(4):1816–1827, 1991.

Rebecca L Gómez. Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436, 2002.

Esther Grabe and Ee Ling Low. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, 7(515-546), 2002.

Bernard Harmegnies and Dolors Poch-Olivé. A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech Communication*, 11(4): 429–437, 1992.

James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111, 1995.

Derek M Houston and Peter W Jusczyk. The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5):1570, 2000.

Aren Jansen and Ken Church. Towards unsupervised training of speaker independent acoustic models. *Proceedings of Interspeech*, 2011.

Keith Johnson. Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis*, pages 29–54, 2004.

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

Chia-ying Lee and James Glass. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of ACL*, pages 40–49. Association for Computational Linguistics, 2012.

P. Li and Y. Shirai. *The acquisition of lexical and grammatical aspect*. Mouton de Gruyter, New York, 2000.

Björn Lindblom. Economy of speech gestures. In *The production of speech*, pages 217–245. Springer, 1983.

Brian MacWhinney. *The CHILDES Project*. Erlbaum, Mahwah, NJ, 2000.

Jessica Maye, Janet F Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, 2002.

Jessica Maye, Daniel J Weiss, and Richard N Aslin. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11(1):122–134, 2008.

Reiko Mazuka, Yosuke Igarashi, and Ken'ya Nishikawa. Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus. *The Technical Report of the Proceedings of the Institute of Electronics, Information and Communication Engineers*, 106:1115, 2006.

Bob McMurray, Richard N. Aslin, and Joseph C. Toscano. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3): 369–378, 2009.

Parham Mokhtari and Kazuyo Tanaka. A corpus of Japanese vowel formant patterns. *Bulletin of The Electrotechnical Laboratory (ETL), Japan*, 64:57–66, 2000.

Seung-Jae Moon and Björn Lindblom. Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, 96 (1):40–55, 1994.

Michael A Picheny, Nathaniel I Durlach, and Louis D Braida. Speaking clearly for the hard of hearing iiacoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.

Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth

Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech (2nd release), 2007.

Linda Polka and Janet F Werker. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421, 1994.

Carolyn Quam and Daniel Swingley. Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *Journal of Memory and Language*, 62(2):135–150, 2010.

Carolyn Quam, Sara Knight, and LouAnn Gerken. The distribution of talker variability impacts infants word learning. *Journal of the Association for Laboratory Phonology*, 8 (1), 2017.

Gwyneth C Rost and Bob McMurray. Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2):339–349, 2009.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

Rebecca Anne Scarborough. Lexical confusability and degree of coarticulation. In *Annual Meeting of the Berkeley Linguistics Society*, volume 29, pages 367–378, 2003.

Leher Singh. Influences of high and low variability on infant word recognition. *Cognition*, 106(2):833–870, 2008.

Leher Singh, James L Morgan, and Katherine S White. Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, 51 (2):173–189, 2004.

Leher Singh, Katherine S White, and James L Morgan. Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4(2):157–178, 2008.

Janet W Stack, Winifred Strange, James J Jenkins, William D Clarke III, and Sonja A Trent. Perceptual invariance of coarticulated vowels over variations in speaking rate. *The Journal of the Acoustical Society of America*, 119(4):2394–2405, 2006.

Daniel Swingley. Contributions of infant word learning to language development. *Phil. Trans. R. Soc. B: Biological Sciences*, 364(1536):3617–3632, 2009.

Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278, 2007.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. Unsupervised learning of acoustic subword units. *Proceedings of the Association for Computational Linguistics*, pages 165–168, 2008.

Susan Ellis Weismer and Linda J Hesketh. Lexical learning by children with specific language impairmenteffects of linguistic input presented at varying speaking rates. *Journal of Speech, Language, and Hearing Research*, 39(1):177–190, 1996.

Janet F Werker and Richard C Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63, 1984.

Katherine S White and Richard N Aslin. Adaptation to novel accents by toddlers. *Developmental Science*, 14(2):372–384, 2011.