HYPERCORRECTION IN SPEECH PERCEPTION

J. J. Ohala

Department of Linguistics
University of California
Berkeley, CA 94720

E. E. Shriberg

Department of Psychology

ABSTRACT

Listeners hearing distorted speech may be able to "correct" it if they have enough information about the nature of the distortion. We present evidence that this ability to correct a distorted speech signal is bought at a slight cost, namely, hyper-correction. We presented American English listeners with brief (85 msec) samples of 11 vowels that had been 1 kHz low-passed filtered. When presented in a way that gave them no chance to learn the filter characteristics they made 33.3% correct identifications. Many of the errors involved front vowels (those having a high Formant 2) being confused with back vowels (which often have F2 and F1 fused at 1 kHz or less). When the samples were preceded by a redundant precursor sentence filtered in the same way as the vowel samples, listeners raised their correct identifications to 50.1% but there was an increase in "hyper-correction" errors, i.e., where back vowels were identified as front vowels. This may provide clues as to how listeners "correct" distorted signals.

I. INTRODUCTION

As has often been remarked, speech is an extremely robust signalling system. The message generally survives in the face of noise, filtering, speaker variability, and many other distortions.[1] In man-made communication systems robustness is bought with redundancy and the same must be true with speech. There are apparently many types or levels of redundancies and different ways that listeners utilize them. We know from the work of Warren on phoneme restoration [2] that listeners can "fill in" obliterated phonemes and syllables if there is lexical, syntactic, semantic, or pragmatic redundancies elsewhere in the utterance. Phonological redundancy exists, too: speech perception studies show that there are multiple cues to most phonemic (lexically distinctive) contrasts.[3] The channel the signal is heard on may provide useful redundancy: there is evidence that listeners can factor out some of the variability in speech if they can identify the source of the variability [4]. A remarkable tape prepared by Jack Boehm [5] showed that heavily filtered vowels can be perceptually normalized by the simple strategy of hearing them in the context of a short sentence that has been filtered in the same way. This paper was inspired by the Boehm tape. It attempts to discover how this repair or correction of the signal takes place on the basis of cues from a precursor sentence.

II. METHOD

2.1 Speech Samples.

Four adult male native speakers of American English were recorded uttering the sentence "I will now say the word bVb," where V was any of the eleven English vowels [i I eɹ ɛ æ ɚ ʌ uʷ U oʷ ɑ] These samples were digitized at 44.1 kHz after being low-pass filtered at 20 kHz. From each speaker we retained one token of the precursor phrase "I will now say the word..." (henceforth 'precursor') and 85 msec-long segments of the vowels excised digitally from the middle of the bVb syllables. These vowel segments, which contained no discernible trace of the consonantal transitions, had their amplitude envelope ramped (10 msec) at the beginning and end to eliminate abrupt onsets and offsets. There was approximately 250 msec between the end of the precursor and the onset of the vowel. We thus worked with a basic set of 44 stimuli (11 vowels X 4 speakers), each vowel preceded by the precursor in the voice of the same speaker who had uttered the stimulus vowel. As described below, for various conditions the vowels and the precursor sentence were low-pass (L-P) and high-pass (H-P) filtered. The L-P cutoff was 1 kHz and the H-P 1.5 kHz; the attenuation of the filters was better than 60 db/octave. The overall amplitude of all vowel stimuli was adjusted subjectively by the experimenter in order that each vowel stimulus sound more or less equally loud, especially in relation to the precursor.

This set of 44 stimuli, each trial consisting of precursor plus vowel, was presented in 6 blocks, plus an initial training block, as shown in Table 1 where the original blocks are given on the left and modifications of those blocks (if any), on the right. Modified blocks were presented to listeners; the original blocks were the focus of statistical analysis. The rationale for the various blocks and their structure is as follows. The training blocks familiarized the Ss with the format of the test; the deliberate unequal ratios of high and low vowels in these two short blocks was to get Ss used to the idea that certain types of vowels might be disproportionately represented in some of the blocks (as it would no doubt appear to them when they came to Block 5, for example, that back vowels predominated). Block 1 presents one control condition where all parts of the stimuli are presented Hi-Fi; this would also indicate the highest level of intelligibility obtainable given the truncation of the vowels, etc. Blocks 2 and 3 (after being mixed to form Blocks 2' and 3') is

another type of control; they would presumably yield the lowest level of intelligibility of the vowels since listeners would have little or no chance to determine the shape of the filtering applied to the vowels. Blocks 4 and 5 are the experimental conditions and, of these, Block 5 with both precursor and vowel L-P is most important. They would indicate whether Ss could reduce their error rate if they had experience with the way the vowels were filtered. Blocks 6a and 6b (after being randomly mixed to form Block 6) would indicate if it was the blocked presentation of stimuli similarly filtered that permitted reduction of error rate or whether just the precursor sentence itself was sufficient for this. Blocks 6a and 6b were half the size of the others because we were concerned that the experiment was becoming too long.

The training blocks were presented first, then Block 1; the order of presentation of the remaining blocks was varied randomly for each S.

2.2 Subjects.

Ss (listeners) consisted of 16 phonetically-trained native speakers of English, 10 females and 6 males, recruited as volunteers from the students of the Departments of Linguistics or Psychology at University of California, Berkeley. Phonetically-trained subjects were used because they were able to use IPA symbols without requiring training. Subjects were paid for their participation.

2.3 Procedure and Instructions.

The test was administered individually to the subjects as follows. A subject was seated in front of the console of a Macintosh computer equipped with a mouse. Stimuli were presented over high-quality earphones at a comfortable loudness. Subjects viewed a screen with the eleven IPA stimulus vowels arranged as in a traditional vowel space with an arrow whose position could be controlled by a mouse. Ss identified a heard vowel by moving the cursor to the appropriate vowel symbol and depressing the button on the mouse. The computer automatically recorded and tallied their responses and prompted them to start the next block, etc. Ss were instructed orally and via an identical written text. They were told (a) how to identify the vowels at the end of the sentence "I WILL NOW SAY THE WORD...", (b) not to expect that vowels will be equally distributed in each block, (c) that they had 4 sec after hearing the vowel to make a response (after which the computer would proceed to the next trial whether they had responded or not), (d) the length and structure of the experiment.

After going through the practice block, subjects were allowed to perform the rest of the experiment unsupervised. The total time taken to complete the experiment after finishing the practice blocks was approximately 35 min. Due to a software error two subjects, both female, were presented with Blocks 2' or 3' twice and missed the complementary blocks (3' and 2', respectively). Since this resulted in an uneven distribution of vowels presented to them, their data were eliminated in the final analysis, thus reducing the number of subjects to 14.

IV. RESULTS

Confusion matrices were derived automatically from the computer's log of each experimental session). Any off-diagonal response, including 'no response' was counted as an error. Overall there were few 'no responses' (1.7%). Confusions between front and back vowels were counted as any between the vowels [i I eɹ ɛ æ], 'Front Vowels' and [u^{w} U o^{w}], 'Back Vowels'. Table 2 gives the relevant results by blocks: the number of confusions of front vowels as back vowels ('Front > Back'), of back as front vowels ('Back > Front'), other errors, and total correct (also expressed as percentage). A Chi-square test performed on the difference in error rate and error type between Blocks 2 and 4 was significant ($p < .05$) and that between Blocks 3 and 5, highly significant ($p < .001$). Thus the decrease of Front > Back and the increase of Back > Front errors in Block 5 vis-a-vis Block 3 was significant. The slight difference in errors between Blocks 4 and 6a and Blocks 5 and 6b was not significant.

V. DISCUSSION

Many of the confusions due to the artificial shortening of the vowels and the filtering are expected and have been reported by other researchers.[6][7] Many [æ]'s were confused with [ɛ] and [ɑ]'s with [ʌ], due, no doubt to the shortening. [əʴ] was highly distinct under most condition except that it was often given as an incorrect response to other vowels in Block 3, accounting for 16.7% of the errors.

Relevant to our hypothesis, correct identifications increased significantly when listeners had an opportunity to learn how the vowel stimulus was filtered by virtue of hearing the redundant precursor filtered in the same way. Although this improvement in rate of correct identifications was greatest when presented <u>en bloc</u> with other stimuli, block presentation was not significantly better than that where just the filtered precursor by itself was heard (Block 6). The redundant cues regarding filtering characteristics which lead to reduction of errors can be obtained in a very short time span: the approximately 1 sec needed to take in the precursor. (This was true in the Boehm tape, too [5],) More importantly, the error reduction has a cost: an increase in the number of hyper-corrections, that is, errors due to correcting the signal when it didn't require it. This was manifested by the increase from 7 to 16 misidentifications of back vowels as front vowels between Blocks 3 and 5, respectively. The increase in hyper-corrections did little, however, to counter the increase of intelligibility due to cues from the precursor. This result suggests how this correction of the signal is accomplished. The misidentification of back vowels as front vowels is rather remarkable because front vowels have a prominent F2 (and F3) in a mid-to-high frequency region (1.2 to 2.5 kHz) that back vowels lack. Any processing of the speech signal that would <u>increase</u> such an error must also be remarkable. We think there are three essential components to the correction process. First, a listener knows from experience what the acoustic shape of given speech sounds is "supposed" to be. Second,

when hearing the precursor filtered the listener knows what the message is and detects that it is missing energy in certain bands. Third, the listener assumes the vowel stimulus is filtered in the same way and thus imagines or hypothesizes that it may possess acoustic energy (a formant) in the region that has been removed and bases the identification in part on this hypothesized formant. We should not conclude, however, that this result mean that the increase in identifiability owes nothing to the acoustic cues present in the vowel stimulus itself. If a L-P filtered /I/, for example, were completely indistinguishable from an /U/, then the "filling in missing energy" strategy would not have the net positive payoff that it was shown to have; rather half the responses to /I/ and /U/ should have been distributed by chance between /I/ and /U/ --which is not what happened. It is probable that a filtered /I/ and /U/ do differ in some detail and the "filling in" strategy alerts the listener to give more weight to the differentiating cues that survive the filtering. When there is no reason to suspect filtering or when the probable effects of the filtering are unknown (Blocks 2 and 3), the listener must take the signal as "face value". In the L-P condition, for example, the lack of energy in the mid-to-high frequency band is taken as weighty evidence that the stimulus is not a front vowel.

The results of this experiment have a bearing, we think, on the claims made for a "direct realist" theory of speech perception as advocated by Fowler [8]. In the direct realist theory, the speech signal conveys in a direct, one-to-one way all the relevant information the listener needs to perceive speech articulations; no inferences or hypothesis-formation about the signal is necessary. The experiment reported above was suggested by the first author as a way to test the claims of the direct realist theory [9]: if listeners' responses indicated that they "imagined" acoustic energy in a frequency band where it was not physically present, then they would not be perceiving speech "directly", i.e., relying entirely on what is physically present in the stimulus. We believe that this is what our results did show. Fowler, however, has already denied that such an experiment would be crucial since, she claims, the signal would supposedly still contain some information that something was missing [10]. The crucial result of our study which we put to Fowler as a challenge to direct realist perception is that apparently the mechanism that led to erroneous identifications (hyper-correction) -- which clearly involved hypothesizing something that wasn't there -- is the same process that underlies the successful reduction of perceptual errors in the vast majority of the remaining stimuli. We are led to the conclusion that speech perception works best when the listener makes hypotheses which enable him to reconstruct a "full" phonetic representation.

Acknowledgements. We thank: Richard Felciano and David Wessel of the Center for New Music and Audio Technology (University of California, Berkeley) for making their facilities available to us for this experiment; Adrian Freed and John Lowe for software used in this study; Mark Anderson for the design of the filters used; and Ned Neuberg for sending us the tape prepared by Jack Boehm (see note 5).

References

[1] C. Stumpf, Die Sprachlaute. Berlin: J. Springer, 1926.
[2] R. M. Warren, "Perceptual restoration of missing speech sounds," Science 167.392- 393, 1970.
[3] L. Lisker, "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees", Language & Speech 29: 3-11, 1986.
[4] J. J. Ohala & D. Feder, "Listeners' identification of speech sounds is influenced by adjacent 'restored' phonemes," Proc. 11th International Congress of Phonetic Sciences, Tallinn, Estonia, USSR. Vol 4, pp. 120-123, 1987.
[5] J. F. Boehm, "An example of relative perception of vowels," J. Acoust. Soc. Am. 51.122. 1972.
[6] I. Lehiste & G. E. Peterson, "The identification of filtered vowels," Phonetica 4:161-177, 1959.
[7] J. M. Pickett, "Perception of vowels heard in noises of various spectra," J. Acoust. Soc. Am. 29:613-620, 1957.
[8] C. A. Fowler, "An event approach to the study of speech perception from a direct realist perspective," J. Phonetics 14.3-28. 1986
[9] J. J. Ohala, "Against the direct realist view of speech perception," Journal of Phonetics 14.75--82. 1986.
[10] C. A. Fowler, "Reply to commentators," J. of Phonetics 14.149-170, 1986.

Table 1. Structure of the experimental conditions.

Original Block No	# of tokens	Pre-cursor	Vowel	Modified Block No	# of tokens	Special Feature
Training	44	Hi-Fi	Hi-Fi	Trn-1	22	3:2 ratio hi:lo Vs
				Trn-2	22	3:2 ratio lo:hi Vs
1	44	Hi-Fi	Hi-Fi	1	44	
2	44	Hi-Fi	H-P	2'	44	
			randomized			
3	44	Hi-Fi	L-P	3'	44	
4	44	H-P	H-P	4	44	
5	44	L-P	L-P	5	44	
6a	22	H-P	H-P			
			randomized	6'	44	
6b	22	L-P	L-P			

Table 2. Results by block.

Block No	Pre-cursor	Vowel	Special Conditions	Front > Back Errors	Back > Front Errors	Other Errors (%)	Correct (%)
1	Hi-Fi	Hi-Fi		7	0	100	509 (82.6%)
2	Hi-Fi	H-P	intermixed with Blk 3	4	3	239	370 (60.1%)
3	Hi-Fi	L-P	intermixed with Blk 2	138	7	266	205 (33.3%)
4	H-P	H-P		2	2	206	406 (65.9%)
5	L-P	L-P		46	16	241	313 (50.8%)
6a	H-P	H-P	intermixed with Blk 6b	1	0	112	195 (63.3%)
6b	L-P	L-P	intermixed with Blk 6a	24	7	126	151 (49.0%)