Phonetica
Editor: K. Kohler, Kiel

Separatum
Publisher: S. Karger AG, Basel
Printed in Switzerland

Phonetica 1994; 51: 111–118

Listeners' Normalization of Vowel Quality Is Influenced by 'Restored' Consonantal Context

John J. Ohala
Deborah Feder

Department of Linguistics,
University of California,
Berkeley, Calif., 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 inducing subjects to 'restore' a noise-obliterated medial consonant in VCə 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 were synthetic steady-state 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, thus indicating that the influence of context need not operate only via physical phonetic features. These results suggest that strict phonetic invariance of phonological units may not be necessary.

Introduction

One of the principal problems in current speech research is the variation in the phonetic realizations of what are supposed to be invariant underlying phonological units [Perkell and Klatt, 1986]. Various claims have been made and approaches taken to deal with this including: invariance can be found in the signal if one makes the right measurements [Stevens, 1989; Stevens and Blumstein, 1978, 1981]; invariance can be found if one looks at the proper point in the speech chain, e.g. not in the acoustic signal *per se* but rather at the articulatory gestures which create the acoustic signal [Fowler, 1986; Fujimura, 1986; Liberman and

Revised version of a paper presented at the 11th International Congress of Phonetic Sciences, Tallinn, August 1987.
The present version corrects some errors of tabulation present in the earlier version.

Received:
November 8, 1993
Accepted:
November 25, 1993

Prof. John J. Ohala
Department of Linguistics
University of California
Berkeley, CA 94720 (USA)

Mattingly, 1985; Lisker et al., 1962]; rather than absolute invariance (e.g. in a simple time vs. spectrum template), one should seek relational invariance or invariance in higher-order parameters, e.g. the relation of formant frequencies to each other or to other measured parameters, weighted averages of formant frequencies, auditorily transformed spectra or dynamic parameters [Bladon and Lindblom, 1981; Fant and Risberg, 1963; Hermansky, 1990; Kluender et al., 1988; Miller, 1989; Sussman et al., 1991, 1993; Strange, 1989]; the failure to find invariance is partly due to the wrong choice of unit: the phoneme may not exhibit invariance, but the phone or diphone may show less variability (though at the cost of requiring a much larger inventory of units) [Klatt, 1980; Ohala, 1992]; the units of speech *are* variable at the level of speech production and the acoustic output, but listeners can accommodate variable signals.

There are, in turn, a variety of proposals regarding how listeners might deal with variation: listeners can learn to lump together physically dissimilar stimuli which are functionally equivalent, as readers do with the upper- and lower-case Roman letters [Mann and Repp, 1980, 1981; Nearey, 1992]; listeners can exploit auditory detection methods which permit them to extrapolate missed target frequencies of speech sounds given truncated formant trajectories [Fujisaki and Sekimoto, 1975]; Lindblom [1990] hypothesizes that much of the variability found in speech reflects the speaker's estimate of the listener's perceptual needs; the degree of precision in articulation varies accordingly. He also replaces the search for invariance with a search for the criteria that make speech sounds *sufficiently discriminable.*

There is considerable overlap in many of these approaches.

In this paper we address the claim that listeners can perceptually compensate for variation in the speech signal. There is, in fact, abundant evidence, both from perceptual experiments as well as phonology (sound change), that listeners identify speech sounds in part by normalizing them with respect to their phonetic context [Ladefoged and Broadbent, 1957; Pickett and Decker, 1963; Lindblom and Studdert-Kennedy, 1967; Mann and Repp, 1980, 1981; Ohala, 1986]. But how is this done? Are the physical phonetic parameters of the context, i.e. the values of spectral peaks, used to adjust recognition thresholds, or is it enough for the listener just to know the (categorized) linguistic identity of the context and use something like a table lookup to figure out how that context would influence an unknown sound? We investigated these questions through a series of perceptual tests involving listeners' identification of synthetic vowel stimuli in isolation and in consonantal contexts.

One of the well-documented forms of contextual variation is the perturbation of vowel formant frequencies by adjacent consonants [Lindblom, 1963; Stevens and House, 1963]. There is evidence that listeners normalize these vowels – or shift their identification thresholds – in such a way as to at least partially compensate for the presumed consonantal influence. Ohala et al. [1978; summarized in Ohala, 1981] found that given steady-state (transitionless) vowels from a synthetic /i/ and /u/ continuum, the crossover between these two vowels as judged by American English listeners was more front when embedded in the context of alveolar consonants (/s_t/) than when flanked by labial consonants (/f_p/). They attributed this effect to listeners' knowledge that an intended /u/ could be fronted (have its F_2 raised) in the environment of alveolar consonants, which also have a high F_2. That is to say, they concluded that this normalization was based on listeners' linguistic experience, not on some sort of computation or ex-

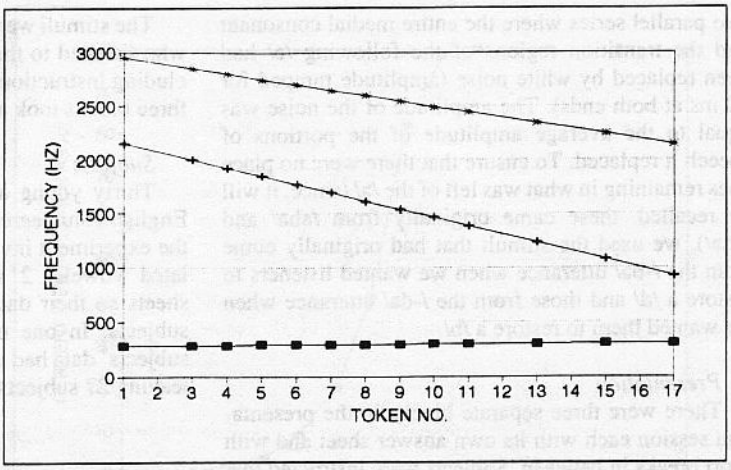

Fig. 1. The /i/-/u/ stimulus continuum. ■ = F_1; + = F_2; $* = F_3$.

trapolation based on the physical parameters in the signal itself. However, this result could still involve a normalization based on some unidentified spectral properties of the flanking alveolar or labial consonants.

In the present experiment we attempted to discover if listeners' normalization of vowel quality as a function of the character of the adjacent consonant could be demonstrated even when the adjacent consonant was not physically present but listeners *thought* that it was present. A new implementation by us of Warren's [1970] phoneme restoration effect, described below, gave us an opportunity to try this.

Method

Stimuli

We prepared five stimulus continua: an isolated vowel /i/-/u/ continuum (henceforth symbolized #V#), an /idə/-/udə/ continuum (Vdə), an /ibə/-/ubə/ continuum (Vbə), and the latter two continua with the medial consonants and some of the /ə/ masked with noise [V(d)ə, V(b)ə].

First, using the Klatt synthesizer [Klatt, 1980] we constructed a 17-step linear stimulus continuum consisting of 100-ms steady-state vowels between and including /i/ and /u/. The continuum endpoints were modeled on the first 100 ms of natural /i/ and /u/ pronounced in isolation by an adult male native speaker of American English. The formant frequencies for /i/ were $F_1=285$, $F_2=2{,}155$, $F_3=2{,}950$ Hz; for /u/, $F_1=305$, $F_2=925$, $F_3=2{,}140$ Hz (fig. 1). Overall sound intensity was incremented 0.5 dB per step from /i/ to /u/ to maintain a subjective impression of equal loudness for all the stimuli. The amplitude of the final 10 ms of the continuum vowels was amplitude ramped to eliminate an unnaturally abrupt termination. Since a pilot experiment showed that the 'crossover' from /i/ to /u/ would happen in the middle of this continuum, some stimuli near the endpoints were omitted from the study, namely, stimuli numbers 2, 12, 14 and 16.

We then constructed two series of VCə stimuli where the V was the /i/-/u/ continuum described above and C was either /b/ or /d/. The -Cə sequences were excised from the same male speaker's natural utterances of /abə/ and /adə/ and digitally spliced after the V from the continuum. The stop closure (fully voiced) was about 40 ms in duration and the /ə/, 60 ms. The fact that there were no consonant transitions in the V preceding the Cs did not markedly affect their naturalness; the burst and transitions into the /ə/ were sufficient to convey convincing percepts of /b/ and /d/. The greater importance of CV (over VC) transitions in conveying stop place cues is well documented [Repp, 1978; Fujimura et al., 1978; Ohala, 1990].

Finally, for the stimulus continua where we wanted listeners to 'restore' a noise-masked medial stop, we took the two VCə continua and constructed two parallel series where the entire medial consonant and the transition regions of the following /ə/ had been replaced by white noise (amplitude ramped for 10 ms at both ends). The amplitude of the noise was equal to the average amplitude of the portions of speech it replaced. To ensure that there were no place cues remaining in what was left of the /ə/ (since, it will be recalled, these came originally from /abə/ and /adə/), we used the stimuli that had originally come from the /-bə/ utterance when we wanted listeners to restore a /d/ and those from the /-də/ utterance when we wanted them to restore a /b/.

Presentation

There were three separate blocks in the presentation session each with its own answer sheet and with short breaks in between. Subjects were instructed that in all three blocks they would be hearing utterances of various types each containing the vowel /i/ or /u/ and that they were to identify it as one of these two (forced choice) and to write that vowel on their answer sheet on the appropriate line, writing 'e' (the vowel whose name is [i] or 'u', if they identified the vowel as /i/ or /u/, respectively. They were told that sometimes there would be a short noise burst next to or on some part of the utterance and that those were potential distracters; we told the subjects (deceitfully) that we were interested in seeing whether these distractions influenced their ability to identify the vowels.

The first block consisted of the isolated vowels, #V#, randomized, each presented 4 times with half of them containing a short extraneous noise burst before or after the stimulus vowel. The interstimulus interval was 3 s. This block then had 13 x 4 = 52 trials.

The second block consisted of the randomized Vdə stimuli, each presented 11 times, as well as the V(b)ə stimuli (those whose medial stops had been masked with noise), each presented twice for a total of (11 + 2) x 13 = 169 trials. The interstimulus interval was again 3 s. The first instance of a V(b)ə stimulus was trial No. 23. Three of the 11 Vdə series had a noise burst [comparable in amplitude and duration to that used in the V(C)ə stimuli] placed before, during or after the VCV utterance. The answer sheet used with this block specified '_da' for each trial. Given that the majority of trials in this block (85%) had medial 'd''s and that the answer sheet specified a medial 'd', we expected the subjects to 'restore' a medial /d/ in the Vbə trials.

The third block was like the second block except that it contained Vbə stimuli and in the 26 tokens the medial consonant masked by noise was V(d)ə. The answer sheet for this block specified '_ba' for each trial.

The stimuli were presented to subjects individually who listened to the taped stimuli over earphones. Including instructions and breaks a session including all three blocks took about 30 min.

Subjects

Thirty young adult native speakers of American English volunteered as listeners for the experiment. In the experiment involving the identification of the isolated vowels, 2 subjects mismarked their answer sheets so their data had to be discarded, leaving 28 subjects. In one of the paired /b/ or /d/ blocks, 3 subjects' data had to be discarded for the same reason, leaving 27 subjects.

Analysis and Results

With the exception of the V(C)ə stimuli none of the trials that had the 'distracter' noise bursts in them were included in the final analysis.

Figure 2 presents the results (in the form of the percentage of /u/ judgments) from block 1 containing the isolated vowels, #V#. Each data point represents 28 subjects x 2 tokens = 56 judgments. As shown in figure 2, the crossover between /i/ and /u/ was, on the average, between tokens 8 and 9. There was considerable individual variation, however, with, in extreme cases, some subjects showing a crossover between tokens 3 and 4 and others between tokens 9 and 10. The fact that the judgments saturate to 0 and 100% at the endpoints demonstrates that the subjects had no trouble associating the stimuli with the target vowels /i/ and /u/ (at least in this forced choice format).

Figure 3 presents the responses to the Vdə and Vbə stimuli from the second and third blocks. Each data point represents 8 x 27 = 216 judgments. There is a clear shift in the /u/ identification function due to the consonantal context. Although the approximate crossover between /i/ and /u/ is still between tokens 8 and 9 for the Vbə stimuli (as in the #V# stim-

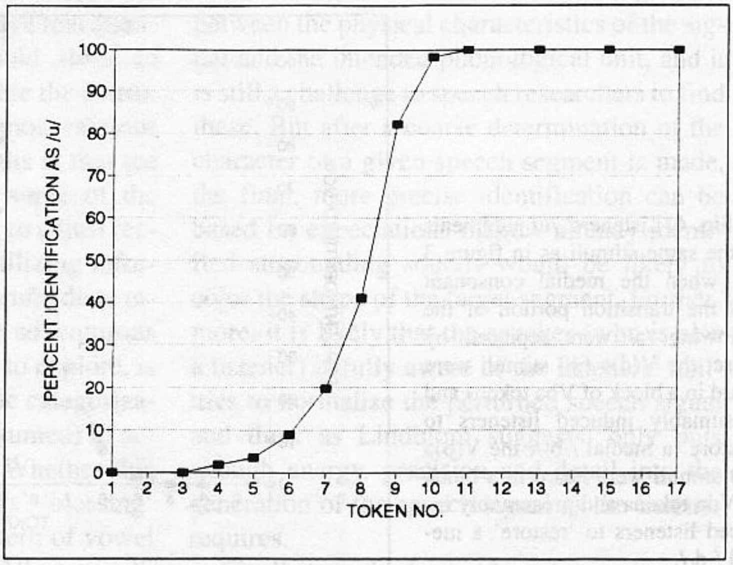

Fig. 2. Listeners' /u/ judgments to the vowel stimuli in figure 1 presented in isolation.

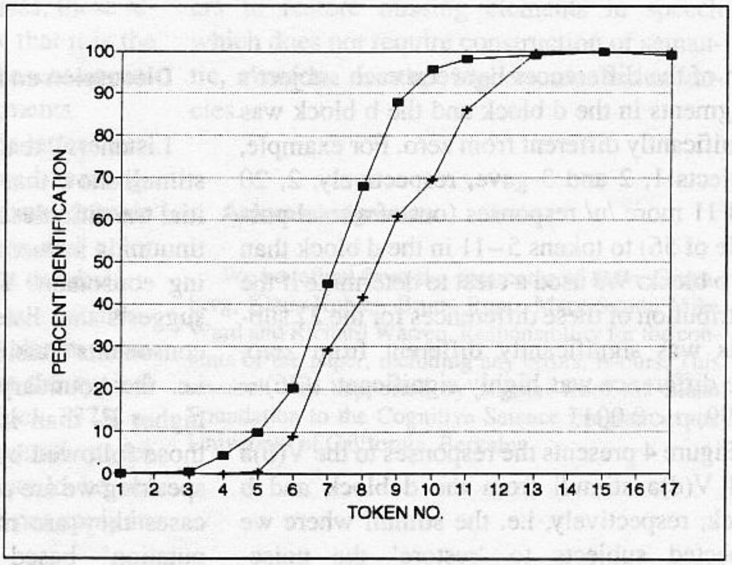

Fig. 3. Listeners' /u/ judgments to the vowel stimuli in figure 1 when followed by spliced-on -də(■) and -bə (+) natural speech syllables.

uli), in the case of the Vdə stimuli it is between tokens 7 and 8. To evaluate the statistical significance of this shift, we first eliminated responses which showed 'saturation' in all relevant conditions, those outside the 5% and 95% response level. This left us with the responses to the 7 tokens from No. 5 to 11. Then, given the great amount of between-subjects variability and since we were just interested in whether the consonantal environment (actual or restored) influenced subjects' judgments on vowel quality, we asked whether the distribu-

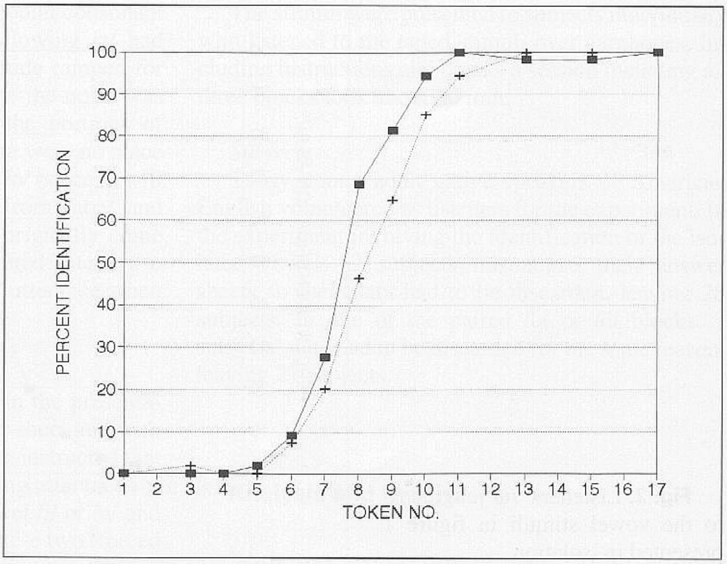

Fig. 4. Listeners' /u/ judgments to the same stimuli as in figure 3 but when the medial consonant and the transition portion of the following /ə/ were replaced by noise; the V(d)ə (+) stimuli were heard in a block of Vbə tokens and presumably induced listeners to 'restore' a medial /-b-/; the V(b)ə (■) stimuli were heard in a block of Vdə tokens and presumably induced listeners to 'restore' a medial /-d-/.

tion of the differences between each subject's judgments in the d block and the b block was significantly different from zero. For example, subjects 1, 2 and 3 gave, respectively, 2, 20 and 11 more /u/ responses (out of a total possible of 56) to tokens 5–11 in the d block than the b block. We used a t test to determine if the distribution of these differences for the 27 subjects was significantly different from zero. The difference was highly significant: $t(26) = 6.299$, $p < 0.001$.

Figure 4 presents the responses to the V(b)ə and V(d)ə stimuli from the d block and b block, respectively, i.e. the stimuli where we expected subjects to 'restore' the noise-masked medial stops. Each data point represents 2 x 27 = 54 judgments. Again, there is a shift in the /u/ identification function in the two blocks, presumably due to the influence of the restored stops. Although the magnitude of this shift is less than in the case where the medial stops were not masked, the difference, evaluated over tokens 5–11, is statistically significant: $t(26) = 3.58$, $p < 0.001$.

Discussion and Conclusion

Listeners' reactions to the Vdə and Vbə stimuli show that their identification of the initial transitionless vowels from the /i/-/u/ continuum is influenced by the perceived following consonant. The nature of this influence suggests that listeners compensate for apical consonants' raising of the F_2 of back vowels, i.e. the boundary between /i/ and /u/ has a higher F_2 than is true of isolated vowels or those followed by a labial consonant. Strictly speaking we are unable to say whether in these cases they are making some kind of 'computation' based on quantitative relations between the unknown vowel's acoustic properties and those of the consonant – e.g. something like a 'locus equation' [Sussman et al., 1991, 1993] – or whether it is simply a matter of identifying (categorizing) the consonant phoneme and then 'looking up' its expected influence on the vowel. But listeners' judgments on the V(b)ə and V(d)ə stimuli where the medial consonant was replaced by noise and embedded in the d blocks and b blocks, respectively, were designed to resolve that ambiguity. Listeners showed threshold shifts to these stimuli similar to those where the consonant was clearly detectable. The most cautious interpretation of these latter results is that the listeners are able to normalize some of the variation in the speech signal, i.e. to adjust recognition thresholds, in part by utilizing information that is not present in the immediate utterance itself. A somewhat more adventurous interpretation, the one we set out to explore, is that listeners refer to the linguistic categorization of the context (present or assumed) in accomplishing this normalization. Whether this is true as opposed to the listeners' 'biassing' their responses, based on the pattern of vowel identifications to the Vdə and Vbə stimuli, would require a further, more elaborate, experiment to determine. Nevertheless, these results are consistent with the view that it is the linguistic categorization of the adjacent consonant which guides listeners' judgments.

There is some precedent for this latter interpretation in other sensory domains: it is recognized, for example, that in the visual domain we achieve a high degree of size and color constancy in part by factoring out the distorting influence of distance and the hue of ambient illumination, respectively, but also in some cases by our knowledgte of what the typical size and colors of objects are [Rock 1975, p. 565]. For example, apples are round, paper and teeth are usually white. It would be remarkable if something similar did not apply in the case of speech perception.

If listeners are capable of integrating linguistic, i.e. categorical, information into their recognition task this implies that absolute – or even relative – invariance in the speech signal corresponding to the intended linguistic units is not necessary to the degree often assumed. Of course, there must be *some* congruence between the physical characteristics of the signal and the intended phonological unit, and it is still a challenge to speech researchers to find these. But after a coarse determination of the character of a given speech segment is made, the final, more precise identification can be based on expectations of how already identified surrounding sounds would be likely to color the shape of the target segment. Furthermore, it is likely that the speaker (who is also a listener) is fully aware of the listeners' abilities to normalize the perturbed speech signal and thus, as Lindblom suggests, only puts enough energy, precision and detail into the generation of the speech signal as the listener requires.

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.

Acknowledgments

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

References

Bladon, R. A. W.; Lindblom, B.: Modeling the judgment of vowel quality differences. J. acoust. Soc. Am. *69:* 1414–1422 (1981).

Fant, G.; Risberg, A.: Auditory matching of vowels with two formant synthetic sounds. Speech Transmission Lab. (Stockholm) Q. Prog. Status Rep. *4:* 7–11 (1963).

Fowler, C. A.: An event approach to the study of speech perception from a direct realist perspective. J. Phonet. *14:* 3–28 (1986).

Fujimura, O.: Relative invariance of articulatory movements: an iceberg model; in Perkell, Klatt, Invariance and variability in speech processes, pp. 226–234 (Erlbaum, Hillsdale 1986).

Fujimura, O.; Macchi, M. J.; Streeter, L. A.: Perception of stop consonants with conflicting transitional cues: a cross-linguistic study. Lang. Speech *21:* 337–346 (1978).

Fujisaki, H.; Sekimoto, S.: Perception of time-varying resonance frequencies in speech and non-speech stimuli; in Cohen, Nooteboom, Structure and process in speech perception, pp. 269–280 (Springer, New York 1975).

Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. acoust. Soc. Am. *87:* 1738–1752 (1990).

Klatt, D. H.: Software for a cascade/parallel formant synthesizer. J. acoust. Soc. Am. *67:* 971–995 (1980).

Klatt, D. H.: SCRIBER and LAFS: two new approaches to speech analysis; in Lea, Trends in speech recognition, pp. 529–555 (Prentice-Hall, Englewood Cliffs 1980).

Kluender, K. R.; Diehl, R. L.; Wright, B. A.: Vowel-length difference before voiced and voiceless consonants: an auditory explanation. J. Phonet. *16:* 153–169 (1988).

Ladefoged, P.; Broadbent, D. E.: Information conveyed by vowels. J. acoust. Soc. Am. *29:* 98–104 (1957).

Liberman, A. M.; Mattingly, I. G.: The motor theory of speech perception revised. Cognition *21:* 1–36 (1985).

Lindblom, B.: Spectrographic study of vowel reduction. J. acoust. Soc. Am. *35:* 1773–1781 (1963).

Lindblom, B.: Explaining phonetic variation: a sketch of the H and H theory; in Hardcastle, Marchal, Speech production and speech modelling, pp. 403–439 (Kluwer, Dordrecht 1990).

Lindblom, B.; Studdert-Kennedy, M.: On the role of formant transition in vowel recognition. J. acoust. Soc. Am. *42:* 830–843 (1967).

Lisker, L.; Cooper, F. S.; Liberman, A. M.: The uses of experiment in language description. Word *18:* 82–106 (1962).

Mann, V. A.; Repp, B. H.: Influence of vocalic context on perception of the [ʃ]-[s] distinction. Percept. Psychophys. *28:* 213–228 (1980).

Mann, V. A.; Repp, B. H.: Influence of preceding fricative on stop consonant perception. J. acoust. Soc. Am. *69:* 548–558 (1981).

Miller, J. D.: Auditory-perceptual interpretation of the vowel. J. acoust. Soc. Am. *85:* 2114–2133 (1989).

Nearey, T. M.: Context effects in a double-weak theory of speech perception. Lang. Speech *35:* 153–171 (1992).

Ohala, J. J.: The listener as a source of sound change; in Masek, Hendrick, Miller, Papers from the parasession on language and behavior, pp. 178–203 (Chicago Linguistic Society, Chicago 1981).

Ohala, J. J.: Phonological evidence for top-down processing in speech perception; in Perkell, Klatt, Invariance and variability in speech processes, pp. 386–397 (Erlbaum, Hillsdale 1986).

Ohala, J. J.: The phonetics and phonology of aspects of assimilation; in Kingston, Beckman, Papers in laboratory phonology I: between the grammar and the physics of speech, pp. 258–275 (Cambridge University Press, Cambridge 1990).

Ohala, J. J.: The segment: primitive or derived? in Docherty, Ladd, Papers in laboratory phonology II: gesture, segment, prosody, pp. 166–183 (Cambridge University Press, Cambridge 1992).

Ohala, J. J.; Riordan, C. J.; Kawasaki, H.: The influence of consonant environment upon identification of transitionless vowels. J. acoust. Soc. Am. *64:* S18 (1978).

Perkell, J.; Klatt, D. H.: Invariance and variability in speech processes (Erlbaum, Hillsdale 1986).

Pickett, J. M.; Decker, L.: Time factors in perception of a double consonant. Lang. Speech *3:* 11–17 (1963).

Repp, B. H.: Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Percept. Psychophys. *24:* 471–485 (1978).

Rock, I.: An introduction to perception (Macmillan, New York 1975).

Stevens, K. N.: On the quantal nature of speech. J. Phonet. *17:* 3–45 (1989).

Stevens, K. N., Blumstein, S. E.: Invariant cues for place of articulation in stop consonants. J. acoust. Soc. Am. *64:* 1358–1368 (1978).

Stevens, K. N.; Blumstein, S. E.: The search for invariant acoustic correlates of phonetic features; in Eimas, Miller, Perspectives on the study of speech, pp. 1–38 (Erlbaum, Hillsdale 1981).

Stevens, K. N.; House, A. S.: Perturbations of vowel articulations by consonantal context: an acoustical study. J. Speech Hear. Res. *6:* 111–128 (1963).

Strange, W.: Evolving theories of vowel perception. J. acoust. Soc. Am. *85:* 2081–2087 (1989).

Sussman, H. M.; McCaffrey, H. A.; Matthews, S. A.: An investigation of locus equations as a source of relational invariance for stop place categorization. J. acoust. Soc. Am. *90:* 1309–3125 (1991).

Sussman, H. M.; Hœmeke, K. A.; Ahmed, F. S.: A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. J. acoust. Soc. Am. *94:* 1256–1268 (1993).

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