Investigating Statistical Approaches to Building a Phonology

Brian Dillon and William Idsardi

September 30, 2009

1 Introduction

In recent years, statistical approaches to language acquisition have generated much enthusiasm, especially in the domain of phonological and phonetic learning. According to some of its proponents, these powerful approaches may supplant the need for much prior knowledge, being able to discover, for example, the number and identity of the vowels of a language. Encouraging experimental results in acquisition and successful computational models make it hard to deny the importance of these emerging techniques, and phonologists ignore this literature at their peril. In this chapter we review some of the most influential work in this area, and suggest ways in which current trends in machine learning and statistical approaches to sound categorization can be integrated with long-standing observations in phonetics and phonology. We present a computer simulation of the acquisition of the Inuktitut vowel space that speaks to the feasibility of this approach as a way of modeling the entire acquisition process, from perception to phonology.

2 Statistical Approaches to Phonological Learning

A good deal is known about the time course of phonological development. Very young infants have famously been termed universal listeners, being able to discriminate amongst a wide range of sounds not present in their input (see, e.g., [1]). A number of studies have shown that these discriminatory abilities soon decay as the infant develops. Declining sensitivity to non-native language vowel contrasts is apparent for vowels as early as 6 months [16], and by 8 months, similar effects are evident in consonant contrasts [33]. A common observation about this arc of development is that it seems to suggest that phonological category learning could in fact drive the building of a lexicon, rather than the other way around as had been previously hypothesized [2],[15]. Given that infants can reliably discriminate minimal pairs only later (around 18 months or so), it appears that at this age, higher-level constructs such as minimal pairs do not drive the development of contrast [7].

Maye and Werker (2002) hypothesized that distributional learning mechanisms are the fundamental building block of phonological development, a hypothesis that is fully compatible with the observed time-course of acquisition [19]. Building on results that show that infants are sensitive to distributional information in other modalities, such as word-learning [28], they aimed to show that infants used distributional cues to bootstrap contrastive categories from the input. Infants at both 6 and 8 months of age were exposed to training sets that contained either bimodal or unimodal distributions over voice-onset times (see Figure 1). When presented with bimodal distributions, infants showed greater sensitivity to differences between the modes in the VOT distribution, whereas for the infants in the unimodal condition, sensitivity was decreased for the same sounds. The authors interpreted these findings as showing that distributional characteristics of the input directly impacted what dimensions of the signal the infants viewed as relevant or contrastive. Werker. Pons, Dietrich, Kajikawa, Fais & Amano (2007) went on to show that, for the vowel space, there are distributions in actual speech that support the acquisition of phonological contrasts. They showed that in infant-directed speech of both Japanese and English speakers, clear distributional cues support the relevant contrasts (i.e. duration cues for the Japanese vowel space), and minimize irrelevant dimensions of variation [32].

In addition to these experimental results, computational models have made increasing inroads in the explanation of how speech categories are acquired using various types of statistically-informed frameworks.

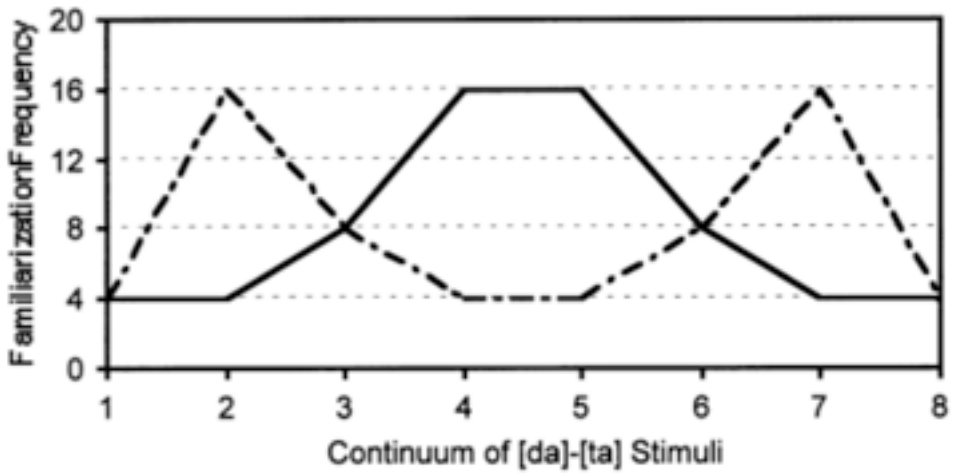

Figure 1: Bimodal and unimodal distributions for familiarization, reproduced from (Maye, Werker and Gerken 2002)

Drawing from the literature on machine learning, many researchers have chosen to characterize the problem of phonological acquisition as a difficult example of an unsupervised clustering problem. Clustering techniques come in many different forms, but all share the notion of carving out subsets of a data set and classifying them together based on a pre-defined distance metric in a given representational space. In Figure 2 we reproduce the famous figure from [24] From this point of view, the phonological learning problem amounts to finding the enclosing ovals given only the unlabelled points.

An example of this cluster analysis approach is the work of de Boer and Kuhl [5], who used a mixture of Gaussians (MOG) model to model the acquisition of vowel categories in the metric space defined by the first two formants (i.e. the typical F1 by F2 vowel plot). MOG models represent category structure as a set of parameterized Gaussian distributions (termed components of the model) in the input space, weighted by a mixing frequency (for further explication and discussion of MOG models, see [21]). In this model, Gaussians were fit with an Expectation-Maximization algorithm (EM, see Demp, Laird & Nan 1977), which is a family of hill-climbing algorithms that seeks to maximize a measure of likelihood for an unobserved category structure. Both the EM algorithm and the MOG approach have received a good deal of attention in the statistical and machine learning literature for the last two decades, and their properties are relatively well-understood (Everitt et al 2001, [11]). De Boer and Kuhl applied these techniques to vowels that were recorded during mother-child interactions, focusing on the vowels at the extreme edges of the vowel space in English (/i/, /u/, and /a/) and limiting the MOG models to three-component mixtures only (that is, they pre-specified the number of components in the model and supplied this prior information to the cluster-fitting procedure). By applying this approach, and clustering within speakers, 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.

Other clustering approaches to learning the vowel space have similarly met with a good degree of success. Coen 2006 analyzed video-recorded samples of American English vowels, and used a cross-modal clustering technique to form and cross-correlate clusters in both acoustic and visual space (i.e. shape of mouth) [4]. Using this technique, he was able to extract a good portion the vowel categories of American English, an important result. This result is perhaps the most dramatic success in the modeling of phonological acquisition: the complete American English vowel space was induced with no prior knowledge of the number of clusters, and without any parameterization of the distributions. Note that this algorithm relies crucially on visual information, a feature that is not obviously relevant to infants learning their phonology, and is obviously not available to blind children [17]. Another important contribution to the modeling of vowel acquisition was the Online EM model of Vallabha, McClelland, Werker, Pons, and Amano 2008 [31]. Using both parameterized and non-parameterized versions of an online clustering algorithm (i.e. one that is updated

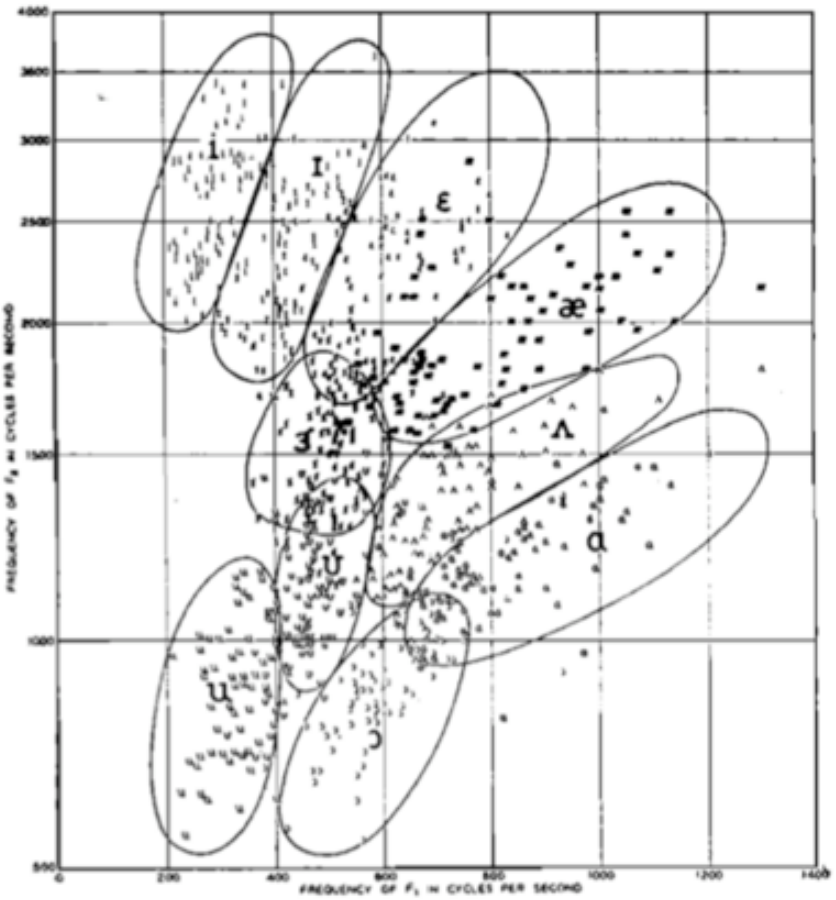

FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

Figure 2: American English vowels, reproduced from Peterson and Barney 1952

after the presentation of each data point), Vallabha et al were able to model the acquisition of the Japanese and English data that was analyzed by Werker et al [32]. Their model fit four-dimensional Gaussians in the raw acoustic space (F1, F2, F3 and duration), updating its weighting for relative weighting of the parameters after each input point. Applying this technique within speakers, they were able to achieve a reasonable rate of success for most speakers, successfully learning the vowel systems 80% of the time for parameterized models, and 60% of the time for non-parameterized models. This model is important in that it is the most successful version of an online algorithm, processing each data point as it occurs, as opposed to the batch algorithms that try to estimate categories based on a large store of exemplars stored in memory in an offline fashion.

In the domain of phonological acquisition of consonants, there have been similar successes in applying machine learning techniques to the problem of category induction. Vallabha and McClelland (2007), for example, modeled experimental results on adult acquisition of a non-native contrast (here, Japanese speakers learning the /r/-/l/ distinction), using neurally-inspired models employing a variant of Hebbian learning.

3 Learning Algorithms: Input and Output

Thus, there are a number of experimental and computational results that suggest that the problem of phonological acquisition relies on statistical generalizations in some form. Some proponents of these techniques may view these successes as obviating the need for extra sources of knowledge in the process. We suggest, however, that instead of using statistical methods to essentially reductionist ends, these techniques can and should be integrated with what linguists take to be important generalizations about phonological representations. As a case in point, consider the acquisition of the vowel space. The vowel space serves as an excellent model, because its representational parameters are well understood, and finding structure in unlabeled vowel data is a notoriously difficult problem.

There are three essential parts to any potential learning algorithm: an input space, a transform (perhaps the identity function) on that space, and an output (the categories acquired). Much of the work in computational modeling of phonological acquisition has focused on refining our notion of a desirable transform that a learner might employ to induce category structure. Here we suggest that this is a myopic way to view the problem, and that careful consideration and refinement of the input and output to a learning model may be essential to successfully characterizing the learning procedure. What constitutes input and output for any given procedure represent what assumptions and knowledge the modelers are comfortable building into the model.

Consider first the input to the learning algorithm, an aspect of all models for acquisition that is fraught with assumptions about what human learners can and do compute from an acoustic signal. The acoustic input is a rich, high-dimensional signal that contains a great deal of extraneous information. Many models either explicitly or tacitly assume that the input dimensions are those that have been assumed to be relevant for speech for half a century or more (e.g., [24]). In the case of the vowels, for example, they generally employ the first three formants, and occasionally duration. While it is true that these dimensions are a reasonable place to begin to classify vowel sounds, and this information is transmitted from the cochlea to the auditory nerve, it is no innocent assumption that the learner can limit their attention to these dimensions and not, say, zero-crossing rate and amplitude, which are also relayed to auditory cortex and could well be relevant to other speech categories, such as fricatives (which as a class have a high zero-crossing rate). It may be the case that the learner employs some variant of feature selection, a family of machine-learning dimensionality-reducing algorithms that aim to strip features that are irrelevant for classification out of the input data set (an overview can be found in [14]). Here, mechanisms for determining the contrastiveness of a feature could be determined from the input data itself, though to our knowledge no model has yet shown that when given a richer set of acoustic data, all and only the correct representational dimensions will be converged upon. Thus it remains possible that models that consider all cochlear dimensions in the input space risk building spurious categories that correspond to any distributional regularities in any given dimension. While there may not be such regularities in the speech of any given adult speaker, if infants modeled these dimensions in their assessment of the distribution of acoustic features, it is all too likely that slight perturbations and random variation on these dimensions would give way to entrenchment of category structure in these dimensions across generations of learners. Indeed, this appears to be exactly the situation that occurs in situations of tonogenesis, where a language slowly develops contrastive use of pitch on lexical items [3]. However, no such emergence of categories built on amplitude and zero-crossing rate appears to ever happen in natural language.

The choice of input representation may also be leveraged to achieve greater success in modeling real-life acquisition situations. As we discuss below, the problem of cross-speaker variation is quite real for the learner. An untransformed formant space spreads the vowel categories of male, female, and child speakers across very different portions of the space [24], and makes it very difficult to find unifying categories. This is not the case with formant space that is built on ratios of formants relative to F3, as discussed in by Monahan and Idsardi [22]. They present neurophsiological (magnetoencephalography, MEG) evidence that the formant ratio F1/F3 (or something substantially similar to this) is computed early in human auditory cortex, by approximately 100ms, thus suggesting a powerful cross-speaker normalization mechanism is available to learners. Given such a representational space, it is reasonable to expect that learners might not be too troubled in forming categories that span different speakers. Here, judicious choices about input data representation may actually significantly impact the ability of models to handle more natural data. Correctly characterizing the input space requires making explicit assumptions about what is and is not represented

in possible phonological grammars, and our understanding of the processing abilities of the human auditory system thus provides an important upper bound on our capacity to model the acquisition process.

On the other hand, one must carefully consider the desired output of a learning algorithm: what exactly is the nature of the system that the learners are attempting to build? A common view is that the categories are represented as parameterized distributions, a view that is supported by experimental evidence on sound categorization [29]. Modeling a category as a general multi-dimensional Gaussian distribution requires estimating the mean and (co)variance for each of the dimensions, therefore a total number of $n+n\hat{2}$ values that need to be estimated for an n-dimensional distribution. Non-parametric estimates typically require a greater number of values to be calculated, and accordingly, require more data for accurate estimation of category structure. Even compact, parameterized distributions may not be appropriate models for phonological categories, however, if one maintains the assumption that the distributions are estimated independently. Recent experimental results suggest that the relevant process is best characterized as acquired sensitivity of a particular feature contrast [18], rather than simple identification of modes in a multidimensional acoustic space. Maye et al were able to show that infants who were given training sets that induced sensitivity on a particular VOT contrast at the coronal place of articulation generalized the same contrast to the velar series, and vice versa. Likewise, Dietrich et al [7] showed that a lexically contrastive distinction in vowel duration present in some Dutch vowels caused children to view duration as a lexically contrastive feature more generally, as opposed to English-speaking children who did not. These findings demonstrate that it is also important to note that there are important generalizations that learners build across distributions that are particular to specific speech categories. That is, once a learner finds that a particular dimension of an acoustic signal is contrastive, they are more willing to apply that contrast to pairs of sounds for which they do have not observed evidence of the contrast being operative. Being able to account for results such as these requires more closely considering the assumptions of the desired end state. More generally, models need to lay plain what the desired end state is: distributions that correspond to single sound categories? Estimates over some kind of feature space?

4 Input and Output in Learning Inuktitut Vowels

Here we consider the effect that input and output can have on one outstanding problem in phonological acquisition: allophony. With a few notable exceptions ([23], [25], among others), this problem has not been addressed by proponents of statistical approaches to phonological acquisition. If one takes the view that phonological contrast is derived from statistical regularities in the input, then it is essential that all and only the correct contrasts are acquired. A case in point is the Inuktitut vowel space, where the acoustic clusters are not isomorphic to the resulting phonological categories.

Underestimating the number of categories in the data (a miss in signal detection terminology, a false negative in medical testing parlance) is the most common way learning algorithms are said to fail at their task. However, failure may just as easily arise from spurious postulation of structure in the vowel space, thereby overestimating the number of categories in the data (false alarms, false positives). This is a very real possibility, as seen in the vowel space of Inuktitut, an Eskimo-Aleutian language spoken in Northern Canada. The vowel inventory of Inuktitut, like many related languages, consists of three vowel phonemes: /i/, /a/, and /u/ ([8]; [30]). In these languages there is a productive process of vowel lowering, whereby vowels are lowered (and backed) in the context of a uvular consonant ([6]; see also [27], [9] for a description of related West Greenlandic processes), plausibly as a result of a retracted tongue root, [RTR]. Thus in a form like /surusiq/ boy, the final phoneme /i/ is pronounced as [e] (similar changes obtain in the same environment for other vowel phonemes /a/ and /u/). This regular, allophonic change has a clearly bimodal reflex in the acoustics, as evident in the clusters around the phones [i] and [e] of the phoneme /i/ in Figure 4 (based on the dataset from [6], see below).

Note that any algorithm that attributes contrastive status to any distinct modes in the input is in danger of being led astray by the presence of modes for both [i] and [e] when they are considered on their own. Put differently, given the current input (independently sampled vowel tokens), we will fail to reach our desired output (three vowel categories) using any number of transforms. Note, however, that the input under discussion incorporates a very strong assumption, that of independence from other tokens and its environment. Given rampant co-articulation effects in language, this may be an unrealistic assumption. If

instead the learner tracks the correlations that hold among tokens in the input, she may be able to undo the effect of being adjacent to a uvular segment by offsetting the spectral characteristics of a given token by a fixed amount when it is in the environment of a uvular. In the simulations below, we consider two types of input for the Inuktitut vowel space: untransformed F1-F2 values, and F1-F2 values that factor out the average effect of a vowels environment. We will see that reconsidering the input space in this manner allows us to immediately overcome the problem of spurious category formation.

4.1 The Model

The transform from input to output that we employ is a mixture of Gaussians, intended as an abstract model of the learner. Such mixture models assign a probability to an observation x_i from a set of n d-dimensional observations $\{x_1 ... x_n\}$ through a combination of constituent probability distributions, as in (1) (for reference, see [21; 13]).

$$f(x_i|\Psi) = \sum_{k=1}^{K} \pi_k \phi(x_i|\theta_k) \tag{1}$$

$$\Psi = (\pi', \theta')' \tag{2}$$

The model acts to assign a probability to any observation x_i by summing the probability of that observation conditioned on each of the K components (i.e. $\phi(x_i|\theta_k)$ for each component k), weighted by that componentś *mixing probability* (π_k). Together, the vector of mixing probabilities π and the parameters for each component θ are designated by Ψ, a vector that represents all unknown parameters. With such a model, there a number of ways in which a mixture model can be used to assign a label to incoming unlabeled data. In what follows, we employ *Maximum A Posteriori* (MAP) Estimation, assigning unlabeled tokens to the single most likely component.

With this model of the output-a set of component categories in a mixture model-the learning problem is simply the estimation of the parameters Ψ. In the case of Gaussian mixture models, this consists of K components, each with its own mean (μ_k) and variance (Σ_k). With models of this sort, the general learning algorithm that is employed in the machine learning literature is Expectation-Maximization (EM) (see [21], chapter 2). The EM algorithm for fitting a mixture consists of two steps. The first, the *E-step*, takes an initial parameter estimate $\hat{\theta}$ and uses it to compute a matrix Z in which each entry z_{ik} corresponds to the conditional probability that observation x_i was generated by the k^{th} component. Z is then used to compute updated parameter estimates which maximize the likelihood (or in this case, the BIC), as a function of changes in μ_k, Σ_k, and π_k. In the simulations reported below, the number of components was estimated prior to fitting mixture models using a bootstrap approach. All simulations were conducted using the R statistical computing environment [26], using the MCLUST package [10; 11; 12].

4.2 The Data

All data considered in this simulation derived from an original data set that consisted of 239 Inuktitut vowel tokens. These tokens were elicited from an adult female speaker of the Cape Dorset dialect by Derek Denis and Mark Pollard [6]. The formant structure was measured by hand, and each token was labeled with F1 and F2 values. In what follows, we refer to the original 239 tokens in F1-F2 space as the *untransformed* data set.

The *uvular-corrected* data set, by contrast, represented a transform that aimed to factor out the coarticulatory effects of uvular segments on vowels (a weakening of the independence assumption). The transform that we applied was a constant spectral correction that was applied to all vowel tokens in the context of a uvular segment, and it was calculated as follows. We pooled all vowel tokens, and marked them with as either adjacent to a uvular segment (in Inuktitut, this means /q/,/ʁ/, or /ɴ/) or not. We then fit two linear regressions that modeled F1 and F2 values of all vowel tokens as a function of the presence or absence of an adjacent uvular segment. The spectral correction for each formant was simply the slope estimate for each of these regression. This correction-applied to all vowels in the context of a uvular segment-was -92.27 Hz for F1 and 374 Hz for F2. Importantly, this was a vowel-general correction, and did not require any knowledge of the category status of individual vowels. Thus, the calculation of this factor can be done before the vowel

space is sub-divided into categories (but it does require the splitting of the consonantal environments into uvular and non-uvular categories). The corrected dataset is plotted in Figure 5. This method clearly does an excellent job of correcting for the uvular influence on /i/ as now [i] and [e] are nearly coincident. The improvement with [o] and [u] is not immediately noticeable, but as we will see, it has a significant effect on the statistical clustering results.

It should be noted that this procedure obviously requires the identification of the [uvular] class of consonants in order to perform the corrected measures. Although even very young infants are quite good at distinguishing classes of plosive consonants, whether or not they can categorize places of articulation across various manners of articulation is very much an open question. The present research suggests that it may be this information may be very useful in later phonological acquisition.

4.3 Bootstrap Analysis for Number of Categories

Rather than allow EM to automatically determine the number of components K that best fit a given data set, we employed bootstrap methods to estimate the level of significance for each K given the data x. In order to do so, we generated bootstrap distributions of the log likelihood ratio, following [20]. Once the bootstrap distribution was obtained, all tests took the form of a simple hypothesis test: the null hypothesis H_0 is that the data is best fit by a K-component mode, and the alternative H_1 represents a *K+1*-component model. By comparing a the log likelihood of the data given a *K+1*-component model to that of a K-component model, we obtain an estimate of the level of significance p relative to the bootstrap distribution of the log likelihood ratio. In other words, if the resultant p-value reaches a preset α, we can determine that the data is significantly better fit by *K+1* components. This procedure is iterated with increasingly larger K until a K is found for which the null hypothesis H_0 cannot be rejected. This value of K is the lowest number of components for which we cannot reject H_0, and the number of components that best fits the data.

K	p
1	0.00
2	0.00
3	0.00
4	0.00
5	0.51

Table 1: Estimated p-values for number of components K, as determined by bootstrap of the log-likelihood ratio (B = 100), using untransformed data.

K	p
1	0.00
2	0.00
3	0.46

Table 2: Estimated p-values for number of components K, as determined by bootstrap of the log-likelihood ratio (B = 100), using uvular-corrected data.

The results are clear. On untransformed data, the data is best described, reliably, by five mixture components, presumably corresponding to the five acoustic phones in the F1-F2 space. In contrast, the simple spectral correction for coarticulatory effects we applied reduced the number of components to three, the number of phonemic categories that are present in Inuktitut. This suggests that the transform has the desired effect of reducing spurious categories in the Inuktitut vowel space. In what follows, we investigate the impact of this reduction on goodness of fit, relative to the desired output of three phonemic categories for Inuktitut.

4.4 Mixture Models of the Inuktitut Vowel Space

The results of the bootstrap analysis suggest that for the untransformed data, five components are the best fit, and that for the uvular-corrected data, three components are the best fit. This suggests that on the untransformed data, each allophone will receive its own mixture component, and for the uvular-corrected data, it is the phonemes themselves that are most immediately evident in the data. In this section, we further break down this result by examining how both three- and five-component mixture models best fit both data sets.

The analysis proceeded as follows: for each data set, a model was fit to the F1 and F2 values for each of the tokens in the two data sets, using EM as described above. The number of components was fixed prior to fitting, in accordance with the results of the bootstrap analysis. The two resulting best fit models are plotted in figures 3-4, showing the alignment of the best-fit model components against the allophonic clusters. To assess how well these models fit the actual labels of the data, we constructed a confusion matrix C from the modelś classification of the data. $C(m,n)$ was the number of tokens from actual category m that were classified as mixture component n. To find which model component best corresponded to each of the phoneme or allophone labelings, rows were reordered to maximize the trace of the confusion matrix. Using the optimal alignment of mixture components and labelings, the precision and recall were calculated for both phonemic classification (for the three-component model), and for allophonic classification (for the five-component model).

	/i/	/a/	/u/
Precision	.963	.718	.972
Recall	.859	.968	.833

Table 3: Precision and recall for phoneme classification, for the best fit three-component models using uvular-corrected data.

	[i]	[e]	[a]	[o]	[u]
Precision	.967	.391	1.00	.560	.484
Recall	.806	.900	.571	.902	.348

Table 4: Precision and recall for allophone classification, for the best fit five-component model using untransformed data.

Visual inspection of the alignment between model components and actual phonemic or allophonic labeling suggests that untransformed data does indeed give rise to spurious allophonic categories. This is most strikingly true for the pair [i]-[e], and this observation is confirmed by the recall scores on each of these categories in Table 4.4. These recall scores suggest that the [i] and [e] tokens were to a large extent classified into distinct model components. Overall, for the five-component model, good fits are obtained with the allophones [i], [e] and [o], but cluster 5 models [a] poorly, and cluster 4 contains a mixture of various tokens, not just [u].

For the three-component model on transformed data, the correct phonemic structure is immediately apparent, as evidenced by the three clusters that align almost perfectly with the actual phonemic categories. As seen in , the correspondence between the induced categories and the actual phones and phonemes is not perfect. Cluster 3 is somewhat higher than actual /a/, and cluster 2 lies between [o] and [u], but models /u/ extremely well.

4.5 Discussion

A most subtle question is whether or not speakers (and listeners) of Inuktitut are aware of the statistical phonetic aspects of both individual allophones and phonemes. The procedure outlined here allows us (with prior knowledge of some special corrective procedures) to go directly to the identification of phoneme categories, without an allophonic way-station. It remains to be seen if this is actually what speakers do, but

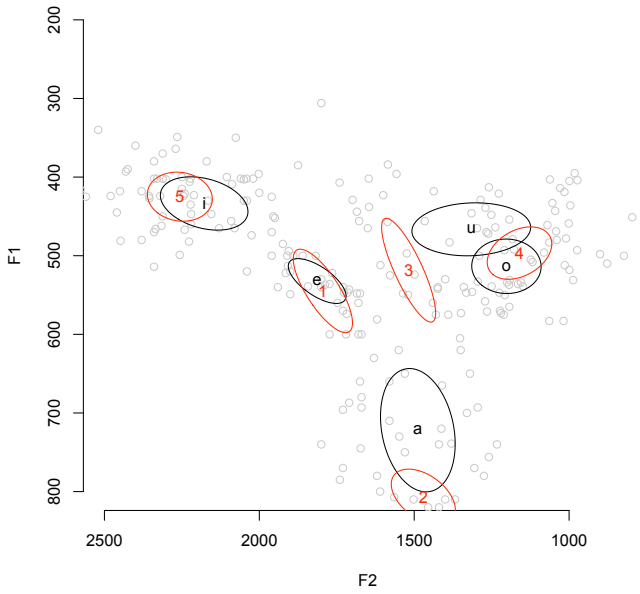

Figure 3: Best fit for untransformed training set: 5 component model

these findings in this small simulation are certainly intriguing.

Inuktitut allophony is a clear example where clustering on the raw data, as suggested by Vallabha et al will not reveal the phonemic categories of the language, though it does approximately reveal the allophones. In addition, a learner that attempts to use statistical modes in acoustic space to build contrast (as in [19]) will induce an incorrect contrast in the case of Inuktitut. Applying some prior phonological knowledge about the general nature of allophony and co-articulation allows us to cluster in a corrected vowel space, which then directly reveals the phonemic structure of the vowel space. An additional benefit of this way of approaching the learning procedure is that not only will the category structure be learned, but some degree of positional information is implicitly present in the learnerś end state. Any learner that learns the allophones separately has to invoke extra procedures to guarantee that [e] will only surface in the context of a uvular segments; the knowledge about positional restrictions is built into the learning process in the current model.

5 Conclusion

In this chapter we have shown that acoustic clusters are not always homomorphic to contrastive categories in a given languages phonology. We suggested a way in which the problem may be avoided, by invoking a mechanism that prevents the spurious postulation of acoustic categories. This amounts to making a more conservative category induction procedure, but this is not the only choice that has been proposed to solve this problem. Another choice is to invoke additional procedures to undo spurious divisions made by the first-pass clustering algorithm (as in [23]). The differences in these approaches make clear predictions about the time course of the acquisition of allophonic relations: the latter predicts a u-shaped curve where allophones are initially treated as phonemes, whereas the predicts that Inuktitut children would never develop a contrastive distinction between [i] and [e] at any point. We hope (and predict) that further simulations, coupled with psycholinguistic and neurophysiological experiments will reveal which of these possibilities is more likely.

More generally, the results presented here demonstrate the importance of considering the input and output in modeling phonological acquisition. We demonstrated that different assumptions about the input lead to drastically different results. In the case of Inuktitut, the desired end state of three phonemic categories is most evident when coarticulatory effects are factored out prior to clustering. Although it is unclear how

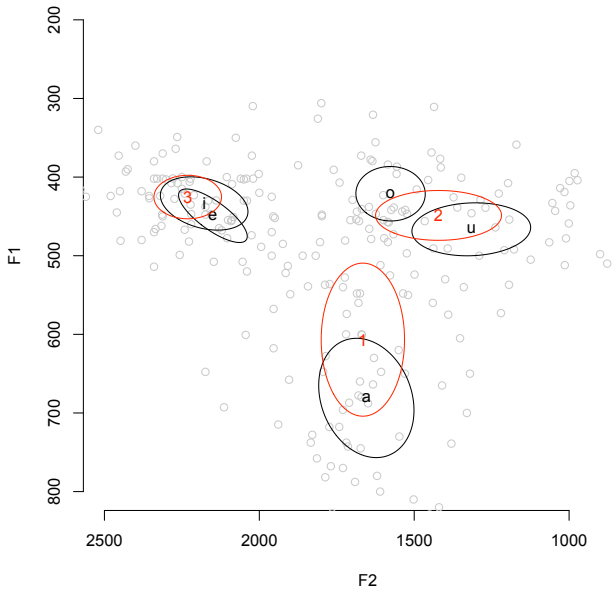

Figure 4: Best fit for uvular-corrected training set: 3 component model

general this particular solution will prove to be, it is clear that more sophisticated notions of the input the learner receives and the output the algorithm produces will play an essential role in furthering computational models of phonological acquisition.

References

[1] Aslin, R. N., Jusczyk, P. W., and Pisoni, D. B. Speech and auditory processing during infancy: constraints on and precursors to language/. New York, NY: Wiley, 1998, pp. 147–254.

[2] Best, C. T. Learning to perceive the sound pattern of English. Norwood, NJ: Ablex, 1995, pp. 217–304.

[3] Bye, P. Evolutionary typology and Scandinavian pitch accent. Kluwer Academic Publishers, 2004.

[4] Coen, M. Self-supervised acquisition of vowels in American English. In Proceedings of the Twenty First National Conference on Artificial Intelligence (AAAI'06). (Boston, MA, July 2006).

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

[6] Denis, D., and Pollard, M. A phonetic analysis of the Inuktitut vowel space. In Inuktitut Linguistics Workshop (University of Toronto, Toronto, Ontario, March 2008).

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

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

[9] Fortescue, M. West Greenlandic. Croon Helm, 1984.

[10] Fraley, C., and Raftery, A. mclust: Model-Based Clustering / Normal Mixture Modeling, 2008. R package version 3.1-10.

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

[12] Fraley, C., and Raftery, A. E. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, September 2006.

[13] Frühwirth-Schnatter, S. Finite Mixture and Markov Switching Models. New York, NY: Springer, 2006.

[14] Guyon, I., and Elisseeff, A. An introductino to variable and feature selection. The journal of machine learning research 3 (2003), 1157–1182.

[15] Jusczyk, P. W. On characterizing the development of speech perception. Hillsdale, NJ: Erlbaum, 1985, pp. 199–229.

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

[17] Landau, B., and Gleitman, L. R. Language and experience: Evidence from the blind child. Cambridge, MA: Harvard University Press, 1985.

[18] Maye, J., Weiss, D. J., and Aslin, R. N. Statistical phonetic learning in infants: Facilitation and feature generalization. Developmental Science 11 (2008), 122–134.

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

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

[21] McLachlan, G. J., and Peel, D. Finite Mixture Models. New York, NY: Wiley, 2000.

[22] Monahan, P. J., and Idsardi, W. J. Early auditory sensitivity to formant ratios in vowel perception: Meg evidence. In Proceedings of the 2008 Annual Meeting of the Cognitive Neuroscience Society (San Francisco, CA, April 2008).

[23] Peperkamp, S., le Calvez, R., Nadal, J.-P., and Dupoux, E. The acquisition of allophonic rules: statistical learning with linguistic constraints. Cognition 101 (2006), B31–B41.

[24] Peterson, G. E., and Barney, H. L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America 23 (1952), 174–184.

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

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

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

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

[29] Smits, R., Sereno, J., and Jongman, A. Categorization of sounds. Journal of Experimental Psychology: Human Perception and Performance 32 (2006), 733–754.

[30] Spalding, A. Inuktitut: A Grammar of North Baffin Dialects. Wurez Publishing: Winnepeg, 1993.

[31] Vallabha, G., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences 104 (33) (2008), 13273–13278.

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

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