

Increasing Realism of Auditory Representations Yields Further Insights into Vowel Phonetics

Randy L. Diehl[†], Björn Lindblom^{†‡} and Carl P. Creeger[†]

[†] University of Texas at Austin, USA

[‡] Stockholm University, Sweden

E-mail: diehl@psy.utexas.edu, lindblom@ling.su.se

ABSTRACT

The ear's remarkable ability to cope with noisy signals is linked to its use of a spatio-temporal mechanism that distributes information about strong spectral components across neural units whose characteristic frequencies (CFs) often span broad frequency ranges. Speech formant information is thus carried not only by channels with CFs near spectral peaks but also by adjacent channels. This paper examines several implications of this mechanism for vowel phonetics: (1) Although there is no evidence of explicit 'formant tracking', spectral peaks are nonetheless granted a special status. (2) There is an auditory warping of the vowel space that enhances contrasts along the open-close dimension relative to the front-back dimension. (3) Incorporating the above mechanism into auditory models helps to unify accounts that appear to take somewhat different approaches to explaining the structure of vowel systems (viz., Stevens's Quantal Theory, the Grenoble group's Dispersion-Focalization Theory and Lindblom's Adaptive Dispersion Theory).

1. INTRODUCTION

There is a curious typological asymmetry in the structure of preferred vowels systems among the world's languages: significantly more contrasts occur along the open-close (sonority) dimension than along the front-back (chromaticity) dimension. For example, the 7-vowel system of Italian exhibits four distinctive levels of sonority but only two distinctive levels of chromaticity. The primacy of the open-close dimension is also reflected in other data including the prevalence of raising and lowering processes in sound change [1], the greater frequency of diphthongs with open-close trajectories (e.g., [aɪ] and [aʊ]) relative to those with back-front trajectories (e.g., [wi] and [ju]), and the strong vowel height component in realizations of tense-lax pairs [2].

The preference for sonority distinctions is all the more surprising given typical acoustic distances separating the

point vowels /i/, /a/, and /u/. The sonority dimension, largely corresponding to variation in first formant frequency (F1), has a much smaller physical extent than does the chromaticity dimension, which corresponds mainly to variation in second formant frequency (F2). F1 spans about 550 Hz, whereas F2 spans about 1500 Hz. Expressing the frequency scales in Mel units reduces this discrepancy but does not eliminate it. On these grounds, one might expect *fewer* vowel distinctions in sonority than in chromaticity rather than the reverse pattern that is actually observed.

2. EARLY SIMULATIONS OF PREFERRED VOWEL SYSTEMS

The asymmetry between sonority and chromaticity contrasts was highlighted in early attempts to model the structure of preferred vowel systems. Liljencrants and Lindblom [3] defined a Mel-scaled F1 x F2 space of possible vowels from which vowel systems of varying sizes were selected using a criterion of maximal intervowel distance (maximal dispersion or contrast). For three or five vowels, the predicted systems closely matched actual systems favored among the world's languages. However, for seven or more vowels, the simulated systems tended to include more high vowels than are typically attested in actual systems (e.g., Italian). This result is, of course, predicted by the relatively large frequency extent of the chromaticity dimension.

In an effort to increase the auditory realism of the simulations, Lindblom [4] adopted a measure of auditory distance based on whole spectra rather than formant frequencies. The spectra were the output of an auditory model incorporating critical-band filtering and appropriate pitch and loudness scaling. Although there was some small improvement in predictive accuracy relative to the formant-based simulations, the problem of too many high vowels remained.

3. RECENT SIMULATIONS: VOWELS IN NOISE

In the early simulations of preferred vowel systems,

auditory distances were always calculated for vowels in the absence of background noise. However, under realistic listening conditions environmental noise is always present, and it is plausible to assume that actual vowel systems have evolved to be perceptually robust with respect to such noise. To evaluate this assumption, Diehl, Lindblom, and Creeger [5] replicated the simulation experiments of Lindblom [4] but with the addition of background noise conditions. The spectral shape of the noise mimicked the long-term average for speech (-6dB/octave), and auditory distances among vowels were calculated at eight different signal/noise ratios, ranging from 10 dB to -7.5 dB. These distances were then averaged to determine the predicted optimal vowel systems for varying inventory sizes. Interestingly, the effect of adding background noise was to yield predicted vowel systems that resembled actual systems more closely than had the earlier simulations. In particular, the problem of excessive high vowels was eliminated. The reason for this may be summarized as follows: (1) Noise eliminates information from the distance calculations. (2) Salient regions of the spectrum (e.g., formant peaks) are more resistant to noise degradation than nonsalient regions. (3) Lower frequency prominences (F1) tend to be more resistant to noise degradation than higher-frequency ones (F2 and F3). (4) This results in a perceptual warping of the vowel space, restricting the size of the chromaticity dimension relative to that of the sonority dimension.

4. RECENT SIMULATIONS: TOWARD A MORE REALISTIC AUDITORY REPRESENTATION OF VOWELS

The filter-bank outputs of the auditory models used in [4] and [5] are whole-spectrum representations corresponding to average firing rate as a function of the characteristic frequency (CF) of the neural channels being stimulated ('excitation patterns'). Such representations account for a variety of psychophysical data, and are thus reasonably well motivated. However, they are incomplete in one important respect: temporal information about stimulus frequency (phase locking) is not included. This may be a serious omission because such temporal information tends to be more resistant to degradation in noise than information contained in excitation patterns [6,7]. In particular, spectral prominences (e.g., formant peaks) are temporally coded not only by neural channels with CFs closest to those prominences but also by adjacent channels with somewhat different CFs. Thus, temporal coding yields a redundant and relatively noise resistant specification of salient regions in the stimulus spectrum. This is illustrated in Fig. 1 (adapted from [8]) which shows the dominant frequency of response to a vowel stimulus as a function of the CF of the auditory neuron.

With these facts in mind, we conducted a new set of vowel system simulations using an auditory model that generated both excitation patterns and dominant frequency representations based on temporal coding. These two types of representation were then combined into a single spatio-temporal measure of auditory distance. The specific

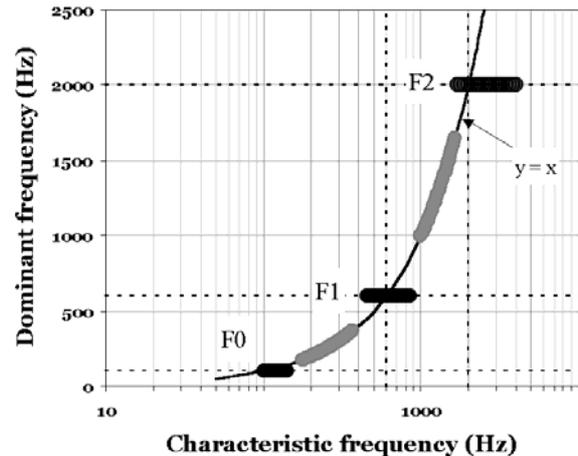


Figure 1. Formant representation in a system using the temporal output of auditory filters (adapted from [8]).

computational steps were as follows:

- (1) Following Lindblom [4], a three-formant space of 'possible vowels' (generated using the articulatory model of Lindblom and Sundberg [9]) was quantized into 19 quasi-cardinal vowels, with roughly equal quantization steps (in Mel) along F1, F2, and F3. The predicted vowel system for a given inventory size was that subset of the quasi-cardinal vowels that maximized the intervowel distances (more precisely, that minimized the quantity $\sum 1/(D_{ij})^2$, where D_{ij} is the auditory distance between vowels i and j).
- (2) An auditory spectrum (excitation pattern) for each vowel was generated as the output of a set of critical-band filters spaced evenly in Bark. The amplitude of each filter output was then rescaled according to equal-loudness contours and converted to Sones/Bark.
- (3) For temporal coding the spectral output of each critical-band filter was subjected to an inverse FFT and the resulting time-domain signal was input to a dominant frequency (DF) detector, which defined the dominant frequency in terms of zero crossings [10]. The output of the DF detectors (one per filter) is the DF representation—a plot of the DF of each filter channel—for a given vowel token.
- (4) In calculating the DF-based distance between two vowels i and j , we first computed the product of DF and spectral loudness (Sones/Bark) channel by channel for each vowel. Then we derived the auditory distance D_{ij} between i and j as the square root of the sum of the product differences squared.

4.1 RESULTS

Figure 2 plots the seven-vowel systems predicted in the formant-based simulations of Liljencrants and Lindblom [3] (left panel) and in the present simulations (right panel). The formant-base approach yields a system with two high vowels between /i/ and /u/, an outcome that occurs rarely if ever among languages with seven-vowel inventories. In

contrast, the system predicted in the current simulations is reasonably similar in structure to the most common seven-vowel systems (e.g., Italian).

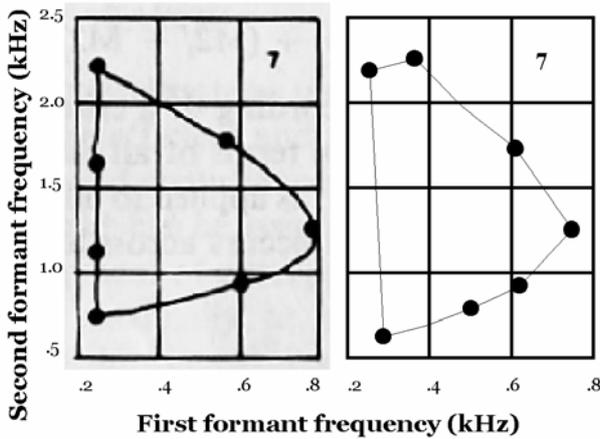


Figure 2. Results of vowel system simulations. Left: The seven-vowel system obtained by Liljencrants and Lindblom [3]. Right: The inventory derived using the present DF-based measure of distance.

4.2 DISCUSSION

Vowel systems predicted on the basis of maximal auditory dispersion have tended to include more high vowels than are attested in actual systems [3,4]. Here we have shown two ways to eliminate this problem. The first way is to calculate auditory distances for vowels in background noise with a spectral shape similar to the long-term average for speech. The second way is to calculate distances using auditory representations that incorporate both spatial (excitation pattern) and temporal (dominant frequency) coding of spectral components. We assume that the latter representations are more complete, and hence more realistic, than representations based on excitation patterns alone.

Earlier we noted how the presence of background noise leads to a perceptual warping of the vowel space—compressing the chromaticity dimension relative to the sonority dimension. It appears that spatio-temporal coding of vowel spectra has an analogous effect on the perceived vowel space. Owing to redundant specification of relatively intense frequency components, salient regions of the spectrum (e.g., formant peaks) contribute disproportionately to auditory representations and hence to auditory distances. Moreover, lower frequency prominences (e.g., F1) contribute disproportionately to auditory distance because of their greater average intensity. This explains why the sonority dimension (corresponding to F1) can accommodate more vowel contrasts than the chromaticity dimension (corresponding to F2 and F3). The perceptual warping of the vowel space is displayed in Fig. 3, which represents distances among the 19 quasi-cardinal vowels in both a formant-based (Mel) scale and a scale based on spatio-temporal (DF-based) coding of vowels.

Notice that the chromaticity dimension (/i/ to /u/) has a greater extent than the sonority dimension (/i/ to /a/, /u/ to /a/) for the formant-base scale, but the reverse is true for the scale derived from spatio-temporal coding.

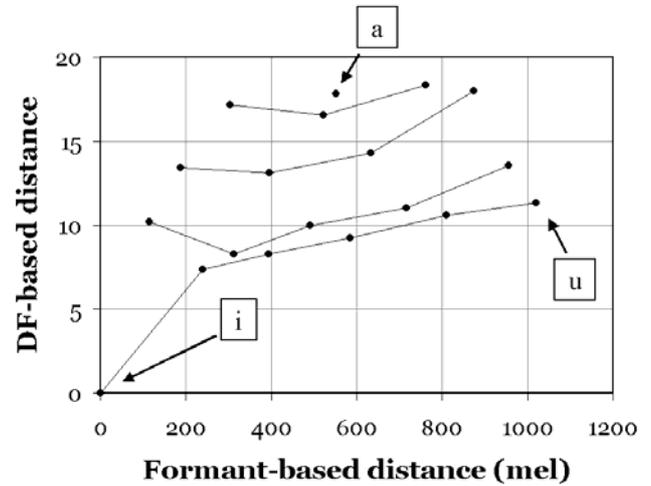


Figure 3. A comparison of two measures of auditory distance applied to a set of 19 quasi-cardinal vowels. Only the distances from the [i] vowel are shown.

5. GENERAL DISCUSSION

The present approach to modeling preferred vowel systems can be motivated on the grounds of its improved auditory realism, relative to earlier approaches, and the resulting gains in predictive accuracy. There are some other advantages as well, two of which will be briefly discussed here. The first concerns a classic issue in phonetics: Is vowel perception better explained on the basis of formant-pattern representations (F1, F2, F3 etc) or on the basis of whole-spectrum representations? The case for formants rests on results of synthetic speech experiments showing that vowel category judgments are much more affected by changes in formant frequencies than by comparable changes in, for example, spectral tilt and formant bandwidth [11]. The case for whole spectra rests on both the positive claim that spectral properties other than formants (e.g., antiformants) *do* have an effect on vowel category judgments and the negative claims that automatic formant tracking algorithms are highly unreliable and that the errors they produce tend not to correspond to those of human listeners [12]. By incorporating a whole-spectrum representation while also granting special status to spectral peaks, the present modeling approach avoids some of the major difficulties associated with an exclusive reliance on either whole spectra or formant patterns.

A second advantage of the present approach is that it helps to unify apparently different accounts of vowel inventory structure. In the Dispersion-Focalization Theory (DFT) of the Grenoble group [13], vowel system predictions are derived by summing two perceptual components: *global*

dispersion and *local focalization*. Global dispersion is abstractly equivalent to our notion of auditory dispersion (although auditory distance in DFT is calculated with respect to a formant-based space similar to that used by [3]). Local focalization refers to intravowel spectral salience related to the proximity of formants. As emphasized by Stevens [14], regions of spectral salience are characteristic of ‘quantal’ vowels (e.g., /i/, /a/, and /u/), which helps to account, on perceptual grounds, for the preferred status of such vowels across languages. In our modeling approach, global dispersion and local focalization (salience) are not treated as separate perceptual components. Rather, salience directly enhances global dispersion as an automatic consequence of spatio-temporal coding. Thus, key elements of Quantal Theory and the Adaptive Dispersion Theory are unified within a single explanatory framework.

REFERENCES

- [1] W. Labov, *Principles of Linguistic Change*, Cambridge, MA: Blackwell, 1994.
- [2] F. Schaeffler, “Typological considerations regarding ‘quantity and ‘tenseness’”, report from the joint PhD program of the Universities of Lund, Stockholm and Umeå, 2002.
- [3] J. Liljencrants and B. Lindblom, “Numerical simulation of vowel quality systems: The role of perceptual contrast,” *Language*, vol. 48, pp. 839-862, 1972.
- [4] B. Lindblom, “Phonetic universals in vowel systems,” in *Experimental Phonology*, J.J. Ohala and J. Jaeger, Eds., pp. 13-44. Orlando, FL: Academic Press, 1986.
- [5] R. Diehl, B. Lindblom and C. Creeger, “Adaptive design of vowel systems,” paper presented at the 4th International Conference on Evolution of Language, Harvard University, March 28, 2002.
- [6] M. Sachs, E. Young and M. Miller, “Encoding of speech features in the auditory nerve,” in *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds, pp. 115-130, 1982.
- [7] S. Greenberg, “Acoustic transduction in the auditory periphery,” *Journal of Phonetics*, Vol. 16, pp. 3-17, 1988.
- [8] B. Delgutte and N. Kiang, “Speech coding in the auditory nerve I: Vowel-like sounds,” *Journal of the Acoustical Society of America*, Vol. 75, pp. 866-878, 1984.
- [9] B. Lindblom and J. Sundberg, “Acoustical consequences of lip, tongue, jaw and larynx movement,” *Journal of the Acoustical Society of America*, Vol. 50, pp. 1166-1179, 1971.
- [10] R. Carlson and B. Granström, “Towards an auditory spectrograph,” in *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds, pp. 109-114, 1982.
- [11] D. Klatt, “Prediction of perceived phonetic distance from critical-band spectra: A first step. IEEE ICASSP, pp. 1278-1281, 1982.
- [12] A. Bladon, “Arguments against formants in the auditory representation of speech,” in *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds, pp. 95-102, 1982.
- [13] J.-L. Schwartz, L.-J. Boë, N. Vallée and C. Abry, “The dispersion-focalization theory of vowel systems,” *Journal of Phonetics*, Vol. 25, pp. 255-286, 1997.
- [14] K. Stevens, “On the quantal nature of speech,” *Journal of Phonetics*, Vol. 17, pp. 3-46, 1989.