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

JOHN J. O'HALLA 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 VCV 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 /d/. Included as V were synthetic vowels from the /i/-/u/ continuum in the environment of physically present /d/ 's. Restore /b/ '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 thresholds or is it enough for the listener just to know the (categorized) linguistic identity of the context? We investigated this question 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 when 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 msecs 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.

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 /i/ or /u/ where the /i/ and /u/ were excised from the same speaker's natural utterances of /i/ and /u/.

In another forced-choice task, subjects 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].

The results from the pilot test which had 14 listeners and a total of 29 responses per stimulus and Fig. 4, the results from the main test with 28 listeners and 8 judgments 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 /d/ in the final, /u/ tokens or the /d/ in the /i/ tokens were present when they weren't. By using the technique called "phonemic 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 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).
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 elements of speech are present in the acoustic signal and can be discovered by listeners without the need for assumptions. 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 discontinuous shift as a function of their identification of an 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 would 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, is it likely that the speaker knows this and only puts enough energy, precision, or information in the speech signal as the listener requires? It is our impression that the speaker often does not in put sufficient linguistic 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 accommodating variation. Stevens [14] has suggested that relative invariance may be more likely than absolute invariance, i.e., a given unit or a distinctive feature characteristic it may be invarient with reference to the phonetic environment it appears in - in his view this notion of “relative invariance” is compatible with our results but only if the phonetic 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 retron 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, although the research was supported by a grant from the Sloan Foundation to the Cognitive Science Program at the University of California, Berkeley.

REFERENCES
