COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 344–377

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12008

A Single-Stage Approach to Learning Phonological Categories: Insights From Inuktitut

Brian Dillon,[a,*] Ewan Dunbar,[b,*] William Idsardi[c]

[a]*Department of Linguistics, University of Massachusetts*
[b]*Department of Linguistics, University of Maryland*
[c]*Department of Linguistics and Program in Neuroscience and Cognitive Science, University of Maryland*

Received 7 July 2010; received in revised form 31 May 2012; accepted 31 May 2012

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 has each in its 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, independently 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 simulations that suggest two problems with this view: First, the learner might mistake the phoneme-level categories for phonetic-level categories and thus be unable to learn the relationships between phonetic-level categories; on the other hand, the learner might construct inaccurate phonetic-level representations that prevent it from finding regular relations among them. We suggest an alternative conception of the phonological acquisition problem that sidesteps this apparent inevitability and acquires phonemic categories in a single stage. Using acoustic data from Inuktitut, we show that this model reliably converges on a set of phoneme-level categories and phonetic-level relations among subcategories, without making use of a lexicon.

Keywords: Phonology; Phonetic categorization; Phonological acquisition; Mixture models

1. Introduction

In recent years, statistical approaches to language acquisition have generated much enthusiasm, especially in the domain of phonological acquisition (Chambers, Onishi, & Fisher, 2003; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996). The problem of

*These authors contributed equally to this work.

Correspondence should be sent to Ewan Dunbar, Department of Linguistics, 1401 Marie Mount Hall, University of Maryland College Park, College Park, Maryland, 20742-7505, USA. E-mail: emd@umd.edu

how human children acquire phonological categories (phonemes) of spoken language presents an ideal model problem for these approaches to language acquisition: We understand a good deal about the time course of phonological development, how the perceptual input to learning is represented, and what the desired end stage of acquisition is. Furthermore, there are very good reasons to model the acquisition process using well-understood methods for statistical inference over perceptual input, as we describe below. However, the general approach to phonological category formation as perceptually driven statistical inference has led to the view that the categorization acquired by the learner is in some sense isomorphic to all and only the distinctions present in the acoustics.

This view leads to a model of acquisition that is incomplete from the point of view of contemporary models of phonological knowledge. This is because it is common for phonological theories to distinguish between phones and phonemes. Phonemes are language-specific, abstract categories used for the purposes of memory encoding in the lexicon. A single phoneme, however, may comprise a set of distinct pronunciations (or phones) that reflect its phonological environment. For example, English is commonly thought to have a single phoneme category /t/ that comprises a number of distinct acoustic realizations (its allophones). An English word like ''sit'' is thought to be stored in the lexicon using this abstract phonemic category as its final segment, typically written using slash notation as in /sɪt/. In pronunciation, the final /t/ phoneme is typically mapped to an unreleased [t˺] phone (where brackets denote phonetic categories) in word-final position. This acoustic realization can vary widely based on phonological context, however: The morphological process of adding -ing to the verb root produces a predictable change in the pronunciation of the root-final /t/, such that in ''sitting'' the /t/ is pronounced as a flap [ɾ], a phone that is phonetically more like Spanish *r* than it is like other pronunciations of /t/. This change is due to the phonological environment created by the addition of ''ing'': Here, the /t/ is flanked by a stressed vowel to its left, and an unstressed vowel to its right, a phonological context that triggers the flap pronunciation. The mapping from /t/ to its allophone in context is referred to as a phonological rule or process. Thus, the /t/ phoneme category in English comprises many distinct acoustic phones. The presence of allophonic alternations of this sort is ubiquitous in the world's languages (see, e.g., Kenstowicz, 1994), and the set of allophonic alternations a language may have is subject to wide cross-linguistic variation. As such, allophonic processes are an important consideration for models of phonological acquisition.

However, statistical models that cast phonological category learning as perceptual clustering imply that the goal of learning is to discover phonetic, rather than phonemic, categories. If the goal of phonological acquisition is to discover the categories used in lexical storage, then phonetic categories are not the desired end state of phonological acquisition. There are varied theoretical approaches to the problem of learning abstract phoneme categories, and the problem of how to learn abstract phonological systems has itself generated a sizeable body of research (Boersma & Hayes, 2001; Goldsmith & Xanthos, 2009; Harris, 1951; Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; Tesar & Smolensky, 1998; among others). However, modeling studies of this kind have typically assumed an input consisting of sequences of phonetic categories, and, in doing so, have tacitly assumed that the learner

is able to reliably identify these categories in a prior stage (see also Lin & Mielke, 2008, who discuss this simplifying assumption).

As we argue, this disconnect between the statistically induced phonetic categories and the phonemic categories that are the target of acquisition has led to an implicit two-stage view of phonological learning. That is, learners first learn phones using statistical inference over acoustic input, and then build phonemes and phonological systems by identifying relations between these phonetic categories. In this article, we suggest that such a view is a consequence of current models of first-stage statistical categorization, because these approaches will converge on phonetic, rather than phonemic, categories. This requires a second stage of acquisition that subsequently builds the relevant phonemic categories from the phonetic categories.

Although the structure of current models of phonetic category formation seems to suggest a two-stage model of phonological categorization, we argue in this article that this two-stage approach is not inevitable. Indeed, the need for a close relationship between phonetic and phonological learning has been noted by a number of researchers investigating the acquisition of phonological systems (Maye, Daland, & Goldrick, 2008; Seidl, Cristià, Bernard, & Onishi, 2009). We present two arguments in favor of an alternative, single-stage approach to the acquisition of phonological categories. First, we present simulation evidence with a data set from Inuktitut that suggests that seemingly inconsequential errors during a phonetic categorization stage impede a second-stage phoneme discovery procedure. Second, we show that with this same data set, the correct phonemic categorization of the data can be obtained with a single-stage categorization model that jointly learns phonemes and processes by factoring out predictable alternations conditioned on environment, rendering subphonemic categories epiphenomenal. The results show the viability of a single-stage conception of phonological category acquisition and suggest that, for the data set examined here, such an approach is in fact more successful than a two-stage approach to phoneme discovery.

1.1. The phonological learning problem

As alluded to above, the two-stage view of phonological acquisition parallels a distinction that linguists have long drawn between *phonetics* and *phonology*. Phonetics refers to the study of perception and production of speech, and phonology is concerned (sometimes implicitly) with the encoding of speech in the lexicon (i.e., long-term memory). Much work in phonetics stems from the observation that phonetic representations are finely detailed and best represented as continuous rather than discrete values (Fant, 1960; Ladefoged, 2001; Ohala, 1976). 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 a discrete rather than a continuous encoding (Chomsky & Halle, 1968; Goldsmith, 1976; Prince & Smolensky, 2004). The inventory of discrete phoneme categories varies from language to language, and an infant acquiring his or her native tongue must identify the phoneme categories that are relevant for his or her language. Part of this task is phonetic in nature, as the infant must determine the distribution of each speech category 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 statistics and machine learning and has led to important new models of phonological acquisition.

One way of modeling this sort of phonological knowledge is with a mixture model (McLachlan & Peel, 2000). Mixture models are statistical models that describe a set of data (e.g., a stream of acoustic observations) as coming from a probability distribution generated by a finite set of component categories (e.g., phoneme segments). On this model of the phonetics–phonology mapping, the listener has an acoustic map that indicates how likely an acoustic token is as a realization of a given phoneme, Pr(acoustics|phoneme). Furthermore, each phoneme also has its own mixing probability Pr(phoneme) of occurring, so that an ambiguous sound will be more likely classified as a more probable phoneme. Cast this way, the task of the learner is to learn the parameters and the mixing probabilities of the components that make up the mixture distribution; in this way, to fit a mixture model to data is to specify these two probability distributions. This is a statistical formulation of the clustering task in machine learning, because the observations form ''clusters'' associated with different mixture components. Fitting such a model is an example of unsupervised learning, because the knowledge of the component assignments that give the phonemic category of any given token is not provided to the fitting algorithm. Instead, this information must be guessed on the basis of the clusters formed by the input. Presumably, the problem faced by the infant in learning phonological categories is an unsupervised clustering problem of this sort, and so phonetic or phonological categorization can be usefully modeled as the search for a mixture model that is optimal for the infant's speech environment. As we detail below, this basic model has formed the basis for a number of successful approaches to the acquisition of phonological categories.

However, this view of phonological category acquisition as unsupervised clustering is complicated by contaminating factors such as environmental noise, speaker variation, and, most important here, the non-trivial mapping between phonemic and phonetic representations because of the existence of phonological processes. In many theoretical approaches that view phonological processes as operations over discrete units, the relation between the phoneme and its pronunciations is stated as a process taking a discrete object (e.g., the phoneme /t/) to another discrete representation (its phonetic realization as unreleased [t˺] or the flap [ɾ]). If it is assumed that more detailed, quasi-continuous phonetic information is filled in after all phonological processes have taken place, then there is a clear distinction between two discrete levels of representation involved in phonological cognition. One is the lexical level (the phoneme level); the other is the discrete ''surface'' level, which is obtained following the application of all of the discrete contextual phonological rules, but none of the phonetic-detail rules that fill in the details of how the segments are pronounced (the phone level).

Although it is a useful (and nearly ubiquitous) theoretical device, it is not clear that there is any independent motivation for assuming that a unique, coherent level of discrete representation follows the application of all contextual rules. Nonetheless, many researchers maintain discrete levels of phonetic and phonemic representation, and this has provided the

implicit theoretical motivation for a two-stage model of phonological acquisition: 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.

1.2. The two-stage approach

The theoretical distinction between discrete phonetic and phonemic encodings is also found in research on phonological acquisition, with research generally focusing either on categorization of a phonetic nature or on the mapping between phonetic and phonological categories. This division of labor has led to an implicit two-stage approach to phonological acquisition. Such a model suggests that phonological acquisition proceeds by first identifying phonetic-level categories, and then using those categories to discover phonemic categories.

The first stage of phonological acquisition, the mapping from acoustics to phones, has been explored in a large body of work on discovering category structure from acoustic data. This work often employs explicit statistical models of inference (de Boer & Kuhl, 2003; Coen, 2006; Feldman, Griffiths, & Morgan, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007). The acquisition target of these models is sometimes cast in a way that is neutral between phones and phonemes: Vallabha et al. (2007), for instance, propose a model for learning ''sound categories'' (p. 13273). On the other hand, Feldman et al. (2009) explicitly note that although their model contains a lexicon, it is more likely to converge on phonetic, rather than phonemic, categories. In both cases, this follows from the structure of the model: Without explicit modeling of the phonological processes, this stage of acquisition is guaranteed to converge on categories that do not abstract out these phonological processes. If the standard understanding of the relation between phones and phonemes is to be preserved, then any one of the resulting categories in such a model is a phone, not a phoneme, whether it happens to cover all, or only one, of the predictable allophones of a given phoneme. For this reason, this line of research implicitly presents itself as one stage in a two-stage process; if the end goal of acquisition is phonemic categories, then models that do not explicitly encode this fact imply the existence of a second stage of acquisition to reach the target state.

The observation that these previous approaches do not reach phonemic solutions is not intended as an argument against statistical approaches to discovering category structure. There is arbitrary variation in acoustic targets for the same phone or phoneme category across languages (Flemming, 2001; Pierrehumbert, 2003), and so it seems that the learner must acquire at least some of the phonetics–phonology mapping. Any fully specified model of phonological acquisition should contain a mechanism for inferring category structure over a perceptual space. Furthermore, given that not all acoustic tokens of a single category within a language will be identical, a learning mechanism is needed that is statistical, in the general sense that it deals with noisy data in some well-defined way. For this reason, the alternative single-stage model we discuss below shares many of the assumptions of these first-stage statistical models.

However, a model of the acoustics-to-category acquisition process that can find only phones requires a way of addressing the second half of the phonological learning problem: the acquisition of the mapping from phones to units of lexical encoding (phonemes) and the phonological grammar. One possibility for this stage is to trivialize this mapping and deny the existence of phonemes, a position that we argue is not desirable on theoretical or empirical grounds. Assuming this is not tenable, it is necessary to develop an explicit theory of how to group phones into phonemes. Ideas about this procedure are implicit in much of theoretical linguistics, including the well-known complementary distribution test (Harris, 1951). More recently, Peperkamp et al. (2006) have proposed solving this problem by comparing the sequence-level distributions of pairs of phones; that is, for each pair of phones p_1, p_2, they examined the probability distribution over phones adjacent to p_1 versus p_2. They proposed that phones with the most dissimilar context distributions are more likely to be variants of the same phoneme, with the probability distributions reflecting a generalization of the traditional notion of complementary distribution, subject to further naturalness constraints on possible phone-to-phoneme relationships. By investigating the context distributions of discrete phones, this algorithm implicitly assumes that phones have been uniquely identified and categorized at a previous stage of acquisition. Note also 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 processes).

An alternative conception of this second stage is seen in work on learning of optimality-theoretic grammars (OT; Prince & Smolensky, 2004). On this approach, the set of phonological processes (the phonological grammar) is a ranked set of well-formedness constraints, which determine the correct pronunciation of a lexical item in its stored (phonemic) form. There are several well-known, computationally explicit algorithms for learning these grammars (Boersma & Hayes, 2001; Hayes, 2004; Pulleyblank & Turkel, 1998; Tesar & Smolensky, 1998). Although they vary in their approaches, they also share the assumption that the input to learning is a set of discrete, phone-level representations, and the grammar is derived once these phones are identified. The work on phonetic learning by Boersma, Escudero, and Hayes (2003) does incorporate low-level phonetic learning into an OT constraint-ranking grammar, but these constraints do not incorporate any contextual or grammatical information at the phonetic level of learning, and thus implicitly ascribe all the systematic contextual pronunciation rules to the mapping between phonemes and phones. By taking the output of a first-pass mapping from acoustics to phones, these approaches thus also implicitly endorse the two-stage view of phonological acquisition. There exist still other approaches to the phoneme-finding problem (Dresher, 2009; Goldsmith & Xanthos, 2009; Jakobson, 1941), but all are formulated under the assumption that a set of phones has already been discovered.

The assumption of two-stage learning is not innocent, however, since, as we will detail below, the success of a two-stage approach to phonological learning crucially depends on the accuracy achieved in the first stage. Errors made in the phone acquisition stage could in principle impair the ability of a second-stage mechanism to extract the correct phonology. Furthermore, although a two-stage view of phonological acquisition appears to be implicit

in the majority of research on phonological acquisition, it is not the only possibility. In the remainder of this article, we explore the feasibility of a single-stage approach to phonological categorization. In particular, we focus on statistical methods of category identification. As noted above, all theories of phonological acquisition must address this mapping from acoustics to discrete categories. Because of this fact, asserting the feasibility of a single-stage approach amounts to asserting the possibility of folding the acquisition of processes and phoneme-level categories into the initial mapping from acoustics to linguistic categories.

1.3. 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. One important and reliable finding is that young infants are initially able to discriminate a wide range of speech sounds, even those not present in their linguistic input (Werker & Tees, 1983, 1984). A number of studies have shown that these discriminatory abilities quickly decay as the infant develops. Declining sensitivity to non-native language vowel contrasts is apparent as early as 6 months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), and by 8 months, similar effects are evident in consonant contrasts (Werker & Tees, 1984). This changing sensitivity is taken to reflect the development of perceptual models of speech sound categories (whether phone-level or phoneme-level). The early onset of this development raises the possibility that this category learning may not be entirely driven by the building of a lexicon, as has been sometimes assumed (Best, 1995; Jusczyk, 1985). Infants at this age know relatively few words (Fenson et al., 1994), and if the relevant lexical knowledge is knowledge of minimal pairs, it may be that these are too rare to have such a reliable effect so early (Dietrich, Swingley, & Werker, 2007).

For these reasons, some researchers have hypothesized that distributional learning mechanisms play an important role in phonological development (Chambers et al., 2003; Maye et al., 2002; see also Vallabha et al., 2007), and there is experimental evidence in support of this hypothesis. For example, Maye et al. (2002) trained infants at 6 and 8 months of age by exposing them to sets of stop consonants with either bimodal or unimodal distributions over voice-onset times (VOTs). When presented with bimodal distributions, infants showed enhanced sensitivity to differences between points at the extremes of the VOT ranges. Infants in the unimodal condition were less able to discriminate endpoints on the VOT continuum, suggesting that they had classified them together on some level. Building on results that show that infants are sensitive to distributional information in other modalities, such as word-learning (Saffran et al., 1996), the authors interpreted these findings as showing that distributional characteristics of the input directly impact the dimensions of the signal the infants view as relevant or contrastive. Werker et al. (2007) went on to show that, for the vowel space, clear distributional cues in the infant-directed speech of both Japanese and English speakers appear to 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 capturing the acquisition of speech categories using various types of statistically

informed frameworks. As discussed above, the acoustics-to-category mapping can be modeled with a range of statistical techniques, with mixture models playing a prominent role. For example, de Boer and Kuhl (2003) used a mixture of Gaussians (MOG) to model the acquisition of vowel categories using the values for the first two formants. MOG models are a form of mixture model that represents category structure as a set of parameterized Gaussian distributions in the input space, each weighted by a mixing probability (see above). The authors fit MOG models to vowels recorded during mother–child interactions using expectation maximization (EM, an algorithm used for maximizing the likelihood of a model with unobserved structure; see Dempster, Laird, & Rubin, 1977). They focused on the vowels at the extreme points of the vowel space in English (/i/, /u/, and /a/), and they fixed the number of components in the model at three when fitting the model. 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.

Vallabha et al. (2007) also used a MOG to model the acquisition of the Japanese and English data that was analyzed by Werker et al. (2007). The models were sets of four-dimensional Gaussians in the raw acoustic space ($F_1 \times F_2 \times F_3 \times$ duration), the parameters of which were updated iteratively after processing each input point online. While similar to standard EM, their method of fitting the MOG acquired categories online, as opposed to batch processing over a corpus of data (as in standard EM). The online nature of the Vallabha et al. model is arguably closer to the procedure used by human infants. Applying this technique on several distinct data sets, each from a different speaker, the model matched the true vowel systems 80% of the time; an alternate model that dropped the assumption of Gaussian components was successful 60% of the time. McMurray, Aslin, and Toscano (2009) also used a version of this online algorithm to model the acquisition of phonetic categories.

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 manifold learning 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 et al. (2009) constructed a hierarchical Bayesian model, including an embedded MOG that jointly solved the problem of inferring categories and a lexicon, allowing for the construction of a base of lexical knowledge that delivered impressive performance in separating highly overlapping categories in English vowel data. The success of the simple MOG approach has thus been shown to benefit from the addition of extra disambiguating information.

Despite these successes, the mapping from acoustics to linguistic categories remains an extremely challenging problem in its own right. Because it appears to vary between languages, this mapping must be learned and is an essential part of any phonological acquisition model. We now turn to a more detailed examination of these models. We first focus on the problem of acquiring phonetic categories using a MOG approach. Using data from Inuktitut, we demonstrate that the sorts of models explored in the literature up to now are at risk of extracting categories that are either insufficiently fine-grained, or too poorly aligned

with the real categories of the language, to enable learners to discover the systematic relations between phones in a second stage of acquisition. With this data set, it does not appear that a two-stage approach to acquiring the Inuktitut phonemes is likely to be successful. In light of this, we consider an alternative conception of the first stage of category acquisition. We show that it is possible to extract categories which correspond better to the phoneme level than the phone level in the acoustic clustering stage (a single-stage approach to phonemic category acquisition), rather than leaving that step for a second stage of acquisition. When compared with a basic MOG approach, the single-stage approach returns a more realistic set of phoneme categories. This provides initial evidence that a single-stage approach to the acquisition of phonological categories is in principle possible, and on the Inuktitut data set considered here, this approach outperforms two-stage approaches by providing a categorization of the acoustic space that better fits the target of acquisition.

2. Experiment 1: Mixture of Gaussians

As noted above, human infants learning speech sound categories may be said to be discovering mixture models of the speech segments they encounter, regardless of the representational level (phonetic or phonemic) of the categories acquired. Most of the models of this learning problem in the literature have assumed each category to be a single multivariate Gaussian in acoustic space; for vowels, this is typically the first two or three formant values, as extracted from the speech spectra. Previous results indicate that, at least in simple settings, a Gaussian mixture model with categories approximating the true categories can be found using the standard techniques applied to this problem in statistics. This has been taken to suggest that our understanding of this part of the infant's learning problem is already fairly clear. This has been demonstrated primarily for simple phonetic category systems; it becomes progressively more difficult to discover the true categories underlying a data set as the clusters become more poorly separated in the input space when the clusters are not actually generated in a way that satisfies the model assumptions (e.g., to the extent that MOG is an inaccurate approximation of the learner's perceptual map). We argue, however, that previous approaches may find significant difficulty for even fairly simple systems, for other reasons.

In what follows, we examine the role of phonological processes in the speech sound category learning problem. We use data from Inuktitut. Inuktitut is an Eskimo-Aleut language spoken in northern Canada. Like many other related languages, Inuktitut has three vowel phonemes: /i/, /u/, and /a/. The quality of a vowel, however, is often affected when followed by one of the uvular consonants (/q/ or /ʁ/; Dorais, 1986). In the dialect of Kinggait (Cape Dorset), uvular consonants lower all vowel tokens to some degree (Denis & Pollard, 2008), suggesting the presence of six phonetic categories (see Fig. 1). Three vowels plus a strong retraction effect before uvular consonants is a fairly common phonological system; similar systems are found in Quechua and Modern Standard Arabic (Kuriyagawa, 1984; Pasquale, 2009). Such a system could potentially make each phoneme acoustically bimodal.

The presence of additional contextually determined subcategories of the three phonemic vowels (for convenience, we will refer to these as [e], [o], and [ɑ], three contextual

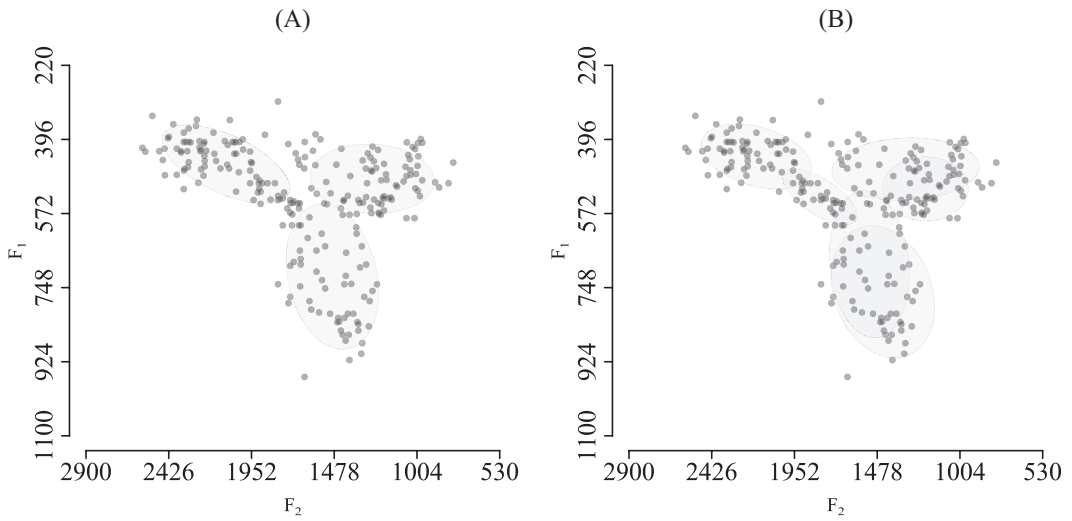

Fig. 1. Plots of Inuktitut vowels, both grouping (A) and splitting (B) predictable allophones, in $F_2 \times F_1$ (backness by height) space. The ellipses mark a 66% confidence region for Gaussians estimated by maximum likelihood on the points from the indicated category.

allophones of /i/, /u/, and /a/, respectively) suggests the presence of six phonetic categories. For such a system, a two-stage model of phonological acquisition must first satisfactorily recover these six categories before any procedure that discovers links between phones can successfully operate. In Experiment 1, we ask whether learning under a simple MOG model does in fact return the set of six adult phonetic categories which could be used as input to a second stage of acquisition. If the result of learning were nicely consistent in giving these six phonetic categories, we would have some indication that a two-stage model is plausible for this data set, although this would still leave open the question of how learners come to have systematic knowledge of phonological processes.

2.1. Materials

The Inuktitut vowel corpus that we employ comes from a study on Inuktitut phonetics (Denis & Pollard, 2008). All vowel tokens were measured from elicited speech of an Inuktitut speaker from Kinggait and were hand-labeled by trained phoneticians for first formant (F_1) and second formant (F_2) values; these measurements were chosen because these two acoustic parameters are known to be highly informative indicators of vowel height and backness, respectively. Two hundred and thirty-nine vowel tokens were measured in this way; we resampled simulated corpora for use in training containing 1,000 and 12,000 tokens from this data set nonparametrically using a two-dimensional kernel density estimate using the *ks* package for R (Duong, 2011; R Development Core Team, 2008), respecting the frequencies of each of the phones in Inuktitut, balanced according to the natural mixing proportions obtained from the Nunavut Hansard corpus (version

2.0; Martin, Johnson, Farley, & Maclachlan, 2003). These proportions differed from the proportions found in the raw phonetic corpus mainly in the relative frequency of the two back phones [u] and [o].[1]

2.2. Methods

There are various methods for optimizing over the set of possible mixtures of Gaussians. We chose a standard Bayesian estimator: a point estimate taken from a sample from the posterior distribution. The posterior distribution of interest was over mixtures of Gaussians given a Dirichlet process prior (an *infinite MOG*: Escobar & West, 1995; Ferguson, 1973). This represents a particular way of stating formally that the hypothesis space is all possible Gaussian mixture models, including models with different numbers of categories, along with a particular way of weighting different mixture models (a Dirichlet process in this context is essentially a certain prior probability distribution over mixture models). After this choice of prior is made, the remainder of the solution is a standard problem in Bayesian statistics.

Bayesian inference makes use of the posterior distribution over hypotheses, that is, the measure of how probable a hypothesis is (in this case how likely any given Gaussian mixture model is) that would be derived by a rational agent under the specified prior distribution (set of modeling assumptions). A ''rational'' agent is simply one that obeys the axioms of probability theory for making decisions under uncertainty. For formal justification, see Cox (1946) and Jaynes and Bretthorst (2003); see also Oaksford and Chater (2001) for empirical justification of the common use of the term ''rational'' for such models.

A Bayesian estimator is useful in this context because joint inference can be done straightforwardly on problems of potentially arbitrary complexity. This is advantageous, for example, when inferring the number of categories in a mixture model (a crucial part of the problem of phonetic category learning). In contrast, frequentist methods (e.g., the traditional EM algorithm) explicitly prohibit the statement of probabilities over model parameters, and this is a serious liability given the inherently hierarchical nature of this problem. There are standard methods available for deriving estimators for the underlying set of mixture components and mixing probabilities justified by the data, assuming some particular fixed number of categories. However, because the learner by hypothesis needs to estimate the number of categories justified by the data, Bayesian estimators that incorporate uncertainty over this part of the model provide a more attractive option for modeling acquisition.

The Bayesian solution to hierarchical problems like this is to treat the parameters as unobserved data and put a prior probability measure on them; the parameters of this prior probability (the hyperparameters) can in turn be learned in exactly the same way, and we can continue to place hyperpriors on the parameters until we have reached a level of model complexity that we believe mirrors that of the human learner relatively well (keeping in mind that adding more learned parameters to the model will not be much better than simply specifying them manually if we do not have enough relevant data). Just as in frequentist estimation, the result will be sensitive to the modeling assumptions, but these assumptions can in principle be as vague (lack of bias) or as precise (strong bias) as desired.

In the case of the number of mixture components, the standard Bayesian solution is to put a prior probability measure on sets of mixture components (in this case, on sets of Gaussians) and associated mixing probabilities and compute the posterior probability over hypotheses given the observed data set. One common probability measure used for this purpose is the *Dirichlet process,* which has as free parameters a *concentration parameter,* α, controlling the a priori tendency to add new categories, and a *base distribution,* G_0, the prior distribution on individual Gaussian components. This can be seen as a method for regularization in which there is a penalty to the likelihood not only for the number of categories but also for having mixture components that do not adhere to some prior expectation about reasonable mixture components (the base distribution).

A posterior sample from a Dirichlet process MOG was drawn using a Gibbs sampler with component parameters drawn from a normal–inverse Wishart distribution with fixed inverse scale matrix and degrees of freedom parameter, and with location parameter M and inverse scale parameter ω; M was itself sampled from a normal distribution centered at zero, and ω from an inverse Gamma distribution; α was sampled from a Gamma distribution (see Escobar & West, 1995; West, 1992; Neal, 2000, for the basic details of the algorithm). To fit each model, a sample of 500 points was drawn from the Gibbs sampler at a lag of 10 after 1,200 burn-in samples. The sample with the highest joint posterior density was used as a point estimate. Hyperparameters were tuned to ensure that they were appropriate to find between one and seven categories on the raw data from which the training corpus was sampled.

Although the use of an informative prior guards against overfitting, we chose to also train each model using 10-fold cross-validation—that is, partitioning each data set into 10 subsets and, for each subset, training on its complement. By testing on the held-out subset, we can verify that the model fits are not overly sensitive to idiosyncracies of the training set. A single chain giving one point estimate was derived from each of the 10 training subsets, for each of the three different sized training sets (raw data, 1,000-point sample, 12,000-point sample). Geweke's z-statistic (Geweke, 1992) was computed on all real-valued parameters and hyperparameters for each chain (comparing the first 10% and second 50% of the chain) to test for stationarity; only runs for which all variables had two-sided normal p-values above 0.001 were retained. Three runs of the 1,000-point model were dropped by this criterion.

2.3. Results

To assess the quality of the fitted phoneme models, we first constructed ideal sets of Gaussian phonetic and phonemic categories using the maximum likelihood estimators for each phoneme (sample means and sample covariances), for each different data set used to train the model (see Fig. 1 for a representative plot). Using these Gaussians as category models, we classified the data sets from which the Gaussians were constructed using a Bayes-optimal decision rule, labeling a point according to the mixture component with the highest posterior probability given that point. This decision rule is optimal in the sense that it minimizes the probability of classification error under the simple zero-one loss function (Duda, Hart, & Stork, 2000).

Table 1
Classification performance for ideal Gaussian models. These values represent the highest possible pairwise F scores (see text) for comparisons between the ideal models' predictions and the data. Two different versions of the true classification are evaluated with this baseline: a three-category phonemic solution (phoneme labels, $K = 3$) and a six-category phonetic solution (phoneme labels plus an indicator for a following uvular, $K = 6$)

	K	F	Prec	Rec
Raw (239 points)	3	0.84	0.83	0.85
	6	0.64	0.66	0.63
1,000 points	3	0.79	0.79	0.79
	6	0.69	0.64	0.76
12,000 points	3	0.78	0.78	0.78
	6	0.68	0.63	0.74

We summarize the baseline levels of performance provided by these optimal classifications in Table 1 using three statistics: *pairwise precision, pairwise recall,* and *pairwise F-measure* (Amigó, Gonzalo, Artiles, & Verdejo, 2009). *Pairwise* refers to the fact that the statistics are constructed by examining every pair of data points and asking whether the two are in the same class (according to either the fitted model or the ideal model). Pairwise statistics are used in clustering evaluation to avoid the issue of constructing a mapping between the model's categories and the true categories; they are still meaningful even if the model finds the wrong number of categories. We obtained the model's predictions about shared class membership, compared them with the true classifications, and computed the precision (percentage of pairs predicted as the same which actually are), recall (percentage of pairs which actually are the same that were predicted as the same), and F-measure (the harmonic

Table 2
Results of Experiments 1–3, 10-fold cross-validation on each of three Inuktitut data sets. Left of table shows distribution over number of resulting categories, and right of table shows pairwise scores at test (see text for discussion of scores). In parentheses is the difference from scores on training data. Comparisons to both three- and six-category (italicized) classifications are shown for Experiment 1

	$K = 1$	2	3	4	5	6	F	Prec	Rec
Experiment 1: general mixture of Gaussians									
1,000	0	0.125	0.625	0.125	0	0	0.70 (+0.02)	0.66 (+0.01)	0.74 (+0.03)
							0.60 (*+0.01*)	*0.50* (*+0.01*)	*0.76* (*+0.02*)
12,000	0	0.1	0.5	0.1	0.2	0.1	0.65 (+0.01)	0.68 (+0.01)	0.63 (+0.02)
							0.58 (*+0.02*)	*0.53* (*+0.01*)	*0.63* (*+0.02*)
Raw	0.1	0.2	0.7	0	0	0	0.65 (−0.03)	0.59 (−0.03)	0.76 (−0.03)
							0.47 (*−0.04*)	*0.34* (*−0.01*)	*0.81* (*−0.02*)
Experiment 2: general mixture of Gaussians, process-corrected data									
1,000	0	0.4	0.5	0.1	0	0	0.73 (+0.02)	0.67 (+0.02)	0.82 (+0.02)
12,000	0	0	0.875	0.125	0	0	0.74 (+0.02)	0.72 (+0.02)	0.76 (+0.02)
Raw	1.0	0	0	0	0	0	0.63 (−0.01)	0.49 (−0.02)	0.88 (−0.01)
Experiment 3: mixture of linear models									
1,000	0	0.111	0.889	0	0	0	0.75 (+0.02)	0.71 (+0.02)	0.80 (+0.02)
12,000	0.143	0	0.571	0.286	0	0	0.69 (+0.02)	0.65 (+0.02)	0.76 (+0.02)
Raw	0.125	0.125	0.75	0	0	0	0.69 (−0.01)	0.64 (−0.00)	0.79 (−0.01)

mean of precision and recall). The same statistics were then computed for each of the models fit for each data set and averaged (geometric mean). Results for models fit in Experiments 1–3 are shown in Table 2.

The results shown in Table 2 show that the MOG model is capable of finding three-category solutions that are not unlike the phonemes of Inuktitut; this is seen in the classification scores for the 1,000-point models: The *F* scores are reasonably close to the *F* scores for the ideal models (compare Table 1) and are reasonably well balanced between precision and recall. See Fig. 2 for a representative example.

More fine-grained phonetic solutions become apparent as the number of data points increases, which is to be expected, partly because of the prior, and partly because the likelihood term, which will come to dominate the prior as the number of data points increases. The likelihood term, all other things being equal, prefers larger numbers of categories (the mixture model with the highest possible likelihood would generally be obtained with as many categories as data points). See Fig. 3 for representative plots of five- and six-category solutions found by this model.

2.4. Discussion

Experiment 1 replicates previous work (de Boer & Kuhl, 2001; Feldman et al., 2009; Vallabha et al., 2007) in showing that a MOG approach to vowel categorization appears to provide a good starting point for modeling the acquisition of language-specific sound

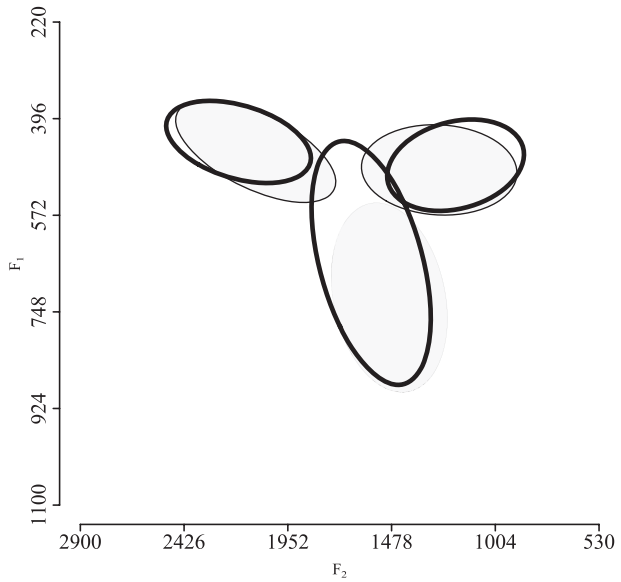

Fig. 2. Plot of a representative three-category model found for the 12,000-point data set in $F_2 \times F_1$ (backness by height) space, in Experiment 1. Outlined ellipses mark a 66% confidence region for the estimated Gaussians. Shaded ellipses mark a 66% confidence region for individual Gaussians estimated by maximum likelihood for the true phoneme categories.

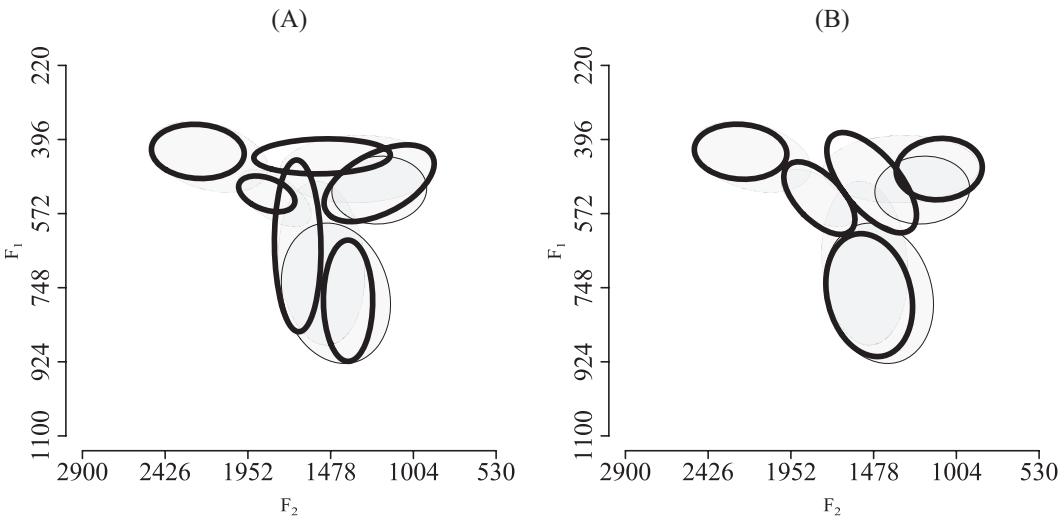

Fig. 3. Plot of representative six-category (A) and five-category (B) models found for the 12,000-point data set in $F_2 \times F_1$ (backness by height) space, in Experiment 1. Outlined ellipses mark a 66% confidence region for the estimated Gaussians. Shaded ellipses mark a 66% confidence region for individual Gaussians estimated by maximum likelihood for the true phonetic categories.

categories. On the Inuktitut data, smaller data sets led to three-category solutions, whereas larger data sets tended to give solutions with more vowel categories. The results of Experiment 1 suggest that a learner assuming Gaussian categories (a simplifying assumption shared with previous research) could come to a phonemic analysis of Inuktitut vowels with the appropriate biases, but also that an analysis with phone-like categories could also be found with the appropriate bias. The role of bias is important here, as these solutions are not ''in the data'': The learning outcome depends on the specific bias implied by the model and its hyperparameter settings, in conjunction with the amount of data the model is given. In particular, as the number of data points increases, models tend to prefer a greater number of mixture components. It should not be surprising that much more research in psychology and speech perception is still needed before we can obtain a complete picture of what that bias looks like.

However, both types of ''complete'' solutions found by the simple MOG in Experiment 1—three-category and six-category solutions—represent at best an incomplete representation of the target phonological knowledge. Consider first the case of the three-category solutions discovered by the learner modeled in Experiment 1. Although it appears that these models arrive at knowledge that approximates the target phonemic categorization, by immediately jumping to phoneme-like phonetic categories, such a model would require a second stage wherein learners rediscover the systematic relationships between particular contexts and the pronunciations of these categories. Importantly, the systematic relationship between a phoneme and its retracted allophone in Inuktitut forms an active piece of knowledge that speakers must acquire: Even in novel words, speakers adapt the pronunciation of the phoneme to its phonological environments. Novel words with the appropriate combination of morphemes are easy to construct, given the complex, polysynthetic nature of Inuktitut

morphology. For example, Inuktitut has a productive process of noun incorporation with certain verbs that allows the direct object and the verb to form a single phonological word (Johns, 2009). Any incorporated noun ending in a vowel will be subject to the effect of a following uvular. Thus, a word like /titi ʁ auti/, ''pen,'' becomes [tite ʁ aute] in the expression /titi ʁ autiqaqtunga/, ''I have a pen,'' which is pronounced as [tite ʁ auteq ʁ qtunga]. If the phonemic three solution category arrived at by the simple MOG model here is correct, then some other mechanism would be needed to allow speakers to recover the knowledge of the allophone subcategories necessary to capture these facts about Inuktitut.

Although a three-category solution is problematic, a six-category, phonetic categorization could potentially provide the necessary first step for a two-stage model of phonological category acquisition. If learners gain knowledge of two separate allophones for a phoneme, they must have the means of learning which of the allophones is to be deployed in any given context. The traditional understanding of this relationship among linguists has been that the phonetic and the phonemic categories occupy different representations in the discrete space of possible lexical representation. This implies that, at the perceptual level, speakers treat [i] and [e] differently, and equate them at a higher, more abstract level of processing. A six-category mixture is consistent with this claim, with the addition of a phonological rule relating the two. Thus, it is natural to ask whether the phone-like mixtures reported above would be amenable to a search for such a phonological rule—a second ''stage'' of learning.

Visual inspection of the resulting five- and six-category models suggests that these models would not provide adequate input to a second stage of learning based on a complementary distribution test. For example, in Fig. 3, it appears that the high back phoneme /u/ has been divided into more fronted and more backed subphonemic categories, rather than the more high and more retracted categories suggested by the ideal model in Fig. 1. In order for five- or six-category solutions to provide the input to a second stage of phonological category acquisition, it should be the case that the acquired categories align with the phonetic categories of the target system; it might, however, be the case that the apparent errors at this stage impede the discovery of higher level phonological categories. To determine this, we examined the five- and six-category solutions obtained by the learner in Experiment 1 and examined the phonetic models obtained in order to ascertain whether a simple distribution-based test would confirm the presence of a rule relating the predictable allophones. Note that we include five-category solutions under the assumption that the phonetic difference between the low allophones of /a/ is subtle, and if learners form a single perceptual category for /a/, one might speculate that this retraction might just be the result of a low-level physiological coarticulation process, not perceived or learned. We do not claim that this is true, but since the five-category solutions were not substantially different from the six-category solution with respect to the high vowels, examining them alongside the six-category solutions simply allowed us to form a better picture of what model performance in a second stage might be.

In order for the phonetic categorization to support a second-stage phoneme acquisition process, it must be the case that the retracted allophones are reliably found in the context of uvular segments, while non-retracted segments are not. To determine this, we applied the symmetrized Kullback–Leibler divergence criterion of Peperkamp et al. (2006).

Symmetrized KL divergence (SKLD) is a standard information-theoretic quantity that is used to ascertain how different two probability distributions are; it can take on any non-negative real-number value, and larger numbers represent more different distributions. Following Peperkamp and colleagues, we examined the probability of observing a uvular consonant versus a non-uvular consonant following each of the five categories constructed by the model, and computed the SKLD for each pair of categories. In Tables 3 and 4, we present the SKLDs for each phone found by the model (for the five-category solutions, the average over the two models along with maxima and minima); the category labels were clear and easy to assign by visual inspection (see Fig. 3).

Peperkamp and colleagues' statistically grounded complementary distribution test attempts to find allophonically related pairs of phones by looking for large values of SKLD; large values suggest more divergent context distributions, and thus a relation closer to complementary distribution. There is no obvious prior notion of ''large'' SKLD in this context, and Peperkamp et al. used a threshold determined from the distribution of SKLD scores. From this point of view, the pattern in the SKLDs is clear: The SKLD for [i]–[e] is consistently among the highest values found, suggesting that complementarity-based metrics for discovering phonemic identity could readily recover the relation between these two phones given this MOG. However, [o]–[u] consistently had some of the lowest SKLD scores. This is consistent with the visual observation that the models did not correctly identify [o]–[u],

Table 3
Symmetrized KL divergences for the distribution of uvulars following each of the phonetic categories, for the six-category solution found among the model solutions in Experiment 1. Phonetic labels were assigned to the categories by visual inspection. Phone pairs that are true allophones in Inuktitut are in bold

	[i]	[e]	[u]	[o]	[a]	[ɑ]
[i]	0	**0.810**	0.033	0.321	0.330	0.846
[e]	–	0	0.478	0.098	0.093	0.000
[u]	–	–	0	**0.138**	0.143	0.504
[o]	–	–	–	0	0.000	0.109
[a]	–	–	–	–	0	**0.104**
[ɑ]	–	–	–	–	–	0

Table 4
Average symmetrized KL divergences (with standard deviation) for the distribution of uvulars following each of the phonetic categories, for each of the two- to five-category solutions found among the model solutions in Experiment 1. Phonetic labels were assigned to the categories by visual inspection. Phone pairs that are true allophones in Inuktitut are in bold

	[i]	[e]	[u]	[o]	[a]
[i]	0	**0.686** ± 0.025	0.304 ± 0.019	0.183 ± 0.011	0.598 ± 0.018
[e]	–	0	0.068 ± 0.016	0.142 ± 0.021	0.003 ± 0.002
[u]	–	–	0	**0.014** ± 0.001	0.043 ± 0.003
[o]	–	–	–	0	0.106 ± 0.002
[a]	–	–	–	–	0

instead splitting the /u/ phoneme in an inappropriate way. The low SKLD values make it unlikely that the five- or six-category MOG solutions found for this data could provide input to a second stage of learning, because the MOG categories do not properly align with the target allophones.

One surprising finding about the MOG model solutions is that they split the /u/ phoneme into front and back variants, rather than the expected [o] and [u]. This suggests that the substantive assumptions implicit in the MOG model are not being met by the data. If the generating categories are not truly Gaussians centered on these phones, then there is no guarantee that an MOG model will converge on the correct classification. In training on raw data and data sampled in a non-parametric fashion, we depart from previous literature that generally trains on data sampled from an ideal MOG. This resampling procedure preserves deviations from the multivariate Gaussian distribution in the raw [o]/[u] data, and so such deviations will make it less likely that the model will be capable of discovering the [o]/[u] categories. The failure of the MOG model to find the correct phonetic categories given these data does suggest that the assumptions of a simple model of the /u/ phoneme as two Gaussians corresponding to [o] and [u] are not being met. This may be because Gaussian phonetic categories are overall a poor model of vowel phones, or it may be because there are additional phonological processes that lead to fronting of the /u/ phonemes (as are attested in the related language Kalaallisut; see Rischel, 1974). The first explanation would imply that human learners do not expect Gaussian phonetic category distributions, but rather make some other distributional assumptions that are not yet understood. Further research is necessary to distinguish these two possibilities.

Thus, Experiment 1 suggests that, although a phonetic category system with enough phonetic categories might be discovered by a learner with a simple Gaussian MOG model, the correct phonemic system would be unlikely to be detected in a second stage of phoneme discovery that uses conventional complementarity-based criteria. The model was able to recover the correct three-category phoneme solution, but we argued that directly accessing phoneme categories in a MOG model creates a problem for the learner: Without the phonetic distinctions between subphonemic categories, it is unclear how the learner could arrive at a full phonological system. The results from Experiment 1 thus suggest that the Inuktitut data presented here provide a challenge to two-stage models of phonological acquisition, as the phonetic categories are not discovered well enough to provide input to a second stage in acquisition. In Experiment 2, we begin to explore an alternative, single-stage model of phonemic category acquisition.

3. Experiment 2: Corrected mixture of Gaussians

In Experiment 1, we showed that the two-stage model of phoneme learning is susceptible to a previously overlooked type of problem: The early phonetic categories must align well with all the phones of the language or else later stages in the acquisition process will be adversely affected. Above we showed that phonetic clustering is likely to pick out systems of discrete categories for Inuktitut that do not align well enough with the phones of the

language for a second stage of learning based on co-occurrences to work, despite a fairly close resemblance to the phones; small differences in the individual category models have serious negative consequences for such distribution-based methods. In Experiments 2 and 3, we develop a model that takes a substantially different approach to solving the same problem by factoring out predictable acoustic variation that arises due to the grammatical rules of the language in the acoustic space, rather than waiting to discover them based on strings of discrete categories.

To illustrate this idea, we briefly present a second MOG model for Inuktitut that implements this idea directly. In Experiment 2, we manually remove the phonetic effect due to following uvulars from all vowel tokens occurring in that context. We then train a MOG model on the resulting transformed data to demonstrate the usefulness of factoring out such transforms; in Experiment 3, we take up the question of how these transforms are acquired. The phonetic category model that results from this procedure is one in which finding phones becomes irrelevant, because the uvular retraction rule has already been handled at the phonetic level. This avoids the problems of a two-stage model in which the phonetic category learning module does not have access to information about which tokens occurred in which contexts and cannot take into account possible effects of grammatical processes when learning categories. In a two-stage model, despite this indifference to the existence of grammatical rules, the category-finding component must nevertheless deliver phonetic categories that will form the basis for finding these rules. Our alternate conception of the learning problem implies that the category learning component does know that uvulars can potentially affect vowel quality, and that it treats the effect of uvulars as a phonetic rule.

3.1. Materials

The materials were the Experiment 1 materials, with one difference. The mean F_1, F_2 value for all the points that occurred before a uvular was computed ($F_{+\mathrm{u}}$); the mean F_1, F_2 value for all the points that did not occur before a uvular was computed ($F_{-\mathrm{u}}$); and the points that occurred before a uvular were corrected for the effect of the following uvular consonant by subtracting ($F_{+\mathrm{u}} - F_{-\mathrm{u}}$) from the formant value. This correction was calculated once for all three vowel phoneme categories, so that all pre-uvular points had the same vector subtracted, regardless of whether they were /i/, /a/, or /u/ tokens.

3.2. Methods

Methods were as in Experiment 1. Application of the Geweke-based criterion for non-stationary chains resulted in the rejection of two runs of the 12,000-point model and two runs of the raw-data model.

3.3. Results

As in Experiment 1, pairwise classification scores were computed for held-out test data. (Note that, as the test data, like the training data, already had the effect of uvularity

removed, it would not have made sense to test the model's classification against the six-way phone classification.) Table 2 shows the results of this classification. It can be seen that across both small and large training sets, three-phoneme solutions are the most common solution reached by the model. The distribution of the phonemes in a three-category solution, as in Experiment 1, line up closely with the target phonemic categories.

The results shown in Table 2 show phoneme classification scores that are higher than any of those seen in Experiment 1 and show a better balance between precision and recall (for the 12,000-point data set). This is because some of the overlap between categories has been removed; this model approximates a listener that can make use of the context in which a segment occurred to adapt its acoustic models (as humans do: see, for example, Nearey, 1990; Whalen, Best, & Irwin, 1997), thus making some regions of uncertainty less ambiguous, and making better phonemic category models available to the learner.

3.4. Discussion

Experiment 2 provided an initial test of a different conception of the phonetic category learning problem than has traditionally been assumed. We removed the effect of a phonological rule before providing the data to the category learning component, and in doing so, we combined two separate stages of the learning problem into the phonetic component. Resulting model fits often returned three-category solutions, and these lined up well with the expected phonemic categorization. By directly linking phonological processes and phonetic categorization in a single space, some of the problems we raised with the results of Experiment 1 are avoided. For three-category solutions, we have effectively coded the knowledge of the phonological process into the learner, resulting in a system that has the phonological process and the phonemic categories, rather than only the undifferentiated categories that were sometimes acquired in Experiment 1. Because contextual rules are directly applied in the phonetic component to undo predictable alternations prior to categorization, the problems with combining phones into higher level phonemic categories are sidestepped altogether.

Although Experiment 2 demonstrates the feasibility of folding together rules and categories during acquisition, this demonstration raises serious questions about acquisition of the rules. We did not require our learner to discover the Inuktitut uvular retraction rule: The knowledge of a uvular rule and the knowledge of the effect of that rule were directly given to the learner to investigate the effect on categorization. A fully specified model of a single-stage approach to categorization should be able to acquire the rules and the categories jointly. In Experiment 3, we address this by presenting a statistical model that jointly estimates a set of categories and a set of phonetic rules. This model can learn a phonetic system while simultaneously taking into account the effect of predictable rules that are not provided in advance.

4. Experiment 3: Mixture of multivariate linear models

Experiment 1 provided a baseline for the performance of statistical learning of phonetic categories on Inuktitut vowel data using a standard MOG model. In Experiment 2, we gave

an initial demonstration of a different, single-stage conception of the problem. We factored out productive rules at the phonetic level, which demonstrates that the traditional two-stage model is not a necessity, and that a single-stage approach that jointly models rules and categories can provide a satisfactory model of phonological acquisition. In Experiment 3, we present a more complete statistical model of the single-stage approach to phonemic acquisition that jointly estimates processes and categories from a set of acoustic inputs.

To accomplish this, we implement one crucial change in the model structure. Rather than constructing a single Gaussian phonetic model for each category (as in Experiments 1 and 2), we model the learner as searching for a set of sets of subcategories, where the subcategories within a set are related by some rule. In other words, each phoneme is defined by a set of Gaussians, in this case, a pair: one for the pre-uvular realizations of that phoneme and another for realizations of that phoneme in other contexts. The idea of a category consisting of a set of subcategories is found in other related areas. For example, in the automatic speech recognition literature, Hidden Markov Models often model an acoustic category as a MOG, rather than a single Gaussian (Jurafsky & Martin, 2000). Another example is the work of Griffiths, Canini, Sanborn, and Navarro (2007), who present Bayesian models of psychological category formation in which each category is modeled by some number of subcategories.

With respect to the current model, these other models are similar in that they would take data that are often modeled as a single Gaussian distribution and instead model the data using multiple Gaussians, to get a more fine-grained, less biased representation of a set of data. An additional constraint we impose in our model is that the data points which are attributed to the two Gaussians need to be in complementary distribution: One Gaussian models the points appearing in a conditioning environment, and the other models the points appearing elsewhere. Any resulting phonemic category generated by the model consists of these two linked Gaussians. Furthermore, to obtain a model that has a straightforward interpretation as ''phonemes plus rules,'' we add an additional constraint of homogeneity of variance. This means that for the allophonic subclusters making up each phoneme, the covariance matrix of the Gaussian (which defines its size, shape, and orientation) must be the same. This should be familiar because it is exactly the constraint that defines a linear model in statistics. The distribution of the response variable is taken to be a Gaussian distribution with a location (mean) that is a linear function of the value of a predictor variable, whether continuous (as in regression) or discrete (as in an ANOVA). This model is important because it gives us a straightforward way of measuring the effect of the predictor. If the only effect of the predictor variable is to shift the mean by some fixed amount, then we can reduce the effect to a single number or, in the present case of multivariate responses, a single vector.

In the model presented below (a multivariate *mixture of linear models,* henceforth *MLM*), the predictor is the presence or absence of the allophonic conditioning environment (one or zero, respectively). The learner must construct a set of categories, each of which is a linear model predicting the phonetic values for the set of segments being categorized (in this case, vowels) from this discrete indicator variable. Because it is a linear model, the learner therefore finds, for each category, an intercept (unperturbed category mean F_1 and F_2), and an

effect of conditioning environment (a shift in phonetic space), in addition to estimating variance. In this way, the model can thus be said to simultaneously discover a set of phoneme categories and a set of associated phonetic rules. In doing so, the model begins to address the problem of learning parts of the phonological system beyond the simple phonetic inventory, and does so in a way which allows the learner to fully leverage the available information.

4.1. Materials

The materials were the same as those for Experiment 1, except that, in Experiment 3, we annotated the data points with a vector of indicator variables marking the presence or absence of a following uvular consonant.

4.2. Methods

The principal difference between this model and the previous model was that each point was modeled as having been drawn from a Gaussian centered at $A^T\boldsymbol{b}$, where A is a 2×2 matrix of regression coefficients and $\boldsymbol{b}$ is an augmented predictor vector, with the pre-uvular indicator (zero or one) as the second element, and one as the first element. The first row of A was thus the intercept (a point in the two-dimensional $F_1 \times F_2$ input space), and the second row the effect of uvularity on the given phoneme. The covariance matrix, Σ, was again learned and was uniform for all the points assigned to a given category in accordance with the homogeneity of variance assumption. The Gaussians centered at the intercept and at the sum of the intercept and the uvularity effect make up the model's representation of the two allophones of a single phoneme. Regression matrices A were drawn from a base distribution that was compound matrix normal–inverse Wishart with fixed inverse scale matrix and degrees of freedom parameter as well as a location parameter (M), and a row covariance matrix (Ω).[2] M was sampled from a matrix normal distribution centered at zero with identity row covariance. Ω was sampled from an inverse Wishart distribution. Note that this model has a simple MOG as a special case, when there are no predictor variables; the Experiment 1 and 2 models were fit using the exact same algorithm, with the only difference being the extra hyperparameters needed in this model. Apart from the introduction of the full matrix of regression coefficients, along with Ω, and the accompanying hyperparameters, the fitting procedure was as before. Again, we ran on three separate data sets and performed 10-fold cross-validation on each. Application of the Geweke-based criterion for non-stationary chains resulted in the rejection of three runs of the 12,000-point model, one run of the 1,000-point model, and two runs of the raw-data model.

4.3. Results

As in Experiment 2, the results shown in Table 2 show that, overall, phoneme classification performance is better than in Experiment 1. In particular, when the classification scores for either of the data sets are compared with the corresponding classification scores from

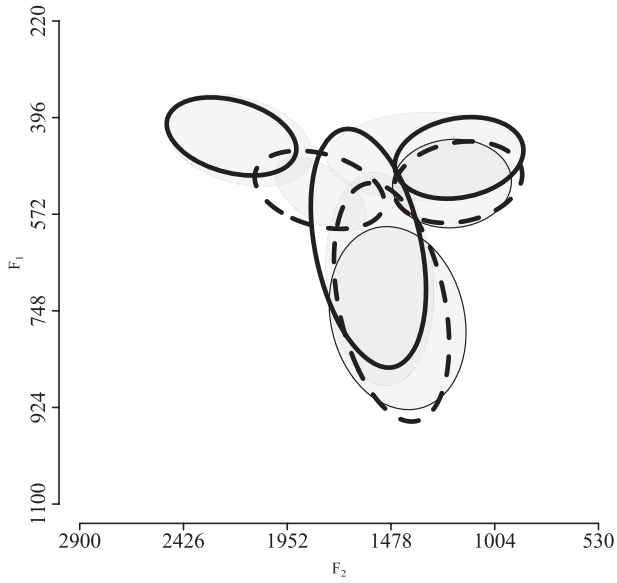

Fig. 4. Plot of a representative three-category mixture of linear models found for the 12,000-point data set in $F_2 \times F_1$ (backness by height) space by the mixture of linear models (Experiment 3). Outlined ellipses mark a 66% confidence region for the sample mean of the estimated Gaussians, each of which is itself part of a linear model which sets up two subcategories for that phoneme; the dotted outlines represent the subcategories shifted by the uvular retraction rule. Shaded ellipses mark a 66% confidence region for individual Gaussians estimated by maximum likelihood for the true phonetic categories.

Experiment 1, they are seen to be higher. This is true for all data sets. A plot of a representative three-category mixture of linear models is shown in Fig. 4.

As in Experiments 1 and 2, pairwise classification scores were computed; in this case, however, predictions about category membership of test points were made in a way that explicitly took into account the presence or absence of a following uvular. That is, for points without following uvular consonants, a decision among the various possible category assignments was made on the basis of the density given by the Gaussian centered at the intercept, and for points with following uvular consonants, the decision was made on the basis of the density given by the Gaussian centered at the intercept plus the effect of uvularity. In other words, for the purposes of this evaluation, the classifier was not asked to assign points to one of the phonetic subcategories induced by the model, but to reconstruct the phoneme from the segment plus the context. Table 2 shows the results of this classification.

As in Experiment 1, the model finds three categories fairly reliably, with a slight shift toward larger numbers of categories for larger numbers of data points. However, when the model does find three categories, its classification performance is better than that of the three-category MOG models. For the three-category solutions, all three performance statistics are statistically significantly higher for the mixture of linear models than for the MOG model. Excluding raw data, the average precision for MLM models (a multivariate mixture

of linear models, henceforth MLM) was 0.69 compared with 0.67 for MOG models (t = 3.9, Welch df = 22.92, $p < .001$; arcsine-transform applied to proportions), and the average recall was 0.74 for MLM models compared with 0.69 for MOG models (t = 4.4, Welch df = 22.94, $p < .001$). The average value of the F statistic was 0.71 for MLM models, and 0.67 for MOG (t = 5.0, Welch df = 22.96, $p < .001$). This same pattern of results also obtains for all the solutions taken together (t test for coefficient of main effect of model type in two-way model × data ANOVA on arcsine-transformed data: for F statistic, $t(31) = 2.5$, $p < .05$; for precision, $t(31) = 2.43$, $p < .05$; for recall, $t(31) = 2.48$, $p < .05$; raw data excluded). Note that although the MOG models are less complex than the MLM models in Experiment 3, this does not mean that the better categorization performance for the MLM models reflects overfitting of the training data by a more powerful model. As all models were evaluated on held-out test data, correct performance requires generalization beyond the training set. Therefore, the MLM model appears to better approximate the true structure of the phonemic categories, rather than idiosyncracies of the training data. These results suggest that a mixture of linear models, while more complex, provides a more realistic model of speech perception.

4.4. Discussion

Experiment 3 demonstrates a novel approach to the problem of learning speech sound categories in human language, contrasting with the standard two-stage model in two ways. First, as in the demonstration model presented in Experiment 2, it encodes the processes relating predictable allophones using phonetic-level information. Despite this increased model complexity (i.e., larger search space), the model is able to arrive at phonemic-level categories which are as good as or better at predicting unseen data than those found using a standard MOG model. Second, while the model is similar to conventional two-stage models in invoking notions of phonetic similarity and complementarity to determine whether a lawful process holds between two phones, it differs in that it does not use the complementary distribution test per se, nor the minimal pair test. In fact, it does not support a notion of ''minimal pair'' at all, because it does not assume any sort of lexicon.

Importantly, the mixture of linear models approach explored in Experiment 3 reliably acquires three phonemic categories in addition to rules that relate allophones of those categories to one another in phonetic space. This is a more robust demonstration of the main idea explored in Experiment 2, showing that processes and categories can in fact be acquired in a single stage of acquisition. Because the model represents both processes and categories in the same phonetic space, they can be jointly acquired easily. This led to better categorization than basic MOG models, but it also has the benefit of providing a more complete model of the link between acoustic input and phonemic categorization. One problem with the results of Experiment 1 was that it was unclear how to go from the MOG categorization to the target phonemic categories. Solutions were either uninformative about subcategories (as in the MOG three-category solution), or they returned phone categories that were different enough from the target phone categories to impede second-stage acquisition of higher level phoneme categories.

Note that this model does not make use of complementary distribution directly, but it can easily be shown that, all other things being equal, a linear model will be more likely to appear in a model fit if it increases the KL divergence of the two allophonic subcategories with respect to the predictor. This is mediated, however, by the Gaussian likelihood function that acts as the phonetic category map; if the pair of Gaussians cannot be made to fit the data well, then the model is equally capable of fitting two separate categories, one for each phone, even if the two are in perfect complementary distribution. The constraint imposed by the linear model likelihood is that the covariance of the two phones must be the same, and if this is violated sufficiently severely by two phones, an appropriate phonemic category will not be found. To our knowledge, this is a novel phonetic similarity constraint on allophonic rules, and it is the first that has been explicitly incorporated into a model of phonetic categories. We are not aware of any fine-grained psychophysical data that would suggest that this is unreasonable as a model of phoneme perception, although more work in both production and comprehension of allophonic variants needs to be done to provide further evidence for this constraint.

5. General discussion

We presented three computational experiments examining the ability of statistical models to categorize an unlabeled set of vowel tokens from Inuktitut. We contrasted the simple MOG approach (Experiment 1), which is generally understood as one part in a two-stage process of phonological acquisition, with an alternative approach to categorization that deals with phonological processes and acoustic clustering in the same stage of acquisition (Experiments 2 and 3). By incorporating the process of discovering phonological processes into the process of discovering sound categories, our approach to sound categorization may be said to be a single-stage approach to the acquisition of phonological categories. Rather than learning phones, the model presented in Experiments 2 and 3 settles on abstract, language-specific phoneme categories during the initial process of categorizing perceptual input. The single-stage approach was seen to give a better fit to the target Inuktitut system in Experiment 3, and it had the added benefit of explicitly learning the phonological process associated with uvular retraction in Inuktitut. Simple MOG approaches, on the other hand, appeared to be unable to recover the regular relation between allophonic pronunciations of the Inuktitut vowels for two reasons. For simple MOG models that settled on three phoneme-like categories, it was unclear how information about the subparts of these categories could be extracted from this model. For models that more closely approximated a phone-level categorization of the space, it was seen that the fit with the target phone categories was not close enough to support a second stage of phoneme acquisition based on measures of complementary distribution.

It has long been noted that in both the production and perception of adult speech, language-specific coarticulatory effects are ubiquitous, and the acoustic cues to each phoneme segment's identity may be distributed across multiple segments (see, e.g., Beddor, Harnsberger, & Lindemann, 2002; Manuel, 1990; Nearey, 1990; Öhman, 1966; *inter alia*). A number

of researchers have employed linear regression models to account for these language-specific co-articulatory effects in production and perception (Cole, Linebaugh, Munson, & McMurray, 2010; Nearey, 1990). This work has shown that regressing out predictable effects of phonological context can improve classification (Nearey, 1990), as well as providing greater separation of acoustic clusters (Cole et al., 2010). The models we described here extend this research by examining the impact of these techniques for the problem of language acquisition. It was seen that the resulting mixture of linear models provided a superior categorization of the Inuktitut vowel space, as well as a single-stage model of the acquisition of phoneme categories.

5.1. Two-stage versus single-stage phoneme categorization

We have argued above that there is a widespread, but often implicit, consensus that phonological category learning is essentially a two-stage process: Phone learning is distinct from phoneme learning, and both phones and phonemes constitute separable, discrete levels of categorization. This view does not necessarily entail that infants precisely master all phones before moving on to learning phonology and phoneme categories. It is entirely possible to simultaneously explore the full joint distribution on hypotheses about phonetic and phonemic categorization. The important feature about two-stage models of phonology is that the information made relevant to the two learning problems is different. In models with this property, the first stage cannot make use of all the information available to the second (in this case, information about allophonic environments). In Experiment 1, we provided simulation evidence that this property can severely limit the ability of the learner to recover both the correct phonetic and phonemic categorization of the acoustic space: Errors in one stage are carried through to another and disrupt learning.

In Experiments 2 and 3, we provided simulation evidence of the benefits of treating the problem of learning phonemes as a problem of learning phonetic rules. This view makes the claim that the phonological system includes quasi-continuous phonetic processes in addition to discrete phonological processes, an idea that is not without precedent. The possibility of the coexistence of these two different types of phonological process throughout the stages of phonological processing is alluded to as early as Chomsky and Halle (1968). In their theory, although there is a clear qualitative distinction between the binary, classificatory features used to store morphemes in the lexicon and the scaled numerical features used to represent phonetic information, they write that ‘‘the phonological rules, as they apply to these representations, will gradually convert these specifications to integers’’ (p. 65). However, it is generally the case that research in phonology has been concerned with rules that manipulate binary features only, not scaled phonetic features (exceptions include Cohn, 1990; Dyck, 1995; Sledd, 1966). One result from the simulations in Experiments 2 and 3 is that the choice of whether to treat a process as discrete or continuous can have a significant impact on models of phonological acquisition. Phonetic rules are continuous rather than discrete, and so they have the advantage that they lie in the same representational space as a perceptual phonetic map. In the context of the models we presented, this allowed us to construct a tight dependence between the learning of rules and the learning of categories. These

models perform better than the standard MOG approach in finding phoneme categories, while at the same time capturing the lawful relations between regions of the acoustic space.

There exist alternative interpretations of the results from Experiments 2 and 3. These simulations suggest that, if the bias for small numbers of categories is sufficiently weak, a learner fitting a mixture model might find phonetic categories only, regardless of whether their model allows them to encode phonetic rules. Thus, one might conclude that phonetic-level categories are the only discrete representations in the linguistic system, and that all linguistic encoding is done in terms of phonetic categories. Views like this are sometimes cast as a rejection of the existence of phonemes (Johnson, 1997; Port & Leary, 2005; Silverman, 2006), but they may also be understood as the claim that the lexical level of encoding (traditionally, the phonemic encoding) does not abstract out phonological processes (Kenstowicz & Kisseberth, 1979). There are, however, compelling theoretical and empirical reasons for rejecting this view. A traditional source of evidence for the view of abstract phonemes as the relevant unit of lexical encoding is that a vast majority of languages actively employ alternations of the sort considered here. As suggested above, the productive deployment of allophonic alternations in novel contexts implies that speakers have internalized the knowledge of the lawful relation between segments. If sounds are stored as a single, abstract category that receives its phonetic value only in context, then these basic facts are easily accounted for. In addition, 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 (Kazanina, Phillips, & Idsardi, 2006; Peperkamp, Pettinato, & Dupoux, 2003; Whalen et al., 1997; White, Peperkamp, Kirk, & 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 was only present for speakers for whom it was a phonemic distinction (Russian speakers). In contrast, Korean speakers showed no such discrimination; in Korean, both [t] and [d] occur as regular allophones of a single phoneme. White et al. (2008) obtained related results by studying infants using the head-turn preference task. They showed that infants trained on an artificial language 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 the sequence of phones. These results are also important because the infants did not require meaning to detect the allophonic alternation. These results are 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 more abstract than simple phonetic clusters.

An additional advantage to the single-stage approach to phonological acquisition is that the acquired model provides all the knowledge necessary to deploy the acquired phonological knowledge. This is not true of two-stage models of acquisition. For example, algorithms that cluster phones into phonemes based on distributional facts (as in Peperkamp et al., 2006) give the learner only limited insight into the processes that generate those allophonic distributions. In order to use the phonological system for the

purposes of production or perception, another stage of learning must be invoked to learn the grammatical processes that are responsible for 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 converges on three phoneme categories plus a process that predictably shifts the target pronunciation in front of uvular segments. 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. In models strictly learning phone categorizations, the acquired model would not give learners any insight into the distribution of the phones within the language.

5.2. Extending the single-stage model

The simulations presented in Experiments 2 and 3 provide initial evidence that a mixture of linear models can correctly extract phonological processes and categories in the Inuktitut data set provided. However, there remain a number of limitations to this model that future work will address.

One important issue for the current model concerns the discovery of potential conditioning environments. Although much of the model operated in an unsupervised fashion, the model did not need to determine which tokens were in a uvular context. Instead, the learner was assumed to have knowledge of which tokens occurred in the context of a uvular segment, which was modeled as a categorical contrast collapsing across all uvular phonemes. Furthermore, the model was not required to determine which contextual features were relevant to phonological processes; only uvular environments were considered because they are known to condition retraction in Inuktitut, but presumably this knowledge is not available to the learner and needs to be discovered. This latter problem is easily addressed: Mixtures of linear models are in principle capable of fitting as many contextual effects as there are predictors, and so a more complete model could possibly incorporate predictors for all possible conditioning environments. However, the question of how a categorical conditioning environment is identified in the first place is more difficult. One response to this is that the learner is jointly attempting to categorize all segments in a string, and segments become available as conditioning environments when the learner has categorized them. This view suggests that Inuktitut learners would need to classify their consonant phonemes before they fully arrive at an analysis of the vowel space. Existing experimental evidence, however, suggests that language-specific vowel categories are available slightly earlier than are language-specific consonant categories (Kuhl et al., 1992; Werker & Tees, 1983, 1984). Alternatively, it may be the case that learners are able to assign a categorical feature ‘‘parse’’ to the acoustic string, even if they do not have language-specific consonant representations yet (Hale & Reiss, 2008; Stevens, 1986). If the learner has access to some feature parse of the consonants before they have identified the consonant categories in his or her language, then this information could potentially serve as predictors or conditioning environments in a mixture of linear models. If this is correct,

then learners should be able to acquire the vowel retraction rule in Inuktitut as soon as they categorize the vowel space, possibly prior to identification of language-specific consonant categories. In order to determine which approach is correct, further work exploring the relationship between consonant categorization and cognition of phonological processes is necessary.

The model presented here may also have implications for adult speech perception. If a mixture of linear models is taken as a model of perception, then the predictions differ from those of models that collapse over all subphonemic distinctions. For example, if each phoneme is modeled by a single Gaussian distribution, then speakers should behave as if they have unimodal perceptual ''map'' of an inherently bimodal acoustic surface. On the other hand, models that maintain that each phonetic category is distinct predict that there should be no influence of phoneme identity on cross-allophone perception. That is, in Inuktitut [i] and [e] should be distinguished as easily as [e] and [a]. The mixture of linear models presented here makes the prediction that perceivers should, under ideal circumstances, behave as if the Inuktitut vowel phonemes are a complex, bimodal distribution having two ''good'' exemplar centers (perhaps as measured by the perceptual magnet effect; Iverson & Kuhl, 1996; Kuhl, 1991; Kuhl et al., 1992), but that there should also be an effect of phoneme identity. Without any categorical distinction between [i] and [e], discrimination should be more difficult than for contrasts that differ in categorical identity. Further experimental work is needed to evaluate these predictions.

Finally, the single-stage model presented here could provide a novel way of approaching the problem posed by incomplete neutralization. Phonological rules that collapse distinctions among phonemes are called neutralizing rules; one well-known example is German word-final devoicing of obstruents, which occasionally creates near-homophonous pairs such as *weck* (''wake,'' imperative) and *Weg* (''path''), both ending in a voiceless sound usually transcribed as [k]. The underlying voicing of the obstruent in these words is evident in other morpho-phonological contexts: The plural form *Wege* ''paths'' is pronounced with a voiced velar obstruent [g]. It has been known for some time, however, that this neutralization is not always complete: The final consonant of *Weg*-type words remains phonetically different than *weck*-type words in both production and perception (Port & Crawford, 1989; Port & O'Dell, 1986; Slowiaczek & Dinnsen, 1985), although the effect can be subtle and has at times been controversial (Baumann, 1995; Fourakis & Iverson, 1984). A standard phonological account of word-final devoicing models the process as a categorical change in a voicing feature on the relevant obstruent, which fails to explain the existence of incomplete neutralization. Thus, as suggested by Port and O'Dell, the change appears to be subsymbolic. Under a mixture of linear models account, the incompleteness effect receives a natural explanation: Because the devoicing rule in this model is a shift in the location, but not the scale, of the obstruent's phonetic distribution, the distributions of derived and underlying voiceless obstruents will not overlap completely. The divergence in these distributions may be responsible for better than chance performance at discriminating true from derived word-final voiceless consonants (Port & Crawford, 1989). When specified with a model of actual German speech, a mixture of linear models approach to this process would make firm predictions about which tokens of devoiced consonants

German speakers should be able to perceive as derived. It remains to be seen if this is an accurate model of the phenomenon of incomplete neutralization, and future work will address this issue.

6. Conclusion

In describing a model of phonological category acquisition, the desired end state is a set of phoneme categories: Sound categories used in lexical storage that may include several distinct allophones. Work in phonological category acquisition has tended to focus either on the problem of finding phones in acoustic space or on the problem of finding systematic relationships between phone categories. In this article, we suggested an alternative model that directly acquires phoneme categories by jointly learning acoustic distributions and the relationships that hold between them. Using data from the Inuktitut vowel space, we showed that this model provides a better fit to the data and has the advantage of arriving at the desired phonemic categorization of the Inuktitut vowel space in a single step. This provides initial support for a single-stage model of phoneme acquisition and further demonstrates the usefulness of the mixture model as a model of category acquisition in human language.

Notes

1. Relative frequencies of each phone in an extract from the Inuktitut Hansard corpus (version 2.0; Martin et al., 2003), with orthographic vowel–‘‘q’’ and vowel–‘‘r’’ sequences taken to be tokens of retracted vowels, were [i], 0.31; [e], 0.05; [u], 0.24; [o], 0.04; [a], 0.29; [ɑ], 0.07. In contrast, the relative frequencies in the phonetic corpus were [i], 0.31; [e], 0.08; [u], 0.18; [o], 0.17; [a], 0.15; [ɑ], 0.12. For the low-frequency phones, therefore, we were able to draw on a relatively robust sample to construct our training sets.
2. A matrix normal distribution is a generalization of the multivariate normal distribution (in which each of the elements of a vector are normally distributed) to a matrix in which the columns of the matrix are normally distributed with some column covariance matrix, and the rows are normally distributed with some row covariance matrix. In the current context, the row covariance matrix can be seen as a parameter controlling the dispersion of the category locations (first row) throughout the space and the similarity of the phonetic rules (second row) to a common mean (but the two are not necessarily independent); note that this parameter was learned. For sampling purposes, the matrix normal distribution has the useful property that the vectorization of a normally distributed matrix follows a multivariate normal distribution with covariance equal to the Kronecker product of the two covariance matrices. See Dawid (1981) for details.

Acknowledgments

This work was supported in part by NSF IGERT DGE-0801465 to the University of Maryland, by NIH 7R01DC005660-07 to David Poeppel and William Idsardi, and by SSHRC Doctoral Fellowship 752-2011-0293 to Ewan Dunbar. We extend special thanks to Derek Denis and Mark Pollard for sharing their Inuktitut recordings, and to Alana Johns for further advice on Inuktitut. We are grateful to Jordan Boyd-Graber, Hal Daumé III, Naomi Feldman, Jeff Heinz, Jeff Lidz, Joe Pater, and Colin Phillips for their useful discussion and insight on the issues contained in this paper. The authors take all responsibility for errors.

References

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*, 461–486.

Baumann, M. (1995). *The production of syllables in connected speech*. Unpublished PhD dissertation, University of Nijmegen.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*, 591–627.

Best, C. T. (1995). Learning to perceive the sound patterns of English. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (pp. 217–304). Norwood, NJ: Ablex.

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

Boersma, P., Escudero, P., & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 3–9 August, pp. 1013–1016.

Boersma, P., & Hayes, R. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, *32*, 45–86.

Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*, B69–B77.

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. Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 06), Boston MA, July 16–20, 2006, pp. 1451–1456.

Cohn, A. (1990). *Phonetic and phonological rules of nasalization*. Doctoral dissertation, UCLA.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167–184.

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, *68*, 265–274.

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, Toronto, 22–23 March.

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, UK: Cambridge University Press.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. Hoboken, NJ: John Wiley and Sons.
Duong, T. (2011). ks: Kernel smoothing. R package version 1.8.2. http://CRAN.R-project.org/package=ks (Accessed 20 May, 2011).
Dyck, C. 1995. *Constraining the phonology-phonetics interface with exemplification from Spanish and Italian dialects*. Doctoral dissertation, University of Toronto.
Escobar, M., & West, M. (1995). Bayesian density estimation and interference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.
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, Amsterdam, 29 July–1 August.
Fenson, L., Dale, P., Reznick, J. S., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*, 1–185.
Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230.
Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, *18*, 7–44.
Fourakis, M., & Iverson, G. (1984). On the ''incomplete'' neutralization of German final obstruents. *Phonetica*, *41*, 140–149.
Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith, (Eds.), *Bayesian Statistics 4*. Oxford, UK: Clarendon Press.
Goldsmith, J. A. (1976). *Autosegmental phonology*. Bloomington, IN: Indiana Linguistics Club.
Goldsmith, J., & Xanthos, A. (2009). Learning phonological categories. *Language*, *85*, 4–38.
Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. Proceedings of the 29th Annual Conference of the Cognitive Science Society, Nashville, 1–4 August.
Hale, M., & Reiss, C. (2008). *The phonological enterprise*. Oxford, UK: Oxford University Press.
Harris, Z. (1951). *Methods in structural linguistics*. Chicago, IL: University of Chicago Press.
Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater, & W. Zonneveld, (Eds.), *Fixing priorities: Constraints in phonological acquisition* (pp. 158–203). Cambridge, UK: Cambridge University Press.
Iverson, P., & Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of/r/and/l/. *Journal of the Acoustical Society of America*, *99*, 1130–1140.
Jakobson, R. (1941). *Child language, aphasia and phonological universals*. The Hague, The Netherlands: Mouton.
Jaynes, E. T., & Bretthorst, G. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.
Johns, A. (2009). Additional facts about noun incorporation (in Inuktitut). *Lingua*, *119*, 185–198.
Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *Ohio State Working Papers in Linguistics*, *50*, 101–113.
Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.
Jusczyk, P. W. (1985). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing, confusion* (pp. 199–229). Hillsdale, NJ: Erlbaum.
Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sound contrasts. *Proceedings of the National Academy of Sciences USA*, *103*, 11381–11386.
Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell.
Kenstowicz, M., & Kisseberth, C. (1979). *Generative phonology*. San Diego, CA: Academic Publishers
Kuhl, P. K. (1991). Human adults and human infants show a ''perceptual magnet effect'' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*, 93–107.

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.

Kuriyagawa, F. (1984). The features of/k/and/q/in Cairo Standard Arabic. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics (RILP), University of Tokyo*, *18*, 65–73.

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

Lin, Y., & Mielke, J. (2008). Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics*, *14*, 241–254.

Manuel, S. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, *88*, 1286–1298.

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and using an English-Inuktitut parallel corpus. Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, 31 May.

Maye, J., Daland, R., & Goldrick, M. (2008). Phonological context as a cue to phonetic identity. Paper presented at the 2008 Annual Meeting of the Linguistic Society of America in Chicago, IL, January 3–6.

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

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

McMurray, B., Aslin, R., & Toscano, J. (2009). Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science*, *12*, 369–378.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347–373.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*, 349–357.

Ohala, J. (1976). A model of speech aerodynamics. *Report of the Phonology Laboratory, Berkeley*, *1*, 93–107.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

Pasquale, M. (2009). Phonological variation in a Peruvian Quechua speech community. In J. Stanford & D. Preston (Eds.), *Variation in indigenous minority languages*. Amsterdam: John Benjamins.

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.

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, MA: Cascadilla Press.

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

Port, R., & Crawford, P. (1989). Pragmatic effects on neutralization rules. *Journal of Phonetics*, *16*, 257–282.

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

Port, R., & O'Dell, M. (1986). Neutralization of syllable-final devoicing in German. *Journal of Phonetics*, *13*, 455–471.

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

Pulleyblank, D., & Turkel, W. (1998). The logical problem of language acquisition in Optimality Theory. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, & D. Pesetsky (Eds.), *Is the best good enough? Optimality and competition in syntax* (pp. 399–420). Cambridge, MA: MIT Press.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rischel, J. (1974). *Topics in West Greenlandic phonology: Regularities underlying the appearance of wordforms in a polysynthetic language*. Copenhagen: Akademisk Forlag.

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

Seidl, A., Cristià, A., Bernard, A., & Onishi, K. (2009). Allophones and phonemes in infants' phonotactic learning. *Language, Learning, & Development*, *5*, 191–202.

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

Sledd, J. H. (1966). Breaking, umlaut, and the southern drawl. *Language*, *42*, 18–41.

Slowiaczek, L., & Dinnsen, D. (1985). On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics*, *13*, 325–341.

Stevens, K. N. (1986). Models of phonetic recognition. II. A feature-based model of speech recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal Satellite Symposium on Speech Recognition, Twelfth International Conference on Acoustics* (pp. 67–68).

Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, *29*, 229–268.

Vallabha, G., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273–13278.

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.

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.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. ISDS Discussion Paper 92-A03, Duke University.

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.

White, K., Peperkamp, S., Kirk, C., & Morgan, J. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*, 238–265.