Adaptive dispersion in vowel perception

Keith Johnson
Ohio State University

Abstract The "hyperspace effect" in vowel perception (Johnson, Flemming & Wright, 1993) may be taken as evidence that adaptive dispersion is an active perceptual process. However, JFW tested for adaptive dispersion in isolated vowel stimuli spoken in a voice unfamiliar to the listeners. The experiment reported in this paper addressed both of these potential concerns and found that both consonant context and talker familiarity modulate the hyperspace effect. However, the reductions induced by context and familiarity were slight. Listeners' preferred perceptual spaces remained hyperarticulated relative to the production vowel space.

Adaptive dispersion refers to the hypothesis that the distinctive sounds of a language tend to be positioned in phonetic space so as to maximize perceptual contrast (Liljencrantz & Lindblom, 1972; Lindblom and Engstrand, 1989; Lindblom, 1990). For example, vowel systems across languages tend to utilize the available acoustic vowel space so that maximal auditory contrast is maintained. Three-vowel systems tend to be composed of /i/, /a/ and /u/; five-vowel systems tend to be composed of /i/, /e/, /a/, /o/, and /u/; and so on. Liljencrans & Lindblom (1972) modeled this tendency by treating each vowel category as a repeller in a dynamical system. They argued that the positions of vowels in the acoustic vowel space is influenced by this dynamical repelling force which has come to be called "adaptive dispersion".

The concept of adaptive dispersion has wide applicability in accounting for language sound systems. For example, Padgett (to appear) discusses a case of adaptive dispersion (in his terms "contrast dispersion") in consonants. He observes that the traditional plain and palatalized consonants in Russian are realized phonetically as a contrast between velarized and palatalized consonants (see also Halle, 1959).

Despite the appealing logic of adaptive dispersion as an explanatory principle, there is very little psycholinguistic evidence that adaptive dispersion is an active perceptual biasing pressure on linguistic systems. Such psycholinguistic evidence can be adduced, however, from the results of Johnson, Flemming & Wright (1993 - henceforth JFW, see also Bradlow, 1995; 1996). JFW asked listeners to choose the "best" exemplar of different English vowel categories from an array of synthetic vowels that spanned a range of F1-F2 combinations. The stimuli were natural-sounding 5-formant steady-state vowels produced with a software formant synthesizer. JFW found that listeners chose vowel qualities that were more disperse in the acoustic vowel space than the vowel qualities produced in normal speech (Figure 1a). This "perceptual hyperspace" effect was undiminished under instructions to choose the vowel "as you would say it" or with stimuli that had more natural intrinsic F0 and duration (Figure 1b).

From these results showing that listeners prefer an expanded acoustic vowel space, JFW suggested that the hyperspace effect reflects listeners' production targets that are subject to undershoot in production (Lindblom, 1963). Alternatively, the hyperspace effect could be taken as evidence of adaptive dispersion in vowel perception. It is this latter interpretation that forms the basis for the experiment reported in this paper.

The problem

Though the results of JFW suggest that adaptive dispersion is an active perceptual process, two aspects of the experiments limit their relevance for explaining cross-linguistic sound patterns. First, the stimuli used by JFW were isolated synthetic vowels. This is a methodological weakness because the words used as visual prompts ("heed", "had", "who'd" etc.) illustrated the vowels in consonantal context while the auditory stimuli were isolated vowels. Because consonants have

a large influence on vowel formant values (Lindblom, 1963; Stevens & House, 1963) and listeners perceptually compensate for vowel target undershoot (Lindblom & Studdert-Kennedy, 1967), the hyperspace effect may have been artificially enhanced by the use of isolated vowel stimuli.

The second limitation of JFW is that the stimuli were produced by a computer "voice" that was unfamiliar to the listeners. The impact of adaptive dispersion as a biasing effect on language sound systems may be over-estimated if the effect is a talker-contingent process that is more likely to occur when listening to the speech of an unfamiliar talker. Recent evidence regarding the talker-contingent nature of speech perception suggests that the hyperspace effect may be reduced when the listener is familiar with the talker. We turn now to a brief review of this evidence.

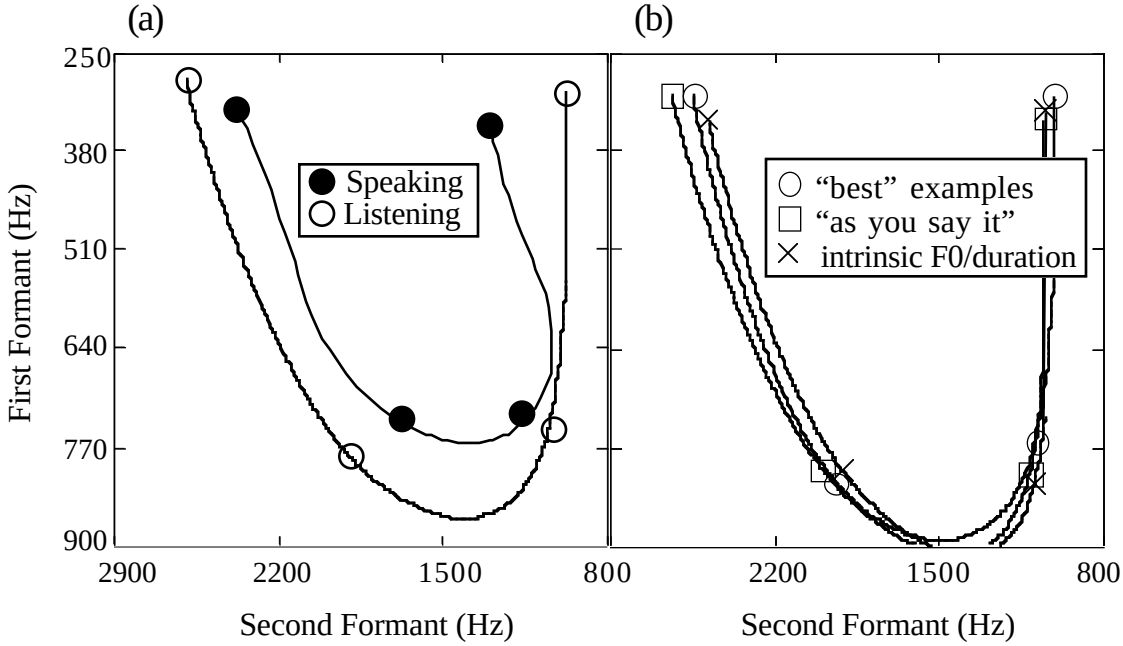

Figure 1. Perceptual hyperspace results redrawn from Johnson, Flemming, & Wright (1993). Results from the corner vowels /i/, /u/, /ˇ/ and /ɑ/ are shown. In panel (a) average formant values for male speakers are compared with the results of the pretest in JFW. Panel (b) shows results from three perception tests using different instructions ("best" versus "as you say it") and different stimuli (controlling for intrinsic F0 and duration). The points within each condition are connected by a spline function.

Talker familiarity effects

Interactions between talker identity and speech processing have been the focus of some recent studies. Mullennix, Pisoni & Martin (1989; Creelman, 1957) found that intelligibility in noise is lower in mixed-talker lists than it is in same-talker lists. This "talker effect" seems to suggest that listeners adapt to talkers over time and that when this process is not interrupted, as it is in the mixed-talker condition, listeners are better able to recognize words. Mullennix & Pisoni (1990; see also Green, Tomiak & Kuhl, 1997) found that talker information cannot be treated as a separable dimension in phonetic perception. Using the Garner paradigm they found that irrelevant talker variation interferes with speeded phonetic classification.

In addition to these studies indicating that the voice of the talker is not "normalized out" during speech processing, there are some studies indicating that the listener's familiarity with the talker has an impact on processing. For example, Walker, Bruce & O'Malley (1995) found that face-voice incongruency reduces the McGurk effect, but only if listeners are familiar with the talker. This paper is important because it shows that a supposedly low-level speech processing mechanism - the integration of visual

and acoustic speech cues - is mediated by personal information. When listeners were familiar with the visual talker shown in the McGurk display, and when the voice of the stimulus was not that of the visual talker but of someone else, the visual/auditory integration usually found with such stimuli was greatly reduced. Visual/auditory integration for these face-voice incongruent stimuli was not reduced (relative to congruent stimuli) for listeners who were not familiar with the talker. In addition to showing that talker identity interacts with speech perception (see also Johnson, Strand & D'Imperio, 1999), this result suggests that speech processing is influenced by familiarity with the talker.

Further evidence of the effect of talker familiarity in speech processing comes from a study reported by Nygaard, Sommers & Pisoni (1994). They found that speech was more intelligible when listeners were familiar with the talker. This result is directly relevant to the hyperspace effect in the following way. If speech produced by an unfamiliar talker is less intelligible than speech produced by a familiar talker, then it stands to reason that listeners would prefer an expanded vowel space for an unfamiliar voice. In this way, the unfamiliar synthetic talker used by JFW may have contributed to the hyperspace effect.

If the hyperspace effect is purely an artifact of JFW's use of isolated vowel stimuli produced by an unfamiliar voice, then their finding cannot be reasonably taken as evidence that adaptive dispersion, as an active perceptual process, has much practical impact in shaping language sound systems. These potentially invalidating factors were tested in the experiment reported below.

Methods

Production data

One male native speaker of American English (KJ) recorded five repetitions of a word list including "heed", "had", "who'd", "hod", and "hud". Each word was read in isolation a careful list-reading style with a falling intonation contour on each word. Frequency values of the first three formants at vowel midpoint were measured from spectrograms of these tokens. The average F1 and F2 frequencies of /i/, /u/, /ˇ/ and /A/ were compared with listener's perceptual choices for F1 and F2. One production of "hud" was selected for use in constructing the perceptual stimuli.

Stimuli

The stimuli for this experiment were modeled on those used by JFW. As in JFW, they were 330 synthetic stimuli that sampled the acoustic vowel space in equal Bark intervals on F1 and F2, producing a grid of F1 and F2 combinations. F4 and F5 were steady-state and had the same values for all of the stimuli. F3 was calculated by rule as in JFW. The stimuli in this experiment differed from those in JFW's studies in that they were given a final /d/ consonant using final formant trajectories calculated by rule, and a naturally produced final closure interval and /d/ release from a production of "hud".

The voice source for the stimuli was also different from the synthetic source used in JFW. In the current experiment the voice source was the LPC residual signal from one token of "hud". This source function, which included both the glottal frication of /h/ and voicing during the vowel, was filtered by a bank of time-varying band-pass filters having formant values as in JFW experiment 1, with the addition of /d/ final formant trajectories. The resulting stimuli sounded like the talker (KJ) who produced the voice source token.

Listeners

Twenty-two listeners served as volunteers in this study. They fell into three groups: 7 naive listeners, 7 students, 8 colleagues. The naive listeners were unfamiliar with the voice of the talker who produced the stimuli. The students were currently enrolled in a course being taught by the talker and so had some familiarity with his voice, and the colleagues had worked with the talker for 1 to 5 years.

Procedure

A grid of boxes was presented to listeners on a computer monitor. Each box was associated with a synthetic vowel stimulus. When the listener clicked on one of the boxes the associated stimulus was played over earphones at a comfortable listening level. Listeners could hear any of the 330 stimuli at any point in a trial by clicking on the associated visual box. They could repeat a stimulus by clicking on a box more than once. There were no constraints on the amount of time that a

listener could take in trial or on the order of stimuli presented during a trial. In each trial one of the test words, “heed”, “who’d”, “had”, or “hod”, was printed in ordinary orthography at the top of the computer screen. The listeners’ task on each trial was to select one stimulus from the grid of 330 stimuli that sounded most like the visually presented word. Each word was presented to each listener five times in random order for a total of 20 trials per listener. The groups of listeners had slightly different instructions.

Naive listeners were instructed to choose the best example of the word. This instruction corresponds to JFW’s “best exemplar” condition. Students and colleagues on the other hand were asked to select a stimulus for each word keeping in mind the question: “Is this what KJ would sound like saying this word?”

At the conclusion of the experiment, colleagues were also asked to rate their familiarity with the speech of the talker (1=highly familiar, 7=not familiar at all), and the success of the synthesis (1=sounded very much like KJ, 7=did not sound at all like KJ).

Results

Figure 2 shows a comparison of the average median formant values chosen by naive listeners in this experiment and comparable data from JFW. (Each listener's formant choices for a word were taken from the median of his/her five trials for that word.) There were several differences between the earlier study and this one: listeners in JFW were from California while the naive listeners in this study were from Ohio, the stimuli in this experiment were synthesized with an LPC residual voice source while the JFW stimuli used a purely synthetic voice source, and the stimuli in JFW were steady-state vowels with no consonant context while the stimuli for this experiment had /h/ frication at the onset and /d/ offset formant transitions and a /d/ closure interval and release burst. In view of these differences it is therefore quite remarkable that the formant values for /i/ and /u/ chosen by listeners in the two studies are nearly identical. For these vowels, the differences between the earlier study and this one seem to have had no effect.

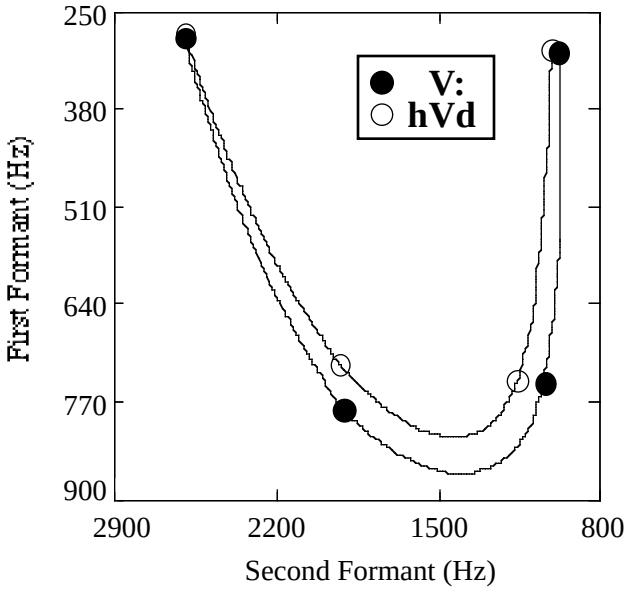

Figure 2. Perceptual results from Johnson, Flemming & Wright (1993) compared with results of the present experiment. Average formant frequencies for the corner vowels /i/, /u/, /ˇ/ and /A/ are shown. Data from JFW are labeled V: and are plotted with filled circles. Data from the naive listeners in the present experiment are labeled hVd and are plotted with open circles. The points in each condition are connected by a spline function.

The formant values chosen for /ˇ/ and /A/ were less extreme in the current study than they were in JFW. The most sensible interpretation of this difference is that lack of consonantal context in JFW led to perceptual overshoot (Lindblom & Studdert-Kennedy, 1967). Listeners seem to expect low vowels in hVd context to have less extreme formant frequencies than they do in isolation. This interpretation is motivated by the lack of an overall effect, as might be expected from the voice-source difference, and the lack of substantial dialectal differences between California and central Ohio for /ˇ/ and /A/.

Figure 3 shows average formant values for speaker KJ compared with the average perceptual results for the naive listeners, colleagues, and the students. As in JFW, naive listeners chose F1 and F2 values that encompass a larger region of the acoustic vowel space than is found for carefully articulated list-reading production. That this vowel space expansion was obtained despite the effect of consonantal context (figure 2) suggests that perceptual overshoot with the isolated vowel stimuli in JFW did not exaggerate the apparent magnitude of the hyperspace effect very much.

The perceptual data were submitted to two analyses of variance with factors stimulus word ("heed", "had", "hod", or "who'd"), and listener group (naive, colleague, student). In the analysis of the F1 data the only reliable effect was that for stimulus word ($F(3,76) = 309.5$, $p < 0.01$). In the analysis of the F2 data there was also a main effect of word ($F(3,76) = 487$, $p < 0.01$). There was also a main effect of listener group ($F(2,76) = 5.0$, $p < 0.01$), but the interaction between group and word did not reach significance ($F(6,76) = 1.36$, $p = 0.24$). Post-hoc comparisons of the groups are mentioned below.

As with the naive listeners, the results shown in figure 3a, suggest that the hyperspace effect also occurred in the responses of the students. However, for these listeners, familiarity seems to have reduced the effect for

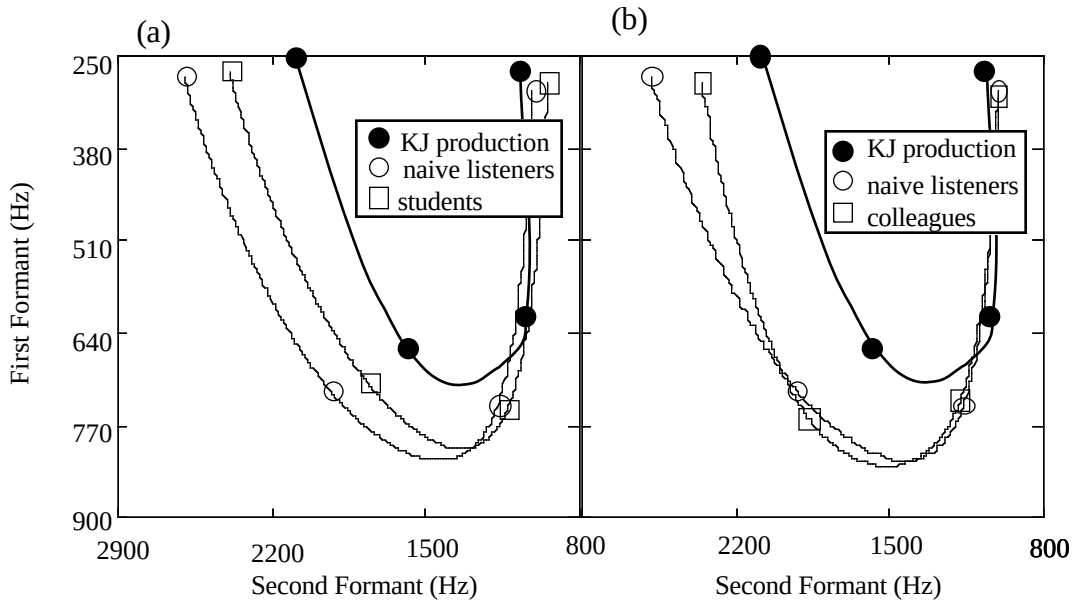

Figure 3. Perceptual results for the three groups of listeners compared with average vowel formant frequencies for talker KJ. Results from the corner vowels /i/, /u/, /ˇ/ and /A/ are shown. In panel (a) average formant frequencies for KJ are compared with the average formant frequencies chosen by naive listeners and students. Panel (b) shows average formant frequencies for KJ compared with average formant frequencies chosen by naive listeners and colleagues. The points within each condition are connected by a spline function.

the front vowels /i/ and /ˇ/. The formant values chosen by students for /i/ and /ˇ/ were more similar to those produced by KJ than were the formants chosen by naive listeners. Pairwise comparisons of F2 means, using Fisher's least significant difference test, found that naive listeners and students differed for both /i/ and /ˇ/. It is important to note, however, that the students knew the purpose of the experiment and the hypothesized role of talker familiarity in reducing the hyperspace effect. So, their reduction of the hyperspace effect may have been cuased by demand characteristics. Interestingly, even with such a bias against producing the hyperspace effect, these listeners chose relatively peripheral vowel formant frequencies.

Figure 3b compares measured vowel formants in the speech of the talker with the average formant frequencies chosen by naive listeners and colleagues. As with the students, the hyperspace effect was found in the responses of the colleagues, and there is a small reduction of the hyperspace relative to the values chosen by naive listeners, but only for /i/. Pairwise comparisons of F2 means, using Fisher's least significant difference test, found that naive listeners and colleagues differed for /i/ but not /ˇ/. This indicates that demand characteristics probably did play a role in the students' responses, but the general pattern of results for the two groups is similar.

The colleagues were also asked to rate their familiarity with the talker and to evaluate the success of the synthetic tokens in mimicking his voice. These ratings are shown in Table 1. The rating data show that this group of listeners felt that they were familiar with the voice of the talker, but that the synthetic tokens were not particularly good examples of his speech.

Table 1. Median ratings (and range) of colleagues' familiarity with the voice of the talker and the success of the synthetic stimuli in mimicking the talker's voice. 1=highest rating, 7=lowest rating.

Familiarity:	2 (1-3)
Synthesis:	3 (2-7)

Discussion

The results of this experiment confirm that the hyperspace effect is very robust. Listeners chose vowels that defined a large acoustic vowel space even though the vowels appeared in a /hVd/ context and, for two groups of listeners, even though they were familiar with the voice of the talker. As predicted though, both consonant context and talker familiarity affected listeners' choices. In /hVd/ context the hyperspace effect was reduced for low vowels. One reviewer suggests that further reduction of the hyperspace effect might be found with variable consonant contexts, which may well be true. The effect of talker familiarity was to reduce the hyperspace effect for front vowels, where naive listeners deviated most from the talker's vowel space. One weakness of the experiment is that the synthetic tokens were not very convincing exemplars of the talker's voice for some of the listeners.

The general result from this study is that vowels that sound 'right' to listeners tend to be more disperse in the acoustic vowel space than natural productions, but listeners' choices are modulated by consonant context and familiarity with the talker.

The perceptual preference, found in JFW and in this study, for a large acoustic vowel space supports Lindblom's theory of adaptive dispersion. Most evidence for adaptive dispersion is indirect, inferring that listeners must prefer maximal contrast because languages tend to roughly obey such a rule more often than they violate it. The robustness of the hyperspace effect is important because it is direct psycholinguistic support for adaptive dispersion. Listeners do prefer maximal contrast.

The modulating effects produced by consonant context and talker familiarity present an interesting theoretical challenge. How are these modulating factors encoded in the perceptual system? Two possibilities suggest themselves for further research. First, the effects of consonant context and talker variation could be due to processes of cue interaction or cue weighting that operate to determine listener expectations and/or preferences in the vowel selection task. This process-based account could be called procedural encoding because talker and consonant variation information is encoded or stored in recognition processes. A second possibility is that speech perception is

based on exemplar storage (Johnson, 1997). In this representation-based account, the perceptual system stores consonant and talker variation in rich exemplar-based category representations. Consonant context might be procedurally encoded because it is largely rule governed, but talker familiarity effects probably require rich exemplar-based representations to the extent that talker variation is arbitrary (Johnson, Strand, & D'Imperio, 2000).

References

Bradlow, A.R.: A comparative acoustic study of English and Spanish vowels. Journal of the Acoustical Society of America 97: 1916-1924 (1995).

Bradlow, A.R.: A perceptual comparison of the /i/-/e/ and /u/-/o/ contrasts in English and Spanish: Universal and language-specific aspects. Phonetica 53: 55-85 (1996).

Creelman, C.D.: Case of the unknown talker. Journal of the Acoustical Society of America 29: 655 (1957).

Green, K.P.; Tomiak, G.R. & Kuhl, P.K.: The encoding of rate and talker information during phonetic perception. Perception and Psychophysics 59: 675-92 (1997).

Halle, M.: The Sound Pattern of Russian. (Mouton, The Hague 1959).

Johnson, K., Flemming, E. & Wright, R.: The hyperspace effect: Phonetic targets are hyperarticulated. Language 69: 505-528 (1993).

Johnson, K., Strand, E.A. & D'Imperio, M.: Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics. xx: xxx-xxx (2000).

Johnson, K.: Speech perception without speaker normalization: An exemplar model. In Johnson, K. & Mullennix, J. (eds). Talker Variability in Speech Processing, pp. 145-166 (Academic Press, New York 1997).

Liljencrants, J. & Lindblom, B.: Numerical simulation of vowel quality systems: The role of perceptual contrast. Language 48: 839-862 (1972).

Lindblom, B.: Spectrographic study of vowel reduction. Journal of the Acoustical Society of America 35: 1773-1781 (1963).

Lindblom, B.: Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W., & Marchal, A. (eds). Speech Production and Speech Modeling. pp. 403-439 (Kluwer, Dordrecht 1990).

Lindblom, B. & Engstrand, O.: In what sense is speech quantal? Journal of Phonetics 17**:** 107-121 (1989).

Lindblom, B. & Studdert-Kennedy, M.: On the role of formant transitions in vowel recognition. Journal of the Acoustical Society of America 30: 693-703 (1967).

Mullennix, J.W., Pisoni, D.B. & Martin, C.S.: Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America 85: 365-378 (1989).

Mullennix, J.W. & Pisoni, D.B.: Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics 47: 379-90 (1990).

Nygaard, L.C.; Sommers, M.S. & Pisoni, D.B.: Speech perception as a talker-contingent process. Psychological Science 5: 42-45 (1994).

Padgett, J.: Contrast dispersion and Russian palatalization. In Hume, E.V. & Johnson, K. (eds). The Role of Speech Perception Phenomena in Phonology. (Academic Press, New York to appear).

Stevens, K.N. & House, A.S.: Perturbation of vowel articulations by consonantal context: An acoustical study. Journal of Speech and Hearing Research 6: 111-128 (1963).

Walker, S.; Bruce, V.; & O'Malley, C.: Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception & Psychophysics 59: 1124-33 (1995).