PERCEPTUAL RESTORATION OF FILTERED VOWELS WITH ADDED NOISE*

ELIZABETH E. SHRIBERG
University of California at Berkeley

Perceptual restoration is a well-known phenomenon for speech segments in context, but less is known about the effect for stimuli that occur without a linguistic context. The current study investigated restoration in the perception of isolated vowels. Vowels excised from natural speech were lowpass filtered (1000 Hz), which removed the high F_2 characteristic of front vowels. This resulted in high rates of front to back vowel confusions; however, these errors were reduced when highpass filtered noise was added to the lowpass filtered vowels. Although the reduction in errors was accompanied by an increase in back to front vowel errors, addition of noise led to improved performance overall. These results suggest that listeners "restored" the high F_2 of front vowels in the noise condition, despite an absence of linguistic context to influence restoration, and that addition of noise led to more effective utilization of cues below 1000 Hz.

Key words: perceptual restoration, vowel perception

INTRODUCTION

We typically perceive speech that is momentarily masked by a louder sound as continuing under the masker. Under normal circumstances we perceive veridically, since in most cases the signal has indeed persisted, rather than disappeared in exactly the regions occupied by the masker. Perceived continuity constitutes an illusion, however, in laboratory experiments in which portions of the signal under the masker have actually been removed.

Illusory perception of deleted speech has been demonstrated in the laboratory for both temporally and spectrally degraded signals. In a classic study of periodically interrupted speech, Miller and Licklider (1950) incidentally observed an increase in perceived continuity (the "picket fence effect") when noise, as opposed to silence, filled the intermittent gaps. A second, striking example from the temporal domain is the "phoneme restoration effect" (Warren, 1970), in which a deleted phoneme is "restored"

---

* The author wishes to thank: John Ohala, for valuable guidance in conducting the research; David Wessel for use of the Center for New Music and Audio Technologies in Berkeley; Adrian Freed, Guy Garnett, Mark Anderson, and John Lowe for aid in stimulus preparation and software design; and Bruno Repp, Jeri Jaeger, Steve Greenberg, Ray Weitzman, and Eric Jackson for many helpful comments on earlier drafts.

Address correspondence to Elizabeth Shriberg, EJ188 SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, U.S.A.

by listeners when a cough is substituted for the excised portion of the signal. Continuity in the spectral domain has been reported by Bashford and Warren (1987) in a study that showed an increase in intelligibility, as well as perceived spectral completion (the speech sounded "wideband"), when oppositely-filtered noise was added to alternating highpass filtered and lowpass filtered portions of a continuous speech signal. While the stimuli and response measures used vary greatly in these and many related experiments, all demonstrate the tendency for listeners to perceive or "restore" missing information when temporal or spectral gaps in the signal have been "covered" by a masker.

Restoration can be explained in part as a general perceptual compensation for masking, which occurs for nonspeech stimuli as well when certain conditions are met. It is necessary, for instance, that the obscuring sound contain components that stimulate the peripheral auditory units that would be stimulated by the signal (were it present), and that there be no perceptible evidence that the signal has been turned off under the masker. Further principles are discussed in detail in Warren (1982) and Bregman (1990). Reports of restoration for nonspeech stimuli include perceived continuity of a steady or gliding tone behind a masker (Thurlow and Elfner, 1959; Houtgast, 1972; Bregman and Dannenbring, 1977) and "contralateral induction" (Egan, 1948; Warren and Bashford, 1976), a lateral compensation for asymmetrical masking in which attribution of noise components to a signal displaces the signal perceptually in space.

The case for speech stimuli is complicated, however, by the presence of linguistic context. Unlike the case for nonspeech signals, in which restoration follows from basic rules of perceptual organization (e.g., Gestalt principles of "good continuation", "frequency proximity", and "closure"; see Wertheimer, 1912), restoration of speech is further influenced by lexical, semantic, and syntactic knowledge. Warren and Sherman (1974), for example, demonstrated the role of sentence context by showing that in hearing the sentence: "It was found that the _eel was on the orange/axle" listeners restored "p" in the context of "orange" but "wh" in the context of "axle" Restoration has been shown to depend on the degree of redundancy in the signal, with a stronger restoration effect occurring for sentences as compared to word lists (Bashford and Warren, 1987), and for sentences read forwards as compared to backwards (Bashford and Warren, 1979). And, in general, restoration is in some sense "driven" by listeners' urge to make sense of what they hear.

A question to ask, then, is whether restoration occurs for speech stimuli when there is no context present to "drive" restoration. The present study investigated this question in an experiment modelled after a suggestion by Ohala (1986), by using spectrally degraded signals (i.e., with part of the spectrum removed) similar to those used by Bashford and Warren (1987). (The experiment is described in detail in Shriberg, 1990; see also Ohala and Shriberg, 1990). In this study, vowels excised from natural speech were lowpass filtered (1000 Hz), so that the high-frequency F_2 characteristic of the front vowels was removed. Subjects were asked to identify these vowels under conditions with and without added highpass filtered noise. An advantage of using these stimuli was that the filtered front vowels constituted familiar (albeit perhaps not exemplary) vowels if taken at "face value" since front vowels that have been degraded in the region of F_2 are frequently misperceived as back vowels (Lehiste and Peterson, 1959), which share

with front vowels a low F_1 frequency. Although front and back vowels differ below 1000 Hz – most notably in that the back vowels have a low-frequency F_2 that is preserved by the filter – studies such as that of Lehiste and Peterson have shown that when the perceptually salient high-frequency F_2 of front vowels is missing, these vowels often lose their perceived "frontness" and are identified as back vowels. Of interest in the present study was whether rates of front-back confusions would be affected when noise was added to the region of the spectrum removed by the filter.

Two hypotheses were tested. The first predicted that addition of highpass noise would induce restoration of the high F_2 of front vowels. Thus it would act to reduce front to back vowel errors relative to the no-noise condition. The second hypothesis was that, in the presence of the noise, *misapplied* restoration of high-frequency information for back vowels might occur, which would increase errors of the opposite type, namely back to front errors. Such errors would be interesting because, unlike the case for restoration of front vowels, misapplied restoration would imply perception of a vowel inconsistent with the acoustic information preserved by the lowpass filter.

METHOD

Subjects

Fourteen native English speakers (11 male, three female; age 21–35) participated in the experiment. They were university students who reported normal hearing and who had some familiarity with IPA symbols and phonetic transcription. Use of these "trained" listeners was justified because pilot tests showed that naive listeners were unable to reliably map from perceived sound to symbol, even after considerable training. Subjects were paid for their participation.

Stimuli

Four adult males, native speakers of American English, pronounced the 11 English vowels ([i, ɪ, e, ɛ, æ, ʌ, ɚ, a, u, ʊ, o]) in the sentence "I will now say the word b_b again" Utterances were lowpass filtered at 20 kHz and digitized at a sampling rate of 48 kHz. Individual files of the quasi-steady-state portions of the vowels were created, centered on the perceived center of the vowel and 85 msec in duration (a duration found appropriate in a pilot experiment). Files excluded bursts or formant transitions of the original flanking consonants and were adjusted to be similar in loudness. File onsets and offsets were ramped in amplitude over a 10 msec region.

Two Finite Impulse Response filters with slopes better than 200 dB/octave were designed using the DSP Designer software package. The lowpass filter began attenuation at 700 Hz and reached maximum attenuation (140 dB) at 1000 Hz. The highpass filter (actually a bandpass filter with an upper cut-off at 10 kHz) was at maximum attenuation at 1000 Hz, passed the signal completely at 1300 Hz, and began attenuation again at 10 kHz. Each filter was applied to the complete set of vowel tokens, producing 44 lowpass vowels (LPV), as well as 44 highpass vowels (HPV), which were used as distractor tokens in the experiment.

A lowpass and a highpass noise stimulus were obtained by applying the same filters in turn to synthesized white noise. Two additional sets of vowel tokens were created by superimposing noise over the filter-reject region of the filtered vowels; i.e., the highpass noise was superimposed over lowpass filtered vowels (LPV/HPN), and the lowpass noise over highpass filtered vowels (HPV/LPN). The noise was 900 msec in duration and set to a level capable of masking or nearly masking the energy above 1000 Hz in the unfiltered vowels. Vowel tokens occurred 400 msec into the noise.

Design

A complete within-subjects design was utilized, with type of filtering, presence or absence of added noise, and block consistency (see below) as factors. The dependent variables were proportions of front to back (F > B) and back to front (B > F) vowel confusions observed in the different conditions. The tokens of interest were the lowpass filtered vowels (LPV), with or without added highpass filtered noise (HPN). Unfiltered vowels (V) provided baseline rates of confusions; highpass vowels (HPV) and highpass vowels with added lowpass masking noise (HPV/LPN) served as distractor tokens.

A "block consistency" factor was included to determine whether repeated presentations of tokens having the same type of filtering could provide a cue to the presence and nature of the degradation, and thereby constitute a second factor influencing perceptual compensation. All subjects received all stimuli, in six presentation blocks. Three blocks were "consistent" containing only V, LPV, or LPV/HPN tokens. The remaining three were "mixed", created by randomly dividing a set containing a mixture of 44 LPV, 44 LPV/HPN, 22 HPV, and 22 HPV/LPN tokens into three blocks of 44 tokens each. The V block was always presented first; the order of the remaining five blocks was counterbalanced over subjects. Order of tokens within blocks was randomized separately for each subject.

Procedure

The subjects were seated in a quiet room, in front of a computer screen on which a vowel chart containing the 11 vowels used was displayed. They were not alerted to the dimension of interest in the study, but were informed that they would be hearing the 11 vowels in random order, not necessarily in equal numbers, and under different degrees of distortion. They were encouraged to respond to all tokens, even if they were unsure of their answers. Training sessions, in which unfiltered, 200 msec vowels were presented, were included to familiarize subjects with the task and to assure they reached criterion in sound-symbol association. The training sessions contained a preponderance of low or high vowels in order to illustrate the possibility of unequal presentation rates without drawing attention to the front-back dimension of interest. The stimuli were presented diotically over Sony MDR–V6 headphones. After each token, subjects used the computer mouse to click on the symbol corresponding the perceived vowel. A "No Response" was recorded if no symbol was selected within four seconds after the end of the presented token. Average total running time, including optional breaks between blocks, was approximately 45 minutes.

TABLE 1

Correct responses and types of errors

Vowel/ Noise	Block	Correct (F,B,C)	ERRORS F>F	F>B	F>C	B>F	B>B	B>C	C>F	C>B	C>C	NR	Total
V	C	452	69	5	2	1	31	26	4	4	16	6	616
LPV	C	188	11	136	108	2	36	51	3	39	36	6	616
LPV	M	165	14	158	90	2	45	43	3	50	41	5	616
LPV/HPN	C	232	52	74	79	29	29	23	17	42	30	9	616
LPV/HPN	M	212	40	81	83	23	31	36	4	51	32	23	616
HPV	M	83	71	7	9	18	15	33	16	10	27	19	308
HPV/LPN	M	79	83	8	6	31	19	8	15	18	20	21	308

Legend:
V = unfiltered vowel, LPV = lowpass filtered vowel, LPV/HPN = lowpass filtered vowel with added highpass filtered noise; HPV = highpass filtered vowel, HPV/ LPN = highpass filtered vowel with added lowpass noise. C/M indicate consistent/mixed presentation blocks. F = front vowel ([i, ɪ, e, ɛ, æ]), B = back vowel ([u, ʊ, o]), C = central vowel ([ʌ, ɚ, a]). ">" = direction of confusion. NR = No Response.

RESULTS

The pattern of responses to each stimulus type in the experiment is shown in Table 1. For conditions of interest, the total number of tokens was 616 (14 subjects X 11 vowels X 4 speakers); for distractor conditions, the total was 308 (14 subjects X 11 vowels X 2 speakers). Error types are shown in the form "presented vowel" > "identified vowel". F > F and B > B categories include within-place errors, but not correct identifications. The research hypotheses were concerned with *relative* rates of errors, since other types of confusions occur for vowels shortened in duration, and since conditions could differ in overall difficulty, affecting absolute errors of all types. Therefore, the relevant measure for errors of interest (F > B and B > F errors) was the *ratio* of these errors to the total number of errors made on vowels of the respective types. Proportions of F > B and B > F errors were thus defined as, respectively, "F > B/(F > F + F > B + F > C)" and "B > F/(B > F + B > B + B > C)".

A two-factor within-subjects ANOVA, with "presence or absence of highpass noise" and "block consistency" (C or M) as factors, was conducted to compare proportions

of F > B errors in the four LPV conditions. Results showed a nonsignificant interaction term, a strong main effect due to addition of noise, $F(1, 13) = 16.71$, $p < 0.01$, and a nearly significant main effect of block consistency, $F(1, 13) = 3.90$, $p = 0.07$. Because the effect of block consistency was stable over subjects, data from the LPV and LPV/HPN conditions were not pooled over the block factor. Analysis of simple main effects due to noise showed a highly significant reduction in the proportion of F > B errors due to the addition of noise for both the consistent and mixed block conditions, $F(1, 13) = 20.41$, $p < 0.01$; $F(1, 13) = 10.98$, $p < 0.01$, respectively. Analysis of simple main effects for the block factor yielded a significantly lower rate of F > B errors attributable to block consistency for the no-noise condition only, $F(1, 13) = 6.63$, $p < 0.05$.

Proportions of F > B errors were significantly higher for the lowpass filtered vowels (both without and with added noise) than for unfiltered vowels, with $F(1, 13) = 109.35$, $p < 0.001$, for the consistent-block LPV *vs.* V comparison, and $F(1, 13) = 24.17$, $p < 0.001$, for the consistent-block LPV/HPN *vs.* V comparison. Proportions of F > B errors in the HPV and HPV/LPN conditions (distractor tokens) were just as low as in the V condition.

Comparison of proportions of B > F errors in another two-factor ANOVA yielded a significant interaction between highpass noise and block consistency, $F(1, 13) = 7.38$, $p < 0.05$. Addition of noise caused an increase in B > F errors for both the consistent and mixed-block comparisons, $F(1, 13) = 31.66$, $p < 0.001$, and $F(1, 13) = 20.11$, $p < 0.01$, respectively. Block consistency was associated with an increase in B > F errors in the noise condition only, $F(1, 13) = 6.53$, $p < 0.05$.

Addition of lowpass noise to highpass filtered vowels caused an increase in B > F errors for distractor tokens as well, $F(1, 13) = 9.21$, $p < 0.05$; however, unlike the case for lowpass filtered vowels, the no-noise condition for distractor tokens showed more B > F errors than the baseline (V) condition, $F(1, 13) = 10.27$, $p < 0.01$.

Table 2 shows confusion matrices for the consistent-block conditions of interest. This table allows for direct comparison of the gain in F > F identifications in the noise condition to the cost of this gain, i.e., B > F errors. Unlike Table 1, F > F and B > B categories here include *both* correct responses and within-place-category errors. The cell values are mean percentages of the total number of vowels of the particular type presented (F or B). As shown, the magnitude of the increase in F > F identifications associated with adding noise (0.31) was larger than the associated increase in B > F errors (0.17) – a reliable difference, $F(1, 13) = 7.12$, $p < 0.05$.

DISCUSSION

The results from this study can be summarized in five major points. First, the results confirmed that eliminating the high frequency F_2 of the front vowels by lowpass filtering caused a dramatic increase in F > B errors. Second, the hypothesis that addition of noise to the region removed by the lowpass filter might lead to restoration of front vowels was borne out: There was a large decrease in F > B errors when noise was added. Third, results were consistent with the proposal that restoration of high-frequency spectral

TABLE 2

Confusion matrices for lowpass filtered front and back vowels

		V		LPV		LPV/HPN	
		identified as		identified as		identified as	
		F	B	F	B	F	B
actual vowel	F	0.97	0.02	0.12	0.49	0.43	0.27
	B	0.01	0.82	0.01	0.68	0.18	0.69

information would also increase the likelihood of B > F errors. Fourth, it was found that despite this increase in B > F errors, restoration resulted in an overall gain in performance accuracy. And fifth, consistency of tokens within a presentation block had a small but reliable effect on both F > B and B > F error rates.

Lowpass filtering caused a disproportionate increase in F > B errors relative to the baseline (no filtering) condition. Although this result was expected, based on reports such as Lehiste and Peterson (1959), it was necessary to confirm the effect for the specific stimuli used. The increase in F > B errors was not observed for highpass filtered tokens, consistent with the notion that it was not simply degradation, but rather selective degradation in the region of F_2, that led to the increase in these errors.

There was a marked reduction in F > B errors, however, when highpass filtered noise was superimposed over the filter-reject region of the lowpass filtered vowels. This result is similar to other examples (involving both speech and nonspeech stimuli) of perceptual compensation for masking; it suggests that the high F_2 of the front vowels could be "restored" by listeners in the presence of noise, but not when the region was left uncovered. In contrast to other studies of restoration for speech stimuli, however, there was in this experiment no linguistic context present to "drive" restoration. Because the stimuli were isolated vowels, there was no basis upon which to prefer one vowel response over another. In addition, unrestored versions of front vowels constituted familiar items (namely, back vowels) that could be given as a response (as indeed they were given, in large numbers, in the no-noise conditions). Thus, despite the lack of context, restoration occurred for lowpass filtered front vowels when noise was added. Again, highpass filtered vowels were not affected; addition of lowpass noise to highpass filtered vowels caused no change in rates of F > B errors.

As predicted, the reduction in F > B errors associated with the added noise was accompanied by a significant increase in B > F errors. This latter type of error is noteworthy for two reasons. First, there is no acoustic basis for perceiving a lowpass filtered back vowel as a front vowel. Whereas lowpass filtered front vowels sound like back

vowels when their high F_2 is removed, lowpass filtered back vowels, which never contained a high F_2 originally, still sound like back vowels. Second, in contrast to the case for restoration (reduction in F > B errors), B > F errors involve a percept inconsistent with the signal in the region preserved by the filter. That is, they involve restoration of a high F_2 for signals already containing a F_2 below 1000 Hz. This suggests that the B > F errors observed in the noise conditions were due to a derived rather than direct perceptual effect, and more specifically that B > F errors constituted instances of misapplied restoration, essentially "mistakes" in which the same process underlying restoration of the high F_2 of front vowels was inappropriately applied to back vowels. One might note that noise caused a significant increase in B > F errors for the highpass filtered tokens (in the noise as compared to the no-noise condition) as well. Misapplied restoration (of F_2) is not an adequate explanation in these cases, however, because: (1) For these highpass tokens, the no-noise conditions showed an increase in B > F errors relative to the baseline condition, thus B > F errors could not be attributed to addition of noise (alone); and (2) since there was no decrease in F > B errors for these tokens in the noise condition, i.e., no accompanying "restoration" effect, it would not make sense to speak of the increase in B > F errors as constituting "misapplied" restoration.

A question that was asked *post-hoc* was whether the increase in B > F errors was equal in magnitude to the increase in within-front identifications (F > F) associated with restoration. It was conceivable that the noise could have simply caused more front vowel responses overall, with listeners indiscriminately restoring front vowels and misrestoring back vowels at proportionally equal rates. Despite the acoustic differences between lowpass-filtered front and back vowels, it was not clear that listeners would be able to determine which vowels to restore in the noise condition, since in the no-noise condition large numbers of front vowels were identified as back vowels. In the present experiment, indiscriminate restoration would imply that for every front vowel correctly restored, there would be a proportionally equal likelihood (after adjusting for the unequal number of front and back vowels presented) of incorrectly restoring a back vowel. As shown in Table 2, however, listeners were significantly better than chance at determining which tokens to restore. The mean increase in the proportion of F > F identifications observed in the noise condition was 0.31, as compared to 0.17 for the increase in B > F errors, a reliable difference. Thus listeners were indeed, to some extent, able to discriminate front from back vowels under filtering, yet they did so much better in the noise condition than in the no-noise condition. Such findings suggest that in the presence of the noise, listeners somehow made better use of the cues to vowel identity present below 1000 Hz.

Although the pattern of responses was similar in consistent and mixed presentation blocks for both the noise and no-noise conditions involving lowpass filtered vowels, there was a significant effect of block consistency on rates of F > B and B > F errors. Block consistency had an effect much like that of adding noise, decreasing F > B and increasing B > F errors. The effect of block consistency was observed for F > B error rates only when noise was not present, perhaps due to the much larger magnitude of the effect of noise. On the other hand, block consistency affected B > F error rates only

when noise was present, suggesting that block consistency alone was unlikely to cause B > F errors, but that it enhanced the effect of noise, when present, on these errors.

CONCLUSION

This study has shown that, in the absence of linguistic context, listeners reduce front to back vowel errors when noise is superimposed over the filter-reject region of lowpass filtered vowels. The restoration effect is strong enough to result in an increase in back to front vowel confusions, implying illusory perception of spectral information never present in the original signal. Despite the increase in these errors, however, restoration leads to improved performance overall, a result which indicates that listeners are able to discriminate front from back vowels under lowpass filtering. That listeners are better at discriminating front from back vowels when the added noise is present suggests that the noise affects the degree to which cues preserved by lowpass filtering are used by listeners in identification.

An important question to be addressed in further work is exactly *how* restoration occurred in the present experiment. One possibility is that listeners were actually abstracting from the noise the components necessary for a high F_2. Another, not necessarily contradictory hypothesis is that the noise caused listeners to avoid taking the lack of an overt F_2 in that region as evidence against a front vowel stimulus. Further study could address this question by varying the masking level of the noise, or by using notched noise from which components necessary for construction of F_2 have been removed.

REFERENCES

BASHFORD, J.A., and WARREN, R.M. (1979). Perceptual synthesis of deleted phonemes. In J.J. Wolf and D.H. Klatt (eds.), *Speech Communication Papers* (pp. 423–426). New York: Acoustical Society of America.

BASHFORD, J.A., and WARREN, R.M. (1987). Effects of spectral alternation on the intelligibility of words and sentences. *Perception & Psychophysics*, **42**, 431–438.

BREGMAN, A.S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

BREGMAN, A.S., and DANNENBRING, G.L. (1977). Auditory continuity and amplitude edges. *Canadian Journal of Psychology*, **31**, 151–159.

EGAN, J.P. (1948). The effect of noise in one ear upon the loudness of speech in the other. *Journal of the Acoustical Society of America*, **20**, 58–62.

HOUTGAST, T. (1972). Psychophysical evidence for lateral inhibition in hearing. *Journal of the Acoustical Society of America*, **51**, 1885–1894.

LEHISTE, I., and PETERSON, G.E. (1959). The identification of filtered vowels. *Phonetica*, **4**, 161–177.

MILLER, G.A., and LICKLIDER, J.C.R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, **22**, 167–173.

OHALA, J.J. (1986). Against the direct realist view of speech perception. *Journal of Phonetics*, **14**, 75–82.

OHALA, J.J., and SHRIBERG, E.E. (1990). Hypercorrection in speech perception. In *Proceedings of the International Conference on Spoken Language Processing*, **1**, 405–408. Kobe, Japan: Acoustical Society of Japan.

SHRIBERG, E.E. (1990). Phantom formants: Hypercorrection in vowel identification as evidence against the direct realist view of speech perception. M.A. thesis, University of California at Berkeley.

THURLOW, W.R., and ELFNER, L.F (1959). Continuity effects with alternately sounding tones. *Journal of the Acoustical Society of America*, **31**, 1337–1339.

WARREN, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, **167**, 392–393.

WARREN, R.M. (1982). *Auditory Perception: A New Synthesis.* New York: Pergamon.

WARREN, R.M., and BASHFORD, J.A. (1976). Auditory contralateral induction: An early stage in binaural processing. *Perception & Psychophysics*, **20**, 380–386.

WARREN, R.M., and SHERMAN, G.L. (1974). Phonemic restoration based on subsequent context. *Perception & Psychophysics*, **16**, 150–156.

WERTHEIMER, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, **61**, 161–265.