TEMPORAL CUES FOR VOWELS AND UNIVERSALS OF VOWEL INVENTORIES

Carrie E. Lang & John J. Ohala

Phonology Laboratory
Department of Linguistics
University of California
Berkeley, CA 94720
clang@trill.berkeley.edu
ohala@cogsci.berkeley.edu

ABSTRACT

Stevens suggested that certain features of consonants are auditorily robust. Such features are abrupt; manifest in 10-30 msec, e.g., [voice], [nasal], [continuant]. Other features such as [palatalized], [pharyngealized] are less robust and are carried on top of (and presumably require more time to be manifested than) the robust features. Robust features are used first--and sometimes exclusively--by languages in constructing a consonant inventory. The less robust features may not be used at all but if they are, the language has already used features from the robust set. We sought to test a similar hypothesis regarding vowels: within the first few tens of msec. only as many vowel contrasts can be differentiated as are attested most commonly in languages of the world -- something like (IPA) /i e a o u/. Analysis of the confusion matrices made by listeners identifying end-gated versions of 11 N. Am. English vowels lends support to the hypothesis.

1. INTRODUCTION

At first glance the segment inventories of languages of the world reveal daunting variety both in size and in the features utilized. From Maddieson [5] we find that segment inventories vary in size from 11 to 141, although 70% range between 20 and 37. The number of features used to classify these distinctive sounds are smaller but still show considerable variation as to how they combine with each other: few proposals for feature systems exceed 40 or so (IPA) and Jakobson, Fant, & Halle [1] make do with 12 binary acoustic features. Nevertheless a closer inspection of these inventories and the features utilized reveals some regularities. As many researchers have noted [3] [4] [5], certain features and feature combinations recur widely in the languages of the world, whereas certain other features or combinations are much less common. Specifically, speech sounds such as [k], [t], [b], [s], [m], differentiated by [±nasal], [±voice], [±continuant], etc. are present in most languages. But speech sounds such as [p^j], [k’], [b^h], differentiated by [±palatalization], [±glottalization], and [±aspiration], etc. are less common. Moreover, those languages that do utilize the less common features and feature combinations tend to have relatively large segment inventories (< 40) that already include the more common sounds and feature combinations.

What is it about some features and feature combinations that account for their being more common or less common in languages’ segment inventories? We should not ignore a purely historical factor in the shaping of languages’ segment inventories: less common sounds such as glottalized stops, aspirated stops, etc. may evolve from--split off from--the more common ones [8]. For example, distinctive aspiration on stops in Ikalanga, a southern Bantu language spoken Botswana, evolved from plain, unaspirated stops in specific phonetic environments [6]. But the question remains why some sounds are more basic and thus prior in the evolution of segment inventories. Lindblom and Maddieson [4] proposed that “Consonant inventories tend to evolve so as to achieve maximal perceptual distinctiveness at minimum articulatory cost.” At present, though, it is difficult to evaluate this hypothesis because of uncertainty over how to measure articulatory cost. As for what makes sounds or features auditorily salient, Stevens [9] offered the following idea:

> “In an acoustic representation of connected speech we find certain regions where there are rapid (10-30 msec) changes in a number of acoustic parameters, e.g., amplitude, periodicity, and spectrum. A hypothesis that has emerged from our and Chistovich’s research, is that the attention of the listener is drawn to these regions, more so than to other regions where changes are less rapid. These regions are, first of all, markers of consonants, but additional information can also be packaged in them along several orthogonal dimensions. We believe languages therefore tend to ‘select’ a consonant inventory that uses up most of these dimensions. These primary dimensions are: [±voice] (presence/absence of periodicity), [±nasal] (presence / absence of low-frequency murmur), [±continuant] (unbroken / interrupted sound), [±grave] (low- / high-frequency tilt to the spectrum), [±compact] (energy spread out / concentrated). After processing the information in these regions of rapid change (= high rate of information transfer), the listener’s attention may focus on the remaining regions and here lie the cues for such dimensions as palatalization, pharyngealization, clicks, etc. It logically follows that the learning of (or introduction of) such distinctions will *follow* the learning of distinctions coded in regions to which primary attention is directed.”

These ideas were further elaborated by Stevens and Keyser [10]. Certain well known tendencies in sound structure and phonological universals lend support to Stevens’ notion that the commonly used features are auditorily robust because they are abrupt and can be detected in a short time window. A length contrast on consonants or vowels, which is among the 'less common' distinctive features, obviously takes longer to detect. Furthermore many of the less common features differentiating consonants, e.g., aspiration, glottalicness, affrication, voice quality distinctions, are manifested on the *release* of consonants, i.e., several 10s of msec. after the manifestation of the more robust features of voicing, continuancy, etc. In vowels, most diphthongs begin with a vowel nucleus similar to another non-diphthongized vowel and then terminate in a distinctive glide, e.g., in English (ARPABET) [ay], [aw]. In such cases it is clear that these diphthongs cannot be differentiated from the other vowels without requiring a long time window.

We sought to evaluate Stevens' notion as it might apply to vowels. Similar phonological generalizations apply to vowel inventories as were made about consonants, above. The modal vowel inventory is something like [i e a o u]. Moreover even if a language has more than five vowels, it still has these vowels. These, then, following the logic of Stevens' idea, are probably differentiated by features that are auditorily robust. The features needed to enlarge a vowel inventory, e.g., vowels with "opposite" rounding (e.g., front rounded [y]), or [±nasal], [±tense], [±long], [±diphthongized], etc., should be auditorily less robust, i.e., slower. We sought, therefore, to see if CV syllables, where the V was one of 11 N. Am. English vowels, which had been gated in varying degrees, would show (a) a reduction of contrastiveness with decreasing duration (greater truncation from the end of the syllable) and (b) the reduced contrastiveness at short durations correlating with the presumed universal vowel contrasts drawn from a set similar to [i e a o u].

2. METHOD

2.1 Stimuli

Two similar studies were conducted. The first involved an adult female speaker of California English who recorded C1V syllables where C1 = [d], [g], and V = (ARPABET) [iy ih ey eh ae aa ah er oh uh uw]. The second involved a male speaker of California English recording [C1VC2] syllables where C1 = /h/ or /q/ (glottal stop) and V = the same set as above and C2 = /d/. These utterances were digitized at 22 kHz and the vowels were gated at regular increments beyond the offset of the initial consonant. In the first study the regular increment was 25 msec and, in the second study, 20 msec. (Henceforth we refer to a given set of 11 stimuli by specifying the initial consonant, C1, and the gate point as measured from the consonantal offset, e.g., **g**50 and **h**40.) At the gate point the speech amplitude was ramped down over 10 msec at the same time as white noise was ramped up to approximately -2 dB below the peak amplitude of the loudest sample. The white noise was extended so that each token was 350 msec long. This led the listeners to imagine that part of the syllable had be masked by the noise while the gate itself did not add any spurious consonantal cues.

2.2 Subjects

Both studies employed fifteen volunteer listener-subjects. All were native speakers of English, students staff and visitors at the Department of Linguistics at UC Berkeley. Some of the same subjects served in the two studies.

2.3 Task

Subjects hear the randomized stimuli presented via a microcomputer over headphones and were required to identify their vowel qualities by clicking on one of the eleven vowels which were arranged in a traditional vowel quadrilateral and labeled with pseudo-orthographic symbols as well as words exemplifying the vowel, e.g., ARPABET /iy/ was labeled 'ee' and with the word 'beet'. Subjects' responses were recorded automatically and written to a text file for subsequent analysis. The text was self-paced and took 20 minutes or less.

3. ANALYSIS & RESULTS

Subjects' responses were converted to confusion matrices, one for each C1 and each gate point. Table 1 gives a raw confusion matrix for **g**50. These matrices, which often showed asymmetries ofconfusion (where vowel *x* was confused with *y* more often than the reverse) were then converted to a similarity matrix showing the similarity, *s*, of each pair of vowels, *x*, *y*, using the formula [2]:

$$s\,(x\,,y) = \frac{f\,(x\,,y)}{f(x\,,x)} + \frac{f(y\,,x)}{f(y\,,y)}$$

where *f(x,y)* is the frequency with which the vowel *x* was heard as the vowel *y*, etc. Table 2 presents the similarity matrix derived from the data in Table 1.

	ae	aa	eh	ey	iy	ih	oh	uh	er	uw	ah
ae	2	1	12								
aa		15									
eh		1	9	3		2					
ey		1	9	3		2					
iy					10	3		1			1
ih			4	2	1	8					
oh		1					9	4			1
uh		1					1	12			1
er		3					1	2	4		5
uw								7		8	
ah		2					3				10

Table 1: Raw confusion matrix for **g**V syllables at 50 msec gate; vertically: presented, horizontally: identified.

3.1 Variation in Contrast

The values in the matrices like those in Table 2 are a reflection of the degree of contrast of a given set of stimuli, i.e., for each C1 and each gate point. The larger the average value of the cells in such a matrix (that is, the greater is the similarity between vowels), the less contrast there is. Fig. 1 shows how this value changed between two gate points, 50 msec. and 100 msec. There was always a decrease in confusion, hence more

	ae	aa	eh	ey	iy	ih	oh	uh	er	uw	ah
ae		0.5	6	0	0	0	0	0	0	0	0
aa			0.11	0.33	0	0	0.11	0.08	0.75	0	0.2
eh				3.33	0	0.72	0	0	0	0	0
ey					0	0.92	0	0	0	0	0
iy						0.43	0	0.1	0	0	0.1
ih							0	0	0	0	0
oh								0.53	0.25	0	0.41
uh									0.5	0.88	0.08
er										0	1.25
uw											0
ah											

Table 2: Similarity matrix for **g**V syllables at 50 msec gate (derived from data in Table 1).

contrastiveness, with the greater gate points. As such, this is not surprising, since, with more of the signal being present, more information should be conveyed. However this finding does have interest for the following reason: vowel quality is often thought of as something which can be characterized by steady-state resonances. Diphthongs constitute exceptions to this but except for /ey ow/ we avoided the recognized diphthongs in our stimuli set. But is was not only contrasts such as /ey eh/, where one was a clear diphthong, which improved with a later gate point; contrasts increased across the board. This reinforces the discovery by Nearey and Assmann [7] that most North American English vowels have some degree of diphthongization which, though slight in some cases, is perceptually important.

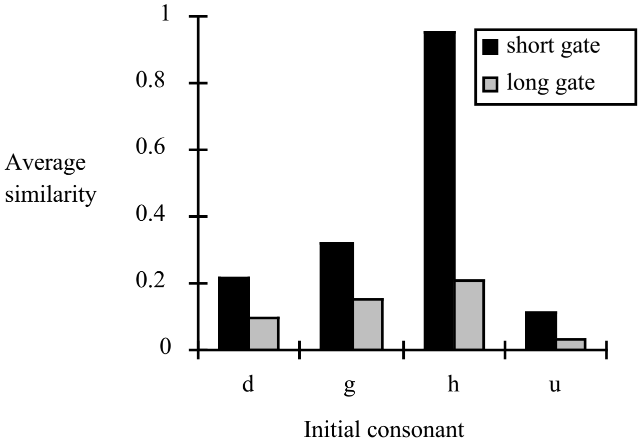

Figure 1: Average similarity between vowels at two gate points: for C1 = [d], [g], 50 and 100 msec; for C1 = [h], [q] (glottal stop), 40 and 80 msec.

3.2 Clustering

In order to determine whether the most robust contrasts at the shorter gate points corresponded roughly to the "universal" vowel inventory, i.e., similar to (IPA) /i e a o u/ (plus or minus two, perhaps), we performed a hierarchical clustering of the similarity matrices [2]. Using the computer program SYSTAT to compute values, we calculated the "distances" between vowels using the Percentage metric, and the linkages (clusters) among these distances using the Average Linkage method [11].

Figs. 2 and 3 give two such cluster analyses for **h**40 and **h**80 respectively. These analyses showed that when C1 = [d] or [g], the confusions at gate 25 were heavily influenced by the consonantal transition. We therefore disregarded the values at the shortest gate points. Although it is difficult to quantify, in general the analyses showed that the strongest clusters at the 50 msec gate corresponded roughly to the "universal" vowel distinctions and that these clusters became weaker at longer gates, i.e., the vowels became individually more distinct. This is evident in Figs. 2 and 3. The first clustering was generally into front (/iy ih ey eh ae/) and back (/uw uh ow ah aa/) vowels. /er/ sometimes patterned separately from these two clusters (as in Fig. 2) or patterned with the back vowels. Then common clusters were /iy ih/, /ey eh ae/, /uw uh ow/, /aa ah/. /iy/ and /aa/ often separated from the other clusters at an early gate. These clusters became weaker at longer gates although /uw uh/ and /ey eh/ frequently continued to be confused even at gates ≥ 100 msec.

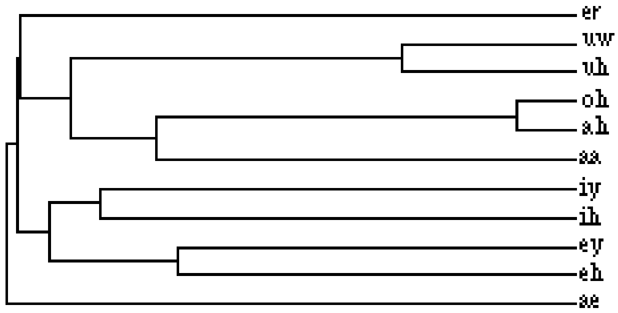

Figure 2: Hierarchical clustering tree for **h**40 msec.

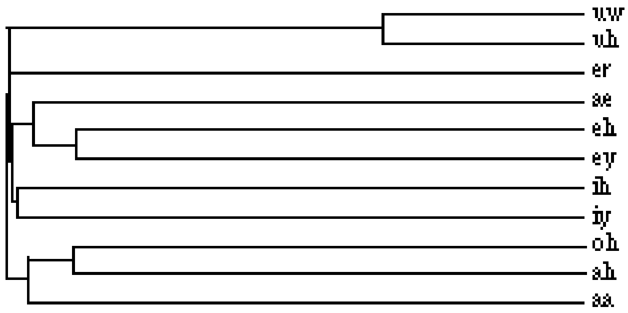

Figure 3: Hierarchical clustering tree for **h**80 msec.

4. DISCUSSION AND CONCLUSION

We are aware that in order to be confident in the conclusions we draw from this study we need more listeners and we need to do similar studies with other languages, especially those whit large vowel inventories making use of other distinctive features than those needed for English vowels, e.g., nasalization and opposite rounding. Such studies are planned for the future. Pending that, however, we tentatively conclude that our results lend some support to our original hypothesis that: the 11 n. Am. English vowels we studied are differentiated by features, some of which are auditorily robust and which can be detected in a relatively short time window (≤ 50 msec) and others which are not robust and which require a longer time window (≥ 50 msec) to detect. The robust features are those found in the modal vowel inventory in languages of the world, i.e., something close to the vowel set (IPA) /i e a o u/.

5. REFERENCES

1. Jakobson, R., Fant, C. G. M., Halle, M., *Preliminaries to speech analysis. The distinctive features and their correlates.* [Acoustic Laboratory, MIT, Technical Report No. 13] Cambridge: Acoustic Laboratory, MIT, 1952.

2. Johnson, S. C., "Hierarchical clustering schemes," *Psychometrika* 32.241-254, 1967.

3. Ladefoged, P., Maddieson, I., *The sounds of the world's languages*, Oxford, Blackwell, 1996

4. Lindblom, B. & Maddieson, I., "Phonetic universals in consonant systems," in L. M. Hyman & C. Li (eds.), *Language, speech, and mind*, Routledge, London. 62-78, 1988.

5. Maddieson, I., *Patterns of sounds*, Cambridge University Press, Cambridge, 1984.

6. Mathangwane, J., *Phonetics and phonology of Ikalanga: a diachronic and synchronic study.* Doctoral Dissertation, University of California, Berkeley, 1996.

7. Neary, T. M. and Assmann, P., "Modeling the role of inherent spectral change in vowel identification," *The Journal of the Acoustical Society of America* 80:1297-1308. 1986.

8. Ohala, J. J., "The perceptual basis of some sound patterns," In B. Connell & A. Arvaniti (eds.), *Papers in Laboratory Phonology IV.* Cambridge: Cambridge University Press. 87-92,. 1995.

9. Stevens, K. N., "Discussion," *Proceedings of the 9th International Congress of Phonetic Sciences, Copenhagen, 1979*, Vol. 3.185-186, 1980.

10. Stevens, K. N. & Keyser, J., "Primary features and their enhancement in consonants," *Language* 65:81-106. 1989.

11. Wilkinson, Leland. *SYSTAT: The system for statistics.* Evanston, IL: SYSTAT, Inc., 1989.