LISTENER-ORIENTED CONSTRAINTS ON ARTICULATORY ORGANIZATION

Randy L. Diehl

Department of Psychology
University of Texas at Austin

ABSTRACT

A good many phonetic regularities reflect a strategy of talkers to maintain sufficent auditory distinctiveness among speech segments. In defending this claim, I discuss a number of converging methods for evaluating the notion of "auditory distance" in a language-independent way.

1. INTRODUCTION

It is often asserted that many phonetic and phonological regularities reflect a kind of trade off between the needs of the talker (minimal necessary effort) and the needs of the listener (sufficient perceptual contrast). Although most phoneticians would not object in principle to this general claim, in practice they have tended to treat it as more of a slogan than as a basis for genuine phonetic explanation. (Of course, there are notable exceptions to this tendency, as the other papers in this symposium effectively demonstrate. See also [15].) Various reasons no doubt exist for this state of affairs, but two come to mind immediately. First, there is the long tradition within linguistics of viewing the study of phonetics as largely devoted to pure physical description of speech sounds. Within this descriptivist tradition, the study of functional aspects of speech communication is often seen as secondary or even irrelevant. Second, there has been considerable skepticism that notions such as "minimal effort" or "sufficient contrast" can be formulated with adequate precision and validity to have significant explanatory content. For example, Ladefoged [14] recently dismissed the possibility of devising a language-neutral measure of auditory distinctiveness on the grounds that such a measure is inevitably confounded by the language bias of the observer.

I wish to offer a rather more optimistic view of the prospects for a program of functional explanation in phonetics and phonology. Although both talker- and listener-oriented selection pressures are involved in the shaping of sound systems and "on-line" speech behavior, the focus in this brief discussion will be on phonetic regularities that appear to reflect mainly the requirements of listeners.

2. THE AUDITORY ENHANCEMENT HYPOTHESIS

It is a striking fact about articulatory organization that the phonetic properties of vowels and consonants tend to covary in a highly regular manner. Across languages, back vowels are usually rounded while front vowels are usually unrounded, high vowels tend to be produced with a higher fundamental frequency (F0) and with a higher velar position than low vowels, voiced consonants are usually preceded by longer vowels and followed by a lower F0 than voiceless consonants, and so on. For many such regularities, phoneticians have sought explanations in terms of putative physical or physiological constraints on production. Thus, for example, the F0 correlate of tongue height has been attributed to a passive mechanical coupling between the tongue body and larynx (hence the phrase "intrinsic vowel pitch") [8, 13].

The problem with many of these explanations is that, although they may be superfically plausible, they fail to account in detail for the relevant facts. For example, the F0 correlate of tongue height shows up even in the esophageal speech of laryngectomized patients [7]. Because such speakers obviously lack laryngeal cartilages as well as a hyoid bone, the F0 effect cannot be explained by the coupling hypothesis. Moreover, recently several groups of investigators have reported that higher vowels are associated with increased levels of cricothyroid activation [9, 18], which suggests that the F0 correlate is under active control by the talker.

My colleagues and I [5, 6] have been investigating an alternative hypothesis concerning the origin of many significant types of phonetic covariation. We claim that the phonetic properties of vowels and consonants covary as they do largely because language communities tend to select properties that have mutully enhancing auditory effects. The obvious result of such a selection strategy is to produce segments that are more distinctive auditorily. Of course, in communication situations involving low noise and relatively high redundancy, talkers can afford to trade away some distinctiveness for greater ease of articulation. But the potential for greater distinctiveness must be built into the sytem to be exploited when necessary.

How might this auditory-enhancement hypothesis explain, for example, the F0 correlate of tongue height? A principal acoustic correlate of high vowels is a low-frequency first formant (F1). However, evidence suggests that the best predictor of *perceived* vowel height is not F1 per se, but rather the difference in Bark units between F1 and F0, with smaller differences yielding perception of a higher vowel [17]. A possible auditory basis for this effect has recently been suggested by Beddor [1]. On the basis of the notion of a 3.5 Bark spectral integrator [3], she hypothesized that when F0 is raised sufficiently close to F1, the spectral center of gravity associated with F1 shifts downward, contributing to a perceived raising of the vowel. If this hypothesis is correct, then the so-called "intrinsic vowel pitch" may actually reflect a strategy of enhancing the auditory distinction between high and low vowels.

Elsewhere [6] we have argued more generally that the auditory-enhancement hypothesis can account for the salient phonetic properties of the most common vowels. In the case of canonical productions of the vowel /u/, for example, almost every independently controllable articulatory parameter is set to enhance the distinctive lowering of the first two formant frequencies. In fact, virtually all of the theoretical options for lowering the first two resonant frequencies of a tube-like configuration appear to be exploited. These options include vocal-tract lengthening through lip protrusion and larynx lowering, constriction near the antinodes of the standing volume-velocity waveforms corresponding to these resonances (i.e., at the lips and velopharyngeal area), and dilation near the nodes of the same standing wave patterns (i.e., at the midpalate and lower pharynx). Of these, only palatal dilation can properly be argued to be a by-product of other vocal-tract gestures, in this case, tongue-body retraction. The rest of the distinctiveness-enhancing gestures appear to be actively selected.

3. METHODS OF EVALUATING AUDITORY DISTINCTIVENESS

Let us now return to Ladefoged's argument that any attempt to devise a metric of auditory distinctiveness will inevitably be entangled in the language biases of the observer. A preliminary answer to this is that one might use measures of acoustic distance in place of an auditory metric, and such physically defined measures would presumably be language independent. For example, the above auditory-enhancement rationale for the detailed articulatory properties of the vowel /u/ is framed entirely in terms of acoustic distinctiveness. In the final analysis, however, acoustic distance metrics are theoretically insufficient. The auditory system is nonlinear in many respects, and a fully adequate distance metric must incorporate these nonlinearities. So the question remains: How can we circumvent the problem of observers' language bias in evaluating the notion of auditory distinctiveness? I will suggest some possible approaches that have been used in our laboratory and elsewhere.

3.1. Speech/Nonspeech Comparisons

In attempting to test the auditory-enhancement hypothesis, my colleagues and I have often compared listeners' categorization performance on speech stimuli to their performance on nonspeech stimuli that are acoustically analogous in certain relevant respects. Consider the following example. In most languages, [+voice] consonants in medial position are distinguished from [-voice] consonants in having *inter alia* a shorter constriction interval and significant glottal pulsing during the constriction interval. We [16] hypothesized that the presence of glottal pulsing makes the short constriction appear even shorter, thus enhancing the distinctiveness of the constriction-duration correlate of [+voice] consonants.

To evaluate this claim, we first had listeners identify two series of /aba/-/apa/ stimuli varying in closure duration. In one series, the items contained a segment of glottal pulsing during closure, while in the other series, the closure interval contained only silence. As expected, the presence of pulsing shifted the /b/-/p/ category boundary, yielding more /b/ responses. A second group of listeners were asked to identify two corresponding series of nonspeech stimuli, each consisting of two square-wave segments separated by a medial gap of varying duration. Like the speech stimuli, one series contained a segment of glottal pulsing during the gap, while the other did not. These stimuli are highly nonspeechlike and they do not correspond to any obvious natural categories, so it is necessary to provide training in labeling the stimuli. Initially, the subjects learn by means of feedback to press one response key when they hear the series-endpoint stimulus with the shortest medial gap and a different key when they hear the series-endpoint stimulus with the longest medial gap. After this training, the listeners are asked to identify the entire series on the basis of similarity to either end-point stimulus. (We refer to this task as end-point similarity matching.) Interestingly, for these nonspeech stimuli, the presence of pulsing during the gap shifted the category boundary in the same direction as observed for the /aba/-/apa/ stimuli. We took this as evidence that glottal pulsing does in fact make the medial gap appear shorter. Notice that the parallel between the speech and nonspeech results suggests that the effect is not to be explained in terms of linguistic experience but rather in more general auditory terms.

3.2. Adult/Infant Comparisons

In a related study [4], we found that prelinguistic infants showed categorical discrimination of the same /aba/-/apa/ stimuli used in the adult experiment just described. In addition, there was evidence that the presence of glottal pulsing during closure caused a category boundary shift comparable in magnitude to that of the adult subjects. The average age of the infant subjects (7 1/2 months) was younger than the age at which effects of linguistic experience are typically first observed. Therefore, the results tend to confirm our earlier conclusion that the effect of glottal pulsing on the perception of the closure-duration cue is of a general auditory nature rather than being a product of linguistic experience.

3.3. Cross-Native-Language Comparisons

Mandarin syllables carrying a mid-high-rising F0 contour (Tone 2) tend to be shorter than those carrying a low-falling-rising contour (Tone 3). We [2] hypothesized that talkers use length differences to enhance the perceptual contrast between Tones 2 and 3. A primary difference between the two tone categories is that Tone 2 has a relatively short initial period of nonrising F0 prior to a rising interval, whereas Tone 3 has a relatively long period of nonrising F0 before its rising interval. The effect of proportionally lengthening one of these contours is to increase the absolute duration (and hence detectability) of the initial nonrising F0, making the contour perceptually more like Tone 3. We tested this hypothesis by comparing perceptual judgments by native Mandarin and native English speakers on various synthetic series of Mandarin tones ranging incrementally from a Tone 2 contour to a Tone 3 contour, with stimuli from the different series varying in length. (The Mandarin-speaking subjects performed lexical labeling, whereas the English-speaking subjects were trained and tested on the end-point similarity matching task, described earlier.) Both groups of subjects had very similar category boundaries, and both showed the predicted effect of syllable length (i.e., longer stimuli being more likely to be assigned to the Tone 3 category). The cross-native-language similarity suggests that language experience is not a main factor in the length effect. Rather, general auditory factors seem to be responsible.

3.4. Human/Animal Comparisons

Along with others such as Kuhl [12], we have conducted a series of experiments comparing speech categorization performance of humans and animals. Most of the results to date indicate rather striking cross-species similarities. For example, in his dissertation work, Kluender [11] had both humans and Japanese quail categorize a /g/-/k/ stimulus set that varied orthogonally in both voice onset time and F1 onset frequency. For both groups of subjects, a lower F1 onset frequency produced a reliable shift in the labeling boundary, yielding more responses corresponding to the [+voice] category. This similarity suggests that the low-frequency F1 typical of voiced stops enhances the perception of voicing for reasons that have little to do with linguistic experience.

3.5. Auditory Modeling

Recently, a highly realistic model of mammalian auditory-nerve response [10] has been made available to us. This model, together with standard distance metrics applied to sets of its output representations, can provide language-independent estimates of auditory distances among speech stimuli. These estimates can in turn be used to evaluate auditory-enhancement accounts of regularities involving phonetic covariation.

4. SUMMARY

There are reasons to be optimistic that functional explanatory notions such as "auditory distance" can be formulated with considerable precision and validity. In the context of discussing evidence for the auditory-enhancement hypothesis, I outlined a number of convergent approaches to the evaluation of auditory distance, each of which apparently avoids the confounding effects of observer bias.

5. REFERENCES

[1] Beddor, P.S. (to appear). "On predicting the structure of phonological systems", *Phonetica.*

[2] Blicher, D.L., Diehl, R.L. & Cohen, L.B. (1990). "Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement", *J. Phonetics*, 18, 37-49.

[3] Chistovich, L.A. & Lublinskaya, V.V. (1979). "The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustic study of perception of vowel-like stimuli", *Hear. Res.*, 1, 185-195.

[4] Cohen, L.B., Diehl, R.L., Oakes, L.M. & Loehlin, F.C. "Infant discrimination of /aba/ versus /apa/", unpublished manuscript.

[5] Diehl, R.L. & Kluender, K.R. (1989). "On the objects of speech perception", *Ecol. Psych.*, 1, 121-144.

[6] Diehl, R.L., Kluender, K.R. & Walsh, M.A. (1990). "Some auditory bases of speech perception and production", *Advances in speech, hearing and language processing*, 1, 243-267.

[7] Gandour, J. & Weinberg, B. (1980). "On the relationship between vowel height and fundamental frequency", *Phonetica*, 37, 344-354.

[8] Honda, K. (1981). "Relationship between pitch control and vowel articulation", in D. Bless & J. Abbs (eds.), *Vocal-fold physiology* . San Diego: College-Hill.

[9] Honda, K. & Fujimura, O (1989). "Phonological vs. biological explanation--intrinsic vowel pitch and phrasal declination", Paper presented at the 6th Vocal Fold Physiology Conference, Stockholm.

[10] Jenison, R.L., Greenberg, S., Kluender, K.R. & Rhode, W.S. (to appear). "A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers", *JASA*.

[11] Kluender, K.R. (1988). "Auditory constraints on phonetic categorization: trading relations in humans and nonhumans", Unpublished doctoral dissertation, University of Texas at Austin.

[12] Kuhl, P.K. (1988). "Auditory perception and evolution of speech", *Hum. Evolution*, 3, 19-43.

[13] Ladefoged, P. (1968). "*A phonetic study of West African languages*", Cambridge: Cambridge University Press.

[14] Ladefoged, P. (1990). "Some reflections on the IPA", *J. Phonetics*, 18, 335-346.

[15] Lindblom, B. (1986). "Phonetic universals in vowel systems", in J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology*, Orlando, Fl: Academic.

[16] Parker, E.M., Diehl, R.L. & Kluender, K.R. (1986). "Trading relations in speech and nonspeech", *Percept. Psychophys.*, 34, 314-322.

[17] Traunmüller, H. (1981). "Perceptual dimension of openness in vowels", *JASA*, 69, 1465-1475.

[18] Vilkman, E., Aaltonen, O, Raimo, I, Arajärvi, P. & Oksanen, H. (1989). "Articulatory hyoid-laryngeal changes vs. cricothyroid activity in the control of intrinsic F0 of Vowels", *J. Phonetics*, 17, 193-203.