ELSEVIER

Neuropsychologia 45 (2007) 598–607

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

Temporal window of integration in auditory-visual speech perception

Virginie van Wassenhove [a,b,*], Ken W. Grant [c], David Poeppel [d]

[a] *Department of Psychology, University of California at Los Angeles, CA 90095, USA*
[b] *Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA*
[c] *Audio-Visual Speech Recognition Lab, Walter Reed Army Medical Center, Army Audiology and Speech Center, Washington, DC 20307, USA*
[d] *Department of Biology and Department of Linguistics, University of Maryland, College Park 20742, USA*

Received 23 August 2005; received in revised form 28 November 2005; accepted 12 January 2006
Available online 10 March 2006

Abstract

Forty-three normal hearing participants were tested in two experiments, which focused on temporal coincidence in auditory visual (AV) speech perception. In these experiments, audio recordings of/pa/and/ba/were dubbed onto video recordings of /ba/or/ga/, respectively (A_pV_k, A_bV_g), to produce the illusory "fusion" percepts /ta/, or /da/ [McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–747]. In Experiment 1, an identification task using McGurk pairs with asynchronies ranging from −467 ms (auditory lead) to +467 ms was conducted. Fusion responses were prevalent over temporal asynchronies from −30 ms to +170 ms and more robust for audio lags. In Experiment 2, simultaneity judgments for incongruent and congruent audiovisual tokens (A_dV_d, A_tV_t) were collected. McGurk pairs were more readily judged as asynchronous than congruent pairs. Characteristics of the temporal window over which simultaneity and fusion responses were maximal were quite similar, suggesting the existence of a 200 ms duration asymmetric bimodal temporal integration window.

Keywords: McGurk illusion; Multisensory; Time; Psychophysics; Analysis-by-synthesis

In natural conversational settings, visual speech information typically supports the perception of auditory speech. For instance, a congruent facial display (articulating the audio speech) provides critical cues in noisy environments (Helfer, 1997; MacLeod & Summerfield, 1990; Sumby & Pollack, 1954) and benefits hearing-impaired listeners (Grant, Walden, & Seitz, 1998). Additionally, visual speech information can also alter the expected perceptual interpretation of clear audio signals. The 'McGurk effect' demonstrates that adding conflicting (incongruent) visual information to an audio signal alters the auditory percept. The presentation of an audio /p/ (bilabial) with a synchronized incongruent visual /k/ (velar) often leads listeners to identify what they hear as /t/ (alveolar), a phenomenon referred to as 'fusion' (McGurk & McDonald, 1976).

The nature of the information provided by each sensory modality and the conditions under which multisensory integration occurs in speech remain unclear. In the McGurk fusion, the nature and the informational content of auditory-visual (AV) speech are fundamentally different, yet sensory inputs converge on a unique percept clearly differing from the initial unimodal percepts. This observation is the basis for the conservative definition of AV speech integration used here, where a unitary integrated percept emerges as the result of the integration of clearly differing auditory and visual informational content at some stage of the speech-processing pathway. Hypothetically then, the McGurk illusion (audio /p/ with a visual /k/) permits one to quantify the degree of integration that has taken place by evaluating the rate of illusory /t/ responses. Hence, a /pa/ or a /ka/ answer corresponds to no integration, and a /ta/ answer to integration. In the present experiments we take advantage of the McGurk fusion to explore the temporal boundaries of AV integration in speech.

Two factors that can vary in AV integration are the spatial and temporal relationships of the stimuli. When considering spatial information, one finds that the McGurk effect remains robust under AV spatial disparities (Jones & Munhall, 1997) as well as under spectral manipulations such as filtering of facial information (Campbell & Massaro, 1997; MacDonald, Andersen, &

* Corresponding author at: University of California, Los Angeles, Department of Psychology, VMPL, 1285 Franz Hall, Los Angeles, CA 90096-1563, USA. Tel.: +1 310 267 2141.
E-mail address: vvw@caltech.edu (V. van Wassenhove).

0028-3932/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2006.01.001

Bachmann, 2000) or a points-of-light facial display (Rosenblum & Saldaña, 1996). On balance the findings suggest that visual speech processing entails distributed, dynamic facial kinematics that are effective for AV integration under various conditions. The contribution of facial kinematics in AV speech finds further support in a study reported by Jordan, McCotter, and Thomas (2000) indicating that luminance distribution may enhance information drawn from facial articulatory movements.

Taken together, these studies suggest that the richness of visual input can be reduced to canonical facial information without significantly impacting AV integration. The main temporal cues appear to be preserved through visual dynamic patterns, as suggested by Summerfield (1987). One of the most observable patterns involves the lip movements, which correlate with the corresponding overall amplitude of the acoustic signal (e.g. Grant & Greenberg, 2001; Rosen, 1992). If such AV correspondence is essential to the speech integration process, tolerance to AV asynchrony may be governed by one's ability to correlate acoustic amplitude fluctuations of the speech signal and facial kinematics, especially in the 3–4 Hz range (i.e., periods of 250–350 ms).

When considering temporal relationships, both for synthetic (i.e. talking heads) and natural speech inputs, an interesting profile has emerged when the temporal coincidence of AV events is manipulated: AV integration of speech does not seem to require precise temporal alignment (Conrey & Pisoni, 2003; Dixon & Spitz, 1980; Massaro, Cohen, & Smeele, 1996; McGrath & Summerfield, 1985; Pandey, Kunov, & Abel, 1986), and accurate AV speech recognition is maintained over a large temporal window ranging from approximately −40 ms audio lead to 240 ms audio lag. This range of asynchrony tolerance in AV speech suggests that such AV correlation may be a possible attribute driving the AV speech integration process. Thus, if AV correlation is decreased by using incongruent AV stimuli (such as in the McGurk pairs), a decrease in 'AV asynchrony tolerance' may be expected and directly reflected in a simultaneity judgment profile, for instance with a reduction of the temporal window width. It is noteworthy that while AV speech recognition remains robust for ∼300 ms, the detection of asynchrony between simpler auditory (tone) and visual (flash) stimuli is on the order of 25 ms to 50 ms (e.g. Hirsh & Sherrick, 1961; Zampini, Guest, Shore, & Spence, in press; Zampini, Shore, & Spence, 2003). Perception of AV simultaneity for complex spectro-temporal signals such as AV speech may involve auditory-visual correlation over a larger temporal window than would be expected on the basis of simpler AV stimuli, for which the comparison of onsets may suffice. Furthermore, AV correlation bears interesting resemblances with the spatio-temporal coincidence rule described for multisensory neurons (Stein & Meredith, 1993). While such coherent multisensory inputs facilitate the binding of multimodal information into a unitary sensory event, additional cortical mechanisms have been proposed by Poeppel (2003) that may account for (cortically based) longer temporal integration constants underlying perceptual unit formation.

The specific goal of this study builds on results by Munhall, Gribble, Sacco, and Ward (1996). In the first of the reported set of experiments, Munhall et al. looked at the effect of asynchrony and vowel context in the McGurk illusion. Their stimuli consisted of audio utterances /aba/ and /ibi/ dubbed onto a video of a face articulating /aga/. The range of asynchronies tested spanned from −360 ms auditory lead to +360 ms auditory lag in steps of 60 ms. Two main results reported by Munhall et al. are of particular interest for the present study. First, the response distributions obtained over all ranges of asynchronies show that the fusion rate (/d/ percepts) remained close to or below 10%.

Auditorily-driven responses dominated for most asynchronies, and visually driven responses (/g/) were prominent from −60 ms audio lead to +240 ms audio lag. These results raise a crucial question regarding the effective occurrence of AV integration in this experiment. A conservative definition of AV speech integration entails that a unitary integrated percept emerges as the result of the 'combination' of auditory and visual information at some stage of the speech-processing pathway. When the percept /g/ (visually-driven response) dominates near AV synchrony, it remains unclear whether a case of visual dominance or a manifest integrated percept is being instantiated. Although we recognize that most AV speech studies have considered any deviation between the percept in audio alone and in bimodal condition to be a case of AV integration, we here consider that the rate of unimodal error-particularly the error in the visual alone condition which may be identical to the fusion percept (such as a /d/ response to a visual alone /g/ and a combined audio /b/ and visual /g/)-needs to be taken into consideration to enable one to distinguish between AV integration and unimodal error. Hence, the reported asynchrony function may be an example in which visual information dominates the auditory percept. If this was the case, it is the temporal resolution of visual speech processing that is being shown, not the influence of asynchronies on AV integration per se.

Secondly, Munhall et al. reported a V-shaped function for auditory driven responses (/b/) with a minimum around 60 ms, suggesting that synchronous auditory and visual stimuli may not be optimal for AV integration. Pair-wise comparisons of the proportion of /b/ responses across the different temporal conditions revealed that the responses at synchrony were significantly different from those at −60 ms (auditory lead) and at 240 ms (visual lead). However, because temporal asynchronies were only tested in steps of 60 ms, it is unclear whether temporal misalignments between 0 ms and −60 ms or between 180 ms and 240 ms would also significantly impact AV integration.

A similar, yet less marked distribution was described by Massaro, Cohen, and Smeele (1996) who used synthetic and natural speech tokens dubbed onto congruent and incongruent synthetic animated facial displays. In their study, two groups of participants were tested on their ability to identify AV speech tokens in a factorial design including syllables /ba/, /da/, /ða/ and /va/. As pointed out previously, it is important to note that no AV combinations of the chosen stimuli in the Massaro et al. study induced a clear McGurk illusion, where incongruent auditory and visual inputs result in a unitary percept distinct from either unimodal percepts. The first group was tested with AV

asynchronies of 67, 167 and 267 ms, and the second group with AV asynchronies of 133, 267and 500 ms.

Congruent AV identification was overall less affected by asynchronies than the incongruent pairs. Although no clear boundary for AV integration could be drawn from the psychophysical measurements, a fit of the Fuzzy Logical Model of Perception (FLMP) to asynchronous AV tokens indicated that a slight decrease of the model performance occurred at 267 ms, while a significant breakdown occurred at 500 ms. From this model fit, Massaro et al. proposed that the boundary of AV integration occurs at about 250 ms of AV asynchrony.

While suggesting possible temporal limitations of AV integration, both studies used a fairly coarse temporal granularity to investigate the effects of asynchronous AV input on speech recognition, and therefore were not able to establish clear boundaries for the influence of visual information on auditory speech processing. As we will point out in the discussion, a fine-grained profiling of the effects of asynchronies on AV integration is necessary to connect perceptual effects with possible neurophysiological mechanisms that underlie this multisensory integrative process.

Motivated by these considerations, two experiments were conducted which explored the tolerance of the McGurk effect to a broad range of AV temporal disparities. The first experiment investigated the effect of AV asynchrony on the identification of incongruent (McGurk) AV speech stimuli. The second experiment focused on the subjective simultaneity judgments for congruent and incongruent AV speech stimuli tested in the same asynchrony conditions. The present studies extend the results of Munhall, Gribble, Sacco, and Ward (1996) by using smaller temporal step sizes, increasing the range of tested asynchronies, and determining the boundaries for subjective audiovisual simultaneity. Specifically, we addressed the following questions: (i) For which onset asynchronies (SOAs) was the fusion response dominant over the auditory or visual driven response? (ii) Is the temporal window for subjective audiovisual simultaneity (explicit temporal percept) equivalent to the temporal window for perceptual fusion (implicit speech percept), and (iii) is this the same for congruent (matched audio and visual stimuli) and incongruent (mismatched auditory and visual stimuli) speech input?.

1. Materials and methods

1.1. Participants

Participants (native speakers of American English) were recruited from the University of Maryland undergraduate population and provided informed consent. Two groups of participants took part in this study. The first group included 21 participants (11 females, average 21 years) who were run in the voiced A_bV_g condition (A_bV_g: audio /b/ and video /g/). The second group consisted of 22 participants (8 females, average 22.5 years) who were run in the voiceless A_pV_k condition (A_pV_k: audio /p/ and video /k/). No participant had diagnosed hearing problems and all had normal or corrected-to-normal vision. The study was carried out with the approval of the University of Maryland Institutional Review Board.

1.2. Stimuli

1.2.1. Video and audio processing

Movies drawn from a set of stimuli used in Grant et al. (1998) were digitized from an analog tape of a female speaker's face and voice. An iMovie file was created unchanged from the original with an iMac computer (Apple Computer, CA). The iMovie was then segmented into each token (A_bV_b, A_dV_d ...) and compressed in a Cinepak format. Each stimulus was rendered into a 640 × 480 pixels movie with a digitization rate of 29.97 frames/s (1 frame = 33.33 ms). The soundtracks were edited using Sound Edit (Macromedia, Inc.). Each soundtrack was modified to produce a fade-in and fade-out effect over the first and last 10 ms. Stereo soundtracks were digitized at 44.1 kHz, with 16-bit amplitude resolution.

1.2.2. Generation of McGurk pairs

Audio /ba/ and /pa/ were extracted from natural utterances produced by the same female speaker and then dubbed onto video /g/ and /k/, respectively, to form the McGurk pairs. Both voiced and voiceless McGurk pairs were tested, in order to insure generalizability. For each McGurk pair, the consonantal burst of the digitized audio file (e.g. /b/) was aligned with the consonantal burst of the underlying audio portion of the video file (e.g. /g/) to within ±5 ms.

1.2.3. Audiovisual alignment in asynchrony conditions

Audio-visual asynchronies were created by displacing the audio file in 33.33 ms increments (frame unit) with respect to the movie file. This process resulted in the creation of stimuli ranging from (+) 467 ms of auditory lag to (−) 467 ms of auditory lead. Thus, a total of 29 stimulus conditions (28 asynchrony conditions and 1 synchrony condition) were used in the study.

1.3. Procedure

Both identification (Experiment 1) and simultaneity judgment (Experiment 2) were designed using Psyscope (version 1.1) together with QuickTime extension (QT OS 8). Responses were recorded using a button box connected to a Mac G4 through a USB Keyspan adapter (28X). Individual responses were recorded on-line.

Identification and subjective simultaneity experiments took place in a dimly lit, quiet room. Participants were seated at about 65 cm from the visual display, with the movie subtending a visual angle of 7.5° in the vertical plane and 9.5° in the horizontal plane. Videos were displayed centered on a 17″ G4 monitor on a black background. The luminance of the video display was suprathreshold for all stimuli insuring that no difference in response latency was artificially induced due to low luminance contrast. Sounds were presented through headphones (Sennheiser, HD520) directly connected to the computer at a level of approximately ∼70 dB SPL.

The average duration of the AV stimuli used in both experiments was 2590 ms, including video fade-in (8 frames) and fade-out (5 frames). Interstimulus intervals (ITIs) were randomly selected among 5 values (500 ms, 750 ms, 1000 ms, 1250 ms and 1500 ms). For both voiced and voiceless conditions, the identification task (Experiment 1) and simultaneity judgment task (Experiment 2) were run separately. Each participant took part, successively, in the identification experiment (e.g. A_bV_g) followed by the subjective simultaneity judgment experiment with the same McGurk pair and congruent counterpart (e.g. A_bV_g and A_dV_d). The task requirements were given prior to each experiment; importantly, participants were unaware of AV asynchronies prior to the identification task. For both identification and simultaneity judgment tasks, no feedback was provided and no training was given prior to testing.

1.3.1. Experiment 1: identification task

The identification task contained 10 presentations of each timing condition (29 timing conditions × 10 repetitions/condition in both A_bV_g and A_pV_k blocks for a total of 290 trials per block). In addition, for A_pV_k identification, 10 trials each of audio-alone /pa/ and visual-alone /ka/ were included to obtain an estimate of unimodal identification performance. Thus, for A_pV_k identification there was a total of 310 trials per subject. A single-trial 3 alternative-forced choice (3AFC) procedure was used. Participants were asked to make a choice as to "what they hear while looking at the face". Three choices were given for each

AV pair. In the A_bV_g pair, participants could answer /ba/, /ga/, or/da/ or /ða/ (voiced, as in "that" or "there"). The options /da/ or /ða/ were mapped onto a single response button. In the A_pV_k pair, participants could answer /pa/, /ka/, or /ta/. Note that for both AV stimuli the first response category corresponds to the auditory stimulus, the second to the visual stimulus, and the third to the fused McGurk percept.

1.3.2. Experiment 2: subjective simultaneity judgment task

The simultaneity judgment task contained 6 repetitions of each timing condition for either McGurk pair (A_bV_g and A_pV_k) and for either natural congruent pair (A_dV_d and A_tV_t), for a total of 696 trials per subject. Stimuli were pseudo-randomly intermixed. A single-trial 2 alternative-forced choice (2AFC) procedure was used.

Following Experiment 1, participants were asked to give their impressions of the difficulty of the task. All participants reported being aware of some cases in which A and V stimuli were not aligned in time. Participants were informed that AV synchrony was, in fact, manipulated and that in a second experiment participants' sensitivity to AV asynchrony was explored. Participants were thus asked, in Experiment 2, to determine if the time alignment of A and V stimuli was accurately rendered during the dubbing process and whether the auditory and the visual utterances were synchronized. Participants were told not to pay attention to the identity of the stimuli but rather to focus on the temporal synchrony of the stimuli. Participants were given two choices: "simultaneous" or "successive". They were told that the order did not matter in the 'successive' case and that they should press this button whether the auditory or the visual appeared to come first.

1.4. Analysis

Responses were sorted and averaged for each participant and each timing condition. A grand average of each possible response per timing condition was then computed across participants. The analysis of participants' performance for each timing condition revealed substantial individual differences across participants. Participants showing a constant average fusion rate lower than 40% regardless of asynchrony were not considered for further analysis (three participants in the A_bV_g condition and one participant in the A_pV_k condition showed an average of 22% fusion rate for all asynchronies).

2. Results

2.1. Experiment 1: identification task

2.1.1. Voiced McGurk pair A_bV_g

Fig. 1 shows the distribution (in percent) of each of the three possible response categories (/ba/, /ga/, /da/ or /ða/) as a function of SOA ($N = 18$). Auditory-visual "ga" responses (visually driven responses) were seldom given, whereas /ba/ (auditorily driven responses) and /da/ or /ða/ fusion responses formed the majority of responses. The overall trend shows that as the asynchrony between the AV utterances increases, /ba/ judgments increase, whereas /da/ or /ða/ judgments (fusion responses (FR)) decrease. An analysis of variance across SOAs shows a significant influence of asynchrony on fusion rate ($F(1, 28) = 9.242$, $p < 0.0001$). Unimodal stimuli were not collected for this pair, and therefore correction of fusion rates could not be calculated (cf. Results Section 1.2). A Fisher's PLSD test applied to uncorrected fusion rate across SOAs showed a range of non-significantly different SOAs between −133 ms and +267 ms. The temporal boundaries of the fusion rate plateau (SOAs at which fusion was maximal) were calculated on the basis of an asymmetric double sigmoidal (ADS) curve fitted to the average fusion rate function. A confidence interval of 95% was chosen to determine the asynchrony values at which the fusion rate was significantly different from that obtained at synchrony. Using an ADS fit ($r^2 = 0.94$) and a 95% confidence limit, a fusion rate plateau was determined to be from −34 ms auditory lead to +173 ms auditory lag. Moreover, the ADS fit confirms the asymmetrical profile of fusion responses and also suggests an off-centered peak towards auditory lag at about +69 ms (cf. Table 1).

2.1.2. Voiceless McGurk pair A_pV_k

Fig. 2 shows the proportions (in percent) of each of the three possible response alternatives (/pa/, /ka/, or /ta/) as a function of SOA ($N = 21$). Comparable to the A_bV_g condition, auditory-visual /ka/ (visually-driven) responses have the lowest probability of occurrence, whereas /pa/ (auditorily-driven responses) and /ta/ judgments (fusion) occur frequently and are clearly affected by audio delay. As the AV asynchrony increases, /pa/ judgments (auditorily driven responses) increase while /ta/ judgments (fusion responses) decrease.

In interpreting the bimodal responses to incongruent audio-visual stimuli, it is important to consider the errors that might be made by audio-alone and visual-alone processing. This is particularly relevant for visual-alone processing, where error rates can

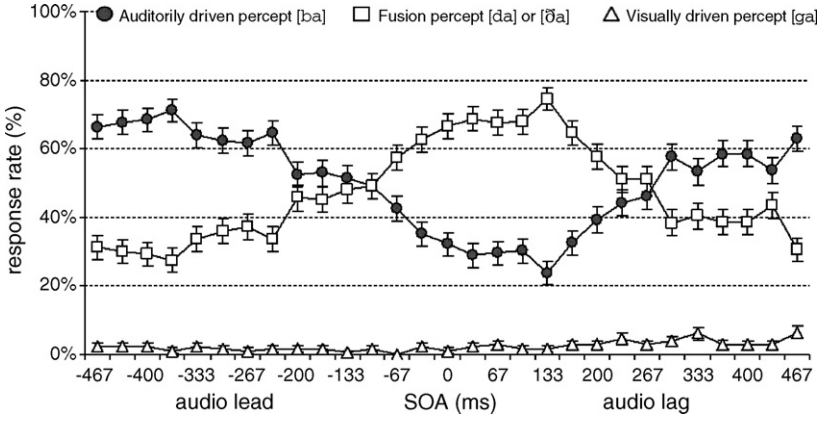

Fig. 1. Response rate as a function of SOA (ms) in the A_bV_g McGurk pair. Mean responses ($N = 18$) and standard errors. Auditorily driven responses (filled circles) are /ba/, visually driven responses (open triangles) are /ga/ and fusion responses (open squares) are /da/ or /ða/.

Table 1
Temporal integration windows parameters across conditions and stimuli

Stimulus	Task	A Lead Left Boundary (ms)	A Lag Right Boundary (ms)	Plateau Center (ms)	Window Size (ms)
A_pV_k	ID	-25	+136	+56	161
	S	-44	+117	+37	161
A_tV_t	S	-80	+125	+23	205
A_bV_g	ID	-34	+174	+70	208
	S	-37	+122	+43	159
A_dV_d	S	-74	+131	+29	205

Measures extracted from ADS fits ($r^2 > 0.9$) and a 95% confidence limit on fusion or simultaneity rate at synchrony condition (SOA = 0 ms). ID is the identification experiment (Experiment 1). S is the subjective simultaneity experiment (Experiment 2).

be quite high. Thus, since visual /ka/ is sometimes perceived as /ta/ it is possible that /ta/ responses to the audio-visual token A_pV_k may in fact be visual-alone driven responses rather than a fusion response representing true bimodal processing. One method for dealing with this potential confound is to use the unimodal error rates to normalize the bimodal fusion response rates. This procedure will generate a more conservative estimate of fusion. In the A_pV_k condition, audio alone and visual alone identifications were collected. Individual fusion rates for the A_pV_k condition were corrected on the basis of the individual's confusions in unimodal conditions (especially in the visual domain) in order to insure the bimodal nature of the fusion response. For example, consider an individual who has a fusion rate of 90% at synchrony. This same individual perceives an audio /pa/ as /ta/, 2% of the time (audio error) and a video /ka/ as /ta/ 30% of the time (video error). The corrected fusion rate (CFR) based upon of the individual's unimodal error rates becomes 58% (measured fusion rate minus audio error and visual error). The corrected fusion rates for each asynchrony value were averaged across participants and compared with the averaged rate of /ta/ responses that would be expected solely on the summation of unimodal error responses /ta/ to an audio alone /pa/ (average of 0.05, $N = 21$) and a visual alone /ka/ (average of 0.48, $N = 21$). If the fusion rate is superior to the sum of error rates in unimodal conditions (i.e. superior to 0.53 (0.05 + 0.48)), unimodal error rates do not suffice to account for /ta/ responses in the bimodal condition.

Fig. 2 illustrates that participants reported perceiving the 'fused' /ta/ over a wide range of audio-visual asynchronies. Auditory-visual /ta/ responses were compared to the unimodal

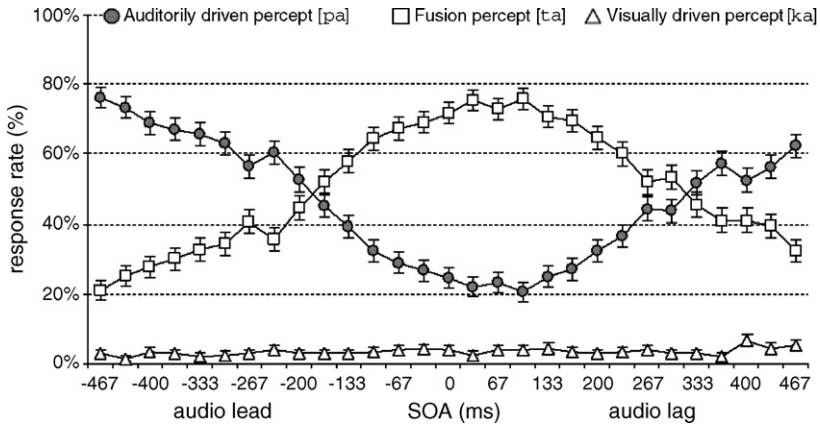

Fig. 2. Response rate as a function of SOA (ms) in the A_pV_k McGurk pair. Mean responses ($N = 21$) and standard errors. Auditorily driven responses (filled circles) are /pa/, visually driven responses (open triangles) are /ka/, and fusion responses (open squares) are /ta/. The sum of unimodal responses /ta/ to auditory alone /pa/ or visual alone /ka/ equals 53%. Fusion rates lower than 53% cannot be accounted for by unimodal errors. Fusion rates exceeding 53% constitute the true bimodal responses and can be observed from −167 ms of audio lead to 267 ms of audio lag.

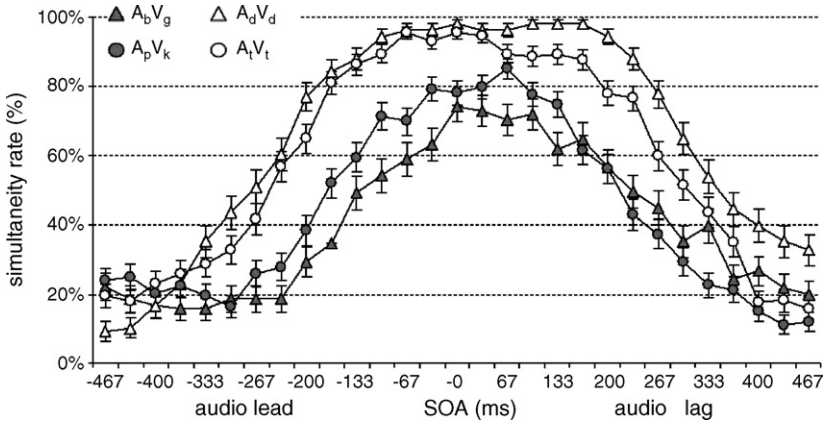

Fig. 3. Simultaneity judgment task. Simultaneity judgment as a function of SOA (ms) in both incongruent and congruent conditions (A_pV_k and A_tV_t $N=21$; A_bV_g and A_dV_d $N=18$). The congruent conditions (open symbols) are associated with broader and higher simultaneity judgment profile than the incongruent conditions (filled symbols). See Table 1 and Fig. 4 for further analysis of integration constants across conditions.

/ta/ occurrences in auditory-alone and visual-alone conditions. The resulting values therefore indicate true bimodal responses. An analysis of variance across SOAs shows a significant influence of asynchrony on fusion rate ($F(1, 28) = 4.336$, $p < 0.0001$). SOAs at which the fusion rate exceeds the averaged summation of error rate value (constant) correspond to the limits at which unimodal error responses /ta/ to an auditory /pa/ (5%) or to a visual /ka/ (48%) may account for the /ta/ response in bimodal condition A_pV_k. According to this definition, true bimodal fusion responses were observed from −167 ms of auditory lead to +267 ms of auditory lag. These same limits were obtained by applying a Fisher's PLSD at 95% confidence to the effect of SOAs on fusion rate ($p < 0.0001$).

Fitting results ($r^2 = 0.98$) showed that the fusion rate (FR) at SOAs ranging from −25 ms of auditory lead to +136 ms of auditory lag did not significantly differ from the fusion rate obtained in the synchrony condition. The ADS fit also confirms the asymmetrical profile of fusion responses and suggests an off-centered peak, towards auditory lag of about +55 ms (cf. Table 1).

2.2. Experiment 2: simultaneity judgment task

2.2.1. McGurk pair A_bV_g-congruent pair A_dV_d

Fig. 3 shows that the rate of simultaneity judgments for both the McGurk pair A_bV_g and the congruent pair A_dV_d decreased as the asynchrony between audio and video stimulus components increased. At synchrony (0 ms SOA), the congruent pair A_dV_d was judged 98% of the time to be simultaneous whereas A_bV_g reached a simultaneity rate of only 74% ($N = 18$). An ADS fit allowed defining the boundaries of the simultaneity plateau in both conditions with 95% confidence. The limits of the plateau, as defined by the ADS fitting procedure, resulted in a temporal window of integration ranging from −73 ms to +131 ms for the congruent pair ($r^2 = 0.98$) and from −36 ms to +121 ms for the incongruent pair ($r^2 = 0.98$).

A paired *t*-test between congruent and incongruent tokens across SOA's revealed a significant difference between the two simultaneity rate profiles ($p < 0.0001$). The incongruent A_bV_g pair was associated with a smaller temporal window and an overall lower rate of simultaneity judgments compared to the congruent profile (cf. Table 1).

2.2.2. McGurk pair A_pV_k-congruent pair A_tV_t

As with the A_bV_g and A_dV_d conditions, Fig. 3 shows that the percentage of simultaneity judgments on both the McGurk stimulus A_pV_k and the congruent stimulus A_tV_t decreased as the asynchrony between audio and video stimulus components increased. At synchrony (0 ms SOA), the congruent pair A_tV_t was judged 95% of the time to be simultaneous whereas the incongruent A_pV_k reached a maximum simultaneity rate of only 80% ($N = 21$). Using the ADS fitting procedure and a 95% confidence limit to define the boundaries of the simultaneity plateau for each stimulus condition resulted in a range from −80 ms of auditory lead to +123 ms of auditory lag for the congruent pair ($r^2 = 0.99$) and −44 ms to +117 ms for the incongruent pair ($r^2 = 0.98$).

A paired *t*-test between the simultaneity rate for congruent and incongruent tokens across SOAs revealed a significant difference between the two data series ($p < 0.0001$). Similar to the trend observed for the A_bV_g McGurk pair, the incongruent simultaneity profile revealed a smaller temporal window and an overall lower rate of simultaneity judgments as compared to the congruent profile (cf. Table 1).

3. Discussion

Two experiments were conducted to examine the effects of audiovisual temporal asynchrony on syllable identification and simultaneity judgment. The major finding was that AV speech inputs are extremely tolerant to bimodal asynchrony, and that bimodal information separated in time by as much as 200 ms is usually perceived as simultaneous. Specifically, both the identification experiment and the subjective simultaneity judgment experiment revealed temporal windows of maximal AV integration of about 200 ms. Information-processing windows of similar duration have been suggested as a basis for perceptual unit formation in the auditory cortices (Loveless, Levänen, Jousmäki, Sams, & Hari, 1996; Näätänen, 1992; Poeppel, 2003;

Winkler, Czigler, Jaramillo, Paavilainen, & Näätänen, 1998; Yabe, Koyama et al., 2001; Yabe, Tervaniemi, Reinikainen, & Näätänen, 1997; Yabe, Winkler et al., 2001) and recent psychophysical and neurophysiological studies, suggests that perceptual unit formation more generally has such a window to organize sensory information within and across modalities (Boemio, Fromm, Braun, & Poeppel, 2005).

Our data is overall consistent with the observations of Munhall, Gribble, Sacco, and Ward (1996) and extend their findings by sampling many more asynchrony values. Although the percentage of fusion remains resilient outside the plateau of integration established by ADS fitting (e.g. from −167 ms to 267 ms for A_pV_k), maximal true bimodal fusions cluster within ∼200 ms. The higher rate of fusion that was obtained here may result from our specific set of stimuli but a comparison of fusion rates across different McGurk stimuli may be desirable.

Both the integration window and the larger range of true bimodal interaction remain well below the estimated 500 ms breakdown of the FLMP fit suggested in an earlier study by Massaro, Cohen, and Smeele (1996). One possible difference may result from the conservative approach that was taken here, first in our choice of stimuli, by considering that an integrated AV percept results from two distinct unimodal inputs, and second, by our correcting the measured fusion rate, insuring that unimodal errors could not account for the integrated percept. Some methodological differences, such as our choice of variable inter-trial intervals, may also contribute to the discrepancies in the estimate of the temporal integration window boundaries, although we feel it is unlikely that these values have a significant impact on our results.

Interestingly, bimodal speech (congruent or incongruent) appears to tolerate much larger asynchronies than has been reported for non-speech stimuli (e.g. Dixon & Spitz, 1980) and argues for temporal integration far beyond the classical notion of simultaneity and temporal order threshold established with simpler stimuli (e.g. Hirsh & Sherrick, 1961; Zampini et al., 2003; Zampini et al., in press).

The subjective simultaneity judgment experiment comparing incongruent (A_bV_g and A_pV_k) and congruent (A_dV_d and A_tV_t) syllables allows one to evaluate the processing of illusory versus real speech percepts. Our interpretation is that in the case of incongruent AV stimuli, the spatio-temporal coincidence of the auditory and visual information is not as tight as in the congruent case. Specifically, differences in rate of perceived simultaneity for incongruent and congruent speech tokens could result from a correlation-like processing of facial kinematics and audio envelope dynamics. The degree of correlation between the two input channels could be used in both simultaneity judgment and token identification. The perceived degree of correlation might be considered directly (or explicitly) in judging simultaneity, and indirectly (or implicitly) in the identification of speech. According to a recent study by Grant and Greenberg (2001), the level of coherence between area of mouth opening and acoustic amplitude envelope can play a significant role in AV speech integration. Thus, the acoustic dynamic envelope and facial kinematics are correlated to a greater degree in the congruent than in the incongruent case. Based on these results, one would predict that, for equivalent SOAs, simultaneity judgments would be different for congruent speech conditions than for incongruent speech conditions. This is indeed what was found. The congruent tokens-A_dV_d and A_tV_t- were more readily considered 'simultaneous' than the incongruent tokens (∼95%). In the McGurk case, simultaneity judgments never exceeded 80%, and remained maximal within a plateau narrower than the congruent tokens. The AV incongruency of the speech tokens impinges, as predicted, on the subjective simultaneity judgment.

What are the implications of an integration window of ∼200 ms? It is noteworthy that the temporal window of integration for AV speech shown in both experiments (Fig. 4) corresponds to average syllable duration across languages (Arai & Greenberg, 1997). A similar window was also found in a temporal order judgment task, where the subjective temporal ordering of audio and visual speech inputs was assessed (van Wassenhove, 2004). Insofar as the syllable is considered a basic and critical unit for the perceptual analysis of speech, temporal

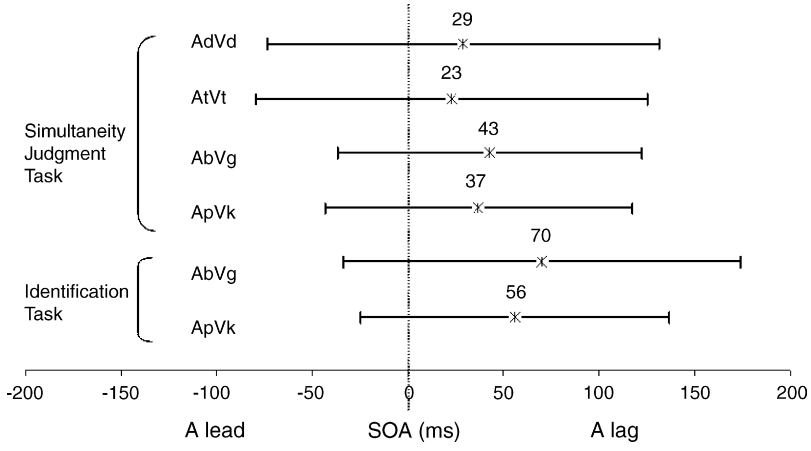

Fig. 4. Temporal integration windows across conditions and stimuli. Temporal integration windows obtained across conditions show similar characteristics in width (∼200 ms) and in the existence of a displacement towards auditory lag. The plateau observed in the simultaneity judgment tasks for congruent tokens (A_dV_d and A_tV_t) is larger than for incongruent tokens (A_bV_g and A_pV_k). The cross marks the center of the plateau defined by the fitting.

analysis on the syllabic scale is desirable and quite probably necessary (Greenberg, 1996, 2005). Moreover, the dynamics of AV utterances in production are on the syllabic scale and an important aspect of a possible supramodal speech code (Liberman & Whalen, 2000). Here, the overall decrease in simultaneity rate, the narrowing of the simultaneity plateau in incongruent AV syllables and the width of the fusion rate plateau suggest that the temporal evaluation mechanism of auditory and visual information streams is also essential to the processing of AV syllabic speech.

The flexibility for tolerating substantial amounts of AV asynchrony is supported by neurophysiological data. Single-unit recordings have shown that multisensory (AV) neurons have permissible windows of integration up to 1500 ms at the subcortical level (Stein & Meredith, 1993). In particular, some multisensory cell populations identified in the superior colliculus show response profiles that converge on optimal response enhancements at ∼100–200 ms of asynchrony (Meredith, Nemitz, & Stein, 1987). Analogous temporal properties could be expected for multisensory cortical neurons.

Data from the present experiments suggest a marked asymmetry associated with the temporal integration window for audio-visual speech input. Leading auditory information decreases integration, while leading visual information tends to enhance it. This trend has previously been reported in connected speech (Grant & Greenberg, 2001) and is typically accounted for by an inherent adaptation of the central nervous system to differences in the speed of light and sound (Massaro, 1996). Recent data suggest that an adaptative mechanism may act upon the perceived synchrony of auditory-visual events (Engel & Ehrenstein, 1971; Kopinska & Harris, 2004; Sugita & Suzuki, 2003) although such perceptual compensation remains controversial (Arnold, Johnston, & Nishida, 2005; Lewald & Guski, 2004). While the speeds of sound and light could be compensated for at the transduction level (acoustic transduction being faster than photo-transduction), the neural conduction speeds vary greatly as a function of processing streams, conveyed information, and levels of processing to name a few. Recent findings also suggest that AV simultaneity of non-speech events can be recalibrated on a minute-scale (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Navarra et al., 2005; Vroomen, Keetels, deGelder, & Bertelson, 2004) and the AV speech integrative mechanism could make use of such plasticity. More importantly, in laboratory experiments of the kind described here, the physical distance between the stimuli sources and the subject are small enough to make any differences in the speed of light and sound negligible (∼5 ms).

For speech inputs, however, the observed asymmetry could derive from the information content carried by the two modalities. Auditory information reaches the primary auditory cortex as early as 10 ms (Celesia, 1976; Howard et al., 1996; Lakatos, Pincze, Fu Javitt, Karmos, & Schroeder, 2005; Liégeois-Chauvel, Musolino, & Chauvel, 1991), and voicing information is already neurophysiologically realized at ∼60 ms (Steinschneider, Schroeder, Arezzo, & Vaughan, 1994; Steinschneider, Volkov, Noh, Garell, & Howard, 1999). Similarly, visual information can be recorded as early as 30 ms in the primary (V1) and motion (MT/V5) visual cortex (Buchner et al., 1997; Ffytche, Guy, & Zeki, 1995; Schroeder, Mehta, & Givre 1998). Although visual information (preparatory movements of face during articulation) is usually available prior to auditory information, the information content provided by speechreading constrains the categorization level to viseme representation. In contrast, voicing information provided by the auditory signal can single out a phoneme within the activated viseme class. Thus, a decision process receiving auditory information first might only be subject to cross-modal influences for the first ∼60 ms following signal onset. Further along into the acoustic signal, the quality of the auditory information suffices for the decision to be made without any need for additional information from other sources (e.g. visual). In the visual lead case, however, the auditory information could intervene more efficiently and for a longer delay because the visual information is ambiguous throughout most of the signal duration. This hypothesis is inherently consistent with our recent 'analysis-by-synthesis' model of AV speech integration in which the dynamics of auditory and visual speech inputs constrain the integration process (van Wassenhove, Grant, & Poeppel, 2005).

The different temporal profile of congruent versus incongruent tokens also complements previous neurophysiological evidence that a congruent A_dV_d or A_tV_t is not equivalent to an illusory A_bV_g or A_pV_k (e.g. Sams & Aulanko, 1991). Despite the apparent perceptual equivalence in labeling, the decrease of AV coherence in signal dynamics for incongruent pairs might be detected by the neural system while at the same time remaining sufficiently high to permit fusion. Accordingly, very similar simultaneity and fusion profiles for the incongruent stimuli were obtained, while a larger permissible temporal window of integration was found for congruent AV stimuli.The present findings support the existence of a ∼200 ms duration temporal integration window in AV speech perception, but shorter temporal integration windows are necessary to allow the extraction of modality-specific information. Early AV interaction presumably involves pre-perceptual processing (e.g. in classic multisensory neural populations such as the superior colliculus), at which a first level of AV information could be extracted. A second stage of interaction might occur pre-phonetically at the auditory sensory store, as suggested by Möttönen, Krause, Tiippana, and Sams (2002), or subsequent to unimodal processing via back projections as argued by Calvert, Campbell, and Brammer (2000). In both proposals, however, it remains unclear whether the involvement of auditory cortices in the processing of AV speech corresponds to the unimodal auditory processing or hosts the AV integration process. In light of recent evidence on the specificity of AV speech perception (Tuomainen, Andersen, Tiippana, & Sams, 2005), a more elaborate network involving both unimodal-processing streams and multisensory areas need to be considered that also account for the reported temporal resolutions.

Acknowledgments

This work was supported by grants NIH DC 0463801 and NIH DC 05660 to DP. A preliminary report of this work was presented at the 31st Annual Meeting of the Society for Neu-

rosciences, San Diego, November 15, 2001 and the 9th Annual Meeting of the Cognitive Neuroscience Society, San Francisco, April 14th, 2002. The opinions or assertions contained herein are the private views of the authors [KG] and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

References

Arai, T., & Greenberg, S. (1997). The temporal properties of spoken Japanese are similar to those of English. In *Proceedings of Eurospeech* (pp. 1011–1014).

Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research*, *45*, 1275–1284.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*(3), 389–395.

Buchner, H., Gobbelé, R., Wagner, M., Fuchs, M., Waberski, T. D., & Beckmann, R. (1997). Fast visual evoked potential input in to human area V5. *Neuroreport*, *8*, 2419–2422.

Calvert, G. A., Campbell, R., & Brammer, M. (2000). Evidence from functional magnetic resonance imaging of Crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.

Campbell, C. S., & Massaro, D. W. (1997). Perception of visible speech: influence of spatial quantization. *Perception*, *26*(5), 627–644.

Celesia, G. G. (1976). Organization of auditory cortical areas in man. *Brain*, *99*, 403–414.

Conrey, B. L., & Pisoni, D. B. (2003). Audiovisual asynchrony detection for speech and nonspeech signals. In *Proceedings of the audio-visual speech processing workshop (AVSP-2003)* (pp. 25–30).

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, *9*, 719–721.

Engel, G. R., & Ehrenstein, W. H. (1971). Visual-auditory distance constancy. *Nature*, *234*, 308–308.

Ffytche, D. H., Guy, C. N., & Zeki, S. (1995). The parallel visual motion inputs into areas V1 and V5 of human cerebral cortex. *Brain*, *118*, 1375–1394.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audio-visual simultaneity. *Nature Neuroscience*, *7*(7), 773–778.

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *Proceedings of the audio-visual speech processing workshop (AVSP-2001)* (pp. 132–137).

Grant, K., 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*, 2677–2690.

Greenberg, S. (1996). Understanding speech understanding: towards a unified theory of speech perception. In *Proceedings of the ESCA workshop on the "Auditory Basis of Speech Perception"*, Keele University, United Kingdom, 1–8.

Greenberg, S. (2005). A multi-tier theoretical framework for understanding spoken language. In S. Greenberg & W. A. Ainsworth (Eds.), *Listening to speech: an auditory perspective* (pp. 411–433). Mahwah, NJ: Lawrence Erlbaum Associates.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language and Hearing Research*, *40*, 432–443.

Hirsh, I. J., & Sherrick, C. E., Jr. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, *62*, 423–432.

Howard, M. A., III, Volkov, I. O., Abbas, P. J., Damasio, H., Ollendieck, M. C., & Granner, M. A. (1996). A chronic microelectrode investigation of the tonotopic organization of human auditory cortex. *Brain Research*, *724*, 260–264.

Jones, J. A., & Munhall, K. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*, 13–19.

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Perception and Psychophysics*, *62*(7), 1394–1404.

Kopinska, A., & Harris, L. R. (2004). Simultaneity constancy. *Perception*, *33*, 1049–1060.

Lakatos, P., Pincze, Z., Fu, K. M., Javitt, D. C., Karmos, G., & Schroeder, C. E. (2005). Timing of pure tone and noise-evoked responses in macaque auditory cortex. *Neuroreport*, *16*(9), 933–937.

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

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*, 187–196.

Liégeois-Chauvel, C., Musolino, A., & Chauvel, P. (1991). Localization of the primary auditory cortex area in man. *Brain*, *114*, 139–153.

Loveless, N., Levänen, S., Jousmäki, V., Sams, M., & Hari, R. (1996). Temporal integration in auditory sensory memory: neuromagnetic evidence. *Electroencephalography and Clinical Neurophysiology*, *100*, 220–228.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated? *Perception*, *29*(10), 1155–1168.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationals, evaluation, and recommendations for use. *British Journal of Audiology*, *24*, 29–43.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, *100*, 1777–1786.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, *77*, 678–684.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–747.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience*, *7*(100), 3215–3229.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, *13*, 417–425.

Munhall, K., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception and Psychophysics*, *58*, 351–362.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Charles Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, *25*, 499–507.

Näätänen, R. (1992). *Attention and brain function*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pandey, P. C., Kunov, H., & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research*, *26*, 27–41.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*, 245–255.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory, and linguistic aspects. *Philosophical Transactions of the Royal Society of London B*, *336*, 367–373.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception,. *Journal of Experimental Psychology, Human Perception and Performance*, *22*(2), 318–331.

Schroeder, C. E., Mehta, A. D., & Givre, S. J. (1998). A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. *Cerebral Cortex*, *8*(7), 575–592.

Sams, M., & Aulanko, R. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141–147.

Stein, B. E., & Meredith, A. M. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan, H. G., Jr. (1994). Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Electroencephalography and Clinical Neurophysiology*, *92*, 30–43.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard, M. A. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of Neurophysiology*, *82*, 2346–2357.

Sugita, Y., & Suzuki, Y. (2003). Implicit estimation of sound-arrival time. *Nature*, *421*, 911.

Sumby, W., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: the psychology of lip-reading* (pp. 3–51). London: Erlbaum.

Tuomainen, J., Andersen, T., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13–B22.

van Wassenhove, V. (2004). Neural correlates of auditory-visual speech desynchronization. In *'Analysis-by-synthesis' in auditory-visual speech perception: a forward model of multisensory integration* (pp. 119–158). PhD Thesis, University of Maryland, College Park.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of United States of America*, *102*(4), 1181–1186.

Vroomen, J., Keetels, M., deGelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Brain Research. Cognitive Brain Research*, *22*(1), 32–35.

Winkler, I., Czigler, I., Jaramillo, M., Paavilainen, P., & Näätänen, R. (1998). Temporal constraints of auditory event synthesis: evidence from ERPs. *Neuroreport*, *9*, 495–499.

Yabe, H., Koyama, S., Kakigi, R., Gunji, A., Tervaniemi, M., Sato, Y., et al. (2001). Automatic discriminative sensitivity inside temporal window of sensory memory as a function of time. *Cognitive Brain Research*, *12*, 39–48.

Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *Neuroreport*, *8*, 1971–1974.

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., et al. (2001). Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. *Brain Research*, *897*, 222–227.

Zampini, M., Shore, D. I., & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research*, *152*, 198–210.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (in press). Audiovisual simultaneity judgments.