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

Phonological theories often assume that features are innate and universal, and that they characterize possible natural classes. Although this assumption is certainly useful for phonological analysis, it leaves several unresolved issues in other fields: first, from the perspective of a learner, acquiring feature-based representations of words is assumed to be trivial, and establishing distinctive features requires the availability of minimal pairs. This perspective is largely inconsistent with the facts of language development (Charles-Luce and Luce, 1990). Second, phoneticians have long questioned the psychological reality of features, partly because no proposal has emerged about how hearers/learners can recover symbolic phonological knowledge from processes of speech perception and production (Ladefoged, 2001). These problems may be remedied by an alternative view of phonological features as being the result of learning from phonetic data.

Inspired by the behavioral study of (Maye and Gerken, 2000) on the learning of sound categories, the present study explores an inductive approach to feature discovery without resorting to minimal pairs. Our basic assumption is that learners have access to a wide variety of phonetic information, and that feature induction can proceed through recursive clustering of the appropriate data. In particular , natural classes are discovered in a coarse-to-fine manner, and features are discovered as a by-product of a hierarchy of natural classes, similar to some representational schemes in phonology (Clements, 2001; Dresher, 2003). Two kinds of data were used in our clustering experiment: the acoustic data consist of a set of segments from the TIMIT acoustic phonetic database (Garofolo, 1988). Phonetic labels of those segments have been removed prior to the experiment and are only used later in evaluation. The articulatory data are obtained with ultrasound imaging. A set of monosyllabic words are used in the ultrasound data collection, containing a set of English consonants occurring in word-initial positions. One image was taken from the maximal constriction during the initial consonant of each word, and vocal tract is characterized by a vector of cross distances from the tongue body to an estimated position of the palate, as obtained from a palate reconstruction technique.

The clustering in the current study is based on a kind of probabilistic model called mixture model (McLachlan and Basford, 1988), and recursive clustering is realized by embedding a hierarchical structure within each mixture model. Two probabilistic, generative models are employed as the respective components of the mixture models for acoustic and articulatory data. Since acoustic speech data vary in duration and spectral properties, a mixture of Hidden Markov Models (HMM) is used to cluster the acoustic speech. For the articulatory data, since the dimensions of the vocal tract are highly correlated, a mixture of Probabilistic Principle Component Analyses (PPCA) is used to cluster the cross distances in the vocal tract. This is based on the assumption that speech articulation can be seen as perturbations along vocal tract configurations that serve as orthogonal modes, and thus can be projected into a lower dimensional space using the composite measures as new coordinates (Story and Titze, 1998). Our experiments using acoustic and articulatory data suggest that these two sources of information lead to the discovery of manner and place features, respectively, thus bringing support to the view of feature learning as a process of mapping from phonetics to phonology.

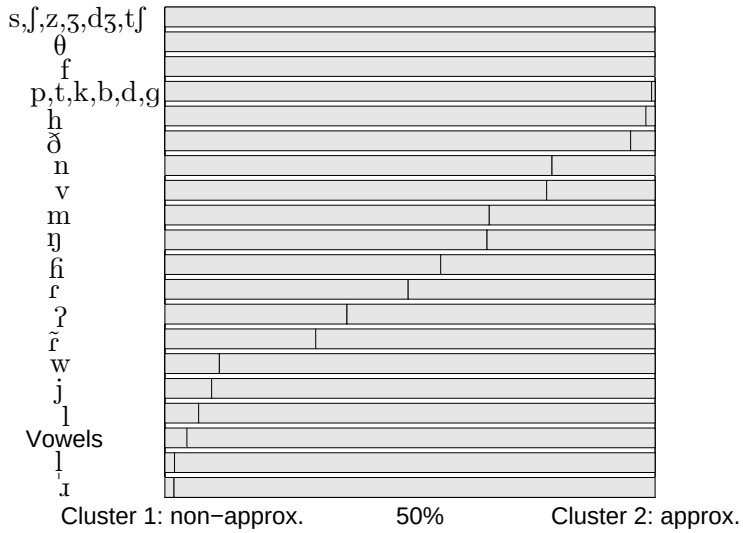

Figure 1: The first-level clustering of acoustic data from the TIMIT database. The vertical bars indicate the percentage of the segments that fall into each of the two clusters, which are interpreted as *approximants* and *non-approximants*, respectively.

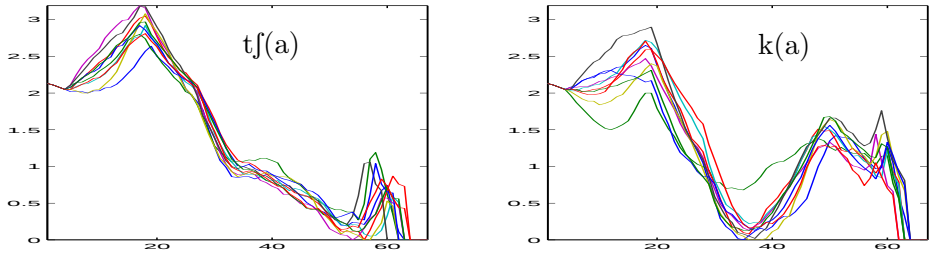

Figure 2: Cross distance values within the vocal tract, calculated from ultrasound images of two consonants: [tʃ] and [k]. Multiple tokens are shown for each consonant in the same context. The left side of each token is aligned with the pharynx, and the right side aligned with the lips.

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.

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

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

Ladefoged, P. 2001. *A Course in Phonetics*. 4th edition. Harcourt Brace.

Maye, J. and L. Gerken. 2000. Learning phoneme categories without minimal pairs. In *Proceedings of the 24th BUCLD*, p. 522–533, Somerville, MA. Cascadilla Press.

McLachlan, G.J. and K.E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

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