A single stage approach to learning phonological categories: Insights from Inuktitut

Brian Dillon*
Department of Linguistics, University of Maryland, College Park, MD 20742, USA

Ewan Dunbar
Department of Linguistics, University of Maryland, College Park, MD 20742, USA

William Idsardi
Department of Linguistics, Program in Neuroscience and Cognitive Science,
University of Maryland, College Park, MD 20742, USA

*Corresponding Author:

Brian Dillon
Department of Linguistics
University of Maryland
1401 Marie Mount Hall
College Park, MD 20742
bdillon2@umd.edu

Abstract

To acquire one's native phonological system, language-specific phonological categories and relationships must be extracted from the input. The acquisition of the categories and relationships have each in their own right been the focus of intense research. However, it is remarkable that research on the acquisition of categories and the relations between them has proceeded, for the most part, independent of one another. We argue that this has led to the implicit view that phonological acquisition is a 'two-stage' process: phonetic categories are first acquired, and then subsequently mapped onto abstract phoneme categories. We present computational simulations that suggest that this view is almost obligatory given current statistical models of category acquisition: they both under- and over-estimate category structure depending on the language. Solving the former problem incompletely will invariably exacerbate the latter problem, and so these approaches are guaranteed to converge on phonetic, rather than phonemic categories. We suggest an alternative conception of the phonological acquisition problem that sidesteps this apparent inevitability, and acquires phonemic categories in a single stage. Using actual acoustic data from Inuktitut, we show that this model converges on a set of phoneme-level categories once predictable acoustic variation is factored out.

Introduction

In recent years, statistical approaches to language acquisition have generated much enthusiasm, especially in the domain of phonological acquisition. On the face of it, the problem of how human children acquire phonological categories (*phonemes*) of spoken language presents an ideal model problem: we understand a good deal about the time course of development, how the perceptual input to learning is represented, and what the desired end stage of acquisition is. There are very good reasons to model the acquisition process as a form of statistical inference over perceptual input, as we document below. However, the general approach to phonological category formation as perceptually driven statistical inference has led to the view that the initial categorization posited by the learner is in some sense isomorphic to all and only the distinctions present in the acoustics. That is, this approach implicitly suggests that such statistical approaches are meant to discover *phonetic* rather than *phonemic* categories.

Phonetic categories are not the desired end state of phonological acquisition, however. Instead the learner must identify language-specific phonemes, phonologically relevant, abstract sound categories that may consolidate several distinct phonetic realizations into equivalence classes for the purpose of lexical, morphological and syntactic operations. There are varied theoretical approaches to this problem, but the problem of how to map phonetic categories to a higher-level phonological system of phonemes or relationships has itself generated a large body of work (Harris, 1951; Tesar & Smolensky, 1998; Boersma & Hayes, 2000; Peperkamp, Le Calvez, Nadal & Dupoux, 2007, Goldsmith & Xanthos 2009, among others). By using phonetic categories as input to various learning algorithms, work on the phonetics-phonology mapping often tacitly assumes that the learner was able to reliably identify these categories in a prior stage.

As we will see, this disconnect between the statistically induced phonetic categories and the target phonemic categories has led to an implicit 'two-stage' view of phonological learning: learners first learn phones, and then build phonemes and phonological systems by identifying relations between them. In this paper, we present simulation evidence that such a view is a necessity given current models of first stage statistical categorization, because current approaches are guaranteed to converge on phonetic, rather than phonemic categories. This demands a second acquisition stage that subsequently builds the relevant phonemic categories. We then argue that a two-stage approach is not inevitable, and present an alternative, 'single-stage' model that learns phonemes directly by factoring out predictable alternations conditioned on environment.

The phonological learning problem

The two-stage view of phonological acquisition mirrors a traditional distinction that linguists have long drawn between *phonetics* and *phonology*. Phonetics refers to the study of audition and production, and phonology is concerned (sometimes implicitly) with the encoding of speech in the lexicon (i.e. long-term memory). Much work in phonetics turns on the observation that phonetic representations are finely detailed, and thus best represented as continuous rather than discrete values (Fant, 1960; Ohala, 1976; Ladefoged, 2001; Silverman, 2006). The phonological level is instead thought to abstract away from the detailed properties of the phonetic representations to varying degrees, and

it is almost always taken to be in a discrete rather than a continuous encoding (Chomsky & Halle, 1968; Goldsmith, 1976; Prince & Smolensky, 2004). The inventory of phonemes varies from language to language, and an infant acquiring her native tongue must identify the phoneme categories that are relevant for her language. Part of this task includes determining the distribution of each phone or phoneme in acoustic and/or articulatory space. Determining which acoustic realizations (or articulatory movements) map to which phonemes is a prime example of an unsupervised learning problem. This characterization of the problem has allowed researchers to make direct contact with a vast literature in machine learning, and has led to important new models of phonological acquisition.

The view of phonological category acquisition as unsupervised clustering is complicated by the non-trivial mapping between phonemic and phonetic representations because of the existence of *phonological processes*, systematic adjustments that affect the pronunciation of sounds in certain environments. For example, the realization of Spanish /b/ varies between fricative and obstruent in an entirely predictable fashion: roughly speaking, the fricative pronunciation occurs between two vowels, and the obstruent pronunciation occurs elsewhere (Harris 1969). Native speakers of Spanish have mastered this alternation, and it is productively deployed across the entire language. Learners must thus acquire this alternation and a single phoneme category /b/ (or equivalent knowledge), but the alternation appears to disrupt any straightforward mapping from acoustics to phonemes.

A wide variety of theoretical approaches have viewed phonological processes as operations over discrete units (Chomsky & Halle, 1968; Prince & Smolensky, 2004). Thus, the relation between the two pronunciations of Spanish /b/ is a phonological process taking a discrete object (the phoneme /b/) to another discrete representation (e.g. the fricative [β]). It is often assumed that more detailed phonetic information is filled in after all phonological processes have taken place (Chomsky & Halle, 1968). This means that there are at least two discrete levels of representation involved in phonological cognition. One is the phonemic level. The other is the level of the phonetic categories (*phones*). Phones are representations that result from the application of phonological processes, and which serve as the input to the mapping to acoustic detail (Hockett, 1942; Hayes, 2009).

Although it is a useful (and nearly ubiquitous) theoretical device, it is not clear that there is independent motivation for a discrete level of representation for phones, an 'intermediate' level of abstraction from acoustics. In fact, the view that phonological representations contain rich acoustic information is gaining much attention in the theoretical literature; see Pierrehumbert (2003) and Silverman (2006) for an overview. Nonetheless, many researchers maintain discrete levels of phonetic and phonemic representation, and this has provided the theoretical backdrop for the two-stage model of phonological acquisition that we are concerned with here. Having two discrete levels of representation allows for a view of phonological acquisition in which the mapping between discrete phones and detailed phonetic information is learned before the mapping between phones and phonemes. Below we will argue that, although it is largely implicit, this view of phonological learning is pervasive in the language acquisition literature.

The two-stage approach

The theoretical distinction between discrete phonetic and phonemic encodings is echoed in research on phonological acquisition. As there are two distinct levels of representation to be acquired, research has focused on either one or the other of these levels. This division of labor has lead to an emergent two-stage approach to phonological acquisition.

The first 'stage' of the two-stage position focuses on the mapping from acoustics to phones. This is represented by a large body of work on discovering category structure from acoustic data, often using explicit statistical models of inference (de Boer & Kuhl, 2001; Coen, 2006; Vallabha, McClelland, Werker, Pons, & Amano, 2007; Feldman, Griffiths, & Morgan, 2009). The aim of these models is sometimes cast in a theory-neutral way: Vallabha et al (2007), for instance, propose a model for learning 'sound categories' (pp. 13273). On the other hand, Feldman et al (2009) explicitly note that their model is likely to acquire phonetic, rather than phonemic, categories. We will demonstrate below that Feldman's characterization is exactly correct: without explicit modeling of the phonological processes, this first stage is guaranteed converge on phonetic rather than phonemic categories. In other words, this line of research, as presented, must be one stage in a two-stage process; it will not converge on phonemic categories by itself.

It is important to note, however, that the mapping from acoustics to categories that these approaches attempt to model is in some sense an irreducible problem for phonological acquisition. There is arbitrary variation in acoustic targets for the 'same' phone or phoneme category across languages (Pierrehumbert, 2003), and so the learner must have a way of acquiring this mapping. It is not clear that we can build an explicit model of phonological acquisition without a mechanism for inferring category structure over a perceptual space. As such, these first-stage models will form the backbone of the alternative model we are considering here, and we will discuss them in some detail below.

If, as is the case, our model of the acoustics-to-category acquisition process is guaranteed to converge only on phonetic categories, we must separately address the second half of the phonological learning problem: the phone to phoneme mapping, or the phonological grammar. One solution is to deny that there is any such mapping. We will not pursue this approach here; there are compelling theoretical and empirical reasons for rejecting this view, and we address these in the discussion. The only other option left is to develop an explicit theory of how to group phones into phonemes. This was the route taken by Zellig Harris, who formalized a now famous complementary distribution metric (Harris, 1951). More recently, there has been much interest in exploring alternative conceptions of this process using more sophisticated statistical measures such as Kullback-Leibler (KL) divergence (see Appendix C) (Peperkamp et al, 2006). Peperkamp and colleagues proposed a solution to this problem that compared the (statistical) distribution of different phones. Phones that had significantly divergent distributions were plausible candidates for variants of the same phoneme, subject to constraints on possible phone-to-phoneme relationships. By looking at distributions of different phones, this algorithm implicitly assumes that the phones have been uniquely identified and discretely categorized at some level. In other words, it assumes a successful first stage of acquisition, of mapping from acoustics to phones. Note that in this approach,

phonological learning is still not complete once the phonemes have been identified. The learner must still learn the grammatical mapping between the phones and phonemes (i.e. the form of the relevant phonological rule). This approach also suffers from an inability to recover certain allophonic relations, as discussed by Dunbar and Idsardi (2009).

An alternative conception of this second stage is seen in work on learning of Optimality-theoretic grammars (Prince & Smolensky, 2004). On this approach, the grammar is treated as a ranked set of well-formedness constraints, which determine the correct surface pronunciation when combined with underlying forms for morphemes. There are several computationally explicit, well-known algorithms for learning these grammars (Tesar & Smolensky, 1998; Pulleyblank & Turkel, 2000; Boersma & Hayes, 2001; Hayes, 2004). Though they vary in their approach, they share the common assumption that the input to learning is a set of discrete, phone-level representations, and the grammar (and phoneme inventory) is derived only once these phones are identified. By taking the output of a first-pass mapping from acoustics to phones, these approaches also implicitly endorse the two-stage view of phonological acquisition. There exist still more approaches to the phoneme-finding problem (Jakobson, 1941; Goldsmith & Xanthos, 2009; Dresher, 2009), but all assume that a set of phones has already been discovered.

As we will detail below, however, a successful mapping from acoustics to phones is far from given, and has been the focus of intensive research from both experimental and computational approaches. The success of a two-stage approach to phonological learning crucially depends on the accuracy achieved in the first stage, as errors made in the first stage mapping could fatally impair the ability of a second-stage mechanism to extract the correct phonology. Though we have sketched a two-stage view of phonological acquisition, and suggested that it appears to be implicit in the majority of research on phonological acquisition, it is clearly not the only possibility. In the remainder of this paper, we explore the feasibility of a single-stage approach to phonological categorization. Our approach is to focus on what we have termed here the first stage of phonological acquisition; in particular, we focus on statistical methods of category identification. All theories of phonological acquisition must address this mapping from acoustics to discrete categories, and there are good experimental and theoretical reasons for assuming it is some form of statistical inference over the input. Because of this fact, the feasibility of a single-stage approach turns on the possibility of folding the acquisition of processes and phoneme level categories into the initial mapping form acoustics to linguistic categories.

Mapping from acoustics to linguistic categories in acquisition

The relation between acoustic variation and linguistic categorization has been the subject of much research in psycholinguistics. Very young infants have famously been termed universal listeners, being able to discriminate amongst a wide range of sounds not present in their input (Aslin, Jusczyk, & Pisoni, 1998). A number of studies have shown that these discriminatory abilities soon decay as the infant develops. Declining sensitivity to non-native language vowel contrasts is apparent for vowels as early as 6 months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), and by 8 months, similar effects are evident in consonant contrasts (Werker & Tees, 1983). A common observation about this

arc of development is that it seems to suggest that phonological category learning could in fact drive the building of a lexicon, rather than the other way around as had been previously hypothesized (Jusczyk, 1985; Best, 1995). Given that infants can reliably discriminate minimal pairs only later (around 18 months or so), it appears that at this age, higher-level constructs such as minimal pairs do not drive the development of contrast (Dietrich, Swingley & Werker, 2007).

Maye, Werker & Gerken (2002) hypothesized that distributional learning mechanisms are the fundamental building block of phonological development, a hypothesis that is fully compatible with the observed time-course of acquisition. Building on results that show that infants are sensitive to distributional information in other modalities, such as word-learning (Saffran, Aslin & Newport, 1996), they aimed to show that infants used distributional cues to bootstrap contrastive categories from the input. Infants at both 6 and 8 months of age were exposed to training sets that contained either bimodal or unimodal distributions over voice-onset times. When presented with bimodal distributions, infants showed greater sensitivity to differences between the modes in the VOT distribution, whereas for the infants in the unimodal condition, sensitivity was decreased for the same sounds. The authors interpreted these findings as showing that distributional characteristics of the input directly impacted what dimensions of the signal the infants viewed as relevant or contrastive. Werker. Pons, Dietrich, Kajikawa, Fais and Amano (2007) went on to show that, for the vowel space, the distributions in actual speech support the acquisition of relevant phonological contrasts. They showed that in infant-directed speech of both Japanese and English speakers, clear distributional cues support the relevant contrasts (e.g. duration cues for the Japanese vowel space), and minimize irrelevant dimensions of variation.

In addition to these experimental results, computational models have been more and more successful in modeling the acquisition of speech categories using various types of statistically informed frameworks. As an unsupervised learning problem, the acoustics-to-category mapping can be modeled with a range of clustering techniques from machine learning. Clustering techniques come in many different forms, but all share the notion of carving out subsets of a data set and classifying them together based on a pre-defined distance metric in a given representational space (Everitt, Landau & Leese, 2001). An example of this cluster analysis approach is the work of de Boer and Kuhl (2001), who used a mixture of Gaussians (MOG) model to model the acquisition of vowel categories using the values for the first two formants. MOG models represent category structure as a set of parameterized Gaussian distributions (termed components of the model) in the input space, weighted by a mixing probability (for further explication and discussion of MOG models, see below and Appendix A). The model was fit with an Expectation-Maximization algorithm (EM, see Dempster, Laird & Rubin, 1977; see also Appendix B), which is a family of hill-climbing algorithms that seeks to maximize a measure of likelihood for an unobserved category structure. Both the EM algorithm and the MOG approach have received a good deal of attention in the statistical and machine learning literature for the last two decades, and their properties are relatively well-understood (McLachlan & Peel, 2000; Frühwirth-Schnatter, 2006). De Boer and Kuhl applied these techniques to vowels that were recorded during mother-child interactions, focusing on the vowels at the extreme edges of the vowel space in English (/i/, /u/, and /a/) and limiting the MOG models to three-component mixtures only (that is, they pre-specified the

number of components in the model and supplied this prior information to the cluster-fitting procedure). By applying this approach, and clustering separately for each speaker, they showed that the model was better able to acquire the categories on infant-directed speech than on adult-directed speech, suggesting one possible utility of infant-directed speech.

Another important contribution to the modeling of vowel acquisition was the Online Mixture Estimation (OME) model of Vallabha and colleagues (2007). Using both parameterized and non-parameterized versions of OME, Vallabha et al were able to model the acquisition of the Japanese and English data that was analyzed by Werker et al (2007). Their model fit four-dimensional Gaussians in the raw acoustic space (F1, F2, F3 and duration), updating its weighting for relative weighting of the parameters after each input point. Applying this technique within speakers, they were able to achieve a good rate of success for most speakers, successfully learning the vowel systems 80% of the time for parameterized models, and 60% of the time for non-parameterized models. This model is important in that it is the most successful version of an online algorithm that processes each data point as it arrives. This is opposed to batch algorithms that try to estimate categories based on a large store of exemplars stored in memory in an 'offline' fashion.

Still other modeling attempts have tested different assumptions about what information is recruited to solve the phonological learning problem: Coen (2006) analyzed video samples of American English vowels, and used a cross-modal clustering technique (not the MOG approach assumed elsewhere) to form and cross-correlate clusters in both acoustic and visual space (i.e. shape of mouth). Feldman, Griffiths, and Morgan (2009) presented a hierarchical Bayesian model that used lexical information to separate highly overlapping categories. The already-impressive success of the basic MOG approach has thus been shown to benefit from the addition of extra disambiguating information.

The mapping from acoustics to linguistic categories proves to be an extremely challenging problem in its own right, and it forms an irreducible part of any phonological acquisition model. We now turn to a presentation of several computational simulations of this stage of acquisition, examining data from Turkish and Inuktitut. One primary result of these simulations is that the generic MOG approach is guaranteed to converge on phones, rather than phonemes. Even extracting the correct set of phones, however, proves extremely difficult. By considering an alternative conception of the first stage clustering process, however, we will show that it is in principle possible to extract phoneme level categories in the first stage clustering. When compared with the generic MOG approach on the same data, the single-stage approach will be see to converge more reliably on the language relevant categories.

The MOG approach: Bounding the problem from above and below

We adopt a mixture of Gaussians (MOG) approach to modeling the acquisition and representation of the vowel space, following recent successful attempts (de Boer & Kuhl, 2001; Vallabha et al, 2007). The mixture of Gaussians approach represents the vowel space as a *mixture* of normal distributions or *components*. Used as a model of the

vowel space, each MOG component is viewed as a phoneme category, and the component describes the location and spread of that phoneme is acoustic space.

In this section we present MOG simulations that effectively bound the learning problem from above and below. The MOG approach, as we will see, is both too powerful and not powerful enough, as a function of the particular vowel system under consideration. In particular, we will show that the crowded front vowel space of Turkish obscures the true phonemic category structure from our MOG model. However, the very same model overestimates the category structure of the Inuktitut vowel space in response to predictable allophonic processes. It is this fact that makes the two-stage approach seem like a conceptual necessity: we need more and more sensitivity to separate out categories, but this distances us from our goal of phoneme identification. In order to separate the phoneme categories for Turkish, we are guaranteed to converge on phonetic categories for Inuktitut. If we are assured of converging on phonetic categories, then complete phonological acquisition requires a second mechanism to identify phonemes given the phonetic inventory.

Turkish

We first consider the application of the MOG approach to the Turkish front vowel space. The Turkish front vowel space contains four vowel phonemes: /i/, /e/, /y/ and /ø/. This is similar in organization to the English and Japanese front vowel spaces considered by Vallabha et al (2007), but rather than length, it is vowel quality (namely, height and rounding) that distinguishes vowel phonemes (see Figure 1). The Turkish space under consideration provides an interesting test case because of the discrepancy in frequency among the vowel categories. Previous approaches (Vallabha et al, 2007; but cf. Feldman et al, 2009) have modeled the vowels as equi-probable in the input, an assumption that can be quite misleading. Consider Table 1, which presents the ratios of occurrence of acoustically close contrasts in child-directed speech in a number of languages. Notably, Turkish front vowel contrasts are among the more equally distributed. The Japanese length contrast, on the other hand, provides a much more skewed distribution. As seen in Figure 1, the Turkish vowel phonemes considered here are quite close in acoustic space. The frequency discrepancy among these phonemes compounds the learning problem. The simulations on Turkish presented here examine whether or not the generic MOG approach can overcome these obstacles to reliably separate out the four phoneme categories present in the front vowel space of Turkish.

Contrast	Ratio
Estonian /e/-/ø/	38:1
Japanese /e/-/e:/	15:1
Japanese /i/-/i:/	14:1
Estonian /i/-/y/	10:1
Swedish /i/-/y/	9:1
Swedish /e/-/ø/	8:1

Turkish /e/-/ø/	6:1
Turkish /i/-/y/	5:1

Table 1: Approximate ratios of occurrence of acoustically similar contrasts across a sample of languages. Ratios are based on occurrence in child-directed speech found in CHILDES corpora (McWhinney, 2000; Plunkett & Strömqvist, 1992; Strömqvist, Richthoff, & Andersson, 1993; Kohler, 2004; Slobin, 1982; Miyata, 1992; Miyata, 1995; Miyata, 2000; Noji, 1973-77).

All vowel data presented here were drawn from the METU speech corpus (Salor, Ciloglu & Demirekler, 2006). Speakers varied in age, gender, dialect region, and educational background, although for the simulations reported below, only data from female speakers was modeled. For each vowel in the corpus, formants were automatically extracted using the Snack package for TCL. For this, we used a 12-pole LPC analysis, using 30 ms time bins with Hamming windowing. Each token was labeled with an F1, F2, and F3 value drawn from the center of the vowel. In order to avoid errors in formant tracking or mislabeling of formants, tokens for which any formant value was more than 2 standard deviations from the mean for a given vowel were excluded. In all, 51,687 vowel tokens were collected, and 1298 (2.5%) were excluded according to these criteria, resulting in 50,389 vowel tokens.

The parameters for each of the front vowel categories (F1, F2, F3, and duration) were estimated from this set of tokens on a speaker-by-speaker basis. For ease of comparison with other work, we only analyzed data from female speakers. Of the 42 female speakers represented in the corpus, any speaker who had less then 10 productions for any given front vowel was excluded. According to this criterion, 15 speakers were excluded. Of the remaining 27 speakers included in the simulation, the average number of vowel tokens used in estimating the Gaussian parameters of the vowel category was 45. These categories were subsequently resampled to provide the actual training and test sets, following the method used by Vallabha et al (2007) and Feldman et al (2009).

Simulations in this experiment were performed by fitting mixtures of Gaussians using the expectation maximization algorithm (sees Appendices A and B). This was done using the R statistical computing environment (R Development Core Team, 2008), using the package MClust (Fraley & Raferty 2002; Fraley & Raferty, 2006). On any given run, a training set containing a total of 2,500 tokens was produced by sampling from the estimated 4-dimensional Gaussians, with values for F1, F2, F3 and duration. On runs for which the mixing frequency of the vowel categories was equal (*uniform distribution* training sets), there were 625 tokens of each vowel category. On runs for which the mixing frequencies matched the frequencies of each phoneme in the corpus (*empirical distribution* training sets), there were 1125 /e/ tokens (~45%), 975 /i/ tokens (~39%), 175 /ø/ tokens (~7%) and 225 /y/ tokens (~9%). As in Vallabha et al (2007), 10 independent runs were conducted for each speaker in the simulation. A 'run' consisted of independently generating both an empirical distribution training set and a uniform distribution set, and fitting a mixture model to both. In what follows, we report the number of categories in the best-fit model. A 'successful run' was defined as a run for

which the resultant mixture model had four components that had a combined mixing probability of at least 98%.

In order to assess the goodness-of-fit for successful runs, we employed the Kullback-Leibler distance (see Appendix C) as a measure of how similar the estimated category distribution was to the actual distribution. In order to calculate this value for each vowel category, we first assigned each estimated component in the mixture model to an underlying vowel category. This was done by calculating the Euclidean distance between the Gaussian centers of all estimated components and all actual vowel distributions, forming a distance matrix. The trace of the resultant matrix was then minimized, finding the best alignment of mixture model categories to vowel categories by minimizing the distance between their centers. Once the alignment was obtained between mixture components and vowel categories, the KL distance between the estimated category and its corresponding vowel distribution was calculated.

[INSERT FIGURE 1 HERE]
[INSERT FIGURE 2 HERE]

The performance on the Turkish front vowel space simulations is summarized in Tables 1-2. Table 1 shows the number of speakers who had at least one successful run in the 10 trials. Almost all speakers achieved at least one successful run with the uniform training set, but almost 20% of speakers fail to achieve this landmark in the empirical training set case. For each speaker who had at least once successful run, we also present the average number of training runs that were considered successful. A majority of runs were successful with uniform training sets; empirical training sets, on the other hand, were far more often failures than successes. A second measure of performance is given in Table 3. Table 3 presents the KL distances between the estimated categories and the actual generating vowel categories, for successful runs only. The results show that the results from uniform training runs were far better matched to the generating categories than were the results obtained from empirical training sets. For reference, Figure 2 shows examples of the underlying distributions and the estimated categories for a single representative speaker, along with the corresponding KL values for each vowel.

Training set	# of speakers with successful runs	Average # of successful runs (± 1 SE)
Uniform	26 / 27	6.96 ± 2.31
Empirical	22 / 27	2.2 ± 1.12

Table 2: Performance on individual runs in the Turkish front vowel space.

Training set	/i/	/e/	/y/	/ø/
Uniform	.04 ± 0.06	.03 ± 0.04	.05 ± 0.09	.05 ± 0.07
Empirical	.31 ± 0.69	.40 ± 1.31	.63 ± 1.50	.49 ± 1.05

Table 3: Average KL distances (with standard deviations) between estimated categories and actual vowel categories, for successful runs in the Turkish front vowel space.

These results show that on the uniform training problems, the vowel space was readily recovered for a majority of speakers and trials. However, there was a significant decline in performance when the actual empirical mixing probabilities for the front vowels were used, and the correct vowel categories were only rarely obtained. These simulations reported here suggest that the problem of statistical clustering of phonetic categories in acoustic space in some ways more robust and in other ways more difficult than previously reported. On the one hand, our results suggest that adult-directed speech contains cues to category structure as reliably as infant-directed speech. When the training data contained a uniform distribution over the vowel category mixing probabilities, our success was comparable to that of Vallabha et al (2007), who used infant-directed speech elicited in the laboratory. The MOG approach was able to reliably converge on the four target categories for all speakers, and did so on approximately 69 % of runs. In those cases where it did converge, an analysis of the KL distance between actual and estimated categories revealed an extremely tight fit between them.

However, the results suggest a dramatic decrease in success once the vowel tokens were distributed according to the empirical mixing probabilities of their respective categories. Only 80% of speakers achieved a single successful run with the empirical training sets, and among those who did, only 22% of runs were successful on average. The KL distances in these cases were much larger than in the uniform case (for reference, see the KL distances in Figure 2), suggesting much less reliable category estimates were produced for empirical distribution training sets.

As mentioned above, such discrepancies in mixing probabilities are rampant in natural language, and the Turkish case does not present a particularly extreme example. The current results suggest that learners of Japanese or Estonian cases face a particularly difficult task in acquiring the less frequent categories of long vowels (for Japanese) or front rounded vowels (for Estonian). Available experimental evidence shows that difficult contrasts do not entirely impede learning in the first year of life. Kuhl et al (1992) show that Swedish 6-month olds show language-specific category prototype effects for Swedish /y/, despite the fact that it occupies a position in the vowel space that could obscure its presence. As seen in Table 1 above, the Swedish /i/-/y/ contrast may in fact be more unbalanced than the Turkish case presented here. However, the current results are in line with experimental findings that show that infants show decreased discrimination for difficult acoustic contrasts. Sabourin, Werker, Bosch, and Sebastian-Galles (submitted) show that for infant learners of Canadian English, the contrast /e/-/ɪ/ is distinguished less reliably than /ɪ/-/ɛ/, despite roughly similar acoustic distances. In spoken English, /ɪ/ is about 2.3 times as frequent as /e/, but occurs approximately as often as /ɛ/ does (Mines, Hanson & Shoup, 1978). This may be taken as evidence that the category structure of /e/ as distinct from /ɪ/ is less robustly represented during the stage of acquisition. The current modeling results suggest that this could be due in part to the differential rates of occurrence in the input, a point noted by Sabourin et al.

The most crucial outcome of the Turkish simulations is that whatever the solution employed by human learners or modelers to overcome this particular problem, it will be one that requires a more powerful category-detecting apparatus. To learn Turkish, we need to be able to reliably find more categories than we presently do. As we will see, however, using the exact same learning mechanism on Inuktitut data will force exactly

the opposite conclusion: we already are too sensitive to category structure to recover phoneme categories. In the context of the general MOG approach to category acquisition, this will demand a second stage in phonological acquisition that acquires higher-level phoneme categories.

Inuktitut

We observed that for the present MOG approach, certain common cross-linguistic distributional patterns obscure category structure during acquisition. We now focus on the Inuktitut vowel system to ask if the reverse situation ever obtains: does the MOG approach ever reliably overestimate the phonemic structure of the language under consideration? Inuktitut is an Eskimo-Aleut language spoken in northern Canada. Like many other Inuit languages, Inuktitut has three vowel phonemes: /i/, /u/, and /a/. The quality of a vowel, however, is often affected by a following uvular consonant (either /q/ or /r/), in a way that varies across different dialects (Dorais, 1986). In the dialect under consideration, uvular consonants lower all vowel tokens to some degree (Denis & Pollard, 2008), suggesting the presence of six phonetic categories (see Figure 3), though this effect is often not reported with the low vowel /a/. An important question for the MOG approach is whether or not this predictable allophonic variation leads to 'spurious' category formation during the acquisition of phoneme categories.

The Inuktitut vowel corpus that we employed comes from a study on Inuktitut phonetics performed by Derek Denis and Mark Pollard (2008). All vowel tokens were measured from elicited speech of a speaker from Cape Dorset (Kinngait), and were hand-labeled by trained phoneticians for F1 and F2 values. 239 vowel tokens were measured in this way. Note that although Experiment 1 used reconstructed data to provide larger, more well-behaved training sets, no such technique was used in Experiment 2. This was because resampling the data required making an assumption as to the actual modes or categories contained in the data. Raw data was used to ensure that the result was not biased towards any certain number of categories. An additional benefit of using raw data is that we make no assumptions about the mixing probabilities of the categories: the relative frequency of the different phones in the Inuktitut data set roughly mirrors the frequency of the phones in the language more generally. The Inuktitut vowel data is presented visually in Figure 3, with Gaussian estimates of relevant categories present in the data.

Because of the relatively small number of tokens relative to the Turkish data, and the fact that there was only one speaker, a slightly different analysis method was employed. Recall our question of interest--in both the Turkish and Inuktitut cases--is whether the number and identity of components in the best fitting MOG correspond to the adult phonology. To answer this for Inuktitut, we assessed the number of components in a MOG approach that best fit the Inuktitut data using a bootstrap analysis for number of categories was performed (McLachlan, 1989; see Appendix D). This analysis proceeds by assuming the null hypothesis that the data is best fit by K categories, starting at $K = 1$. We can estimate whether or not adding a new component into the mixture significantly improves the fit by first calculating the observed log-likelihood ratio between K- and $K+1$-component models on the original data. The observed log-likelihood ratio is then compared with a bootstrap estimate of the null distribution of that ratio to see if there is a

significant increase in log-likelihood with the addition of an extra component. If there is, then the procedure is repeated with K+1 categories. The K for which no significant increase in log likelihood is gained by adding a new component is the minimum number of components that best fits the data set. This procedure was done over the raw data with no resampling so as not to bias the solution towards the number of Gaussians that were sampled. As before, simulations were performed by fitting mixtures of Gaussians using the expectation maximization algorithm (sees Appendices A and B).

The results of the simulations are clear. The Inuktitut data is reliably fit by 4 components, as summarized in Table 4. For each component, the gain in the log-likelihood statistic that occurred when an additional component was included was assessed relative to the null-distribution for a model of similar size and complexity (see appendix D). Since there is no significant increase in this statistic when a four-category model is given an extra component to model the data, we can conclude that this data reliably contains four categories. A four-category solution implies that the model converges on neither phonetic nor phonemic categories that are present in the data.

Once the number of best-fitting categories was determined, a mixture model with that number of components was fit to the data. The fitting procedure was constrained to only consider models with unconstrained, full covariance structures (for a discussion on model-based fitting of mixture models, see Fraley & Raftery, 2002). The fit of the model against the underlying allophonic categories is shown in Figure 4. It can be seen that there is tight registration between two estimated categories and the surface phones comprising /i/ (i.e. [i], [e]), but the two acoustic phones are collapsed into one component for other phonemes (/u/, /a/). Thus, the model does not uniformly recognize phonetic or phonemic categories.

# of components	p
1	.00
2	.00
3	.04
4	.30

Table 4: p-values for the addition of a new vowel component on the raw Inuktitut data, given the existing number of components, as determined by bootstrap of the log-likelihood statistic (McLachlan, 1989; see Appendix D).

This straightforward result demonstrates that for the Inuktitut data considered, the MOG approach reliably overestimates the categories present in the input by attributing extra components to allophonic variants of some vowel phonemes. The spurious categories that were estimated aligned very closely with the allophones of the phonemic categories. In other cases, it did not appear to be sensitive enough to reliably distinguish the phonetic variants of a single phoneme. Instead, in these cases, the allophonic variants

were collapsed into a single component. The MOG approach--the same applied to Turkish above--now appears to be too sensitive to be an adequate model of the phoneme-learning process. Allophonic variation, which is extremely common across languages, warps the distribution of the input data in a way that obscures the underlying phonemic category structure. In the case of the Inuktitut data under investigation here, this meant that our model was unable to capture either the phonemic or phonetic categorization of the space.

The positing of spurious phonetic categories is exactly what is predicted under the two-stage view: each important phonetic variant *should* be separately clustered, as only then can they serve as effective input to the second stage of phoneme recognition. Note, however, the phonetic categories estimated with the current Inuktitut data would not support a second-stage phoneme learning procedure, because it does not converge on the correct number of phonetic categories. For both the Turkish and the Inuktitut problems, a failure to correctly identify the first-stage phones would foil any attempt to build phoneme-level categories from phone-level information.

The simulations from Turkish and Inuktitut above serve to effectively bound the learning problem. Our current approach is not able to recover the category information for Turkish; this demands more sensitivity to category structure. However, the same approach is already too sensitive to the category structure of the vowel space: the Inuktitut simulation shows that it overestimates the number of phonemes in the language. If we increase our sensitivity to category structure to capture the Turkish space, we will be guaranteed to converge on phonetic categories for Inuktitut. If one is dedicated to maintaining a general, MOG approach to the acquisition of sound categories, then there are a number of possible answers to this situation. One approach is to reject phonemes as an object of acquisition, which we do not consider here (see discussion in detail below). The other approach is to invoke a second stage of phonological acquisition, calling upon extra mechanisms to account for the distribution of phones and to detect the underlying phonemes. If the first-stage is as noisy as our simulations suggest, the second stage could well require mechanisms to group allophones into phonemes and also to split apart incorrectly agglomerated categories. In what follows, we show that such extra mechanisms are not necessary, and that an alternative, single-stage conception of the phonological acquisition problem provides empirical and conceptual advantages to models of acquisition.

The single-stage approach

When jointly considered, the Turkish and the Inuktitut simulations present a particular challenge to the recovery of phonemes using general statistical approaches. On the one hand, the present approach under-estimates the relevant Turkish vowel phonemes. The very same approach, however, over-estimates the number of vowel categories in a language like Inuktitut. If one attempts to address one or the other problem within the same general machine-learning framework, it will exacerbate the problem for the other. That the Turkish problem could eventually be solved by more clever machine learning methods is entirely reasonable: Feldman et al (2009), for example, present an alternative vowel category acquisition technique based in a hierarchical Bayesian framework that outperforms the generic MOG approach. But with increased sensitivity

comes more non-phoneme categories, compounding the phoneme-learning problem that is already evident for a language like Inuktitut. If the number of distinct categories grows with increasing power of the learning algorithm, then we simply push the phoneme-learning problem into the second stage.

Given the trading relation between the two levels of representation in the two-stage problem, a re-conception of the phonological learning problem is desirable. In what follows we sketch the outline of one way to do this, by suggesting a single-stage approach. We then present a simple proof of concept that this approach not only works, but in fact fares much better on the Inuktitut data than does the phone-clustering attempt presented above.

Factoring out processes

In order to map from acoustics to phonemes in a single acquisition stage, the primary difficulty that the learner faces are the multiple processes that obscure that relation; these processes range from co-articulatory dependencies between segments to abstract phonological mappings. If the learner can estimate these effects while learning the categories, and factor them out, then she will be left with the irreducible clusters from which the variety of surface forms are generated: the phonemes. At a more abstract level, the general MOG approach we have been considering maintains an independence assumption: each vowel token is sampled from its category without regard to the environment. The single-stage presented below jettisons this assumption by explicitly modeling the effect of context on vowel quality.

In order to accomplish this, we can recast the learner's 'phonology' in a formal statistical model. In the general MOG approach, the learner's task is simply to estimate the parameters of the components in the mixture. In order to account for context effects, we may instead model the learner is trying to estimate a regression model of the following form:

$$\vec{F}_{observed} = \vec{F}_k + \beta_{1k} I_1 + \beta_{2k} I_2 + ... + \beta_{nk} I_n$$

This model simply states that any observed acoustics for a given vowel phoneme *k*-F_{observed}-is a linear combination of an 'underlying' set of formant values F_k and *n* conditioning environments. Each conditioning environment serves to offset the original underlying value of the formants by some amount B. In equation (1) above, *I* represents an indicator variable, which serves to weight the contribution of the relevant conditioning environment. This formalization is a simple but powerful foundation to give a learner, and we will see that this simple model gives us a framework within which a learner can learn vowel phonemes directly from the acoustics.

In what follows, we present a proof of principle that having knowledge of the regression coefficients makes learning more reliable and robust, and allows the learner to converge on the underlying phonemic categories. At this point, it is worth introducing some modeling assumptions we make in order to make the argument. As with any modeling endeavor, we make a number of simplifying assumptions to present the results in a succinct way. Below we describe in some detail these assumptions, but we note here at the outset that almost none are crucial to the success of the model, and in fact, many

will need to be dismantled before a fully specified learning model for this type of phonological acquisition can be presented.

The model specified above is essentially a mixture of regressions: each vowel category has its own intercept (F_k), and its own regression coefficients. In the model we present here, the learner is constrained such that each category must share the same regression coefficients with all the others. In other words, it assumes phonological processes affect all phonemes similarly, as happens to be the case in Inuktitut (and which may be the case for certain classes of co-articulation effects). This model will severely underperform when the vast range of allophonic vowel variations are considered: there are numerous examples of phonological processes that target specific phonemes or phoneme classes. This assumption restricts the possible solutions that the model can consider, and relaxing this assumption will only increase the power of this learner. This assumption conveys benefits in model estimation: for simplicity, we can estimate the regression coefficients using simple techniques given that we know the structure of the correct answer (as we discuss below), and with this estimate, we can ask the question of interest: do the underlying acoustic values (the intercepts in the regression) correspond to phonemic categories? Note that the way in which we estimate the regression coefficients, however, is *not* intended as a model of the coefficient estimation process employed by the learner. To provide this, we must extend this model to include joint estimation of coefficients (allophonic shifts) and intercepts (phonemic cluster means). There exist a number of possible implementations that allow for this sort of estimation procedure (more sophisticated versions of EM and hierarchal Bayesian methods, among others).

A second assumption here is that learners can uniquely identify the conditioning environments, and can determine which ones will be worth modeling. This assumption, too, can be relaxed or dropped with no loss of generalization. This can, in principle, be folded into the estimation problem along with the intercepts and coefficients using the very same estimation procedures. Furthermore, the environments need not be discrete; the structure of the model allows for continuous regressors. For ease of presentation, we assume the regressor environments to be discrete and fixed, but this is not intended as a psychological claim about the learner's capacities. Our aim here is much more modest: we simply aim to ask if the residue from this process corresponds to the phonemic categories of interest in the language, and if this category structure can be reliably estimated. This assumption also raises the empirical question of which categories the infant learner can use to build such a model. If the distribution and identity are estimated in tandem with everything else, then one predicts that the notion of possible phonological rule is quite liberal. An alternative possibility is that infants use pre-existing perceptual biases to form these conditioning environments, perhaps drawing on the fact that infants can distinguish certain consonant classes more readily than others (Narayan, Werker & Beddor, in press). Either option is entirely compatible with the model, although they make different predictions about the generalizations a learner might make.

In what follows we show that by factoring out the predictable effect of a conditioning environment on vowel formant values, we obtain a more robust clustering of the data (i.e. one that corresponds better to the underlying categories). Furthermore, the components of such a model correspond to phonemic, rather than phonetic clusters. This recasting of the learning problem presents a proof of principle that this alternative

conception does convey advantages in acquisition, and opens new lines of inquiry in the modeling of phonological acquisition.

Inuktitut revisited

Recall that the MOG approach we applied to the raw Inuktitut data converged on a vowel space with four, rather than three categories. We noted that this solution is unsatisfactory for a number of reasons, even if one assumes a two-stage acquisition process: the Inuktitut dialect arguably has six, not four, phonetic categories that need to be learned (Denis & Pollard, 2008), and of the ones that were recovered, there was a rather poor fit to the underlying phonetic categories. Here we repeat that simulation with a slightly different data set to ask whether or not the correct set of phonemic categories can be recovered once the predictable processes were factored out.

In order to factor out the predictable vowel lowering process, we estimated the average format values for a vowel that directly preceded a uvular segment (/q/ or /r/; labeled here [+uvular] vowels), and the average formant values for a vowel that did not ([-uvular]). Importantly, all transforms were derived without knowledge of underlying category structure; they were estimated and applied uniformly across all vowel tokens. From these, we calculated the average spectral shift in formant space that occurred as a result of being in the context of a uvular. Once this was obtained, a *process-transformed* data set was derived by subtracting the average spectral shift from every vowel token that occurred in the environment of a uvular. Thus, the subset of vowel points that occurred in the environment of a uvular were shifted by a constant amount, such that the mean of the distribution of [-uvular] and [+uvular] vowels were the same. The spectral shift lowered the first formant of [+uvular] vowels by 92.27 Hz, and raised their second formant by 374.00 Hz. This process is outlined visually in Figures 5 and 6, which plot the estimated Gaussian distributions for [+uvular] and [-uvular] vowels. Prior to the spectral transformation (Figure 5), there is an appreciable distance between the means of these distributions, and after the transformation (Figure 6), they differ only in variance.

[INSERT FIGURE 5 HERE]

[INSERT FIGURE 6 HERE]

Once the process-transformed data set was created, model fitting and evaluation proceeded in the same fashion as before: the best-fitting number of categories was determined by bootstrap test, and the model-fitting procedure estimated an unconstrained mixture model with that number of categories for the data. The results of the bootstrap test for number of components are summarized in Table 5, and the resulting categories are plotted against the underlying vowel distributions in Figure 8. The process-transformed data is best fit by three categories, and there is a clear registration between these categories and the underlying phonemic categories.

# of components	*p*
1	.00
2	.01
3	.49

Table 4: *p*-values for the addition of a new vowel component on the process-transformed Inuktitut data, given the existing number of components, as determined by bootstrap of the log-likelihood statistic (McLachlan, 1989; see Appendix D).

[INSERT FIGURE 7 HERE]

[INSERT FIGURE 8 HERE]

In this example, the MOG approach was able to successfully converge on phonemic categorization of the Inuktitut data by applying a maximally simple spectral correction derived from a consideration of the average shift in formant values in the presence of a uvular environment. This stands in contrast to the results obtained on the raw data with exactly the same approach, which obtained a poor fit to either phonetic or phonemic categories.

Discussion

In this article we argued that there is a widespread, but often implicit, consensus that phonological category learning is essentially a two-stage process. By running simulations on Turkish and Inuktitut vowel data, we determined that this position is practically obligatory given the general MOG approach: the current methods are alternately too sensitive to category structure (Inuktitut), and not sensitive enough (Turkish). Though we did not pursue a solution to the latter problem in the current work, any such solution will require more powerful category learning mechanisms, perhaps in the vein of research currently underway (Coen, 2006; Feldman et al, 2009). However, by increasing the sensitivity to category structure, such models will inevitably converge on phonetic-level categories. Given the general MOG approach, this aggravates the phoneme-learning problem, and requires a 'second-stage' that will generalize these phones into phoneme-level categories. Viewed as a clustering problem (Everitt et al 2001), this is essentially divisive followed by agglomerative clustering procedures.

However, we argued that the problem of spurious phonetic clusters did not, in fact, require this second-stage of learning. We presented an alternative model of the phonological acquisition process that learned phoneme categories in a single stage. If the learner factors out predictable variation in acoustic space during the course of the learning process, she will converge on phonemic categories as the underlying components of the vowel space. By simulating the acquisition process using a slightly modified MOG approach, we found that the single-stage approach more reliably converged on the language specific phoneme categories. Specifically, we found that

when the raw acoustic Inuktitut data was considered, a poor fit to either phonological or phonetic categories was achieved. A simple correction for a regular vowel-lowering process conditioned on uvular segments, however, helped the model to converge on the phonemic solution in one clustering step. There remains much work to be done to further evaluate the viability of such a model of learning. In particular, we have not presented here a model that fully estimates all regression coefficients and relevant environments in tandem with the phonological categories. Instead, we simply demonstrated that once predictable processes were removed, the vowel space is cleanly clustered into phonemic categories.

The very general nature of the Inuktitut lowering rule made this demonstration relatively straightforward, and future work will be dedicated to extending this modeling to more complex systems. For example, consider a closely related language, West Greenlandic. West Greenlandic has a lowering process that is very similar to the Inuktitut process presented above. However, the West Greenlandic lowering process only applies to the high vowels /i/ and /u/, which are pronounced (and written) as [e] and [o] when they precede a uvular. In addition, there is a vowel fronting process that fronts /u/ in the context of a coronal consonant (/t/, /s/, or /n/) (Saddock, 2003). Thus, West Greenlandic has at least 6 distinct surface phones that are generated from three underlying phonemes, and two regular allophonic processes. It is not obvious that these effects can be estimated just considering the general effect of uvular or coronal consonants across the entire vowel space, as was possible in the Inuktitut dialect considered here. Instead, such a system would require the estimation of phonological processes to be truly embedded in the category learning procedure. As mentioned above, there are a number of possible computational solutions to this problem, including more sophisticated versions of EM and hierarchical Bayesian modeling. In order to acquire the more difficult West Greenlandic vowel space, the single-stage model must be augmented with a more sophisticated estimation procedure that allows category-restricted transformations.

Of course, an alternative response to the finding that MOG approaches are guaranteed to return phonetic level categories is that phonetic levels categories are the only discrete representations in the linguistic system, and that all linguistic encoding is done in terms of phonetic categories. This view rejects the existence of phonemes as phonologically active representations, a view that has been espoused by a number of authors (Johnson, 1997; Port & Leary, 2005; Silverman, 2006). There are, however, compelling theoretical and empirical reasons for rejecting this view. One type of empirical evidence for this view is that alternations of the sort considered here are psychologically active in virtually every language. Recall that in Inuktitut, /i/ is approximately pronounced as [e] whenever it appears before /q/. This alternation is active in the sense that speakers will reliably pronounce /i/ sounds as [e] even in words they have never encountered before. This is quite easy to see in Inuktitut, because of its extremely productive morphology. Consider the past-tense suffix /qqau/, which begins with a uvular segment /q/. Any verb stem this attaches to will be subject to the phonological processes associated with the uvular; thus, the verb /kii/ “bite” suffixed with /qqau/ gets pronounced as [keeqqau] and not *[kiiqqau]. Like other phonological processes, this reliably occurs with new, unseen words (a fact made famous by Jean Berko-Gleason’s “wug” test, Berko-Gleason (Berko, 1958)). In order to become a competent speaker of Inuktitut, the learner must master this active alternation. If one

chooses to deny the phoneme as an account of these alternations, then one is faced with the theoretical problem of devising an alternative account for this set of facts.

In addition to these observations, there is experimental evidence from infant and adult speech perception that suggests that phoneme-level distinctions, rather than phonetic level distinctions, are implicated in common measures of discrimination (Whalen, Best & Irwin, 1997; Peperkamp, Pettinato & Dupoux, 2003; Kazanina, Idsardi, & Phillips, 2006; White, Griffiths & Morgan, 2008). For example, Kazanina et al (2006) used magnetoencephalography to show that one neural signature of sound discrimination (the mismatch field, MMF) to a [t]-[d] distinction only obtained for speakers for whom it was a phonemic distinction (Russian speakers). Korean speakers, whose language also has these two phones but only as allophonic variants of the same phoneme, showed no such discrimination. White and colleagues presented similar results by studying infants using the head-turn preference task. They showed that infants trained on an artificial language with no associated meaning were able to generalize across regular allophonic variation to extract phonemes. At test, infants treated strings of sounds that contained the same sequence of phonemes as one word, regardless of their phonetic content. These results are also important because the infants did not require meaning to detect the allophonic alternation. These results are entirely compatible with the model presented here, but run against the predictions of models that rely on similarity in meaning to explain allophonic variation (Silverman 2006). Thus there is convergent evidence from linguistics, speech perception and acquisition research that points to a level of sound categorization above the level of the phone.

Although we presented empirical results that suggested that the single-stage approach would fare better than the two-stage approach on the Inuktitut data, there are also conceptual advantages to a single-stage approach. Even if one assumes a two-stage process for learning categories, the phonological acquisition process is still not complete. An algorithm that clusters phones into phonemes based on distributional facts (as in Peperkamp et al, 2007), gives the learner only limited insight into the processes that generated those allophonic distributions. Yet another stage of learning must be invoked to learn the rules (or constraint rankings) that generated the observed patterns. In exploiting the processes in the category acquisition stage, however, the single-stage approach returns a much more deployable set of phonological knowledge: a set of phonemes, and the processes that relate them to their allophones. In the case of Inuktitut, the learner converged on three phoneme categories, and a process that predictably shifted the target pronunciation in front of uvular segments. This subsymbolic shift in acoustic space represents the knowledge that might be described by a phonological rule. Together with the phoneme categories, this knowledge gives the language user all the knowledge necessary to produce an appropriate vowel token given a phonological environment.

The single-stage model also makes a claim about possible levels of representation in phonology. Just as the two-stage model learned two distinct, discrete representations, the single-stage model learns only one. As discussed in the introduction, the assumption of distinct phonetic and phonemic levels of representation has been the hallmark of generative approaches to phonology for more than half a century. The single-stage model rejects that view in favor of a system in which the phoneme is the only discrete level of categorization in the system. The surface pronunciations are only that: acoustic or articulatory targets with no discrete or categorical status. All variations in the

pronunciation of a phoneme category--from abstract allophonic processes to hard co-articulatory facts--are described in a sub-symbolic manner as predictable perturbations in the underlying phoneme's acoustics. This predicts that purely allophonic variants of a given phoneme--ones that do not have phonemic status in their own right elsewhere in the language--should simply fail to show effects of categorical status. This generates predictions for phonological processing and phonological typology. For processing tasks that are sensitive to categorization, a single-stage approach predicts that the allophonic variants should not necessarily convey the same perceptual advantage that fully phonemic contrasts do (a prediction borne out by Kazanina et al, 2006). If this representation is coupled with the assumption of categorical conditioning environments, then this model predicts the impossibility of phonological processes that are conditioned on allophonic variants alone. So-called *feeding* orders of allophonic phonological processes, where one allophonic change creates the environment for another, are not predicted to be possible in the sub-symbolic domain. Instead, the single-stage much must instead rely on phoneme-to-phoneme changing processes (i.e. changing the categorical label; this corresponds approximately to morpho-phonemic processes in Harris 1951 or to structure-preserving lexical processes in Kiparsky 1982) to implement such an order. As such, this formulation of the model predicts the impossibility of purely allophonic feeding processes, where an allophone (with no categorical status, by hypothesis) triggers another phonological process.

The single-stage model presented here also provides a novel way of approaching the problem posed by *opacity* in phonological acquisition (so-called *counterfeeding* and *counterbleeding* orders). Opacity results when a generalization is not "surface true" (Kiparsky, 1973; Bakovic, 2007). This occurs when one phonological process destroys the environment for a second process, but that second process applies anyway. Consider a well-known case from Canadian English. Canadian English famously has a process that raises the diphthongs /ai/ and /au/ before voiceless consonants, causing "write" /rait/ to be pronounced as [rəit]. Like many other North American Englishes, Canadian English also has a rule that causes /t/ and /d/ to be 'flapped' between vowels, causing the final consonant in "beat" /bit/ to be pronounced as a flap in "beating" [biɾɪŋ]. Importantly, in an environment where both of these processes could apply, they do: "writing" is pronounced as [rəiɾɪŋ]. This interaction of processes is said to be opaque because the generalization that the diphthong /ai/ is raised before voiceless segments is not true in this case. This dialect produces the raised variant of /ai/ before a segment that is voiced on the surface. In this case, it is the 'voicelessness' at the level of the underlying phoneme that appears to condition the application of the rule. Opacity is readily accounted for in serial rule-based models of phonology, as in Chomsky & Halle (1968). The rule that raises the vowel simply applies before the rule that turns the /t/ into a flap (which could be modeled as a transform effecting an extreme reduction in the closure duration). However, it is somewhat more difficult to state these generalizations in surface-based theories like Optimality Theory (Mielke, Armstrong & Hume, 2003; Bakovic 2007). From the point of view of the model presented here, however, this interaction is straightforwardly captured without any explicit notion of rule ordering (and bears a striking resemblance to the account given by Harris 1951: 71). If one assumes that only categorical information can trigger (serve as the conditioning environment for) phonological processes, then the surface pronunciation of /t/ as a flap by reduction of its closure duration is expected to

have no effect on the effects it has on neighboring segments. As the categorical identity of the flap remains /t/ regardless of surface pronunciation, the opaque interaction results without rule ordering. The two transforms (vowel raising and closure shortening) can apply independently and simultaneously. This is a surprising result that stems directly from limiting the possible levels of discrete representation in phonology. It remains to be seen if limiting phonological theorizing to a single level of discrete representation will produce theories that have the same empirical purchase as current theories do. Our main concern here has been instead to argue that this restriction conveys advantages in modeling the acquisition process.

Conclusion

In this article, we examined the general MOG approach to phoneme category acquisition. By considering data from Turkish and Inuktitut, we bounded the learning capacity of the MOG approach from above and below: in order to achieve the sensitivity to extract phonemes from some languages (Turkish), we are guaranteed to converge only on phonetic categories in others (Inuktitut). This fact apparently demands a two-stage solution to phonemic category acquisition: first, infants must acquire their language-specific phonetic categories, and only then can they determine the phoneme categories.

We argued, however, that the phoneme acquisition process can instead be conceived of as a single-stage process if learners actively factor out predictable variations in the acoustics. By comparing single- and two-stage approaches to a sample of Inuktitut data, we showed that the single-stage reliably converged on phoneme categories in one clustering step, while the two-stage approach struggled to extract either phonetic or phonemic categories on the same data. This simple proof of concept demonstrates the utility of this alternative conception of the phonological acquisition problem, which we also argued has conceptual and theoretical advantages over a two-stage approach.

Acknowledgments

This work was supported in part by NSF IGERT DGE-0801465 to the University of Maryland and NIH 7R01DC005660-07 to David Poeppel and William Idsardi. We would like to extend special thanks to Derek Denis and Mark Pollard for sharing their Inuktitut recordings. We are grateful to Jordan Boyd-Graber, Naomi Feldman, Jeff Lidz, Jeff Mielke, and Colin Phillips for their useful discussion and insight on the issues contained in this paper. All remaining errors are our own.

Appendix A: Mixture models

A mixture model assigns a probability to an observation x_i from a set of n d-dimensional observations $\{x_1, x_2,\ldots, x_n\}$ through a combination of constituent probability distributions, as in (4) (for reference, see McLachlan & Peel, 2000 and Frühwirth-Schnatter, 2006).

$$f(x_i \mid \Psi) = \sum_{k=1}^{K} \pi_k f_k(x_i \mid \theta_k)$$
$$\Psi = (\pi^T, \theta^T)^T$$

A mixture model Ψ containing K component distributions is parameterized by a vector of component probability (π) and component parameters (θ; for normals, this consists of a mean μ and covariance Σ). This model acts to assign a probability to any observation x_i by summing the probability of that observation conditioned on each of the K components (i.e. $p(x_i|\theta_k)$ for each component k), weighted by that component's mixing probability π_k. In order to fit a mixture model, the parameter vector Ψ (containing mixing probabilities and component parameters) must be estimated.

Appendix B: Expectation-Maximization for mixture models

There is no closed form expression for the maximum-likelihood estimator (MLE) for a mixture of normal distributions, and so estimation of mixtures often proceeds by employing iterative computational methods (McLachlan & Peel, 2000). One popular option is the expectation-maximization algorithm (Dempster et al, 1977).

In estimating a mixture model with the EM algorithm, the data is viewed as incomplete: given our observations $\{x_1, x_2, \ldots, x_n\}$ there are associated, unobservable component labels $\{z_1, z_2, \ldots, z_n\}$. The EM algorithm treats these labels as missing data, and proceeds in two steps: the expectation (E) step, which computes the conditional expectation of the labels and the maximization (M) step, which employs those expectations to find new model parameters that maximize the likelihood of the data.

The algorithm is initialized by selecting random, valid starting values for all parameters (where valid means that probabilities must sum to 1 where appropriate). The E-step evaluates the 'responsibility' that each component has for each observed value, where the responsibility the probability of a given label for that observation.

$$\gamma(z_{nk}) = p(z_{nk} = 1 \mid x_n) = \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)}$$

Thus the E-step results in a matrix Z, where each entry Z_{nk} corresponds to $\gamma(z_{nk})$: the conditional expectation (given the current estimated model parameters) of the missing data label k given a data point x_n. Using these expectations (or 'responsibilities'), the M-step re-estimates the parameters in Ψ using the following ML estimators:

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

$$\hat{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)^T$$

$$\hat{N}_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\hat{\pi}_k = \frac{\hat{N}_k}{N}$$

Where N_k refers to the expected number of observations in each category, and N is the total number of observations. This procedure is iterated until convergence, and is guaranteed to increase the log-likelihood of the model with each iteration (see proof in Dempster et al, 1977).

Appendix C: Kullback-Leibler distance

Kullback-Leibler (KL) distance (also known as *KL divergence* or *relative entropy*) is an information-theoretic measure that is often used to quantify the distance or similarity between two probability densities. The general form of the KL distance between two densities *f* and *g* is:

$$D_{KL}(f \parallel g) = \int f(x)\log\frac{f(x)}{g(x)}dx$$

For two Gaussian distributions N_0 and N_1, this can be expressed in closed form as the following (Hershey & Olson, 2007):

$$D_{KL}(N_0 \parallel N_1) = \frac{1}{2}(\log_e(\frac{\det\Sigma_1}{\det\Sigma_0}) + tr(\Sigma_1^{-1}\Sigma_0) - d + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0))$$

Where *d* is the dimensionality of the Gaussians. The KL distances reported in this paper were computed with the closed form expression, which uses a logarithm with base *e*. Thus all reported KL distances are in nats, rather than bits (another common unit for KL distances).

Note that the KL distance is not a true metric because it is not symmetrical: $D_{KL}(f \| g)$ is not generally equal to $D_{KL}(g \| f)$. In practice, one distribution is designed as 'true' and the other distribution the 'target'. We calculated all KL distances by setting the truth to be the parameters of the underlying vowel category (N_0), and the target distribution (N_1) was specified by the parameters of the estimated category.

Appendix D: Bootstrap test for number of model components

Rather than allow EM to automatically determine the number of components *K* for a model, we employed bootstrap methods to estimate level of significance for each *K*

given the data $X = \{x_1, x_2, \ldots, x_n\}$. In order to do so, we generated bootstrap distributions of -2 log λ, following McLachlan (1989). All tests take the form of a simple hypothesis test: the null hypothesis H_o is that the data is best fit by a K-component mode, and the alternative hypothesis H_1 is that the model is instead best fit by a K+1-component model. To proceed, we first fit a K-component model Ψ_B to the original data X. From this estimated model (with K components), a sample X' of n observations is produced. The value of -2 log λ is then determined by fitting both a K component model and a K+1-component model to X', and measuring the difference in the log-likelihoods of the models This process is iterated independently B times, and doing so generates a boostrap approximation to the null distribution of -2 log λ. Once we obtain the null distribution of our test statistic, we may then evaluate the significance of the actual observed value of -2 log λ on the data.

To do this, we compare a K+1-component model to a K-component model fit to the original data X. The difference in the likelihood scores is used to estimate the test statistic -2 log λ. From this, we obtain an estimate of the level of significance p of the observed increase in model likelihood (i.e. the probability that the null hypothesis is correct}. In other words, if the resultant p-value reaches a preset α, we determine that the data is significantly better fit by K+1 components. This procedure is iterated with increasingly larger K until a K is found for which the null hypothesis H_o cannot be rejected. This value of K is the lowest number of components for which we cannot reject H_o; in other words, it is the number of components that best fits the data.

References

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: constraints on and precursors to language. In *Handbook of child psychology: Volume 2: Cognition, perception, and language* (Damon William, ed.), pp. 147-198. New York, NY: Wiley.

Bakovic, E. (2007). A revised typology of opaque generalizations. *Phonology*. 24, 217-259.

Berko, J. (1958). The child's learning of English morphology. *Word*. 14, 150-177.

Best, C.T. (1995). Learning to perceive the sound patterns of English. In *Advances in Infancy Research*, (C Rovee-Collier, LP Lipsitt, eds), pp. 217–304. Norwood, NJ: Ablex.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Coen, M. (2006). Self-supervised acquisition of vowels in American English. In *Proceedings of the TwentyFirst National Conference on Artificial Intelligence* (AAAI'06).

de Boer, B., & Kuhl, P. K. (2001). Infant-directed vowels are easier to learn for a computer model. *Journal of the Acoustical Society of America*. 110 (5), 2703.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 39(1), 1-38.

Denis, D. & Pollard, M. (2008). A phonetic analysis of the Inuktitut vowel space. Inuktitut Linguistics Workshop. University of Toronto, Toronto, Ontario.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceeding of the National Academy of Sciences* 104, 454–464.

Dorais, L.-J. (1986). Inuktitut surface phonology: A trans-dialectal survey. *International Journal of American Linguistics.* 52(1), 20–53.

Dresher, B. E. (2009). *The contrastive hierarchy in phonology*. Cambridge: Cambridge University Press.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis*. 4th ed., London: Arnold.

Fant, C. G. M. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Fraley, C. & A. E. Raftery. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*. 97, 611-631.

Fraley, C. & A. E. Raftery. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. *Technical Report No. 504, Department of Statistics, University of Washington*. http://CRAN.R-project.org/package=mclust

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York, NY: Springer.

Goldsmith, J. A. (1976). *Autosegmental phonology*. Bloomington, IN: Indiana Linguistics Club.

Goldsmith, J. & Xanthos (2009). Learning phonological categories. *Language.* 85, 4-38.

Harris, J. (1969) Spanish Phonology. MIT Press.

Harris, Z. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hayes, B. (2009). Malden, MA: Wiley-Blackwell.

Hockett, C. (1942). A system of descriptive phonology. *Language* 18, 3-21.

Hershey, J. & Olsen, P. (2007). Approximating the Kullback-Leibler divergence between Gaussian mixture models. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii.

Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *Ohio State Working Papers in Linguistics*. 50, 101-113.

Jusczyk, P. W. (1985). On characterizing the development of speech perception. In J. Mehler & R. Fox, *Neonate cognition: beyond the blooming, buzzing, confusion*, pp. 199–229. Hillsdale, NJ: Erlbaum.

Kiparsky, Paul. (1973). Abstractness, opacity, and global rules. In *Three Dimensions in Linguistic Theory*, ed. by Osamu Fujimura, 57–86. Tokyo: TEC.

Kiparsky, Paul. (1982). Lexical Phonology and Morphology. In In-Seok Yang (ed.), *Linguistics in the Morning Calm*. Seoul.

Kohler, K. Erwerb der frühen Verbmorphologie im Estnischen. PhD thesis, University of Potsdam, 2004.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. *Science.* 255, 606–608.

Ladefoged, P. (2001). *Vowels and consonants: An introduction to the sounds of languages*. Oxford: Blackwell.

McLachlan, G. J. (1989). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics.* 36 (3), 318–324.

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. New York, NY: Wiley.

MacWhinney, B. (2000). The CHILDES project: tools for analyzing talk, third ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition.* 82, B101–B111.

Mielke, J., Armstrong, M., & Hume, E. (2003) Looking through opacity. *Theoretical Linguistics.* 29.1-2, 123-139.

Mines, MA, Hanson, B, & Shoup, J. (1978). Frequency of Occurence of Phonemes in Conversational English. *Language and Speech.* 21 (3), 221-241.

Miyata, S. (1992). Wh-questions of the third kind: The strange use of wa-questions in Japanese children. *Bulletin of Aichi Shukutoku Junior College*. 31, 151–155.

Miyata, S. (1995). The Aki corpus longitudinal speech data of a Japanese boy aged 1.6-2.12. *Bulletin of Aichi Shukutoku Junior College*. 34, 183-191.

Miyata, S. (2000). The TAI corpus: Longitudinal speech data of a Japanese boy aged 1;5.20 - 3;1. *Bulletin of Shukutoku Junior College.* 39, 77–85.

Narayan, C., Werker, J., & Beddor, P. (in press). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*.

Noji, J. (1973-1977). Yooji no gengo seikatsu no jittai i -iv. *Bunka Hyoron Shuppan*.

Ohala, J. (1976). A model of speech aerodynamics. *Report of the phonology laboratory, Berkeley.* 1, 93-107.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceedings of the 27th Annual Boston University Conference on Language Development*. *Volume 2* (pp. 650- 661). Somerville, Mass.: Cascadilla Press.

Peperkamp, S., Le Calvez, R., Nadal, J.-P. & Dupoux, E. (2006). The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* 101, B31-B41.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper, eds., *Introduction to frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, 46(2-3), 115-154.

Plunkett, K., & Strömqvist, S. (1992). The acquisition of Scandinavian languages. In D. Slobin, *The crosslinguistic study of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 457–556.

Port, R. and Leary, A. (2005). Against formal phonology. *Language.* 81, 927-964.

Prince, A. & Smolensky, P. (2004). Optimality Theory: Constraint interaction in generative grammar. Oxford: Basil Blackwell.

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sabourin, J., Werker, J.F., Bosch, L., & Sebastián-Gallés, N. (submitted). Perceiving vowels in a tight vowel space: Evidence from monolingual infants. *Developmental Science.*

Saddock, J. M. (2003). *A grammar of Kalaallisut (West Greenlandic).* Lincom Europa.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science.* 274, 1926–1928.

Salor, O., Ciloglu,T., and Demirekler, M. (2006). ODTU Türkçe mikrofon konusması veritabanı. In *Signal Processing and Communications Applications*, IEEE. Antalya, Turkey.

Silverman, D. 2006. *A critical introduction to phonology: Of sound, mind and body*. New York: Continuum.

Slobin, D. (1982). Universal and particular in the acquisition of language. In L. Gleitman & E. Wanner (Eds.), *Language acquisition: The state of the art*. Cambridge: Cambridge University Press, 128-170.

Strömqvist, S., Richthoff, U., and Andersson, A.-B. (1993). Strömqvist's and Richthoff's corpora: A guide to longitudinal data from four Swedish children. *Gothenburg Papers in Theoretical Linguistics* 66.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition* 103, 147–162.

Whalen, D., Best, C., & Irwin, J. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics*, *25*, 501-528.

Vallabha, G., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*.104 (33), 13273–13278.

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology.* 37, 278–286.

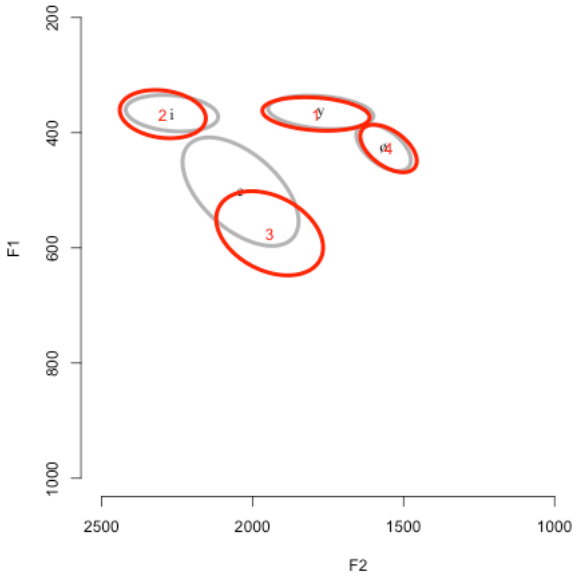

Figure 1: Underlying vowel categories (grey) and estimated mixture model components (red) for a uniform training set a representative Turkish speaker plotted in F2-F1 space. Ellipses delimit .20 of the Gaussian's probability mass. KL distances between mixture components and their corresponding vowel distributions are .14 (/i/-2), .03 (/y/-1), .39 (/e/-3), and .07 (/ø/-4).

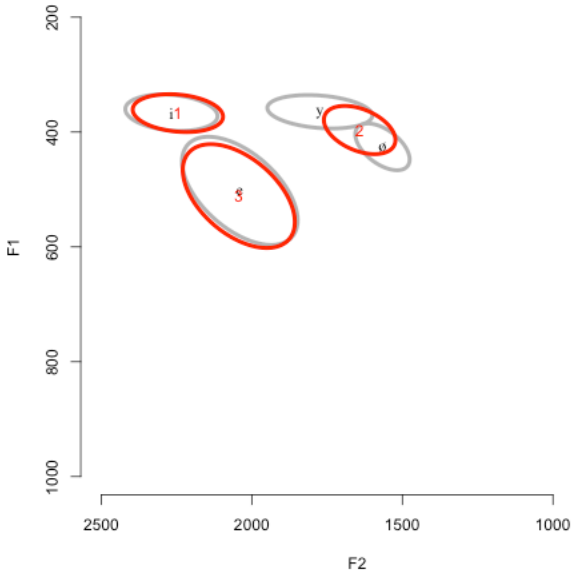

Figure 2: Underlying vowel categories (grey) and estimated mixture model components (red) for an empirical training set for a representative Turkish speaker plotted in F2-F1 space. Ellipses delimit .20 of the Gaussian's probability mass.

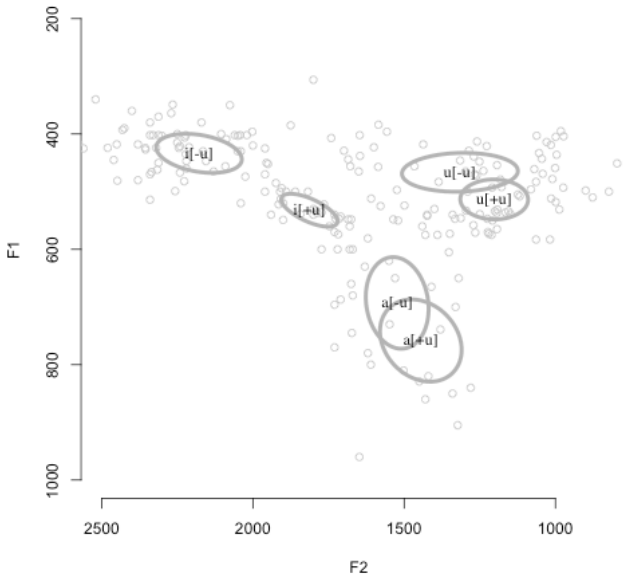

Figure 3: Plot of Inuktitut vowel data in F2-F1 space. Ellipses delimit .20 of the probability mass of a Gaussian estimate of the distribution of vowel productions in uvular environments ([+u]) and in non-uvular environments ([-u]).

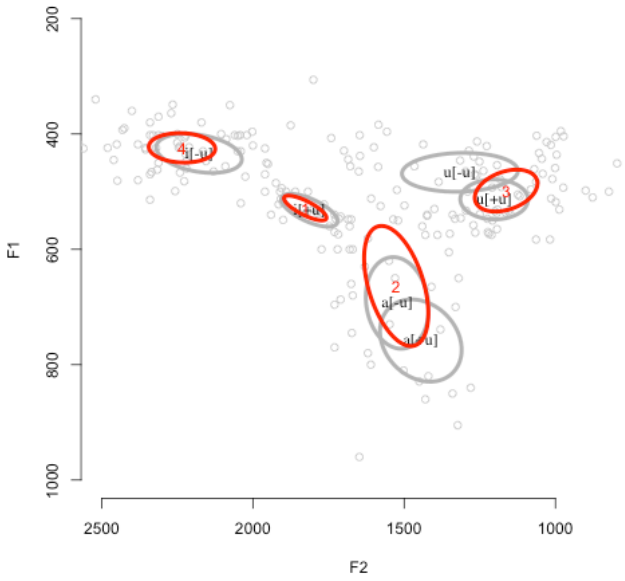

Figure 4: Plot of Inuktitut vowel distributions (grey) in F2-F1 space, along with best-fit four-category solution (components in red). Ellipses delimit .20 of the probability mass of the corresponding Gaussian.

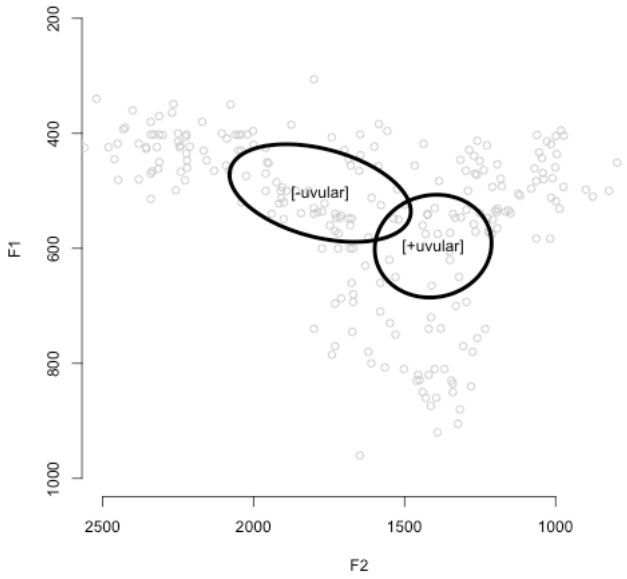

Figure 5: F2-F1 plot of distributions of Inuktitut vowels in uvular contexts ([+uvular]) and in non-uvular contexts ([-uvular]) prior to process-transformation. Ellipses delimit .20 of the probability mass of the corresponding Gaussian estimate.

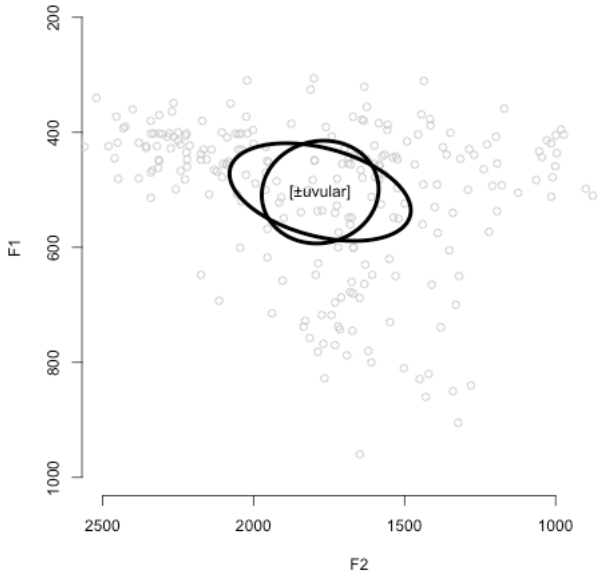

Figure 6: F2-F1 plot of distributions of Inuktitut vowels in uvular contexts ([+uvular]) and in non-uvular contexts ([-uvular]) after process-transformation. Ellipses delimit .20 of the probability mass of the corresponding Gaussian estimate.

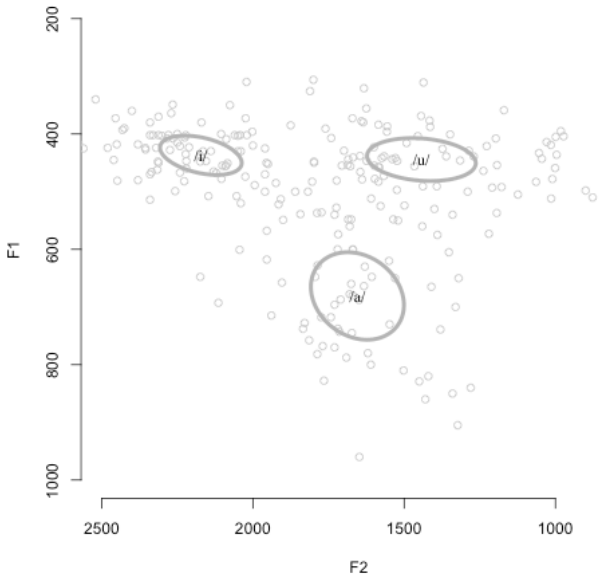

Figure 7: Plot of process-transformed Inuktitut vowel data in F2-F1 space. Ellipses delimit .20 of the probability mass of a Gaussian estimate of the distribution of vowel productions for each phoneme category.

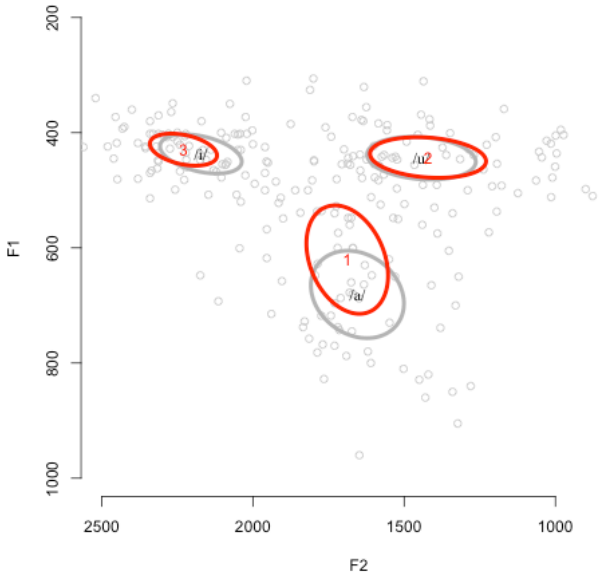

Figure 8: Plot of process-transformed Inuktitut vowel distributions (grey) in F2-F1 space, along with best-fit three-category solution (components in red). Ellipses delimit .20 of the probability mass of the corresponding Gaussian.