Higher-order structure for vowel variation is specific to the culture and to the individual listener

Andrew R. Plummer

Resolving differences in acoustic parameter space representations of categorically similar sounds produced by different talkers (and thus by different vocal tracts with potentially very large differences in shape and size) in order to facilitate categorization tasks and analysis is a long-standing issue in the development of models of speech perception. Methods for addressing the issue have taken on different names depending on the type of categorization task – e.g., "speaker adaptation" for building automatic speech recognition systems (see Lee & Rose, 1996; Gerosa et al., 2007, inter alia), "vowel normalization" for doing sociophonetic analysis of vowel changes in progress (see Clopper, 2009, inter alia). With few exceptions, such methods for doing "vowel normalization" focus on differences between men and women and most of them attempt to equate adult male vowel spaces with adult female vowel spaces via (relatively) fixed mappings over parameter space representations. Hindle (1978) characterized these kinds of methods as "technical" solutions to the normalization issue, i.e., those designed to aid in carrying out an engineering task or technical analysis, and contrasted them with what he called the "psychological aspect" of normalization, which addresses the question "[w]hat is a speaker doing when he equates two vowels spoken by different speakers and having different formant values?" (162).

Use of technical solutions in a categorization task or analysis brings along with it the logical consequence that vocal tract size and shape effects are assumed to be irrelevant to that task or analysis. For example, Labov (2006) describes gender-related variation as "a mixture of the effects of vocal tract length differences and social factors" and casts normalization as an answer to the question of how to "separate the two types of influence, and arrive at a scaling factor that eliminated only the differences due to vocal tract length without removing the effect of social factors?" (502-3). While potentially expedient, this assumption seems to be undesirable in a number of areas where technical solutions are applied, especially vowel categorization at the initial stage of phonological acquisition. At this stage of rapid physical, cognitive, and social development, applying the necessary assumption underlying the technical solution requires making further assumptions about universality that are counter-factual. Rather, culture-specific effects on infants' acquisition of vowel categories that begin to take shape during the first eight months of life (making them pre-linguistic and hence independent of the possible "top-down" influence of language forms) suggest looking carefully at the psychological aspect of normalization in the emergence of vowel systems in ontogeny.

In this presentation, we argue that body size types, age, gender, and so on are socially interpreted categories and that learning the culture-specific social interpretations of categories that are based on natural variation in vocal tract size and shape plays a crucial role in the emergence of vowel systems during early infancy. We propose a conceptual framework for modeling the emergence of vowel systems based on the following guiding proposition: during ontogeny, vowel systems emerge as a set of culture-specific mappings that infants use to relate sensory space representations of vocalizations from different talkers with affective and affiliative information. Moreover, these mappings facilitate the emergence of sound categories, which in turn act as the building blocks of a phonological grammar and an emerging lexicon. In this light, the emergence of vowel systems in ontogeny is fundamentally incommensurate with a fixed technical solution to vowel normalization. In addition to the conceptual framework, we present a corresponding computational modeling architecture in which infants generate structures, called "manifolds," over sensory space representations of their productions and those of their caretakers, and "align" the structures based on affiliative information in vocal interaction with caretakers, in order to generate the culture-specific mappings that relate the differing sensory space representations. If essentially on the right track, the framework and architecture suggest that higher-order structure for vowel variation is not only culture-specific, but specific to an individual listener. Moreover, during ontogeny infants invent (rather than discover) idiosyncratic higher-order structures for processing the vowel variation in their serially expanding speech communities.

References

Clopper, C. (2009). Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass*, *3*(6), 1430–1442.

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*, 847–860.

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (Ed.) *Linguistic Variation: Models and Methods*, (pp. 161–171). New York: Academic.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, *34*, 500–515.

Lee, L., & Rose, R. C. (1996). Speaker normalization using efficient frequency warping procedures. In *Proceedings of ICASSP-96*.