Perception & Psychophysics
1991, 50 (6), 524-536

Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect

KERRY P. GREEN
University of Arizona, Tucson, Arizona

and

PATRICIA K. KUHL, ANDREW N. MELTZOFF, and ERICA B. STEVENS
University of Washington, Seattle, Washington

Perception & Psychophysics
1991, 50 (6), 524-536

Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect

KERRY P. GREEN
University of Arizona, Tucson, Arizona

and

PATRICIA K. KUHL, ANDREW N. MELTZOFF, and ERICA B. STEVENS
University of Washington, Seattle, Washington

Studies of the McGurk effect have shown that when discrepant phonetic information is delivered to the auditory and visual modalities, the information is combined into a new percept not originally presented to either modality. In typical experiments, the auditory and visual speech signals are generated by the same talker. The present experiment examined whether a discrepancy in the gender of the talker between the auditory and visual signals would influence the magnitude of the McGurk effect. A male talker's voice was dubbed onto a videotape containing a female talker's face, and vice versa. The gender-incongruent videotapes were compared with gender-congruent videotapes, in which a male talker's voice was dubbed onto a male face and a female talker's voice was dubbed onto a female face. Even though there was a clear incompatibility in talker characteristics between the auditory and visual signals on the incongruent videotapes, the resulting magnitude of the McGurk effect was not significantly different for the incongruent as opposed to the congruent videotapes. The results indicate that the mechanism for integrating speech information from the auditory and the visual modalities is not disrupted by a gender incompatibility even when it is perceptually apparent. The findings are compatible with the theoretical notion that information about voice characteristics of the talker is extracted and used to normalize the speech signal at an early stage of phonetic processing, prior to the integration of the auditory and the visual information.

Over the past four decades, extensive research has been done on the psychological processes underlying the perception and production of spoken language. Much of this research has focused on how the listener processes the acoustic structure of speech in order to arrive at the intended meaning of an utterance.

Although speech perception has primarily been considered an auditory process, recent studies have shown that visual information provided by movements of a talker's mouth and face strongly influences what an observer perceives (Green & Kuhl, 1989, 1991; Green & Miller, 1985; MacDonald & McGurk, 1978; Massaro & Cohen, 1983; McGurk & MacDonald, 1976; Reisberg, McLean, & Goldfield, 1987; Summerfield & McGrath, 1984). A particularly convincing demonstration of the effects of vision on speech perception is provided by the stimulus situation in which the separate auditory and visual inputs seem to fuse or blend into a new percept—one that has not been presented to either modality alone. For example, in the McGurk effect (McGurk & MacDonald, 1976), the auditory syllable /ba/ is presented in synchrony with a videotape of the talker pronouncing the syllable /ga/; the resulting syllable is perceived as /da/, a syllable that has not been presented to either modality and that represents a combination of both.

This fusion effect is somewhat surprising, because it was long assumed that visual information in the form of lipreading was effective only when the auditory signal was degraded (Sumby & Pollack, 1954). The McGurk effect demonstrated that visual information is potent even when the auditory signal is clear and unambiguous. From the standpoint of theory, we have to explain how such diverse acoustic and optic information—frequency transitions indicating /b/ in the auditory domain and mouth movements indicating /g/ in the visual domain—are combined to produce /d/ by the perceptual system. Although the phenomenon itself has been well-documented (Green & Kuhl, 1989, 1991; MacDonald & McGurk, 1978; Manuel, Repp,

This research was supported by National Institutes of Health Grant NS-26475 to Kerry P. Green and National Institutes of Health Grant HD-18286 to Patricia K. Kuhl. We would like to thank Virginia Mann and an anonymous reviewer for helpful comments on a previous version of the manuscript. A portion of these data were presented at the spring meeting of the Acoustical Society of America, State College, Pennsylvania, 1990. Correspondence concerning the article should be addressed to Kerry P. Green, Cognitive Science, Psychology Building, Room 312, University of Arizona, Tucson, AZ 85721.

Studdert-Kennedy, & Liberman, 1983; Massaro & Cohen, 1983; McGurk & MacDonald, 1976; Mills & Thiem, 1980; Roberts & Summerfield, 1981), the boundary conditions of the phenomenon and the particular circumstances that affect when the auditory and visual information are integrated remain to be charted (Summerfield, 1987).

It has been suggested (e.g., by Welch & Warren, 1980) that studying how perceptual systems deal with intermodal discrepancies will inform us about the intermodal organization that underlies normal perception. For example, Welch and Warren have proposed a model that describes factors affecting the integration of information in multimodal situations. One important assumption of their model is that of *unity*. The perceiver forms an assumption about whether he or she is observing a single or a multiple event. If the information from the two modalities is perceived as consistent, then a high-unity assumption is produced and the information is treated as belonging to a single event. Under these conditions, the information is combined, even though it is actually discrepant. Alternatively, if the information from the two modalities is perceived as discrepant, a low-unity assumption is produced and the observer treats the information from the two modalities as separate and belonging to two different events. Under these circumstances, the information is not combined.

The ventriloquism effect is an example of a multimodal situation in which the unity assumption holds. When there is a spatial discrepancy between the visual and auditory locations of a sound's source, observers typically hear the sound as emanating from the spatial location of the visual source. Studies of this phenomenon have shown that it is susceptible to a number of factors, including the congruence or *cognitive compellingness* of the auditory and visual information (Jack & Thurlow, 1973; Jackson, 1953; Warren, 1979; Warren, Welch, & McCarthy, 1981). Congruence, or cognitive compellingness, derives from factors such as the temporal congruence between the auditory and visual signals and the extent to which the two streams of information appear to go together.

Regarding temporal congruence, Jack and Thurlow (1973) demonstrated that when a puppet's mouth movements are temporally synchronized with ongoing speech sounds, there is a greater displacement in the perceived localization of the sound than there is when speech is not temporally coincident with the mouth movements. Thus, the temporal synchrony of the auditory and visual information produced a more compelling event, leading to greater displacement of the auditory sound source. Warren et al. (1981) performed a similar experiment by using a videotape of a talker reading a passage. Again, when the auditory signal from the videotape was temporally congruent with the video signal, observers perceived greater displacement of the location of the auditory source than they did when the auditory signal was temporally displaced.

Warren et al. (1981) also demonstrated that cognitive congruency had an influence on the ventriloquism effect. When the video signal of the talker's face was replaced by a dot on the video monitor, there was very little displacement in perceived localization of the sound source. Finally, Jackson (1953) found that the amount of displacement in perceived localization was increased when the characteristics of the auditory signal (a whistle sound) matched the characteristics of the visual signal (a steam kettle with a puff of steam coming out of it vs. a steam kettle with no steam). Taken together, these several studies demonstrate that a reduction in either the temporal or the cognitive congruency between the auditory and visual signals dramatically reduces the magnitude of the ventriloquism effect.

A question that arises is whether the McGurk speech effect, like the ventriloquism effect, is also influenced by the unity or congruency of the auditory and visual signals. The unity question has not been adequately addressed for the speech case. To date, studies of the impact of temporal asynchrony on the McGurk effect have been inconclusive. Cohen (1984) reported that temporal asynchronies of up to 200 msec have little influence on the magnitude of the McGurk effect. However, it is not clear how aware the subjects were of a discrepancy between the auditory and visual signals. Dixon and Spitz (1980) have shown that observers simply may not be able to detect onset asynchronies between auditory and visual speech information for temporal differences of less than 190 msec, and these are similar to the values used by Cohen (1984).[1] Moreover, other studies indicate that temporal asynchronies of 80–400 msec can disrupt the integration of auditory and visual speech information under some circumstances in the McGurk situation (McGrath & Summerfield,1985) and other situations as well (Dodd, 1977, 1979). Thus, it remains unclear from these studies whether, or how much, the McGurk effect is affected by a reduction in the temporal congruence between the auditory and visual signals.

No studies have directly examined whether changes in the cognitive congruency of the auditory and visual information alter the McGurk effect. The specific purpose of the present study was to manipulate the cognitive congruence between the auditory and visual signals. This was achieved by having perceivers view a novel combination of auditory and visual information—a gender discrepancy produced by combining a male talker's voice with the video of a female talker's face, and vice versa. In most previous experiments on the McGurk effect, the same talker has produced both the auditory and the visual signals.[2] If the McGurk effect is influenced by the cognitive congruency, and the perceptual "unity" of the signals, there ought to be a weaker McGurk effect in the gender-discrepancy condition, wherein the two streams cannot have been produced by the same person (cf. Welch, 1989). Such a finding would suggest that speech fusion effects are similar to other types of perceptual phenomena—that they are sensitive to intersensory discrepancies and are thus appropriately characterized by models such as that proposed by Welch and Warren (1980).

Research on the perception of speech suggests an alternative possibility, however. This work has been directed at how the perceptual system handles the large