Multisensory integration of speech signals: The relationship between space and time

Article *in* Experimental Brain Research · November 2006

DOI: 10.1007/s00221-006-0634-0 · Source: PubMed

2 authors:

Jeffery A Jones
Wilfrid Laurier University

Michelle Jarick
MacEwan University

Exp Brain Res (2006) 174:588–594
DOI 10.1007/s00221-006-0634-0

RESEARCH NOTE

Multisensory integration of speech signals: the relationship between space and time

Jeffery A. Jones · Michelle Jarick

Received: 12 April 2006 / Accepted: 13 July 2006 / Published online: 10 August 2006

Abstract Integrating audiovisual cues for simple events is affected when sources are separated in space and time. By contrast, audiovisual perception of speech appears resilient when either spatial or temporal disparities exist. We investigated whether speech perception is sensitive to the combination of spatial and temporal inconsistencies. Participants heard the bisyllable /aba/ while seeing a face produce the incongruent bisyllable /ava/. We tested the level of visual influence over auditory perception when the sound was asynchronous with respect to facial motion (from −360 to +360 ms) and emanated from five locations equidistant to the participant. Although an interaction was observed, it was not related to participants' perception of synchrony, nor did it indicate a linear relationship between the effect of spatial and temporal discrepancies. We conclude that either the complexity of the signal or the nature of the task reduces reliance on spatial and temporal contiguity for audiovisual speech perception.

Keywords McGurk effect · Speech · Audiovisual · Spatial · Temporal

J. A. Jones
Centre for Cognitive Neuroscience,
Wilfrid Laurier University, Waterloo,
ON, Canada N2L 3C5

J. A. Jones (✉) · M. Jarick
Department of Psychology, Wilfrid Laurier University,
Waterloo, ON, Canada N2L 3C5
e-mail: jjones@wlu.ca

Introduction

Research shows that the Gestalt laws of common fate and spatial proximity are important not only for processing of unimodal information, but also are fundamental to the integration of multisensory information (Welch and Warren 1980; Stein and Meredith 1993). Auditory–visual interactions have been well studied and research has shown that spatially and temporally coincident audiovisual signals are detected more easily (Vroomen and de Gelder 2000; Lovelace et al. 2003; Bolognini et al. 2005) and more quickly (Spence and Driver 1996; Driver and Spence 1998; McDonald et al. 2000) than unimodal stimuli. This behavioral evidence parallels physiological studies of animals showing that several brain structures contain neurons that respond to stimuli from more than one modality (Stein and Meredith 1993). These multisensory cells often have spatially corresponding receptive fields and respond more vigorously to visual and acoustic stimuli that occur in spatial and temporal proximity (King and Palmer 1985; Meredith et al. 1987; Stein and Meredith 1993; Meredith and Stein 1996).

For humans, face-to-face speech communication represents one of our richest multisensory experiences. Much of the research investigating the integration of nonspeech signals also applies to integrating speech signals. For example, listeners' perception of auditory speech is enhanced when they are able to see a speaker's face (Sumby and Pollack 1954; Schwartz et al. 2004). Listeners are also sensitive to temporal asynchronies, although they tolerate much larger temporal discrepancies than those tolerated with nonspeech signals (Dixon and Spitz 1980; Massaro et al. 1996; Munhall et al. 1996; van Wassenhove et al. 2006).

However, audiovisual integration of speech and non-speech differs in at least one important way: integrating speech signals occurs regardless of spatial discrepancies between the auditory and visual sources of the signal (Bertelson et al. 1994; Jones and Munhall 1997; Colin et al. 2001). Why the perceptual system disregards information related to the location of the signal sources for speech remains an important, but unanswered question.

Previous research addressing the role of spatial and temporal factors for speech perception has only manipulated these parameters independently (Bertelson et al. 1994; Massaro et al. 1996; Munhall et al. 1996; Jones and Munhall 1997; Colin et al. 2001). However, many recent investigations examining much ‘simpler’ stimuli have begun to focus predominantly on the interaction between space and time (Slutsky and Recanzone 2001; Lewald and Guski 2003, 2004; Wallace et al. 2004). The interaction has typically been studied using the “ventriloquism effect”. The “ventriloquism effect” is an illusion that arises when visual and auditory signals, such as a tone and a light, simultaneously occur from slightly separated locations; there is a tendency to underestimate the separation distance between the stimuli and perceive them as originating from a single source (for a detailed description see Stein and Meredith 1993). It makes sense to believe that the integration of the audiovisual stimuli would linearly decline with increased spatial disparity. However, more recent reports suggest that integration may occur within a fairly wide spatial and temporal window (Lewald et al. 2001; Slutsky and Recanzone 2001; Lewald and Guski 2003, 2004) and that the brain modulates the temporal window based on the stimuli’s perceived distance from the observer (Sugita and Suzuki 2003; cf. Lewald and Guski 2004). However, until now the spatiotemporal relationship involving speech signals still needed to be explored.

One possible explanation for the perceptual system’s apparent disregard of information about the relative location of sources for speech signals is that the brain relies on the correlation between the dynamic characteristics of the signals. A great deal of research shows that phonetic information is distributed across the face during speech production (Vatikiotis-Bateson et al. 2000) and that this motion along with gross head motion is highly correlated with the acoustics and used by listeners (Munhall et al. 2003). Thus, the perceptual system might process spatiotemporal information regarding the source of the signals but bind these signals based on their dynamic relationship unless spatiotemporal separations exceed some threshold. In this study, we sought to determine this threshold by parametrically investigating the interaction between spatial and temporal disparities of the auditory and visual signals during speech perception. To that end, we presented participants with a video of a speaker mouthing the nonsense syllable /ava/ paired with audio of the speaker producing the nonsense syllable /aba/. This type of incongruent audiovisual pairing typically alters the perception of the auditory speech signal generating an illusion commonly known as the “McGurk effect” (McGurk and MacDonald 1976). The McGurk effect has been widely investigated (Green et al. 1991; Walker et al. 1995; Rosenblum et al. 1997; Sekiyama 1997; Hampson et al. 2003; Windmann 2004) and has served as a valuable tool for examining the integration of audiovisual speech signals. In line with previous research (Bertelson et al. 1994; Massaro et al. 1996; Munhall et al. 1996; Jones and Munhall 1997), we predicted that the temporal asynchronies would affect the strength of audiovisual integration as indexed by the McGurk effect, but would be unaffected by the spatial disparities. However, we hypothesized that an interaction between spatial and temporal coincidence would exist such that temporal asynchronies would be less tolerated with increased spatial separations.

Experiment 1

Method

Participants

Twenty-four Wilfrid Laurier University undergraduates (14 women; mean age = 22.5 years) received an honorarium for their participation. All were native speakers of North American English, had normal or corrected-to-normal vision, and reported no history of language or hearing impairments or disorders. All were right-handed (mean score of 32.8 on the Dutch Handedness Questionnaire; Van Strien 1988). Participants gave their informed consent for participation and the Wilfrid Laurier Research Ethics Board approved all the experimental procedures.

Materials

The audiovisual stimuli consisted of a video of a man articulating the bisyllable /ava/ and the sound of his voice saying /aba/. The signals were synchronized by aligning the acoustic burst onsets of the visual consonant /v/ with that of the auditory /b/ stimuli. According to a previous work, maximal integration for speech occurs when the audio track is delayed by 60 ms with respect to the video and remains unaffected by delays of up to 180 ms (Munhall et al. 1996). In the present study, temporal delays ranged from −360 ms (auditory leading the visual) to +360 ms (visual leading the auditory) in 60 ms increments, resulting in 13 delay conditions.

Videos were recorded on to DVD and presented on a monitor situated with the center of the video at eye level. The sound was presented at an average of 60 dB (SPL) from one of the five loudspeakers in a semi-circular array that was hidden by a black curtain. The loudspeakers were 85 cm above the ground (approximately ear level), and 85 cm from the participant, situated at 0, 45, 90, 135, and 180° in azimuth. Participants responded by pressing keys labeled "B" and "V" on a keyboard.

Procedure

Participants sat in a dimly lit sound booth in front of a video monitor. They placed their chin on a chin-rest to ensure equal distances between their head and the sound sources. Each delay condition was randomly presented three times from each of the five loudspeaker locations, resulting in 195 trials (13 delays × 5 locations × 3 repetitions). Participants were instructed to attend to the visual and auditory stimuli and to indicate whether the consonant they heard was either /b/ or /v/ by pressing the appropriate key on the keyboard. The left–right position of the keys was counterbalanced across the participants. Sessions lasted approximately 30 min.

Results and discussion

A 5 × 13 (speaker location by temporal delay) repeated measures analysis of variance (ANOVA) was used to analyze the data. The dependent measure was the percentage of correct auditory responses (i.e., /b/ responses). Lower proportions of /b/'s indicated a greater visual influence on auditory perception.

A main effect of delay was observed, F (12, 276) = 113.64, $P < 0.0001$. As seen in Fig. 1a, the results for temporal delay produced an asymmetrical U-shaped pattern, where the positive asynchronies (i.e., visual first) elicited more McGurk effects than the negative asynchronies (i.e., auditory first). Particularly robust McGurk effects occurred for delays between −60 and +180 ms. This asymmetrical pattern is consistent with the other studies that investigated speech (e.g., Dixon and Spitz 1980; Munhall et al. 1996; van Wassenhove et al. 2006) and nonspeech signals (e.g., Slutsky and Recanzone 2001; Stone et al. 2001). Moreover, several

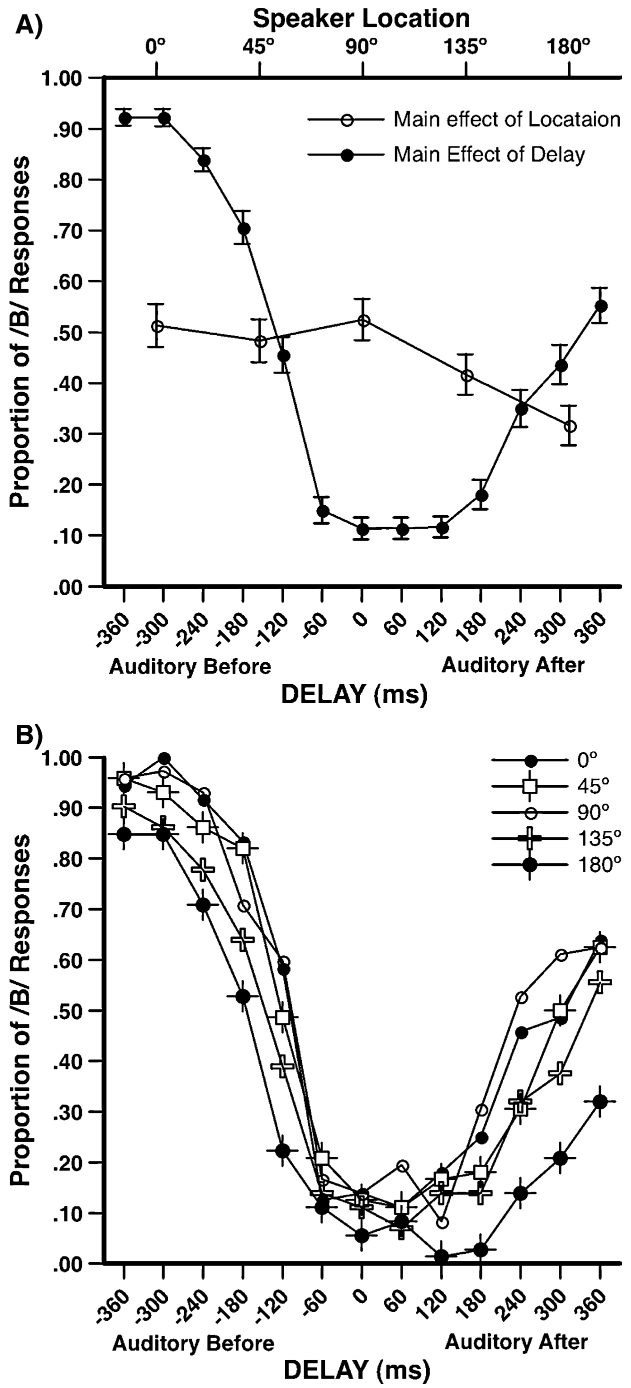

Fig. 1 **a** Mean proportion of /b/ responses as a function of temporal disparity of the auditory stimulus (*closed circles*) and of spatial position of the loudspeakers (*open circles*); fewer /b/ responses suggest stronger McGurk effects. The *error bars* correspond to the standard errors of the means. **b** Mean proportions of /b/ responses across the different time delays and spatial locations, showing the significant two-way interaction

other studies have also shown that the McGurk effect is resistant to delays of up to 240 ms (Massaro 1987; Summerfield 1992; Massaro et al. 1993, 1996; Munhall et al. 1996).

Jones and Munhall (1997) used the largest spatial discrepancies to date and demonstrated that integration persists for locations up to 90°, bilaterally. The current study extended those findings by including spatial discrepancies up to 180° (from directly in front, to directly behind the participant). Contrary to our expectations, we found a reliable main effect of loudspeaker location, $F\ (4, 92) = 22.89$, $P < 0.0001$. Figure 1a shows that there were fewer /b/'s reported when the sound came from the loudspeakers at 135 and 180° than when the sound originated from the loudspeakers located at 0, 45, and 90°. However, consistent with Jones and Munhall (1997), there was no significant difference found in the strength of the McGurk effect for positions up to 90° (LSD, $P > 0.1$).

The increase in the level of visual influence on auditory perception for positions behind the participant was unexpected. Humans tend to be poorer at localizing sound sources that are behind them (Oldfield and Parker 1984) and because of the identical interaural time differences for the front and back positions, it is not uncommon for listeners to localize these sounds as coming from positions in front (Giguere and Abel 1993; Macpherson and Middlebrooks 2000). If participants in our experiment made front–back confusions, this fact, plus relative amplitudes differences due to pinnae effects, could have led to more McGurk effects for positions behind them. This explanation is speculative and can only be explored in future studies where localization responses are gathered. Nevertheless, it is clear that increased spatial discrepancies do not lead to reduced audiovisual integration of speech signals.

Besides the main effects, the results also showed a significant interaction between the spatial and temporal disparities, $F\ (48, 1104) = 2.04$, $P < 0.0001$. Although we did predict an interaction, the linear relationship between the spatial and temporal disparities that we expected did not exist. We envisioned a pattern demonstrating a systematic relationship between the two variables, such that, as the spatial and temporal disparities increased, the strength of the McGurk effect would weaken. But as can be seen in Fig. 1b, no such pattern emerged. The interaction manifested at delays below −60 ms and above +60 ms and is not representative of a linear relationship between space and time.

Experiment 2

The results of our first experiment clearly demonstrated that increasing the spatial separation between the auditory and visual stimuli does not degrade audiovisual integration of speech signals. And although there was statistical interaction between space and time, linear increases in the spatial distance between the auditory and visual source of the signals did not lead to a predictable pattern of effects on delay. The purpose of Experiment 2 was to determine if the pattern of responses in Experiment 1 could be explained by participants' detection of temporal asynchrony in our stimuli.

Method

Participants

Eighteen Wilfrid Laurier undergraduates (12 women; mean = 21.7 years) received an honorarium for their participation. All were native speakers of North American English, had normal or corrected-to-normal vision, and reported no history of language or hearing impairments or disorders. Thirteen of these participants took part in the first experiment.

Materials

The stimuli were identical with those in Experiment 1. The only difference was that the keyboard was labeled with "Y" and "N" instead of "B" and "V".

Procedure

The procedure was also similar to Experiment 1. The important difference was that participants made synchrony judgments, as opposed to identifying the speech sound. They pressed the key labeled "Y" if the movements of the face appeared to be correctly aligned with the sound of the voice (synchronous) and "N" if they were not (asynchronous). The left–right key order was again counterbalanced across participants.

Results and discussion

The analysis was identical to Experiment 1 (5×13 ANOVA); however, the dependent measure for this experiment was the percentage of asynchronous judgments. Unlike the results of Experiment 1, the interaction between space and time for this experiment was not significant, $F\ (48, 816) = 1.37$, $P = 0.052$ (see Fig. 2b). However, the results did reveal a significant effect of delay, $F\ (4, 68) = 6.78$, $P < 0.0001$, with more synchrony judgments elicited for the delays ranging between −60 and 180 ms. As seen in Fig. 2a, an inverted U-shaped pattern similar to that of Experiment 1 occurred across the delays; a bias existed such that more synchrony judgments were made when the face led the voice, than vice versa. This finding is similar to previous research demonstrating that for participants to report a misalignment of a voice with the visible lip movements, the visual stimulus must lead the auditory stimulus by 140 ms more (McGrath and Summerfield 1985). Note that in our experiment, asynchrony judgments never reached 100%, even for the extreme temporal delays of −360 and +360 ms.

We also found a reliable main effect of spatial position, F (4, 68) = 6.78, $P < 0.0001$. Figure 2a illustrates the trend across the loudspeakers. The loudspeakers positioned at 0° (directly in front) and 180° (directly behind) elicited the strongest synchrony judgments. The number of asynchronous judgments increased when the sound emanated from loudspeakers at 45 and 135°, and the smallest number of synchrony judgments was made for the loudspeaker located at 90° (directly beside the participant). To our knowledge, this is the first experiment to examine the perception of simultaneity for speech signals with such extreme spatial locations. Our results suggest that participants tended to perceive an audiovisual stimulus as synchronous when the auditory source was directly in line with the visual source. This finding lends some credibility to the suggestion that front–back confusions might account for why the strength of the McGurk effect in Experiment 1 remained unaffected by the loudspeakers located behind the participants.

Conclusions

Our data support previous conclusions that audiovisual speech perception is impervious to spatial discordances and accommodates large temporal asynchronies. In addition, integration is not reduced by the combination of spatial and temporal disparities. We see several possible explanations for these results. One possibility is that there is specialized processing of speech signals. According to The Motor Theory (Liberman and Mattingly 1985), the perception of speech occurs through the encoding of intended speech gestures used for speech production (e.g., the movements of the articulators and shape of the vocal tract). Therefore, strict spatial and temporal synchrony of the signals might not be necessary for the extraction of the speech gestures. Indeed recent studies have supported a perception–production linkage for speech, showing that motor areas are active during the observation of speech stimuli (Skipper et al. 2005). However, another possibility is that the differential reliance on spatial location observed for speech and nonspeech integration is an artifact of the stimuli used in typical experiments. Audiovisual speech signals are complex and dynamic,

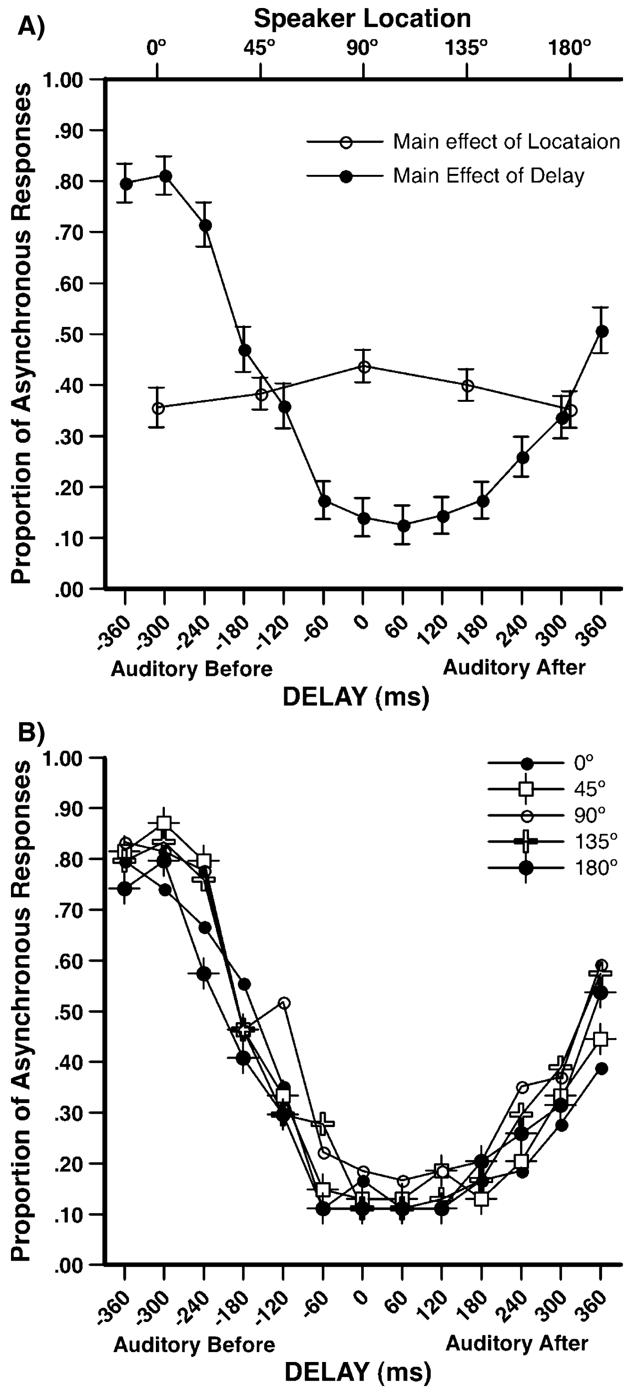

Fig. 2 Experiment 2: **a** Mean proportion of asynchronous judgments as a function of temporal disparity (*closed circles*) and spatial position of the loudspeakers (*open circles*). The *error bars* correspond to the standard errors of the means. **b** Mean proportion of asynchronous responses as across the time delays and spatial positions. As can be seen, asynchronous judgments never reached 100%

and the perceptual system may rely exclusively on the correlation between the time-varying characteristics of the signals (e.g., the rate of change of vocal tract shape observable in the auditory and visual information; see

Summerfield 1987) as the basis for integration rather than the congruency between spatial location and the temporal coincidence of onset and offset times. Thus, similar results may be obtained when equally complex and dynamic nonspeech stimuli are tested.

Furthermore, the relative unimportance of spatio-temporal contiguity may in fact be task related. In a recent PET study, Macaluso et al. (2004) presented a video of a speaker with acoustics that were either in synchrony or 240 ms early and were emitted either from the same location as the visual stimulus or 16.6° to the left. The level of brain activity in the dorsolateral occipital cortex was increased when the audiovisual stimuli were both synchronous and spatially coincident. By contrast, activity in ventral occipital areas and the left superior temporal sulcus increased when the audiovisual stimuli were synchronous, independent of the location of the sound. This pattern of activity suggests that discrete brain regions are involved in identification and localization tasks involving speech. In our study, participants performed identification and simultaneity judgments, and these types of tasks may not have invoked the neural network involved in processing location. It is clear that further testing of more complex, nonspeech signals using various tasks are needed to address these questions.

Acknowledgments This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Bertelson P, Vroomen J, Wiegeraad G, de Gelder B (1994) Exploring the relation between McGurk interference and ventriloquism. In: Proceedings of the international congress on spoken language processing, Yokohama, pp 559–562

Bolognini N, Frassinetti F, Serino A, Ladavas E (2005) "Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs. Exp Brain Res 160:273–282

Colin C, Radeau M, Deltenre P, Morais J (2001) Rules of intersensory integration in spatial scene analysis and speechreading. Psychol Belg 43:131–144

Dixon NF, Spitz L (1980) The detection of audiovisual desynchrony. Percept Psychophys 9:719–721

Driver J, Spence C (1998) Crossmodal attention. Curr Opin Neurobiol 8:245–253

Giguere C, Abel SM (1993) Sound localization: effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. J Acoust Soc Am 94:769–776

Green KP, Kuhl PK, Meltzoff AN, Stevens EB (1991) Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. Percept Psychophys 50:524–536

Hampson MG, FH, Cohen MA, and Nieto-Castanon A (2003) Changes in the McGurk Effect across phonetic contexts. Boston University, Technical Report CAS, Boston

Jones JA, Munhall KG (1997) The effects of separating auditory and visual sources on audiovisual integration of speech. Can Acoust 25:13–19

King AJ, Palmer AR (1985) Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Exp Brain Res 60:492–500

Lewald J, Ehrenstein WH, Guski R (2001) Spatio-temporal constraints for auditory–visual integration. Behav Brain Res 121:69–79

Lewald J, Guski R (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Brain Res Cogn Brain Res 16:468–478

Lewald J, Guski R (2004) Auditory–visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. Neurosci Lett 357:119–122

Liberman AM, Mattingly I (1985) The motor theory revised. Cognition 21:1–36

Lovelace CT, Stein BE, Wallace MT (2003) An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. Brain Res Cogn Brain Res 17:447–453

Macaluso E, George N, Dolan R, Spence C, Driver J (2004) Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage 21:725–732

Macpherson EA, Middlebrooks JC (2000) Localization of brief sounds: effects of level and background noise. J Acoust Soc Am 108:1834–1849

Massaro DW (1987) Psychophysics versus specialized processes in speech perception: an alternative perspective. In: Schouten MEH (ed) Psychophysics of speech perception, Dordrecht, Netherlands, pp 46–65

Massaro DW, Cohen MM, Gesi A, Heredia R, Tsuzaki M (1993) Bimodal speech perception: an examination across languages. J Phon 21:445–478

Massaro DW, Cohen MM, Smeele PM (1996) Perception of asynchronous and conflicting visual and auditory speech. J Acoust Soc Am 100:1777–1786

McDonald JJ, Teder-Salejarvi WA, Hillyard SA (2000) Involuntary orienting to sound improves visual perception. Nature 407:906–908

McGrath M, Summerfield Q (1985) Intermodal timing relations and audio–visual speech recognition by normal-hearing adults. J Acoust Soc Am 77:678–685

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J Neurosci 7:3215–3229

Meredith MA, Stein BE (1996) Spatial determinants of multisensory integration in cat superior colliculus neurons. J Neurophysiol 75:1843–1857

Munhall KG, Gribble P, Sacco L, Ward M (1996) Temporal constraints on the McGurk effect. Percept Psychophys 58:351–362

Munhall KG, Jones JA, Callan D, Kuratate T, Vatikiotis-Bateson E (2003) Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol Sci 15:133–137

Oldfield SR, Parker SP (1984) Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. Percept 13:581–600

Rosenblum LD, Schmuckler MA, Johnson JA (1997) The McGurk effect in infants. Percept Psychophys 59:347–357

Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio–visual interactions in speech identification. Cognition 93:B69–B78

Sekiyama K (1997) Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. Percept Psychophys 59:73–80

Slutsky DA, Recanzone GH (2001) Temporal and spatial dependency of the ventriloquism effect. Neuroreport 12:7–10

Spence C, Driver J (1996) Audiovisual links in endogenous covert spatial attention. J Exp Psychol Hum Percept Perform 22:1005–1030

Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech production. Neuroimage 25:76–89

Stein BE, Meredith MA (1993) Merging of the senses. The MIT Press, Cambridge

Stone JV, Hunkin NM, Porrill J, Wood R, Keeler V, Beanland M, Port M, Porter NR (2001) When is now? Perception of simultaneity. Proc Biol Sci USA 268:31–38

Sugita Y, Suzuki Y (2003) Audiovisual perception: implicit estimation of sound-arrival time. Nature 421:911

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215

Summerfield Q (1987) Some preliminaries to a comprehensive account of audiovisual speech perception. In Dodd B, Campbell R (eds) Hearing by eye: the psychology of lip-reading, Erlbaum, Hillsdale, pp 3–51

Summerfield Q (1992) Lipreading and audio–visual speech perception. Philos Trans R Soc Lond B Biol Sci 335:71–78

Van Strien JW (1988) Handedness and hemispheric laterality. Dissertation, Vrije Universiteit, Amsterdam

van Wassenhove V, Grant KW, Poeppel D (2006) Temporal window of integration in auditory–visual speech perception. Neuropsychologia, in press

Vatikiotis-Bateson, E, Kuratate, T, Munhall, KG, Yehia, HC (2000). The production and perception of a realistic talking face. In: Fujimura O, Joseph BD, Palek B (eds) Proceedings of LP'98, Item order in language and speech 2, Prague, pp 439–460

Vroomen J, de Gelder B (2000) Sound enhances visual perception: cross-modal effects of auditory organization on vision. J Exp Psychol Hum Percept Perform 26:1583–1590

Walker S, Bruce V, O'Malley C (1995) Facial identity and facial speech processing: familiar faces and voices in the McGurk effect. Percept Psychophys 57:1124–1133

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252–258

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638–667

Windmann S (2004) Effects of sentence context and expectation on the McGurk illusion. J Mem Lang 50:212–230