Discovering place and manner features: What can be learned from acoustic and articulatory data

Ying Lin
University of Arizona

Jeff Mielke
University of Ottawa

Discovering place and manner features: What can be learned from acoustic and articulatory data

Discovering Place and Manner Features: What Can Be Learned from Acoustic and Articulatory Data?

Ying Lin and Jeff Mielke

1 Introduction

1.1 Features and Innateness

The history of phonological features stems from the quest for a minimal classificatory system for speech sounds in all human languages (Jakobson et al. 1952). Partly due to the influence of information theory, early feature theory took the position of assigning binary values to features. Feature theory took a decisive turn with the study of sound patterns within the context of generative phonology (Chomsky and Halle 1968). Sounds with similar behavior in phonological patterning were grouped into *natural classes*. In most cases, natural classes could be identified using sets of feature values shared by sounds that are members of the natural class. Moreover, Chomsky and Halle argued that the feature representation of speech sounds must be common to all humans, thus adding the innateness claim to the feature theory.

Although features have proven extremely useful for phonological analysis, the innateness claim has been criticized on several grounds. The first line of criticism comes from the perspective of language acquisition. Although the innateness claim simplifies the input representation to infants, language-specific distinctions of speech sounds still depend on identifying which features are distinctive, which presumably hinges on the availability of minimal pairs (Jakobson 1941). The minimal-pair perspective is largely inconsistent with children's input for language development (Charles-Luce and Luce 1990). Recent work on language acquisition has shown that children are highly individualistic in their order of acquiring sounds and words (Vihman 1996). This result is unexplained by the theory that places a set of innate features at the core of phonological acquisition. Second, phoneticians have long questioned the psychological reality of features, partly because no proposal has emerged about how phonological features can be transmitted through processes of speech perception and production.[1] The third type of criticism arises within the phonological theory itself. For example, Mielke's (2007) survey found that the feature systems that have been proposed do not predict the

[1] For example, in his last paper (Ladefoged 2005), Peter Ladefoged summarized his criticisms of the innate feature theory.

wide range of classes that actually occur. Hence innate features have failed to achieve what they were designed to do: characterizing all the possible natural classes.

1.2 A New Proposal: Feature Induction

In light of these three critiques, we take the position that many of the problems seen in the theory of innate features can be resolved by taking an inductive approach to natural classes and distinctive features. The current paper is mainly concerned with one aspect of this inductive approach: instead of looking at sound patterning, we focus on how phonological features can be learned from distributions within the phonetic input, based on the assumption that human learners can discover categories on the basis of sound distributions alone (Maye et al. 2002).

Although our goal is seeking an alternative to the innate feature theory, it should be noted that any inductive learning scheme cannot be separated from assumptions about the innate bias of the learner. Our work is no exception. In particular, we invoke the following assumptions as regards the nature of phonological features.

First, we assume that the learner is biased towards an inherent hierarchical structure within natural classes. Consequently, we focus on phonological features that can be associated with the contrast between natural classes at the same level of the hierarchy. Although similar representational schemes have appeared in formal phonology (Clements 2001, Dresher 2003), our work has a significantly different emphasis since the hierarchical representations are embedded in an inductive model that is capable of learning directly from phonetic data.

Second, in pursuing feature induction, we have also made the simplifying assumption that the phonetic input can be analyzed in terms of segment-sized units. This assumption is not an essential aspect of the model, since prior work has demonstrated that the segmental structure can also be induced from the phonetic input (Lin 2005). Although the phonetic input aligned with segmental units is fed to the learner, it is crucial that the phonemic categories of these units are not available.

Our third assumption is that a phonological learner has access to a wide variety of phonetic information. In particular, we have provided articulatory and acoustic data to the learner to simulate the environment in which feature induction takes place. One research question being explored in this paper is whether the nature of the phonetic information has an effect on the type of features induced from the data.

Given the above three assumptions, we have argued for *hierarchical clustering* as a supporting mechanism for feature induction. Using a *hierarchical mixture model*, a tool for conducting statistical cluster analysis (section 2), we demonstrate that a hierarchical clustering of the phonetic data reveals robust phonetic distinctions within a hierarchy of natural classes, and that different types of phonetic distinctions emerge from different sources of information. Presumably, such distinctions serve as part of the basis on which feature induction takes place. Perhaps less surprisingly, our results also indicate that manner features are closely tied to acoustics (section 3), while place features are closely tied to articulation (section 4), suggesting that two different types of information are involved in the development of two different types of features.

2 Methodology: Hierarchical Mixture Models

The finite mixture model is a tool of cluster analysis (McLachlan and Peel 2000). In a typical application, the data to be analyzed is generated from two or more separate sources and then "mixed" together without indicating from which source each data point is generated. The goal of mixture modeling is to fit the following probability distribution, indicated by $p(y|\theta)$, to a set of data that exhibits a categorical structure:

$$p(y|\theta) = \lambda_1 f_1(y|\theta) + \lambda_2 f_2(y|\theta) + \cdots + \lambda_M f_M(y|\theta) \quad (1)$$

The symbol θ denotes the collection of unknown parameters in the model, which includes all the categories. The number of mixture components is given by M, a fixed integer. The distribution functions $f_i(y|\theta), i = 1, \cdots, M$ each characterize a separate category, or a component of the mixture. They are usually chosen from the same family of distributions. The probability that a sound is drawn from component m is λ_m, which is also part of the model parameter, and is subject to the constraint $\sum_m \lambda_m = 1$. Alternatively, one may view λ_m as a *prior probability*, which determines the proportion of data that is generated by the m-th component. Seeing λ_m as prior probability reflects the assumption that a larger category is more likely to account for the data than a smaller one, given the same likelihood.

As a special case of the finite mixture model, the hierarchical mixture model allows embedding of smaller categories within larger ones. For example, given three discrete categories, they can either be put into a one-level, flat model, or another model with two levels, with each level distinguishing two categories. These two options are illustrated in Figure 1.

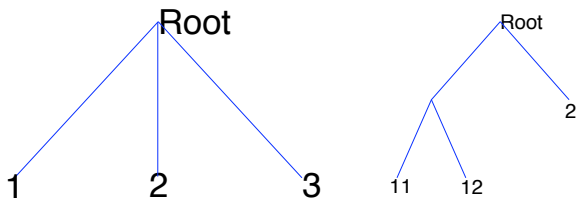

Figure 1: Two ways of modeling three categories with a finite mixture model; left: a flat model; right: with a hierarchical model.

The latter—a hierarchical approach—is better suited for our goal since it has the potential to discover the inherent hierarchical structure within phonetic categories. For the hierarchical model in Figure 1, the probability function is as follows:

$$p(y|\theta) = \lambda_1 \left[\lambda_{11} f_{11}(y|\theta) + \lambda_{12} f_{12}(y|\theta)\right] + \lambda_2 f_2(y|\theta), \tag{2}$$

The prior weights must satisfy $\lambda_{11} + \lambda_{12} = 1$ and $\lambda_1 + \lambda_2 = 1$. Hence, each level in the hierarchy can be regarded as a separate mixture model. In the discussion to follow, all the clusters are labeled using a prefix coding, in which the names of the clusters indicate their positions within the hierarchy. For example, "`11`" denotes that the cluster is the left child of the left node on the top level of the hierarchy. Note, however, that only the clusters on the leaf nodes are associated with individual component models of the mixture. Intermediate nodes (such as "`1`") each represents a collection of models. Thus, the hierarchical model allows us to analyze the result of clustering in a sequence of steps, each focusing on a small number of clusters on the same level.

A fully specified hierarchical mixture model gives rise to clusters that can be organized in a hierarchical manner. In order to obtain the cluster membership of each data point, a gradient categorization is realized through the following quantity:

$$w_y^i = \frac{\lambda_i(\theta) f_i(y|\theta)}{\sum_{i=1}^{M} \lambda_j(\theta) f_j(y|\theta)} \propto \lambda_i(\theta) f_i(y|\theta), \qquad i = 1, \cdots, M \tag{3}$$

Here the weight w_y^i is a direct combination of the prior probability $\lambda_i(\theta)$ and the likelihood $f_i(y|\theta)$. Due to the similarity of (3) and the Bayes formula, some also refer to w_y^i as a *posterior probability*, i.e., the chance that a stimulus y comes from each of the models in the mixture. We can represent the contribution of each component towards the stimulus y as a vector $(w_y^1, \cdots, w_y^M)$,

where $\sum_i w_y^i = 1$. This vector representation of categorization reflects the fact that a stimulus can be ambiguous between different categories. When one of w_y^i, $i = 1, \cdots, M$ is equal to 1 and the rest go to zero, then this gradient representation is equivalent to a categorical one. For example, suppose we are interested in the clusters in the first level: $\lambda_1 [\lambda_{11} f_{11}(y|\theta) + \lambda_{12} f_{12}(y|\theta)]$ and $\lambda_2 f_2(y|\theta)$. The category membership of each data point can be expressed with a pair of numbers between 0 and 1 that add up to one. For example, while the membership of a stimulus can only be either $(0, 1)$ or $(1, 0)$ under a categorical scheme, mixture models allow for category memberships such as $(0.5, 0.5)$ (halfway between the two clusters) or $(0.3, 0.7)$ (a slight preference towards the second cluster).

Since the posterior probability can be interpreted as a "fraction" of the data that belongs to each cluster, and all levels of the hierarchical model have only two clusters ($i = 1$ or 2), we have used the following strategy in visualizing the results of the clustering: first, the fractions w_y^i are calculated for all tokens and all clusters. Then we performed a segment-by-segment analysis by summing together all the fractions that belong to each cluster. For example, for the hierarchical model in the right panel of Figure 1, if we are interested in the top-level split of the clusters, this amounts to comparing the sum of $\{w_y^{11} + w_y^{12}\}$ with the sum of $\{w_y^2\}$, across all the data points y. By placing two subcategories at each level of the hierarchical mixture model, we have biased the learner towards finding a gradient version of binary features. In the general case, three or more clusters can be placed on each level to capture the notion of multi-valued features, although this option is not explored in this paper.

The learning problem of our hierarchical mixture models is solved by the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), a standard tool for fitting such models. Two different instances of this algorithm were used to train the mixture model for the acoustic and the articulatory data, respectively. In actual implementation, the hierarchical mixture model is gradually "grown" in a succession of steps, and finer-grained categories are discovered within the coarse-grained ones discovered from the earlier steps (details of this algorithm are discussed in Lin (2005)). After the algorithm has converged, the posterior probability vector corresponding to each segment is extracted, and a summary of the posterior probabilities is calculated to facilitate interpretations of the clusters and the features that distinguish those clusters within the natural class hierarchy.

3 Clustering Acoustic Data with a Hierarchical Mixture Model

Acoustic data for the clustering experiment is taken from the TIMIT database (Garofolo 1988), a phonetically transcribed corpus that has been used in much speech research. It consists of read utterances recorded from 630 speakers from 8 major regions of American English, of which we use the data from the New England dialect region. Because the goal of TIMIT was to train phone-based systems, the reading materials were chosen to be phonetically-balanced, and all the data was manually transcribed by expert phoneticians. 4,230 consonants from 22 speakers are used in the clustering experiment. The segments are selected from the waveforms according to the time points included in the expert transcriptions, but no labels are used in the learning phase of the model. The phonetic transcriptions are only used later in the interpretation of the clusters.

3.1 Mixture of Hidden Markov Models

Acoustic data impose several constraints on the mixture model used in the clustering analysis. A speech sound is generally not stationary, but contains multiple points of acoustic change. Moreover, a speech sound typically does not have a fixed length, therefore making it difficult to map speech sounds into a space with a fixed dimension. These observations are the major constraints on the mixture model for clustering acoustic data. An important tool for modeling this type of data, used extensively in speech engineering, is the hidden Markov model (HMM). HMM is essentially a Markov model equipped with extra machinery to handle variability. The data is again assumed to be generated in two steps: first, a *state sequence* is generated from the underlying/"hidden" Markov chain by following *transitions* of the chain; then the observed data sequence is generated from the *output distribution* of each state. In speech applications, it is common to specify a *normal mixture* output distribution for each state of the HMM. The present work uses a constrained architecture of HMM that only allows left-to-right transitions, and all output distributions are set to a mixture of two normal distributions[2], each with a diagonal covariance matrix (Rabiner and Juang 1993).

The HMM requires some acoustic modeling of the speech signal, which is also called the *front-end*. The front-end used in this study is of a fairly standard

[2] Although the output distribution is also a mixture, it only models one time slice of speech, and is at a different level than the mixture model of hidden Markov models.

type: the cepstral coefficients in the Mel-frequency domain. Also standard in speech engineering is the use of "delta" features for the purpose of capturing the spectral change between successive analysis windows. The result of such analysis is adding another 13 dimensions to the static cepstral vector, resulting in 26 dimensions for each frame of short time-analyzed speech.

In order to combine the HMMs into a hierarchical mixture model, a learning algorithm that extends the standard HMM training was used (Lin 2005). This algorithm starts by first clustering the average spectral vector of each segment with the K-means algorithm and Itakura-Saito metric (Rabiner and Juang 1993), and then improving the clusters through iterations of the Expectation-Maximization algorithm.

3.2 Results of Clustering Acoustic Data

A three-level hierarchy of five consonant categories is obtained as the result of clustering the TIMIT consonants with a mixture of hidden Markov models. After the category membership is computed for each segment, a level-by-level analysis of the hierarchy of clusters is carried out in the same manner as described in Section 2. Figures 2–4 illustrate the segment-by-segment analysis of these clusters, in a coarse-to-fine manner. This is reflected in the encoding of the resulting clusters using the prefix coding. Limited by space, we only discuss three of the four features that are associated with this hierarchy.

Figure 2 shows that the first clusters to emerge are separated along the sonorant-obstruent dimension. Voiceless and strident obstruents are the most extremely obstruent, while liquids, nasals, and glides are the most extremely sonorant. In the middle are segments that are phonetically more ambiguous, such as glottal consonants [h ɦ ʔ], which are treated ambiguously in phonology as a result of their ambivalent patterning with obstruents and sonorants in different languages. Also in the middle are most of the nonstrident voiced obstruents [ɡ b ð v], which are acoustically more like sonorants and also pattern as such in some languages (Mielke 2007). This partition is more compatible with an acoustic definition of [sonorant] (contra Chomsky and Halle (1968)), which makes sense because it is based on acoustic data. The articulatory definition of [sonorant] would group the glottals with sonorants, but because the acoustic consequences of a supralaryngeal constriction with increased pressure behind it are similar to those of a laryngeal constriction, glottals such as [h] are on the obstruent side of center.

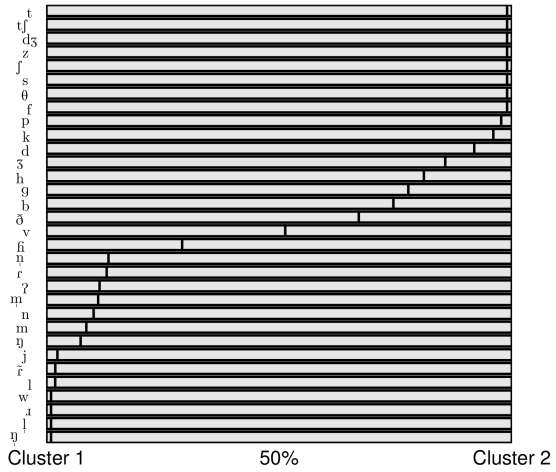

Figure 2: The top-level partition of all consonants.

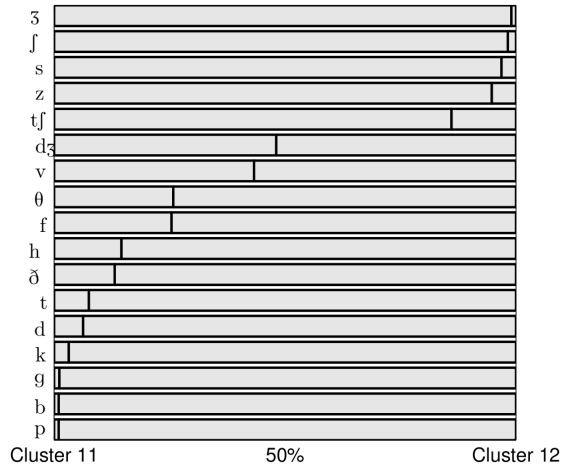

Figure 3: The second-level partition of Cluster 1 (obstruents).

The second-level partition among obstruents involves the strident-mellow dimension (Figure 3). Stops are the most mellow, while sibilant fricatives are the most strident. The middle ground is held by affricates, which are not phonetically strident throughout, and non-strident fricatives, which are less noisy than their strident counterparts but noisier than stops.

The second-level partition among sonorants is shown in (Figure 4). In general this partition contrasts nasals and acute glides with more grave approximants, although the glottal fricative and stop do not fit very well with respect to this division.

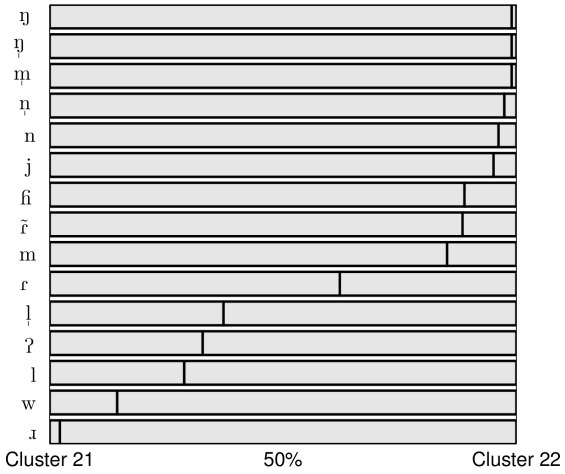

Figure 4: The second-level partition of Cluster 2 (sonorants).

4 Clustering Articulatory Data

4.1 Collection of Articulatory Data

Our articulatory data used in this experiment is derived from simultaneous audio, video, and ultrasound video recordings that were made while subjects produced CVC sequences. The CVC stimuli contained all of the consonant phonemes found in American English, with C1 always equal to C2. For each consonant, there was a stimulus with each of the vowels /a ʌ i/, and an additional set of stimuli contained only the vowels. For /h/ (which does not appear in coda position in English) and /ŋ/ (which does not appear in onset position), phonotactically permissible CV and VC stimuli were used instead. Each stimulus was elicited 10 times, with the exception of /aŋ ʌŋ iŋ/ and /ha hʌ hi/, which were elicited 20 times each, in order to have 60 recordings of each consonant (a fact which is not exploited further in this paper). The subject was a male native speaker of American English.

The palate and tongue surfaces shown in the ultrasound images were

traced using position information contained in the video image to combine the two surfaces after correcting for head and/or transducer movement. A point on each of the upper and lower lips was marked on the video image and transformed into ultrasound head space. On the basis of the lip points and the tongue and palate surfaces, the rest of the mid-sagittal oral cavity and pharynx tract was filled in according to an estimate of the locations of the soft palate, pharynx, and upper and lower teeth. This estimate allows cross-distances to be measured or estimated for the length of the vocal tract (except the nasal cavity) from the larynx to the lips. The result of these estimates is two surfaces (the upper and lower surfaces of the vocal tract). Cross-distances between the two surfaces were measured automatically. In the last step of pre-processing, the cross-distance data is re-sampled to 60 dimensions using linear interpolation. As examples, Figure 5 illustrates the original tongue tracings and the pre-processed cross-distance data for two types of tokens, "k/_a" and "k/_i".

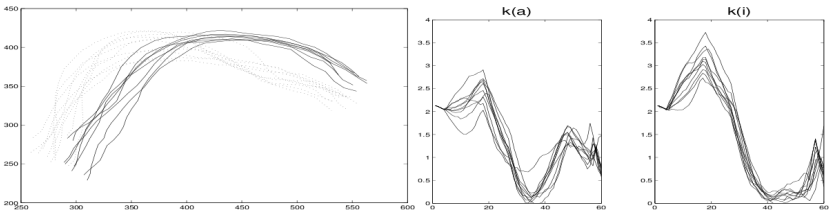

Figure 5: Tracings of the tongue profile from the ultrasound images for [k]/_a and [k]/_i.

4.2 Mixture of Probabilistic Principal Component Analyzers

The articulatory data impose rather different kinds of constraints on the mixture model. Because the preparation of the articulatory data involves significant hand-correction by a human expert, we have decided to associate each consonant with one single frame of the imaging data, instead of a series of images unfolding in time. Although this articulatory representation of consonants eliminates the problem posed by time-series data, it is still necessary to seek a proper way of reducing the dimension of the articulatory data, since the dimensions of the vocal tract are highly correlated. Principal Component Analysis (PCA) is such an approach that seeks a low-dimensional representation of the vocal tract during speech production (Story and Titze 1998): each profile of the vocal tract is expressed as a linear combination of elementary gestures (or "orthogonal modes/bases") that alters the neutral shape of the

vocal tract. By only considering the orthogonal modes/bases that contribute to most of the total variance in the data, we are able to obtain a dimension-reduced representation of the original data.

However, traditional PCA is not associated with a probabilistic function, as required for mixture models. An alternative is the Probabilistic PCA (Tipping and Bishop 1999). In the current experiment, we created a hierarchical mixture of PPCA's by using Probabilistic PCA's as components of the hierarchical mixture model. For a fixed number of categories, the parameters of this model are learned by the same type of EM algorithm (Tipping and Bishop 1999). In addition, the same coarse-to-fine search strategy is also adopted to grow the mixture model to the desired size.

4.3 Results of Clustering the Articulatory Data

The total set of cross-distances data is used in the clustering experiment, based on the hierarchical mixture of PPCA's. The result is another 3-level hierarchy of natural classes. The same analysis as in 3.2 is applied to the analysis of these clusters as well, and part of the results are shown in Figure 6 and Figure 7.

The left panel of Figure 6 shows the split between consonants with a pharyngeal or velar constriction and those with a constriction elsewhere. The right panel provides a breakdown of the distribution of velar consonants with regard to each vowel context, illustrating the fronting effect of [i] (but not [ʌ a]) on velar stops in English. Figure 7 shows the split between consonants with a primary alveolar, postalveolar, or palatal constriction and those without one. Notice that interdentals group with the non-alveolar/non-palatal cluster, and that [l], which involves alveolar and velar constrictions, does not cluster with either group.

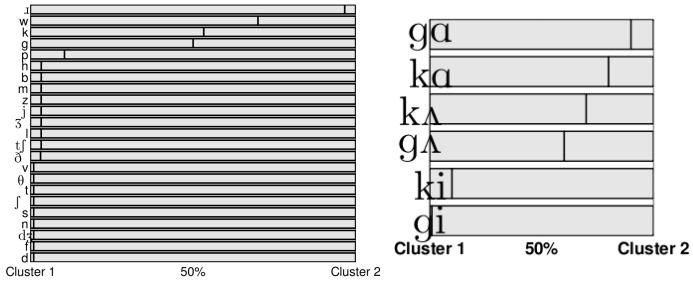

Figure 6: left: The top-level partition of all consonants; right: Distinction within each different context of the velars.

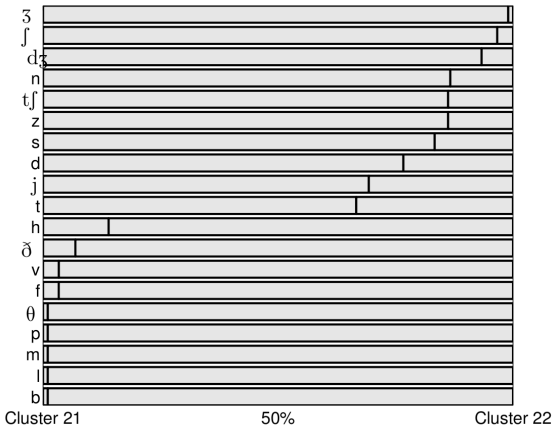

Figure 7: The second-level partition of non-velar consonants.

5 Discussion

In summary, we have presented an inductive approach to features as a revision to the innate feature theory. Feature induction is seen as part of a clustering problem in which the learner discovers a hierarchy of natural classes, a view that is consistent with recent proposals in phonological acquisition. Rather than taking features as innate and universal, the inductive approach depends on the phonetic distributions within the target language, and induced features are seen as specific to the phonetic inventory of each language. In addition, ambiguous segments are assigned gradient memberships within a natural class hierarchy. Hence, the inductive approach offers a solution to the challenges to the innate feature theory.

The clustering of segments according to acoustic and articulatory data results in a number of familiar partitions of segments which have had innate features proposed for them. The fact that phonetic dimensions which frequently are exploited by phonological patterns can be learned from phonetic data alone suggests that these dimensions do not need to be defined in Universal Grammar (contra Chomsky and Halle (1968)). This is consistent with Stevens's (1989) identification of areas of acoustic-articulatory stability that correspond to recurrent feature values (although with a different interpretation of the implications of phonetic stability for the innateness of features).

The connections between sources of data and types of features (place-articulatory vs. manner-acoustic) suggests that other differences can be found

in these types of features, in terms of the types of sound patterns in which they are involved. The observation that place features pattern differently from manner features is well established in phonological theory in the form of Feature Geometry (Clements 1985, Sagey 1986, Halle et al. 2000, Clements and Hume 1995), with place features hierarchically organized, often privative, and patterning together, and manner features located in the root node, where they do not pattern as constituents and are less involved in assimilation. The hierarchical organization of binary splits that are the result of clustering are reminiscent of Feature Geometry hierarchies, but are derived by a model of unsupervised learning. This model does not utilize information about phonological patterning, but only depends on distributions within phonetic data.

References

Charles-Luce, J., and P.A. Luce. 1990. Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language* 17:205–515.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. Cambridge, Mass.: MIT Press.

Clements, G.N. 1985. The geometry of phonological features. *Phonology Yearbook* 2:225–252.

Clements, G.N. 2001. Representational economy in constraint-based phonology. In *Distinctive Feature Theory*, ed. T. Alan Hall, 71–146. Mouton de Gruyter.

Clements, G.N., and Elizabeth V. Hume. 1995. The internal organization of speech sounds. In *The Handbook of Phonological Theory*, ed. John Goldsmith, 245–306. Cambridge, Mass.: Blackwell.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39:1–38.

Dresher, Elan. 2003. The contrastive hierarchy in phonology. In *Toronto Working Papers in Linguistics (Special Issue on Contrast in Phonology)*, ed. Daniel Currie Hall. University of Toronto.

Garofolo, J.S. 1988. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST).

Halle, Morris, Bert Vaux, and Andrew Wolfe. 2000. On feature spreading and the representation of place of articulation. *Linguistic Inquiry* 31:387–444.

Jakobson, Roman. 1941. *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1952. *Preliminaries to Speech Analysis*. Cambridge, Mass.: MIT Press.

Ladefoged, Peter. 2005. Features and parameters for different purposes. Paper presented at the Linguistic Society of America annual meeting.

Lin, Ying. 2005. Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition. Doctoral Dissertation, UCLA.

Maye, J., J.F. Werker, and L. Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82:101–111.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite Mixture Models*. New York, NY: Wiley.

Mielke, Jeff. 2007. *The Emergence of Distinctive Features*. Oxford: Oxford University Press.

Rabiner, L., and B.H. Juang. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

Sagey, Elizabeth. 1986. The Representation of Features and Relations in Nonlinear Phonology. Doctoral Dissertation, Massachusetts Institute of Technology.

Stevens, Kenneth. 1989. On the quantal nature of speech. *Journal of Phonetics* 17:3–45.

Story, Brad H., and Ingo R. Titze. 1998. Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics* 26:223–260.

Tipping, M., and C. Bishop. 1999. Mixtures of probabilistic principle component analyzers. *Neural Computation* 11:443–482.

Vihman, Marilyn May. 1996. *Phonological Development: The Origins of Language in the Child*. Cambridge, Mass.: Blackwell.

Ying Lin
Department of Linguistics
University of Arizona
Tucson, AZ 85719
yinglin@email.arizona.edu

Jeff Mielke
Department of Linguistics
PO Box 210028
Ottawa, Ontario
K1N 5N6
Canada
jmielke@uOttawa.ca