LISTENERS' IDENTIFICATION OF SPEECH SOUNDS IS INFLUENCED BY ADJACENT "RESTORED" PHONEMES

JOHN J. OHALA and DEBORAH FEDER

Department of Linguistics
University of California
Berkeley, California 94720 (USA)

ABSTRACT

When listeners' identifications of speech sounds are influenced by adjacent sounds is it only the quantitative phonetic characteristics of these neighboring sounds that matter or could their qualitative linguistic identity play a role? We tested this by leading subjects to restore or induce the noise-obliterated medial consonant in V_1CV_2 utterances by first presenting them with several prior utterances where this medial consonant could be heard clearly and was consistently the same, either a /b/ or a /d/. Included as V_1 were synthetic vowels from the /i - u/ continuum. More /u/'s were identified out of this continuum in the environment of physically present /d/'s than /b/'s. Restored /d/'s had the same effect (vis-a-vis restored /b/'s), thus indicating that the influence of context need not operate only via physical phonetic features. These results challenge the 'direct realist' theories of speech perception as well as claims that 'invariant' features of speech sounds are to be found by normalizing these features with respect to the physical phonetic characteristics of their surroundings.

INTRODUCTION

There is abundant evidence that listeners identify speech sounds in part by normalizing them with respect to their phonetic context [1, 2, 3, 4, 5]. How is this done? Are the physical phonetic parameters of the context used to adjust recognition threshholds or is it enough for the listener just to know the (categorized) linguistic identity of the context? We investigated this questions through a series of perceptual tests involving listeners' identification of synthetic vowel stimuli in isolation and in consonantal contexts. (In what follows, we collapse descriptions of two of these tests--a pilot study and a main test, which differ in some details. The description is kept general and details and differences given only where essential.)

IDENTIFICATION OF VOWELS IN ISOLATION

First, we constructed a 17-step linear stimulus continuum between the vowels /i/ and /u/; see Fig. 1. The continuum endpoints were modeled on the first 100 msec of natural /i/ and /u/ pronounced in isolation by an adult male native speaker of American English. (Since the 'crossover' from /i/ to /u/ was expected to happen in the middle of this continuum, some stimuli near the end points were omitted

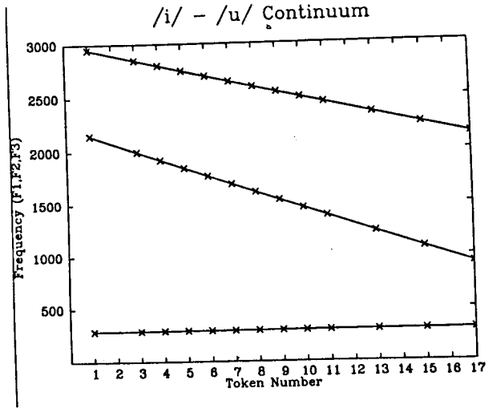

FIGURE 1

from the study, those steps showing absence of 'x''s on the formant parameters.) In a forced-choice identification task, listeners gave the response function shown in Fig. 2, where the ordinate shows percent identification of tokens as /u/ and the abscissa, the /i/ -- /u/ continuum (/i/ at the left and /u/ on the right). These results were obtained from 28 native American English-speaking listeners, each responding twice to each stimulus for a total of 56 responses per data point.

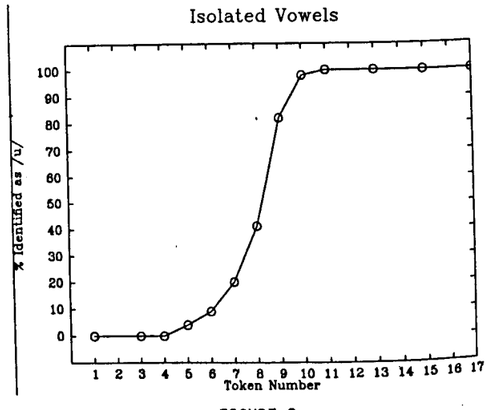

FIGURE 2

IDENTIFICATION OF VOWELS IN CONSONANTAL CONTEXT

Second, we sought to replicate the finding that this function shifts when the vowels are put in certain consonantal contexts [6, 7]. Using digital splicing, we embedded our vowels in nonsense words of the form /ibə/, /ubə/, /idə/, /udə/--where the /bə/ and /də/ were excised from the same speaker's natural utterances of /abə/ and /adə/. In another forced-choice task, listeners exhibited a shift in the earlier response so that more /u/'s were heard in the context of a following /d/--presumably because listeners allowed for and factored out the elevated F2 that alveolar consonants produce on back vowels [8, 9]. Fig. 3 shows the results from the pilot test which had 14 listeners and a total of 28 responses per stimulus and Fig. 4, the results from the main test with 28 listeners and 8 judgements per stimulus per listener for a total of 224 responses per data point.

IDENTIFICATION OF VOWELS IN CONTEXT OF RESTORED CONSONANTS

Third, we asked whether this same shift in the function due to consonantal context would appear even if the consonants were not physically present in the signal but if the listeners instead just imagined that they were. We attempted to make listeners believe that the /b/ in the /ibə/, /ubə/ tokens or the /d/ in the /idə/, /udə/ tokens were present when they weren't, by using the technique called "phoneme restoration" [10] where high redundancy of the message induces the listener to "fill in" missing elements. The redundancy in our case was provided by presenting all our stimuli in two major blocks, one in which the medial consonant was or seemed to be a /b/, and the other in which /d/ was or appeared to be the consistent medial consonant. To enhance this priming, we also began each block with a number of tokens in which the consonant was clearly present. In approximately 15 to 20% of the stimuli in each block we completely replaced the medial consonant by white noise (always equal in intensity to the average intensity of the voicing during the consonantal closure).

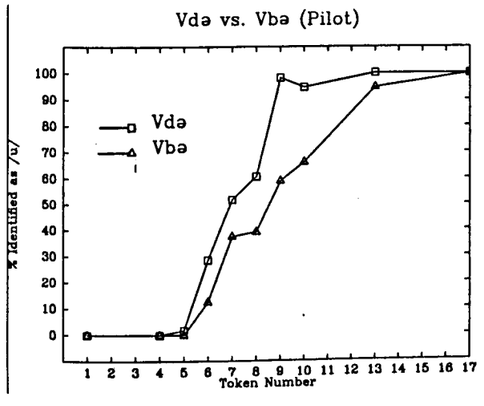

FIGURE 3

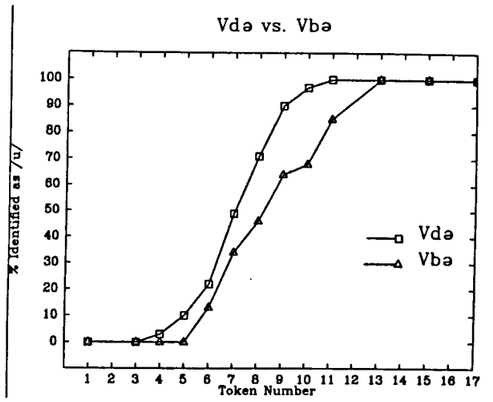

FIGURE 4

The schwa portion of these latter stimuli consisted of several periods from the center of the naturally spoken schwa, i.e., a portion with minimal, if any consonantal 'coloring'. To insure that restoration of the consonants did not derive from any residual cues remaining in the schwa, we used the schwa that had originally followed a /b/ in the stimuli where we wanted a /d/ to be restored by our listeners, and, similarly, a schwa that had originally followed a /d/ where we wanted them to restore a /b/. As foils to prepare listeners for hearing noise, another 15% of the tokens also contained noise bursts at various locations, such as during the intervals between stimuli, over the vowel, or superimposed on (but not replacing) the consonantal closure. Listeners were told that the noise bursts served as distractor to the identification task.

The results from the pilot test are shown in Fig. 5 (where only a fraction of the entire vowel continuum was studied). There was an important difference between the pilot and the main test. In the former, subjects wrote down the entire VCV utterance they heard. The results shown in Fig. 5

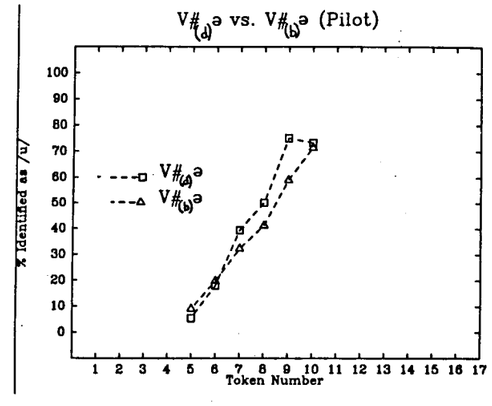

FIGURE 5

are those for which the consonant reported was the one we were trying to get subjects to restore. The shift for these tokens is in the same direction as that for the physically present consonants. Each data point represents the average of 28 responses. For the main test we gave subjects answer sheets with the "b"'s and "d"'s already present; they were only required to fill in the vowel. These results are shown in Fig. 6. Here there were 56 judgements per data point: 28 subjects times 2 judgements per stimulus.

The difference between the two consonantal contexts has been shown to be significant by several preliminary curve-fitting analyses. However, our statistical findings cannot yet be considered conclusive since the irregular shapes of some of our curves have made it difficult to fit enough of them to any single statistical model to make comparisons among them possible and meaningful.

DISCUSSION AND CONCLUSION

We conclude that since the magnitude of these shifts under the restored or imagined phonemes is not as great as the case with physically present phonemes, listeners adjust their identification threshholds in part due to processing of the actual acoustic parameters of the signal and in part--perhaps as a default case--on the basis of the linguistic identity of contextual segments which may in some cases be provided through non-phonetic channels. In the latter case, one imagines that the listener knows from experience the typical effects of one segment on another and uses this information to adjust recognition threshholds.

This result should not be surprising: it is well recognized, for example, that in the visual domain we achieve a high degree of color constancy in part by factoring out the distorting influence of the hue of ambient illumination but also by our knowledge of what the colors of typical objects are and how these colors are modified in various situations. It would be remarkable if something similar did not apply in the case of speech perception.

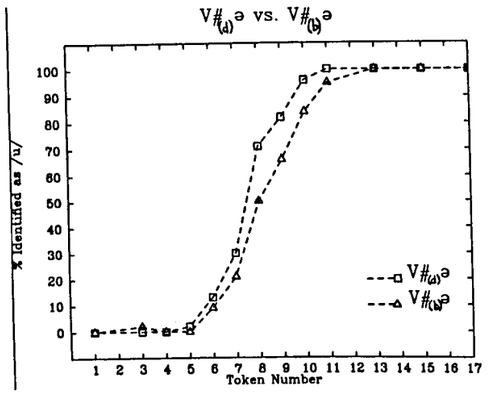

FIGURE 6

We believe that these results present a challenge to the direct realist view of speech perception which holds that all the information needed to identify the intended message elements of speech are present in the acoustic signal and can be discovered by the listeners without the need for inferences and the like; for example, Fowler [11] claims that

> "...[speech] perception must be direct and, in particular, unmediated by cognitive processes of inference or hypothesis testing, which introduce the possibility of error."

Here, listeners showed that their speech sound identification was influenced by entities not physically present in the signal. Specifically, the identity of an ambiguous stimulus was resolved by reference to predicted effects of an assumed, i.e., hypothesized, environment. In this latter respect, our results are compatible with those of Mann and Repp [4], who showed that listeners' identification of one variable stimulus shows a discontinous shift as a function of their identification of an another adjacent ambiguous segment.

These results do not actually refute the direct realist view, though, since in one recent formulation of it [11, 12], it has been allowed that listeners can sometimes operate on what might be called "automatic pilot"--that is, by making assumptions, even unwarranted assumptions, about what is present in the signal. Nevertheless, direct realists would maintain that, in principle, if listeners just paid closer attention to the speech signal, speech perception would be accomplished "directly" and they wouldn't make the kinds of perceptual "mistakes" as they did in our study. Our results challenge this view, too, though, by raising the following question: if listeners are capable of integrating non-phonetic information into their recognition task, isn't it likely that the speaker knows this and only puts enough energy, precision, and detail into the generation of the speech signal as the listener requires? It is our impression that the speaker often does not in fact put sufficient phonetic details into the speech signal to permit decoding of the message in a direct way.

These results also bear on the question of whether there are or should be acoustic invariants of phonemes (or other message units) in speech [13]. Clearly, these results add to the evidence that absolute invariance is not necessary; the listener has ways of accomodating variation. Stevens [14] has suggested that relative invariance may be more likely than absolute invariance, i.e., a given unit or a distinctive feature characterizing it may be invariant with reference to the phonetic environment it appears in--in his view they physical phonetic environment. The notion of "relative invariance" is compatible with our results but only if the linguistic identity of the context, not exclusively its physical properties, are admitted as figuring in the normalizing process.

Finally, we think we have demonstrated a potentially quite useful way of inducing listeners to restore missing elements in speech which does not require construction of semantic, syntactic, or other higher-order redundancies.

ACKNOWLEDGEMENTS

We benefited from the comments of Klaus Kohler, Bruno Repp, Mary Smith, and Richard Warren on earlier versions of this paper. Responsibility for the contents of the paper, including any errors, is ours, though. This research was supported by a grant from the Sloan Foundation to the Cognitive Science Program at the University of California, Berkeley.

REFERENCES

[1] Ladefoged, P. & Broadbent, D. E. 1957. Information conveyed by vowels. J. Acous. Soc. Am. 29.98-104.

[2] Pickett, J. M. & Decker, L. 1963. Time factors in perception of a double consonant. Lang. & Speech 3.11-17.

[3] Mann, V. A. & Repp, B. H. 1980. Influence of vocalic context on perception of the [š] vs [s] distinction. Perception & Psychophysics 28.213-228.

[4] Mann, V. A. & Repp, B. H. 1981. Influence of preceding fricative on stop consonant perception. J. Acous. Soc. Am. 69.548-558.

[5] Fowler, C. A. 1981. Production and perception of coarticulation among stressed and unstressed vowels. J. Speech & Hearing Res. 46.127-149.

[6] Ohala, J.J, Riordan, C. J., & Kawasaki, H. 1978. The influence of consonant environment upon identification of transitionless vowels. J. Acous. Soc. Am. 64.S18.

[7] Ohala, J. J. 1981. The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (eds.), Papers from the Parasession on Language and Behavior. Chicago: Chicago Linguistic Society. 178-203.

[8] Lindblom, B. 1963. Spectrographic study of vowel reduction. J. Acous. Soc. Am. 35.1773-1781.

[9] Stevens, K. N. & House, A. S. 1963. Perturbations of vowel articulations by consonantal context: An acoustical study. J. Speech & Hearing Res. 6.111-128.

[10] Warren, R. M. 1970. Perceptual restoration of missing speech sounds. Science 167.392-393.

[11] Fowler, C. A. 1986a. An event approach to the study of speech perception from a direct realist perspective. J. Phonetics 14.3-28.

[12] Fowler, C. A. 1986b. Reply to commentators. J. Phonetics 14.149-170.

[13] Stevens, K. N. & Blumstein, S. E. 1981. The search for invariant acoustic correlates of phonetic features. In P. Eimas & J. Miller (eds.), Perspectives on the study of speech. Hillsdale, NJ: Erlbaum.

[14] Stevens, K. N. 1986. Models of phonetic recognition II: An approach to feature-based recognition. Proc., Montreal Symposium on Speech Recognition, July 21-22, 1986, McGill University, Montreal. Canadian Acoustical Association. 67-68.