



# Auditory–visual integration of talker gender in vowel perception

Keith Johnson\*, Elizabeth A. Strand and Mariapaola D'Imperio

*Department of Linguistics, Ohio State University, 222 Oxley Hall,  
1712 Neil Avenue, Columbus, OH 43210-1298, U.S.A.*

*Received 4th August 1998, and accepted 26th November 1999*

The experiments reported here used auditory–visual mismatches to compare three approaches to speaker normalization in speech perception: radical invariance, vocal tract normalization, and talker normalization. In contrast to the first two, the talker normalization theory assumes that listeners' subjective, abstract impressions of talkers play a role in speech perception. Experiment 1 found that the gender of a visually presented face affects the location of the phoneme boundary between [u] and [ʌ] in the perceptual identification of a continuum of auditory–visual stimuli ranging from *hood* to *hud*. This effect was found for both “stereotypical” and “non-stereotypical” male and female voices. The experiment also found that voice stereotypicality had an effect on the phoneme boundary. The difference between male and female talkers was greater when the talkers were rated by listeners as “stereotypical”. Interestingly, for the two female talkers in this experiment, rated stereotypicality was correlated with voice breathiness rather than vowel fundamental frequency. Experiment 2 replicated and extended experiment 1 and tested whether the visual stimuli in experiment 1 were being perceptually integrated with the acoustic stimuli. In addition to the effects found in experiment 1, there was a boundary effect for the visually presented word: listeners responded *hood* more frequently when the acoustic stimulus was paired with a movie clip of a talker saying *hood*. Experiment 3 tested the abstractness of the talker information used in speech perception. Rather than seeing movie clips of male and female talkers, listeners were instructed to imagine a male or female talker while performing an audio-only identification task with a gender-ambiguous *hood-hud* continuum. The phoneme boundary differed as a function of the imagined gender of the talker. The results from these experiments suggest that listeners integrate abstract gender information with phonetic information in speech perception. This conclusion supports the talker normalization theory of perceptual speaker normalization.

© 1999 Academic Press

---

\*Correspondence to K. Johnson, Department of Linguistics, Ohio State University, 222 Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210-1298, U.S.A. E-mail: [kjohnson@ling.ohio-state.edu](mailto:kjohnson@ling.ohio-state.edu).

## 1. Introduction

The speech signal conveys a great deal more than can be represented in phonemic transcription. As Ladefoged & Broadbent (1957) emphasized more than 40 years ago, speech conveys social, cultural, and personal information, in addition to linguistic information, in a complex web of interrelating sources of variation. One important observation in this regard is that speakers adopt patterns of pronunciation which are to an extent arbitrary markers of their individuality; they adopt individual speaking styles (Eskenazi, 1993). Given this realization, our view of speech perception and auditory word recognition assumes that the recognition of words and phonemes in the speech signal takes place in an environment of non-invariance across talkers (Johnson, 1997*a, b*; see also Palmeri, Goldinger & Pisoni, 1993; Pisoni, 1993; Sheffert & Fowler, 1995; Goldinger, 1998).

The convenient fiction of an “ideal speaker-listener” proposed by Chomsky (1965)<sup>1</sup> reflects a set of basic assumptions used in most studies of speech production and perception. One of the primary assumptions in Chomsky’s idealization is that speakers and listeners employ an invariant speech code to transmit linguistic messages to each other (Lieberman *et al.*, 1967; Stevens & Blumstein, 1978), and so the apparent lack of acoustic invariance in the speech signal has traditionally been a topic of significant interest (Perkell & Klatt, 1986).

Other researchers have suggested that an invariant speech code is not a logical necessity in a theory of speech perception (Lindblom, 1990; Johnson, 1997*a*). This alternative approach harks back to an older attitude, expressed by Firth (1964),<sup>2</sup> that individual differences in speech are central to how language works in society. Firth’s assertion that there is no “narrow reflex connexion between speaking and hearing” is an assertion that speech production is patterned not only to provide cues for the identification of linguistic units but also serves a social function by conveying socio-cultural and personal information (see Ladefoged & Broadbent, 1957). In this view, speech variability across talkers is structured by socio-cultural factors to which listeners are sensitive.

The experiments reported in this paper explore these factors by investigating the influence of auditory and visual talker information in vowel perception. To place this work in context, we first review the literature on auditory–visual integration in speech perception (in Section 1.1), as well as the literature on perceptual normalization (in Section 1.2).

### 1.1. Auditory–visual integration in speech perception

In face-to-face communication, listeners use both auditory and visual information to recognize speech. For example, Sumby & Pollack (1954) found that the speech reception

<sup>1</sup>“Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance” (Chomsky, 1965, p. 3).

<sup>2</sup>“Your ears are eager to make allowances for considerable variation of voice and pronunciation in the interest of common understanding which it is their business to serve. Indeed it might almost be said that, snobbery apart, the peculiar excellence of the ear for speech is precisely its latitudinarianism. If intelligibility depended on a narrow reflex connexion between speaking and hearing, we should all speak exactly alike and be no better than poultry” (Firth, 1937, 1964, pp. 22–23).

threshold shows a gain of up to 20 dB when speech is seen as well as heard (see also Erber, 1969, 1975; Dodd, 1977; Summerfield, 1979; Grant & Braida, 1991). This increased intelligibility is also available from relatively impoverished point-light facial displays (Rosenblum, Johnson & Saldana, 1996). Indeed, one of the infant's first speech behaviors is to integrate the auditory and visual aspects of speech (Kuhl & Meltzoff, 1984; Kuhl, Williams & Meltzoff, 1991; Bahrick, Netto & Hernandez-Reif, 1998). During the past 20 years, the process of auditory–visual (AV) integration in speech perception has been studied extensively using a method of AV mismatch originally developed by McGurk & MacDonald (1976). In this method, listeners are asked to respond to stimuli for which the visual phonetic information is different from the auditory phonetic information.

#### 1.1.1. *AV integration in consonant place perception*

McGurk & MacDonald (1976) and MacDonald & McGurk (1978) found that visual information can influence the perception of consonant place of articulation in clear and unambiguous auditory signals. In their classic demonstration (often called simply the “McGurk effect”), audio /ba/ was synchronized with a videotape of the talker saying /ga/. Observers watching and listening to these AV mismatches typically heard the talker saying /da/, a blend of the audio and visual syllables. This effect has been replicated numerous times, and occurs whether or not the observer is aware that AV mismatches will be presented. Additionally, the effect occurs even when the display is shown upside down (Massaro & Cohen, 1996); it occurs with both real words and nonsense syllables (Dekle, Fowler & Funnell, 1992); and even infants show the McGurk effect (Rosenblum, Schmuckler & Johnson, 1997).

Consonant place, which is weakly present in the acoustic signal, is much more susceptible to AV blending of this sort than is consonant manner (Green & Kuhl, 1989). Clarity of the audio signal has been shown to effect the magnitude of the McGurk effect as well (Sekiyama, 1998), a fact which is consistent with evidence that the hearing-impaired and elderly use visual information more than do young subjects with unimpaired hearing (Walden, Montgomery, Prosek & Hawkins, 1990; Walden, Busacco & Montgomery, 1993; Tillberg, Roennberg, Svaerd & Ahlner, 1996; Grant & Seitz, 1998; Grant, Walden & Seitz, 1998; Massaro & Cohen, 1999). The McGurk effect has been demonstrated for speakers of a variety of languages, with some evidence for cross-linguistic differences in the amount of visual information used in speech perception (Sekiyama & Tohkura, 1993; Sekiyama, 1997; Sams, Manninen, Surakka, Helin & Katto, 1998).

The perceptual or cognitive nature of the AV integration found in the McGurk effect has been the focus of a number of studies (e.g., Massaro & Friedman, 1990; Fowler & Dekle, 1991; Green & Kuhl, 1991). In one noteworthy analysis, Braida (1991) studied whether the cross-modal integration in the McGurk effect occurs prior to separate identification of the place of articulation in the visual and auditory signals. Braida (1991) found that a “prelabelling” model of integration produced more accurate predictions of the results of five studies of consonant place AV integration than did “postlabelling” models of integration.

#### 1.1.2. *AV integration in vowel perception*

In addition to studies of AV integration in the perception of stop place of articulation, several studies have shown that auditory and visual phonetic information are integrated

in the perception of vowels. For example, Summerfield & McGrath (1984) hypothesized that visual information would influence vowel perception because vowel distinctions involve noticeable articulatory motions of the face (lip rounding and jaw height). They tested this hypothesis by pairing stimuli from three synthetic vowel continua (/i/-/u/, /i/-/a/, /u/-/a/) with videos of a talker saying either /i/, /u/, or /a/. Their results indicated that the visual vowel influenced listeners' labeling performance on the continua (especially the /i/-/u/ continuum), both when listeners were unaware of the AV vowel mismatch as well as when they were told to expect AV mismatches and to respond to the audio signal only.

Lisker & Rossi (1992) also found visual effects in the perception of vowels by a group of speech researchers who were asked to judge the degree of lip-rounding in audio, visual, and AV presentations of vowels. They found, for AV stimuli in which face and voice lip-rounding were mismatched, that the judged degree of lip-rounding was influenced by the visual stimulus. This effect was still present, though weaker and less consistent across listeners, when listeners were told to expect AV mismatches and were instructed to base their responses on the audio stimulus alone. Green & Gerdeman (1995) also found AV integration in vowel perception. Their experiments measured listeners' sensitivity to visual vowel information in terms of the strength of the McGurk effect when the auditory and visual vowels were mismatched (e.g., /i/ vs. /a/). They found that the McGurk effect was weaker when the auditory and visual vowels are mismatched.

### 1.1.3. *AV integration with talker voice/face mismatches*

Some studies of AV integration have explored the effects of matching the face of one talker with the voice of another. For example, Green, Kuhl, Meltzoff & Stevens (1991) found that the magnitude of the McGurk effect was not reduced when the face and voice were mismatched for gender, even though the talker mismatch was perceptually apparent to subjects. From this finding, Green and colleagues argued that perceptual speaker normalization must alter the auditory representation before auditory and visual information are integrated. That is, they argued that their findings were consistent with a kind of "postlabelling" model of AV integration, to use Braida's (1991) term.

Sekiyama (1998) used "cross-talkers dubbing" to combine audio and video recordings of a compelling talker (one whose recorded stimuli seemed to evoke a strong McGurk effect) with recordings of a non-compelling talker. This study aimed to discover what combination of visual and auditory intelligibility would result in strong (or weak) McGurk effects. The author concluded that low audio intelligibility leads to greater reliance on visual cues for stop place of articulation (as would be expected from previous research on the strength of the McGurk effect with hearing-impaired listeners).

Bahrnick, Netto & Hernandez-Reif (1998) found that infants as young as 4 months detected AV mismatches in the age of a talker, preferring displays in which an adult's voice was paired with an adult face, as well as ones in which a child's voice was paired with the face of a child talking. Bahrnick *et al.* (1998) also found an interesting tendency for children to prefer viewing children's faces, but this study may have confounded AV synchronization and talker mismatch.

Walker, Bruce & O'Malley (1995) tested the strength of the McGurk effect "in a situation in which the face and voice identities [mismatched]". One group of their listeners was familiar with the talkers who provided the visual and auditory stimuli, while one group was not familiar with the talkers. Using talker-mismatched AV stimuli with

both gender mismatches and within-gender mismatches, Walker *et al.* (1995) found that the McGurk effect was dramatically reduced for both gender and talker mismatches when listeners were familiar with the talker. Schwippert & Benoit (1997) also employed a talker AV mismatch condition, with similar results.

Strand & Johnson (1996) presented subjects with gender-mismatched AV stimuli in a slightly different type of experiment. Our study looked for a shift in the identification functions of synthetic fricative noises as a function of the gender of the visually presented talker. The data suggested a conclusion that is somewhat different from that described in Green *et al.* (1991). We found that the gender of the visual talker produced a shift in the fricative labeling function, analogous to the shift found as a function of the gender of the voice in the auditory vowel. Therefore, we argued that visual information is used by listeners in a multimodal talker normalization process. The present study extends the results described by Strand & Johnson (1996) with a study of the integration of AV talker information in vowel perception. The experiments were designed to contrast three views of perceptual speaker normalization.

## 1.2. Speaker normalization in speech perception

Theories of speech perception must assume the existence of some mechanism to recover the “same” words or sounds from speech produced by different talkers. Following previous researchers, we will use the term “speaker normalization” to refer to this aspect of speech perception and reserve the term “talker normalization” to refer to one particular account of speaker normalization. Three broad approaches to speaker normalization can be found in the literature, including radical invariance, vocal tract normalization, and talker normalization. These approaches are briefly described in the following sections.

### 1.2.1. Radical invariance

A number of researchers have described the process of speaker normalization in speech perception as an auditory perceptual process (Potter & Steinberg, 1950; Nearey, 1978; Traunmüller, 1981; Syrdal & Gopal, 1984; Sussman, 1986; Miller, 1989; Watkins & Makin, 1994). What these approaches have in common is that they characterize speaker normalization as a purely auditory effect, involving the interaction of auditory dimensions to produce a (relatively) talker-independent auditory representation.

For example, Syrdal & Gopal (1984) proposed that the distances between peaks in the auditory spectrum give a talker-independent representation of vowels. In their approach, then, the relevant auditory dimensions of vowels are the frequency differences  $F1 - F0$ ,  $F2 - F1$ , and  $F3 - F2$ , where these frequencies are expressed on the auditory Bark scale. For Sussman (1986), the auditory dimensions are formant ratios. For Miller (1989), the dimensions are the differences of the log formant frequencies (which are equivalent to the logs of the formant ratios), with the innovation that  $F0$  is entered into these ratios as the time-averaged value (geometric mean) rather than as an instantaneous  $F0$  in the vowel. The unifying perspective in these works is that they assume that in some projection of the acoustic space, there must be a talker-independent—or invariant—representation of vowel categories. They further assume that this invariant representation is the one used by listeners in recognizing vowels.

The radical invariance view seems to have been the one adopted by Green *et al.* (1991) in their account of the AV talker-mismatched McGurk effect, in which they posit that the acoustic signal is normalized prior to AV integration. The Strand and Johnson (1996) finding suggests, however, that perceptual speaker normalization is sensitive to both auditory and visual information. This “visual talker normalization effect” is incompatible with the radical invariance view, which places speaker normalization wholly in the auditory domain.

### 1.2.2. Vocal tract normalization

In contrast to radical invariance, a number of researchers have proposed that speaker normalization is accomplished by reference to estimated vocal tract length (Joos, 1948; Ladefoged & Broadbent, 1957; Fujisaki & Kawashima, 1968; Nordström & Lindblom, 1975; Holmes, 1986; Nearey, 1989). For example, Ladefoged & Broadbent (1957) presented ambiguous test words in carrier phrases that had different formant frequency ranges (see also Ainsworth, 1974, 1975; Dechovitz, 1977; van Bergem, Pols & Koopmans-van Beinum, 1988; Nearey, 1989). Ladefoged & Broadbent (1957) suggested that the range of vowel formant frequencies exhibited in the carrier phrase defines for the listener the length of the speaker's vocal tract, information which is then used to interpret the ambiguous vowel formant pattern in a test word at the end of the phrase.

Fujisaki & Kawashima (1968) interpreted effects of  $F3$  and  $F0$  frequencies in isolated syllables in a similar way. They concluded that these frequency components of an isolated vowel serve as acoustic cues for vocal tract length, which the listener then uses to evaluate ambiguous  $F1$  and  $F2$  values. Gussenhoven and Rietveld (1998) also found that the range of formant values in a phrase influenced listeners' perception of  $F0$  in intonational prominence judgments. In general, proponents of the vocal tract normalization approach have suggested that a talker-independent representation is not a direct result of auditory processing (as in the radical invariance view), but rather results from a somewhat indirect path of inference based on a perceptual estimate of the length of the speaker's vocal tract.

There are a number of direct and indirect acoustic cues that the listener can use to estimate the length of the speaker's vocal tract. Direct cues are aspects of the acoustic speech signal that are directly related to vocal tract length, and include  $F3$  frequency (as in, e.g., the evaluation of the phonetic significance of  $F1$  and  $F2$  in Nordström & Lindblom, 1975) and the range of formant frequencies in a carrier phrase (Ladefoged & Broadbent, 1957). Indirect cues of vocal tract length are aspects of the speech signal that are not causally related to the length of the vocal tract though they may, on average, be correlated with vocal tract length, especially if the correlation includes values for both male and female speakers. Such indirect cues include  $F0$  (Fujisaki & Kawashima, 1968) and mode of vocal fold vibration (Holmes, 1986; Nearey, 1989).

It is important to note also that vocal tract length is likely at least partially available in a visual display of a speaker, so visual information can also provide perceptual evidence about vocal tract length. In this way, the vocal tract normalization approach can also offer an account of Strand & Johnson's (1996) visual talker normalization effect.

The perceptually estimated vocal tract that listeners are assumed to use in this approach is sometimes given a very specific definition. For example, Nordström & Lindblom (1975) defined the perceptual vocal tract length factor as the length of the vocal tract from the larynx to the lips, as estimated from the frequency of  $F3$  in low unrounded

vowels. Estimated vocal tract length is then used to scale the raw formants to a talker-independent vocal tract length.

This explicit formulation of vocal tract normalization reveals a problem, however, in the vocal tract normalization approach: the actual length of the vocal tract is not static, but instead varies constantly during speech. Therefore, “estimated vocal tract length” is necessarily an abstraction as regards most speech sounds. Further emphasizing the abstract nature of the vocal tract length factor is the assumption made by some authors that acoustic features not causally related to vocal tract length (i.e., indirect cues) may be used by listeners in perceptual speaker normalization. Among these factors are *F0* and voice quality, which, though correlated with vocal tract length in samples that include men and women, are not determined acoustically by the length of the vocal tract. The third approach to perceptual speaker normalization, termed “talker normalization”, identifies this abstract vocal tract as connected with the perceived identity of the talker.

### 1.2.3. Talker normalization

Johnson (1990a) asked listeners to rate synthetic speech stimuli on a scale from “most male” to “most female” and found that these ratings of perceived talker identity were good predictors of perceptual vowel normalization in a series of experiments. These experiments were modeled on the Ladefoged & Broadbent (1957) classic experiment which varied the formant range in carrier phrases to induce a labeling change for stimuli embedded in the carrier phrases. Johnson (1990a) used carriers which had different *F0* ranges (evoking different talkers), while the test vowels in the carriers were identical. The relationship between perceived talker identity and vowel identification behavior seems to pose a problem for the radical invariance view, but is not incompatible with the vocal tract normalization view. Johnson (1990a), however, chose to call the latent variable “perceived talker identity” rather than “estimated vocal tract length”.

Several implications follow from this terminological shift.<sup>3</sup> First, an abstract characterization of the listener’s perceptual representation of the talker suggests that perception may be influenced by factors beyond just acoustic or visual cues for vocal tract length, such as familiarity with a particular talker, or general socio-cultural expectations (including expected differences between men and women). This provides a connection with research on gender differences across languages (e.g., Fant, 1975; Bladon, Henton & Pickering, 1984) which shows that the acoustic differences between men’s and women’s speech vary from language to language.

A notion of “perceived talker identity” in a talker normalization approach gives us a way to deal with the implication of these cross-linguistic studies, i.e., that an *accurate* vocal tract normalization process could produce *inaccurate* speech identification. The talker normalization approach is also consistent with the Walker *et al.* (1995) finding that familiarity with the talker reduces AV integration in the McGurk effect for

<sup>3</sup>We are using “talker” to refer to the listener’s perceptual impression of a person’s identity based on hearing or watching the person speak. This contrasts with our use of “speaker” in a more generic sense, which includes objective facts about a person, such as vocal tract length. Similarly, because we are espousing a theory of speech perception that refers to subjective impressions of talkers we also use the term “gender” to refer to sex-based differences between people that are subjective in that “gender” refers to a socio-cultural construct based in part on sex. Thus, when we call our theory “talker normalization” as distinct from “speaker normalization”, we mean to assert that speech perception is influenced by socio-culturally based “gender” expectations.

talker-mismatched stimuli. This finding indicates that talker identity, not just vocal tract length, is involved in speech perception. Similarly, Nygaard, Sommers & Pisoni (1994) found an effect of talker familiarity in speech perception. The results of this work suggest that speech recognition in noise was easier for listeners when the stimuli were produced by familiar voices. In addition, the Daly, Bench & Chappell (1996) finding that gender stereotypes affect speechreading performance suggests that socio-cultural expectations may also influence speech processing.

Second, accepting “perceived talker identity” as a relevant concept in speech perception theory has implications for our view of the cognitive implementation of the normalization process. In vocal tract length normalization, speaker normalization is implemented as an algorithmic process: the listener estimates vocal tract length and then scales the formant frequencies or speech spectrum accordingly. In talker normalization, though, we view speech perception in a system of rich representations of talkers as well as phonetic categories (Mullennix & Pisoni, 1990; Palmeri *et al.*, 1993; Pisoni, 1993; Sheffert & Fowler, 1995; Tomiak & Kuhl, 1997; Johnson, 1997*a, b*; Goldinger, 1998). This reliance on rich representations offers an account for the listener’s ability to learn talker-specific and language-specific patterns of indirect as well as direct talker cues, and it provides a mechanism for gender stereotypes to have an impact on speech perception. We will return to the processing implications of talker normalization theory in the conclusion.

### 1.3. Overview of experiments

Experiment 1 extends the Strand and Johnson (1996) study of fricative perception to address AV integration in the perception of vowels as well. As in our earlier work, male and female acoustic stimuli were crossed with male and female visual stimuli, and the gender stereotypicality of the voices was varied. For these voices, however, gender stereotypicality was, to an extent, independent of  $F_0$ . Experiment 2 replicates the results of experiment 1 and shows that these effects occur when auditory and visual information are integrated in phonetic perception. In experiment 2, we found visual vowel effects similar to those reported in Summerfield & McGrath (1984), as well as Lisker & Rossi (1992). Experiment 3 explores the abstractness of talker representations using the technique of “imagined” faces (Kuhl, Williams & Meltzoff, 1991).

## 2. Experiment 1

Experiment 1 explores the interaction of visual and auditory gender in vowel perception. In this experiment, listeners were presented with movies of talkers saying words that could be identified as *hood* or *hud*. The independent variables were the visual and auditory gender of the talker and the rated stereotypicality of the voice gender, and the dependent variable was the boundary between the vowels [ʊ] and [ʌ] along an  $F_1$  continuum. This experiment allows us to explore in a preliminary way the role of listeners’ gender expectations in speech perception, where gender expectations are manipulated both visually and acoustically.

The four voices used in the experiment spanned a range of rated gender stereotypicality where, for the female talkers, fundamental frequency ( $F_0$ ) was not correlated with perceived talker identity. This manipulation allows us to ask about the specificity of



gender expectations. We also manipulated the visual gender of the stimuli, in one condition pairing each of the voices with a movie of a male talker and in another condition pairing each of the voices with a movie of a female talker. This manipulation allows us to ask whether gender-based expectations evoked by seeing a male or a female face will have an effect on the perception of speech and whether this effect will interact with the gender stereotypicality of the voice. There are two questions that we want to ask here. First, does the gender of a face affect speech perception performance? And, does this effect happen for all, or only some, voices?

## 2.1. Method

### 2.1.1. Participants

There were four groups of participants in this study. The first group of 39 participants (25 women and 14 men) were video- and audio-taped as they read a word list including the words *hood* and *hud*. The second group of 18 participants (7 men and 11 women) watched short video-only movie clips of the first group and rated each one on a 5-point scale from “most male appearing” to “most female appearing”. The third group of 11 participants (6 men and 5 women) rated the endpoints of the resynthesized audio stimuli on a 5-point scale from “most male sounding” to “most female sounding”. The fourth group of 20 participants (5 men and 15 women) viewed the complete digital movies (video clip plus resynthesized audio stimulus) and identified the word produced in each one as either *hood* or *hud*.

Participants in groups 1, 2, and 4 were undergraduate students at The Ohio State University who received 5 dollars each for participating in the study. Group 3 was composed of graduate students in the Ohio State Department of Linguistics who volunteered their time. None of the participants reported any history of speech or language disorder and all were native speakers of English.

### 2.1.2. Visual stimuli

The visual portions of the movies were constructed by digitizing video clips of the head and shoulders of two Ohio State undergraduate students as they pronounced the word *hud*. We chose these two talkers on the basis of viewers’ (in group 2 above) ratings of them at the extremes of the “most male appearing” to “most female appearing” continuum. Video clips of the talkers producing *hud* were digitized on a Macintosh 7100 AV computer at 30 frames per second in 24 bit color. The video signals were edited to show the face of the talker for 1 s prior to the onset of speaking (one frame just prior to the word onset was repeated 30 times), and the final frame after word coda was repeated to pad the clip to a total duration of 2 s. The size of the video stimuli presented to subjects was approximately  $160 \times 140$  pixels ( $3'' \times 2.5''$ ) on a 15" color monitor.

### 2.1.3. Auditory stimuli

The sound tracks of the movies were constructed using naturally produced source functions from four talkers to excite a bank of time-varying filters which corresponded to vowel formants in a continuum from *hood* to *hud*.

For each of four talkers (two male and two female) drawn again from group 1, we extracted the source function from a production of *hud* by inverse filtering the original

speech waveform (digitized at 11.025 kHz, 16 bits) with a time-varying LPC inverse filter (10 ms step size, 25 ms analysis window). The number of LPC coefficients was 10 for the female speakers and 12 for the male speakers. Vowel durations were matched within gender.

We filtered these extracted source functions using a cascaded bank of time-varying band pass filters to produce synthetic stimuli ranging from *hood* to *hud* and having the voice source characteristics of the original talkers. The synthesis parameters for the *hood* to *hud* continuum are shown in Table 1, while schematic representations of the endpoint stimuli formant trajectories are shown in Fig. 1. The four continua produced by this method thus had identical formant trajectories and differed only in voice source.

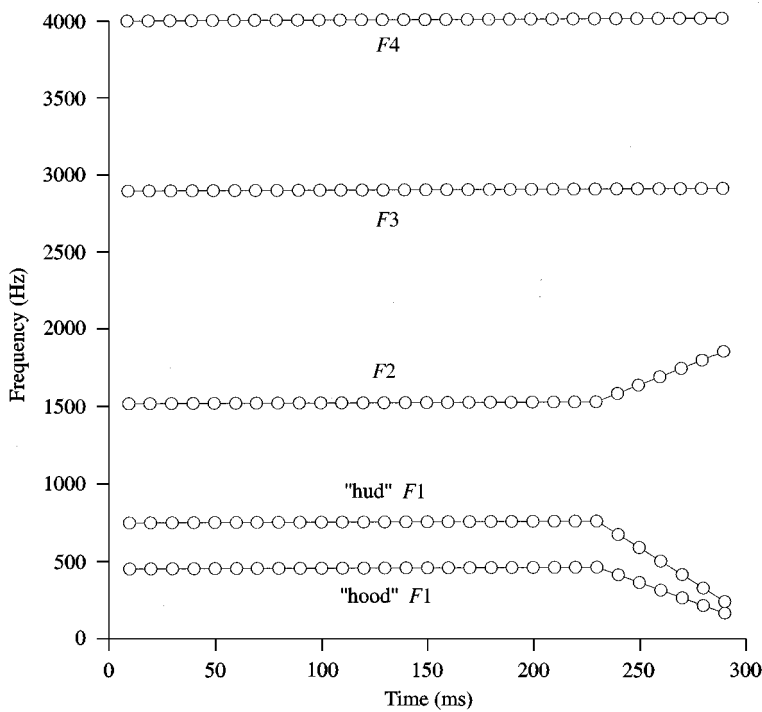
The *hood* and *hud* end-point stimuli in these continua were presented to listeners in group 3 for rating on a 5-point scale from “most male sounding” to “most female sounding”. The voices ranged from an average rating of 1.5 (a “stereotypical male” voice) to 4.15 (a “stereotypical female” voice). The other two voices we will refer to as “non-stereotypical male” (average rating of 2.0) and “non-stereotypical female” (average rating of 4.06), though the two female voices in particular were not rated very differently on this rating scale. Interestingly, the stereotypicality ratings for the female speakers do not correlate with fundamental frequency. The speaker who was rated as marginally less stereotypical in these tokens had a higher fundamental frequency at the midpoint of the vowel than did the stereotypical speaker (213 Hz for the less-stereotypical female voice, vs. 196 Hz for the more stereotypical female voice).

Acoustic analysis of the *hood* end-point stimuli suggests that the bases for listeners’ stereotypicality ratings of the female stimuli were associated with spectral tilt, the acoustic measurements for which are given in Table 2. The difference in amplitude between the loudest harmonic in the *F1* region and the amplitude of the first harmonic (*F1*–*H1*; Ni Chasaide & Gobl, 1997) was larger in the non-stereotypical female voice than it was in the stereotypical female voice. In addition, the absolute difference between the amplitude of *F4* and the amplitude of *F3* was greater for the stereotypical female voice than it was for the non-stereotypical female voice.

So, the female voice that was judged to be (somewhat) more stereotypical had a higher-amplitude first harmonic and lower *F4* amplitude, both of which suggest a difference in the manner of vocal cord vibration between the two speakers such that the stereotypical speaker had a more steeply sloped spectrum (i.e., more breathy phonation). For these two speakers, then, we have a case in which judged stereotypicality follows phonation type rather than fundamental frequency.

TABLE 1. Formant frequencies and bandwidths (in Hz) used in the resynthesized speech stimuli in experiments 1–3. The first seven columns show the steady-state formant frequencies. The eighth column shows the bandwidths of the formants, and the last column shows the target locus frequencies (20 ms beyond the end of vowel voicing) of the formants

	1	2	3	4	5	6	7	BW	Locus
<i>F1</i>	450	495	542	591	642	694	750	100	50
<i>F2</i>	1520	→						130	1950
<i>F3</i>	2900	→						186	2900
<i>F4</i>	4000	→						230	4000



**Figure 1.** Schematic representation of trajectories of the first 4 formants of the synthetic stimuli.

TABLE 2. Acoustic measurements of the synthetic stimuli used in experiments 1 and 2. The measurements were taken from stimulus 1 (the *hood* end-point stimulus) in each of the four synthetic continua. The male and female talkers who were judged by listeners as more “male sounding” or more “female sounding” are labeled stereotypical male or female (St-M or St-F). Likewise, the male and female talkers who were judged to be less “male sounding” or “female sounding” are labeled non-stereotypical male or female (Ns-M or Ns-F). *F*0 and spectral tilt measures (*F*1a-*H*1a and *F*4a-*F*3a) were taken from the midpoint of the vowel (60 ms Hamming window, DFT spectrum)

	St-F	Ns-F	Ns-M	St-M
Duration (ms)	190	220	160	150
<i>F</i> 0 (Hz)	196	213	143	82
<i>F</i> 1ampl <i>H</i> 1ampl (dB)	7.8	13.6	10.9	15.7
<i>F</i> 4ampl- <i>F</i> 3ampl (dB)	− 17.0	− 10.3	− 9.3	− 8.0

The audio stimuli (28 total, 7 values of *F*1 × 4 voices) were added as sound tracks to both the male visual stimulus and the female visual stimulus. This was done on-line as the experiment was conducted (during each response–stimulus interval) by replacing the original time-aligned sound track with the audio stimulus selected for a particular trial. Alignment of the onsets of the video and audio portions of the stimuli was accurate to

within one video frame (33.3 ms), which is well below the JND for AV misalignment in speech (Munhall, Gribble, Sacco & Ward, 1996).

#### 2.1.4 Procedure

Each AV stimulus was presented 10 times to subjects in group 4 for identification as *hood* or *hud* in a two-alternative forced-choice task. The response-to-stimulus interval (RSI) was approximately 2 s. Subjects were seated in a single-walled sound-attenuated booth. They faced a 15-in color monitor which showed both the visual portion of the stimulus and an on-screen response box with buttons labeled *hood* and *hud* (presenting the response box on-screen helped to insure that participants attended to the visual portion of the stimulus). The monitor was between 12 and 16 in from the subject's face. Audio stimuli were presented over earphones at a comfortable listening level and subjects used a mouse to choose their response on the on-screen response box. The order of presentation was randomized separately for each subject in 10 blocks of 56 stimuli, and no practice trials were given.

### 2.2. Results

For each of the eight experimental conditions (2 face genders  $\times$  2 voice genders  $\times$  2 levels of voice stereotypicality), we calculated each subject's identification crossover boundary between *hood* and *hud* by linear interpolation.<sup>4</sup> These data were then entered into a three-way repeated-measures analysis of variance with factors FACE gender (male or female), VOICE gender (male or female) and voice STEREOtypicality (stereotypical or non-stereotypical). The results are shown in Table 3, with Fig. 2 illustrating the FACE main effect and Fig. 3 illustrating the interaction of voice stereotypicality and voice gender.

The visual gender (FACE) main effect was reliable ( $F(1, 19) = 5.82, p < 0.05$ ) and in the direction we expected for a perceptual talker normalization effect. The  $F1$  boundary between *hood* and *hud* for the female-face stimuli (598 Hz) was higher than it was for the male-face stimuli (587 Hz), which corresponds to the fact that women generally have higher formant values than men.

The effect of the gender of the audio portion of the stimuli (VOICE) was also reliable ( $F(1, 19) = 62.6, p < 0.01$ ) and was also in the direction we expected for a perceptual talker normalization effect. The  $F1$  boundary between *hood* and *hud* for stimuli produced by female voices averaged over gender stereotypicality (610 Hz) was higher than was the boundary for the stimuli produced by male voices (575 Hz).

Table 3 and Fig. 3 also show the interaction between voice stereotypicality and voice gender (STEREO  $\times$  VOICE). This interaction was also reliable ( $F(1, 19) = 17.12, p < 0.01$ ). The lowest boundary on the  $F1$  continuum occurred with the stereotypical male voice (564 Hz) and the highest  $F1$  boundary occurred with the stereotypical female voice (617 Hz). The two non-stereotypical voices had boundaries between these extremes (non-stereotypical male at 586 Hz, and non-stereotypical female at 603 Hz).

None of the other factors in the ANOVA were significant.

<sup>4</sup>Interpolation is an adequate measure of boundary locations in these data because: (1) the continuum spanned a wide range of  $F1$  values; (2) listeners responded to each stimulus 10 times, thereby providing an accurate representation of the identification function for each condition; and (3) the identification functions were categorical. By using boundaries as a response measure, we do not intend to claim that phoneme boundaries have any theoretical status; they are merely a convenient measure of the perceptual effects we are studying.

TABLE 3. Face and voice effects in experiments 1 and 2 and the Instruction set effect in experiment 3. The table shows averaged phoneme boundaries (in Hz) on the F1 vowel continuum. Factors are: visual gender (Face Male = M, Female = F), voice gender (Voice Male = M, Female = F), and voice stereotypicality (stereotypical = St, non-stereotypical = Ns), or suggested gender (Instruction condition, Male = M, Female = F)

	Face		Voice			
	F	M	St-F	Ns-F	Ns-M	St-M
Experiment 1	598	587	627	603	586	564
Experiment 2	596	571	632	584	566	554
<i>Instruction condition</i>						
Experiment 3	604	589				

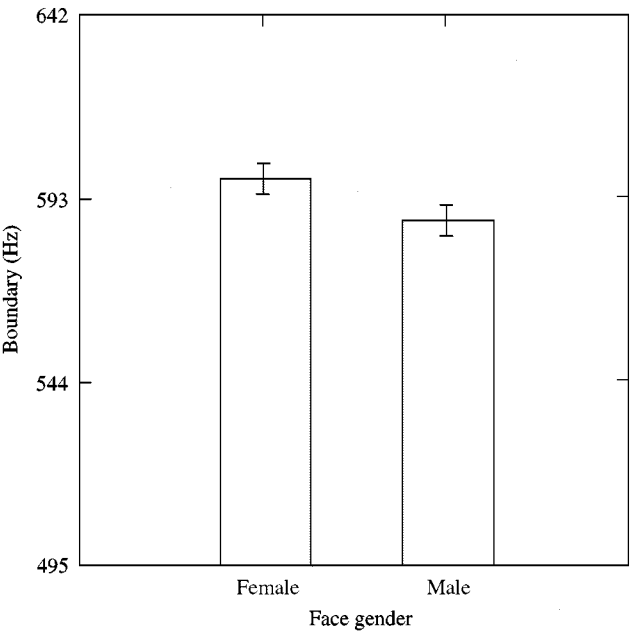
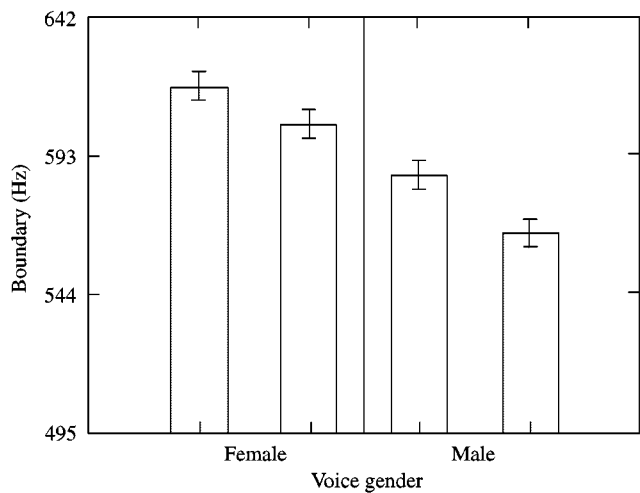


Figure 2. The face gender effect in experiment 1.

2.3. Discussion

The results of this experiment are consistent with previous research on perceptual vowel normalization, which has found that the boundaries between vowel categories differ for male and female voices (Miller, 1953; Fujisaki & Kawashima, 1968; Slawson, 1968; Johnson, 1990*b*). Our experimental results are also consistent with previous findings (Johnson, 1990*a*; Strand & Johnson, 1996) showing that perceptual speaker normalization is correlated with listeners’ ratings of voices on a stereotypicality scale, where the perceptual normalization effect is stronger with “stereotypical” voices than it is with



**Figure 3.** The voice gender  $\times$  voice stereotypicality interaction in experiment 1:   
□, stereotypical; ▒, non-stereotypical.

“non-stereotypical” voices. One important difference between the present results and earlier work is that in this experiment, listeners’ stereotypicality judgments were based on more than just *F*<sub>0</sub>; phonation type (in this case, breathiness) also seemed to affect perceptions of stereotypicality, at least for the female talkers. Unlike the stimuli used in Johnson (1990*a*) or Strand and Johnson (1996), in the present study the most stereotypical female voice did not have the highest fundamental frequency. The vowel boundaries observed in this experiment, then, were better-predicted by rated gender stereotypicality than by fundamental frequency (contra-radical invariance theories such as Syrdal & Gopal, 1984, or Miller, 1989).

Experiment 1 also found that vowel boundaries shifted as a function of the gender of the visually presented talker. This result is consistent with the “visual talker normalization effect” found by Strand & Johnson (1996) in fricative perception. The visual talker normalization effect is incompatible with the radical invariance view of speaker normalization because it shows that speaker normalization is not a purely auditory/perceptual process. This phenomenon also runs counter to the “first normalize, then integrate the audio and visual signals” view assumed by Green *et al.* (1991) in their study of visual talker mismatches in the McGurk effect.

The results of this experiment do not differentiate, however, the vocal tract normalization and the talker normalization theories of speaker normalization. Both of these theories predict that indirect cues for vocal tract length (in one case) or talker identity (in the other) will have an impact on speech perception. The experiment does, however, highlight the role of indirect acoustic cues (*F*<sub>0</sub> and breathiness) in contrasting talkers of the same gender where vocal tract length differences may be less well-correlated with acoustic cue differences.

One problem with this experiment is that the visual syllables were always tokens of the word *hud*. This means that we cannot rule out the possibility that the visual talker normalization effect involved a “post-labelling” mode of AV integration because we have no evidence of phonetic AV integration in listeners’ responses. In experiment 2, we

explore the nature of the AV integration by using visual stimuli that contrast both phonetic information (lip rounding and jaw height) and talker information.

### 3. Experiment 2

Experiment 2 is a replication and extension of experiment 1. In addition to the factors tested in experiment 1, this experiment tests for the integration of visual phonetic information in the vowel perception task. In one condition, we paired the acoustic stimuli with movies of talkers saying *hood* and in another condition with movies of talkers saying *hud*. This manipulation tests for AV integration of phonetic information, visual lip rounding, and jaw height with *F1* frequency in vowel perception. If the visual talker effect found in experiment 1 resulted from post-labelling AV integration, we might expect in experiment 2 to find lack of AV integration of phonetic information. If, however, in this experiment we find a visual word effect, we could conclude that the talker information in the visual stimulus is also being perceptually integrated with the acoustic stimulus (Mullennix & Pisoni, 1990; Green, Tomiak & Kuhl, 1997).

Previous research has found that AV integration of phonetic information can occur despite talker mismatches (Green *et al.*, 1991), and several researchers have found AV integration in vowel perception (Summerfield & McGrath, 1984; Lisker & Rossi, 1992; Green & Gerdeman, 1995), but no study has shown AV integration for vowels with talker-mismatched stimuli.

#### 3.1. Method

##### 3.1.1. Participants

Forty-seven participants (26 women, 21 men) participated in the experiment. They were all native speakers of English who reported normal speech and hearing and were paid 5 dollars for participating. Four participants were excluded from further analysis because they failed to identify the end-point stimuli accurately (this behavior was markedly different from that of the other listeners who treated the continuum categorically). Three participants were randomly excluded to equate the number of listeners in each group. Therefore, the data reported below are based on responses from 40 participants.

The participants were assigned to one of four groups, as described in the procedures below. Ten (6 women, 4 men) were in group 1, 10 (7 women, 3 men) were in group 2, 10 (6 women, 4 men) were in group 3, and 10 (5 women, 5 men) were in group 4.

##### 3.1.2. Stimuli

The AV stimuli used in the experiment were composed of the resynthesized audio stimuli used in experiment 1, which were then dubbed onto digitized video recordings of male and female talkers saying either *hood* or *hud*.

*Audio stimuli.* The 28 (seven *F1* steps  $\times$  two genders  $\times$  two levels of gender stereotypicality) audio stimuli from experiment 1 were used again in this study.

*Visual stimuli.* We prepared video clips of *hood* and *hud* as produced by the two talkers who were used in experiment 1. These talkers were rated in a pretest as being

stereotypically male or female in appearance, and the preparation of the video clips was the same as in experiment 1.

*Auditory-visual stimuli.* One hundred and twelve stimuli resulted from crossing the four visual stimuli with the 28 audio stimuli. As in experiment 1, the AV test stimuli were created on-line during the experiment. During the inter-stimulus interval, the original sound track of the selected digital movie was replaced by the selected audio stimulus.

### 3.1.3. Procedure

The experiment had four factors: voice stereotypicality, visual word, face gender, and voice gender. Due to the large number of stimuli, we decided to treat two factors (voice stereotypicality and visual word) as between-subjects factors, and treat the remaining two factors (face gender and voice gender) as within-subjects factors. Listeners in group 1 responded to 10 repetitions of each of the tokens composed from the non-stereotypical male and female voices and stereotypical male and female faces saying *hood*. Listeners in group 2 responded to 10 repetitions of each of the tokens composed from the stereotypical voices and stereotypical male and female faces saying *hood*. Listeners in group 3 were given the stimuli composed of the non-stereotypical voices and stereotypical male and female faces saying *hud*, and listeners in group 4 were given the stereotypical voices with stereotypical male and female faces saying *hud*.

All other aspects of the procedures were the same as in experiment 1.

## 3.2. Results

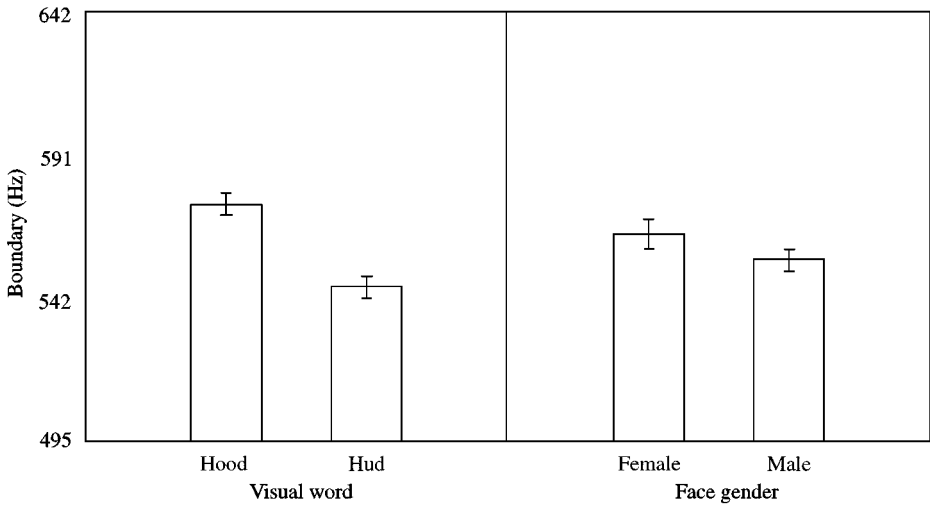
For each of the experimental conditions, we used linear interpolation to calculate each subject's 50% identification crossover boundary between *hood* and *hud*. These data were then entered into a four-way repeated-measures analysis of variance with factors FACE gender (male or female), VOICE gender (male or female), voice STEREotypicality (stereotypical or non-stereotypical), and visual WORD (*hood* or *hud*). Voice stereotypicality and visual word were between-subjects factors, while face gender and voice gender were within-subjects factors.

The visual word main effect (WORD) was significant ( $F(1, 36) = 26.9, p < 0.01$ , indicating that listeners' identification of the F1 *hood-hud* continuum was influenced by the word being produced in the visual display (the visual word main effect is illustrated in Fig. 4). When the visual display was a production of *hood*, there were more *hood* identification responses. The average phoneme boundary on the F1 continuum was 596 Hz when the visual word was *hood* and 577 Hz when the visual word was *hud*. This result indicates that phonetic AV integration occurred (Summerfield & McGrath, 1984; Lisker & Rossi, 1992; Green & Gerdeman, 1995).

The face gender main effect (FACE), also illustrated in Fig. 4, was also significant ( $F(1, 36) = 21.6, p < 0.01$ ). When the face was female, listeners were more likely to identify the stimulus as *hood* than when the face was male. The average phoneme boundary for female face stimuli was 596 Hz, while the average boundary for the male face stimuli was 571 Hz. See Table 3 for a comparison with the results of experiment 1.

The voice gender main effect (VOICE) was also significant ( $F(1, 36) = 81.4, p < 0.01$ ). Stimuli produced by female talkers were more often identified as *hood* than were stimuli produced by male talkers. The phoneme boundaries averaged over gender





**Figure 4.** The visual word and face gender effects in experiment 2.

stereotypicality in Table 3 were 607 Hz for female talkers and 560 Hz for male talkers. The boundary difference mirrors the typical difference in male and female formant values.

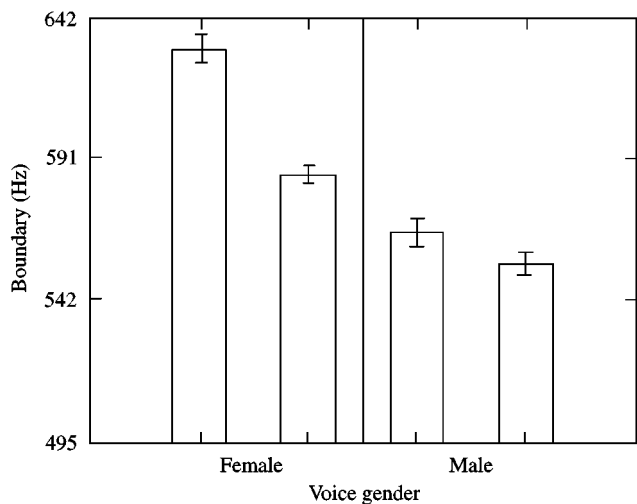
Also, as in experiment 1, gender stereotypicality interacted with voice gender (STEREO  $\times$  VOICE) ( $F(1, 36) = 31.91, p < 0.01$ ). The average phoneme boundaries for this interaction are shown in Table 3, and results are illustrated in Fig. 5. The voice gender effect was larger for the stereotypical voices than for the non-stereotypical voices. The difference between the female voices was larger than the difference between the male voices, and contributed to the stereotypicality main effect (STEREO), which was also significant ( $F(1, 36) = 13.0, p < 0.01$ ). The average boundary for the stereotypical voices was 591 Hz, while the average boundary for the non-stereotypical voices was 574 Hz.

Two others interactions were reliable and are shown in Table 4. The face gender by voice gender interaction (FACE  $\times$  VOICE) was significant ( $F(1, 36) = 5.42, p < 0.05$ ). The visual face effect was smaller with male voices ( $< 10$  Hz difference) than with female voices ( $\sim 20$  Hz difference). This interaction is related to the three-way interaction of face gender, voice gender, and voice stereotypicality (FACE  $\times$  VOICE  $\times$  STEREO), which was also significant ( $F(1, 36) = 7.4, p < 0.05$ ). The face gender effect did not occur for the stereotypical male voice but did occur for the other three voices.

### 3.3. Discussion

The weak three-way interaction of FACE  $\times$  VOICE  $\times$  STEREO suggests that the visual talker normalization effect may vary depending on the voice presented in the audio portion of the stimulus. This is a topic that warrants further research.

The main result of experiment 2 was that we found a visual word effect and a visual talker normalization effect, suggesting that listeners perceptually integrated AV information for both vowels and talkers. The visual word effect replicates earlier research on AV integration in vowel perception (Summerfield & McGrath, 1984; Lisker & Rossi,



**Figure 5.** The voice gender  $\times$  voice stereotypicality interaction in experiment 2:   
□, stereotypical; ▒, non-stereotypical.

**TABLE 4.** The three-way interaction of visual gender, gender stereotypicality, and voice gender in experiment 2. The table shows average phoneme boundaries (in Hz) on the *F1* vowel continuum. The visual gender effect (Face F *vs.* Face M) did not occur for the stereotypical male talker

	Voice F		Voice M	
	Face F	Face M	Face F	Face M
Stereotypical	643	619	554	554
Non-stereotypical	590	576	573	558
Average	617	598	564	556

1992; Green & Gerdeman, 1995) with stimuli that had talker gender mismatches, as in the Green *et al.* (1991) study of the McGurk effect with talker mismatches.

Several additional aspects of this experiment deserve comment. Johnson (1990*b*) and Green *et al.* (1997) found that when stimuli are blocked by talker (such that for any block of trials only one talker is presented), perceptual normalization effects largely disappear. In modeling this phenomenon, Johnson (1990*b*) concluded that it is due to ‘talker contrast” in mixed-talker lists. This conclusion calls into doubt the existence of stable talker representations independent of contrast effects within blocks. It is interesting to note, then, that the relative order of the vowel boundaries for the four talkers in this experiment (where talkers were blocked by stereotypicality) was the same as in experiment 1 (where all four talkers were presented in every block).

Table 3 indicates that the visual talker normalization effect was larger in this experiment than it was in experiment 1. This could be related to the suggestion by Diehl, Lindblom, Hoemeke & Fahey (1996) that female talkers tend to slightly hyperarticulate

vowels as compared with male talkers. Thus, if the female talker for our visual stimuli produced a more rounded *hood* than did the male talker, then the increased visual talker normalization effect in this experiment could reflect a gender difference in speech production. This explanation seems unsupported by our data, as the interaction of face gender and visual word was not significant. The visual word effect was just as large for the male visual talker as it was for the female visual talker.

The results of experiments 1 and 2 are clearly incompatible with the radical invariance view of speaker normalization as well as with audio-only versions of the vocal tract normalization view (Nordström & Lindblom, 1975). However, they do not distinguish between a “visually enhanced” version of the vocal tract normalization view and the talker normalization view. Experiment 3 aimed to distinguish these theories by exploring the abstractness of the talker representation used in speech perception. The vocal tract normalization view tends to be more compatible with a somewhat concrete, stimulus-driven conception of talker information (vocal tract length), while the talker normalization perspective is more compatible with an abstract, listener-subjective conception of talker information.

#### 4. Experiment 3

Experiment 3 was designed to explore the effects of an abstract, subjective talker representation in the perceptual identification of gender-ambiguous vowels. The experiment contrasts the vocal tract normalization theory with the talker normalization theory by focusing on the main difference between them: talker normalization assumes that listeners use an abstract, subjective representation of the talker during perception, while vocal tract normalization assumes that listeners perceive speech by reference to a more concrete parameter, the talker’s vocal tract length. In this experiment, we asked listeners to imagine the talker in an audio-only vowel identification experiment (see Kuhl, Williams & Meltzoff, 1991). The acoustic stimuli were gender-ambiguous; however, we told one group of listeners that the talker was female, while we told another group of listeners that the talker was male.

##### 4.1. Method

###### 4.1.1. Listeners

Thirty-three listeners (12 men, 21 women) participated in the experiment. The listeners were undergraduate students at the Ohio State University who were paid 5 dollars for participating in the experiment. None of the listeners reported any prior history of speech or hearing disorder and all were native speakers of English.

###### 4.1.2. Stimuli

A seven-step *F1* continuum from *hood* to *hud* was produced by using the voice modification technique of Qi (1990), which employs an implementation of the Liljencrants-Fant (LF) voice source model (Fant et al., 1985). The *F0* in the stimuli for this experiment was higher than the typical male voice and lower than the typical female voice. We modified the voice quality by changing the voice source parameters in the LF model so that the talker sounded gender-ambiguous to us.

The voice source from a naturally produced token of *hud* was calculated using LPC inverse filtering as in experiments 1 and 2, and this voice source was replaced by an LF model voice source which had a fundamental frequency of 134 Hz in the center of the vowel and the original intonation contour. The modified voice source was played through a set of time-varying band-pass filters as in experiments 1 and 2 to produce a seven-step continuum from *hood* to *hud*. The formants and bandwidths were the same as those used in experiments 1 and 2. The total duration of each stimulus was 310 ms.

#### 4.1.3. Procedure

The listeners were randomly assigned to one of two groups. The first group was told that the talker was female and that they should imagine the female talker while performing the vowel identification task. The second group was told that the talker was male, and was encouraged to imagine a male talker throughout the task. To motivate the listeners to imagine the talker, we showed each listener a questionnaire which asked about physical characteristics of the talker such as age, weight, and height, and we told them that they would be asked to complete the questionnaire at the end of the listening experiment.

After receiving these instructions, listeners were asked to perform two-alternative forced-choice identifications. The seven stimuli in the *hood-hud* continuum were presented seven times to each listener. The order of presentation was randomized separately for each listener. Stimulus presentation and response collection was done on-line.

After participating in the forced-choice labeling task and completing the questionnaire, the listeners were asked to rate a set of 39 natural tokens and five synthetic tokens of *hood* on a seven-point scale ranging from “most male sounding” (1) to “most female sounding” (7). The *hood* end-point stimulus of the continuum used in the present experiment was judged to be ambiguous between a male and female voice, with an average rating of 3.04.

## 4.2. Results

We measured the 50% identification cross-over boundary between *hood* and *hud* separately for each listener by linear interpolation. Table 3 (above) shows the average *F1* boundary values for the male and female instruction conditions. Listeners were more likely to label an ambiguous vowel *hood* if they were told that the speaker was female ( $t(31) = 2.24, p < 0.05$ ). The average location of the boundary when listeners were told that the speaker was female was 604 Hz, while the boundary was at 589 Hz when listeners were told that the speaker was male. The magnitude of this boundary difference is comparable to the boundary differences seen in experiments 1 and 2 in the visual talker normalization effect.

Table 5 shows average phoneme boundaries for each of the seven blocks of trials in this experiment, while Fig. 6 presents these results graphically. We were concerned that as the experiment progressed, listeners might disregard the secondary task (to envision the male or female talker) and focus only on the vowel labeling task. The data shown in Table 5 and represented in Fig. 6 indicate that this happened. The instruction manipulation produced large boundary differences only in the first and last blocks of the experiment; that is, the suggested gender of the talker only had a substantial impact

TABLE 5. Results of experiment 3 by trial blocks. The table shows phoneme boundaries (in Hz) of response functions tabulated over listeners separately for each block of trials during the experiment

Block	1	2	3	4	5	6	7
Female instructions	612	609	609	599	604	594	614
Male instructions	571	597	605	586	601	578	584
Difference	41	12	4	13	3	16	30

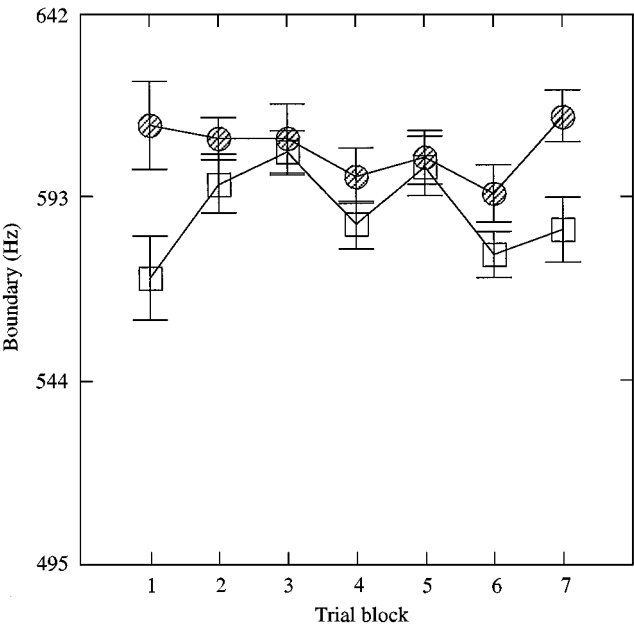




Figure 6. The instruction set effect in experiment 3 as a function of trial block: Instruction set , female; , male.

immediately after the instructions were given and just before the post-experiment questionnaire was to be completed.

4.3. Discussion

This experiment found that the boundary between *hood* and *hud* was sensitive to instructions about the identity of the talker. The voice of the talker in these stimuli was ambiguous between a male and female talker, and one group of listeners was told that the talker was female while another group of listeners was told that the talker was male. This manipulation resulted in a difference in the vowel phoneme boundaries, both averaged across blocks and especially in blocks in which listeners were most likely to attend to the instructions. In light of previous work using carrier phrases or visual images to convey the gender of the talker to listeners, this result seems to suggest that at least a part of the

boundary shift associated with “talker normalization” is due to expectations regarding male and female voices that listeners bring to speech perception, quite apart from the sensory information presented in the stimuli. This evidence for the perceptual impact of imagined talkers supports the “talker normalization” view of perceptual speaker normalization.

## 5. General discussion

We found a small but reliable visual talker normalization effect (experiment 1), which occurs when auditory and visual phonetic information are integrated in phonetic perception (experiment 2), and which can be evoked when listeners are instructed to imagine a male or female talker in the absence of a visual display (experiment 3). These results suggest that speaker normalization in speech perception is based on abstract, subjective talker representations and that listeners perceive talker identity from the totality of information available in the listening situation, including direct acoustic cues for vocal tract length (formants), indirect cues such as *F0* and mode of vocal fold phonation, visual cues, and even imagined talker characteristics. In experiments 1 and 2, we nullified direct vocal tract length cues by controlling the formant frequencies and found that listeners’ vowel identification responses were sensitive to the gender of the visual talker, as well as the *F0* and breathiness of the talker’s voice. In experiment 3, we allowed no variation in any acoustic or visual cue for talker identity and found a talker normalization effect when we asked listeners to imagine the gender of the talker.

In our view, listeners bring to bear experience-based expectations about talkers in the process of perceiving speech. When listeners identify a talker as either female or male, they access gender expectations for what the talker should sound like, and employ these expectations in speech perception. One way that a talker normalization theory can be implemented is as an exemplar-based perceptual system for speech (Mullennix & Pisoni, 1990; Pisoni, 1993; Palmeri et al., 1993; Sheffert & Fowler, 1995; Green, Tomiak & Kuhl, 1997; Johnson, 1997*a, b*; Goldinger, 1998). Exemplar models of category structure have a long history in cognitive psychology (Semon, 1923; Hintzman, 1986; Nosofsky, 1984, 1986; Schacter, Eich & Tulving, 1978; Estes, 1993; Reber, 1993) and are particularly well-suited to account for the perception of the overlapping multimodal (as opposed to Gaussian) acoustic categories used in speech. In the model we have been developing (Johnson, 1997*a, b*), linguistic categories such as the words *hood* and *hud* are composed of detailed instances in memory; likewise, the perceptual representation of a talker is composed of detailed instances in memory (Schacter, 1987). In this kind of system, to “bring experience-based expectations to bear in speech perception” means that the exemplars activated by a particular stimulus are similar to that stimulus, and because the activated exemplars define the system’s response to the input, the perceptual linguistic response emerges on-line from past instances. The listener’s experience of different talkers then produces the abstract, subjective talker representation that characterizes the talker normalization approach. The talker is represented not by one or two defining parameters, but is rather an amalgam of diverse details from the listener’s experience.

We gratefully acknowledge helpful comments from Christian Abry, Marie-Agnes Cathiard, Randy Diehl, Michael Cohen, and especially Terry Nearey, on an earlier version of this paper. Experiment 1 was first presented at the 1998 Annual Meeting of the Linguistic Society of America in New York. Experiment 3 was first

presented at the Fall 1997 meeting of the Acoustical Society of America at Penn State. We are grateful to the conference participants for their comments and to our colleagues at OSU for their support and feedback. This material is based upon work supported by the National Institute on Deafness and Other Communication Disorders under Grant No. 7 R29 DC01645-06.

## References

- Ainsworth, W. A. (1974) The influence of precursive sequences on the perception of synthesized vowels, *Language and Speech*, **17**, 103–9.
- Ainsworth, W. A. (1975) Intrinsic and extrinsic factors in vowel judgments. In *Auditory analysis and perception of speech*. G. Fant, & M. Tatham, editors. London: Academic Press.
- Bahrick, L. E., Netto, D., & Hernandez-Reif, M. (1998) Intermodal perception of adult and child faces and voices by infants, *Child Development*, **69**, 1263–75.
- Bergem, D. R. van, Pols, L. C. W., & Koopmans-van Beinum, F. J. (1988) Perceptual normalization of the vowels of a man and a child, *Speech Communication*, **7**, 1–20.
- Bladon, A., Henton, C., & Pickering, J. B. (1984) Towards an auditory theory of speaker normalization, *Language Communication*, **4**, 59–69.
- Braida, L. D. (1991) Crossmodal integration in the identification of consonant segments, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **43**, 647–77.
- Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Daly, N., Bench, J., & Chappell, H. (1996) Interpersonal impressions, gender stereotypes, and visual speech, *Journal of Language and Social Psychology*, **15**, 468–479.
- Dechovitz, D. (1977) *Information conveyed by vowels: a confirmation*. Haskins Laboratories: Status Report on Speech Research **SR-51**|**52**, pp. 213–9.
- Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992) Audiovisual integration in perception of real words, *Perception & Psychophysics*, **51**, 355–62.
- Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996) On explaining certain male-female differences in the phonetic realization of vowel categories, *Journal of Phonetics*, **24**, 187–208.
- Dodd, B. (1977) The role of vision in the perception of speech, *Perception*, **6**, 31–40.
- Erber, N. P. (1969) Interaction of audition and vision in the recognition of oral speech stimuli, *Journal of Speech and Hearing Research*, **12**, 423–5.
- Erber, N. P. (1975) Auditory, visual and auditory–visual recognition of consonants by children with normal and impaired hearing, *Journal of Speech and Hearing Research*, **15**, 413–22.
- Eskenazi, M. (1993) Trends in speaking styles research. *EuroSpeech93*, pp. 501–9. Berlin.
- Estes, W. K. (1993) Concepts, categories, and psychological science, *Psychological Science*, **4**, 143–53.
- Fant, G. (1975) *Non-uniform vowel normalization*. Speech Transmission Laboratory — Quarterly Progress Report, Vols. **2–3**, pp. 1–19. Royal Institute of Technology, Stockholm.
- Fant, G., Liljencrants, J., & Lin, Q. (1985) *A four-parameter model of glottal flow*. Speech Transmission Laboratory — Quarterly Progress Report, Vol. **4**, pp. 1–13. Royal Institute of Technology, Stockholm.
- Firth, J. R. (1937, 1964) *The tongues of men*. London: Oxford University Press.
- Fowler, C. A., & Dekle, D. J. (1991) Listening with eye and hand: cross-modal contributions to speech perception, *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 816–28.
- Fujisaki, H., & Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels, *IEEE Transactions on Audio and Electro-Acoustics*, **AU-16**, 73–7.
- Goldinger, S. D. (1998) Echoes of echoes? An episodic theory of lexical access, *Psychological Review*, **105**, 251–79.
- Grant, K. W., & Braida, L. D. (1991) Evaluating the articulation index for auditory–visual input, *Journal of the Acoustical Society of America*, **89**, 2952–60 [published erratum appears in *Journal of Acoustical Society of America*, **90**, 2202].
- Grant, K. W., & Seitz, P. F. (1998) Measures of auditory–visual integration in nonsense syllables and sentences, *Journal of the Acoustical Society of America*, **104**, 2438–50.
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998) Auditory–visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory–visual integration, *Journal of the Acoustical Society of America*, **103**, 2677–90.
- Green, K. P., & Gerdeman, A. (1995) Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels, *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1409–26.
- Green, K. P., & Kuhl, P. K. (1989) The role of visual information in the processing of place and manner features in speech perception, *Perception & Psychophysics*, **45**, 34–42.
- Green, K. P., & Kuhl, P. K. (1991) Integral processing of visual place and auditory voicing information during phonetic perception, *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 278–88.

- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991) Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect, *Perception & Psychophysics*, **50**, 524–36.
- Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997) The encoding of rate and talker information during phonetic perception, *Perception & Psychophysics*, **59**, 675–92.
- Gussenhoven, C., & Rietveld, T. (1998) On the speaker-dependence of the perceived prominence of F0 peaks, *Journal of Phonetics*, **26**, 371–80.
- Hintzman, D. L. (1986) "Schema abstraction" in a multiple-trace memory model, *Psychological Review*, **93**, 411–28.
- Holmes, J. N. (1986) Normalization and vowel perception. In *Invariance and variability in speech processes* (J. S. Perkell, & D. H. Klatt, editors), pp. 346–59. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnson, K. (1990a) The role of perceived speaker identity in F0 normalization of vowels, *Journal of the Acoustical Society of America*, **88**, 642–54.
- Johnson, K. (1990b) Contrast and normalization in vowel perception, *Journal of Phonetics*, **18**, 229–54.
- Johnson, K. (1997a) Speech perception without speaker normalization: an exemplar model. In *Talker variability in speech processing* (K. Johnson, & J. Mullennix, editors), pp. 145–66. New York: Academic Press.
- Johnson, K. (1997b) The auditory/perceptual basis for speech segmentation, *OSU Working Papers in Linguistics*, **50**, 101–13.
- Joos, M. (1948) Acoustic phonetics, *Language*, **24** (Suppl. 2), 1–136.
- Kuhl, P. K., & Meltzoff, A. N. (1984) The intermodal representation of speech in infants, *Infant Behavior and Development*, **7**, 361–81.
- Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991) Cross-modal speech perception in adults and infants using non-speech auditory stimuli, *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 829–40.
- Ladefoged, P., & Broadbent, D. E. (1957) Information conveyed by vowels, *Journal of the Acoustical Society*, **29**, 98–104.
- Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967) Perception of the speech code, *Psychological Review*, **74**, 431–61.
- Lindblom, B. (1990) Explaining phonetic variation: a sketch of H&H theory. In *Speech production and speech modeling* (W. J. Hardcastle, & A. Marchal, editors), pp. 403–40. Dordrecht: Kluwer Academic Publishers.
- Lisker, L., & Rossi, M. (1992) Auditory and visual cueing of the [ ± rounded] feature of vowels, *Language and Speech*, **35**, 391–417.
- Massaro, D. W., & Cohen, M. M. (1996) Perceiving speech from inverted faces. *Perception & Psychophysics*, **58**, 1047–1065.
- Massaro, D. W., & Cohen, M. M. (1999) Speech perception in perceivers with hearing loss: synergy of multiple modalities. *Journal of Speech, Language and Hearing Research*, **42**, 21–41.
- Massaro, D. W., & Friedman, D. (1990) Models of integration given multiple sources of information. *Psychological Review*, **97**, 225–52.
- MacDonald, J., & McGurk, H. (1978) Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253–7.
- McGurk, H., & MacDonald, J. (1976) Hearing lips and seeing voices, *Nature*, **264**, 746–8.
- Miller, J. D. (1989) Auditory-perceptual interpretation of the vowel, *Journal of the Acoustical Society of America*, **85**, 2114–34.
- Miller, R. L. (1953) Auditory tests with synthetic vowels, *Journal of the Acoustical Society of America*, **25**, 114–21.
- Mullennix, J. W., & Pisoni, D. B. (1990) Stimulus variability and processing dependencies in speech perception, *Perception & Psychophysics*, **47**, 379–90.
- Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996) Temporal constraints on the McGurk effect, *Perception & Psychophysics*, **58**, 351–62.
- Nearey, T. M. (1978) *Phonetic feature systems for vowels*. Bloomington, IN: IU Linguistics Club.
- Nearey, T. M. (1989) Static, dynamic, and relational properties in vowel perception, *Journal of the Acoustical Society of America*, **85**, 2088–113.
- Ni Chasaide, A., & Gobl, C. (1997) Voice source variation. In *The handbook of phonetic sciences* (W. J. Hardcastle, & J. Laver, editors), pp. 427–61. Oxford: Blackwell Publishers.
- Nordström, P.-E., & Lindblom, B. (1975) A normalization procedure for vowel formant data. *Proceedings of the 8th International Congress of Phonetic Sciences*, Leeds, England.
- Nosofsky, R. M. (1984) Choice, similarity, and the context theory of classification, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 104–14.
- Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship, *Journal of Experimental Psychology: General*, **115**, 39–57.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994) Speech perception as a talker-contingent process, *Psychological Science*, **5**, 42–5.



- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993) Episodic encoding of voice attributes and recognition memory for spoken words, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, 309–28.
- Perkell, J., & Klatt, D. (1986) *Invariance and variability of speech processes*. Hillsdale, NJ: Lawrence Erlbaum.
- Pisoni, D. B. (1993) Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning, *Speech Communication*, **13**, 109–25.
- Potter, R., & Steinberg, J. (1950) Toward the specification of speech, *Journal of the Acoustical Society of America*, **22**, 807–20.
- Qi, Y. (1990) Replacing tracheoesophageal voicing sources using LPC synthesis, *Journal of the Acoustical Society of America*, **88**, 1228–35.
- Reber, A. S. (1993) *Implicit learning and tacit knowledge: an essay on the cognitive unconscious*. Oxford: Oxford University Press.
- Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996) Point-light facial displays enhance comprehension of speech in noise, *Journal of Speech and Hearing Research*, **39**, 1159–70.
- Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997) The McGurk effect in infants, *Perception & Psychophysics*, **S9**, 347–57.
- Sams, M., Manninen, P., Surakka, V., Helin, P., & Katto, R. (1998) McGurk effect in Finnish syllables, isolated words, and words in sentences: effects of word meaning and sentence context, *Speech Communication*, **26**, 75–87.
- Schacter, D. L. (1987) Implicit memory: History and current status, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 501–18.
- Schacter, D. L., Eich, J., & Tulving, E. (1978) Richard Semon's theory of memory, *Journal of Verbal Learning and Verbal Behavior*, **17**, 721–43.
- Schwippert, C., & Benoit, C. (1997) Audiovisual intelligibility of an androgynous speaker. In *Proceedings of the ESCA workshop on audio-visual speech processing (AVSP'97) cognitive and computational approaches*, Rhodes, Greece (C. Benoit, & R. Campbell, editors), 26–27 September 1997, pp. 81–4.
- Sekiyama, K. (1997) Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects, *Perception & Psychophysics*, **59**, 73–80.
- Sekiyama, K. (1998) Face or voice? Determinant of compellingness to the McGurk effect, *Audio-Visual Speech Processing (AVSP'98)*. Terrigal.
- Sekiyama, K., & Tohkura, Y. (1993) Inter-language differences in the influence of visual cues in speech perception, *Journal of Phonetics*, **21**, 427–44.
- Semon, R. (1923) *Minemic psychology* (B. Duffy, Trans.). London: Allen & Unwin. (Original work published 1909).
- Sheffert, S. M., & Fowler, C. A. (1995) The effects of voice and visible speaker change on memory for spoken words, *Journal of Memory and Language*, **34**, 665–85.
- Slawson, A. W. (1968) Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency, *Journal of the Acoustical Society of America*, **43**, 87–101.
- Stevens, K. N., & Blumstein, S. E. (1978) Invariant cues for place of articulation in stop consonants, *Journal of the Acoustical Society of America*, **64**, 1358–68.
- Strand, E. A., & Johnson, K. (1996) Gradient and visual speaker normalization in the perception of fricatives. In *Natural language processing and speech technology: results of the 3rd KONVENS conference*, Bielefeld, (D. Gibbon, editor), October, pp. 14–26. Berlin: Mouton de Gruyter.
- Sumby, W. H., & Pollack, I. (1954) Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, **26**, 212–5.
- Summerfield, Q. (1979) Use of visual information for phonetic perception, *Phonetica*, **36**, 314–31.
- Summerfield, Q., & McGrath, M. (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **36A**, 51–74.
- Sussman, H. M. (1986) A neuronal model of vowel normalization and representation, *Brain and Language*, **28**, 12–23.
- Syrdal, A., & Gopal, H. (1984) A perceptual model of vowel recognition based on the auditory representation of American English vowels, *Journal of the Acoustical Society of America*, **79**, 1086–100.
- Tillberg, I., Roennberg, J., Svaerd, I., & Ahlner, B. (1996) Audio-visual speechreading in a group of hearing aid users: The effects of onset age, handicap age, and degree of hearing loss, *Scandinavian Audiology*, **25**, 267–272.
- Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, **69**, 1465–75.
- Walden, B. E., Busacco, D. A., & Montgomery, A. A. (1993) Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons, *Journal of Speech and Hearing Research*, **36**, 431–6.
- Walden, B. E., Montgomery, A. A., Prosek, R. A., & Hawkins, D. B. (1990) Visual biasing of normal and impaired auditory speech perception, *Journal of Speech and Hearing Research*, **33**, 163–73.

- Walker, S., Bruce, V., & O'Malley, C. (1995) Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, **S9**, 1124–33.
- Watkins, A. J. (1991) Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, **90**, 2942–55.
- Watkins, A. J., & Makin, S. J. (1994) Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, **96**, 1263–82.