Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 44 (2004) 5–18
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Auditory speech detection in noise enhanced by lipreading ☆

Lynne E. Bernstein [a,b,*], Edward T. Auer Jr. [a], Sumiko Takayanagi [a]

[a] Department of Communication Neuroscience, House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA
[b] National Science Foundation, Arlington, VA 22230, USA

Received 1 March 2004; received in revised form 8 October 2004; accepted 13 October 2004

Abstract

Audiovisual speech stimuli have been shown to produce a variety of perceptual phenomena. Enhanced detectability of acoustic speech in noise, when the talker can also be seen, is one of those phenomena. This study investigated whether this enhancement effect is specific to visual speech stimuli or can rely on more generic non-speech visual stimulus properties. Speech detection thresholds for an auditory /ba/ stimulus were obtained in a white noise masker. The auditory /ba/ was presented adaptively to obtain its 79.4% detection threshold under five conditions. In Experiment 1, the syllable was presented (1) auditory-only (AO) and (2) as audiovisual speech (AVS), using the original video recording. Three types of synthetic visual stimuli were also paired synchronously with the audio token: (3) A dynamic Lissajous (AVL) figure whose vertical extent was correlated with the acoustic speech envelope; (4) a dynamic rectangle (AVR) whose horizontal extent was correlated with the speech envelope; and (5) a static rectangle (AVSR) whose onset and offset were synchronous with the acoustic speech onset and offset. Ten adults with normal hearing and vision participated. The results, in terms of dB signal-to-noise ratio (SNR), were AVS < (AVL ≈ AVR ≈ ASR) < AO. That is, AVS was significantly easiest to detect, there was no difference among the synthesized visual stimuli, and all audiovisual conditions resulted in significantly lower thresholds than AO. To determine the advantage of the AVS stimulus, in Experiment 2, a preliminary mouth gesture was edited from the video speech token. This manipulation defeated the advantage for both the original and the edited AVS stimulus, while the audiovisual detection enhancement persisted. Overall, the results showed enhanced auditory speech detection with visual stimuli but no advantage for a fine-grained correlation between acoustic and optical speech signals.

Keywords: Audiovisual speech processing; Speech detection in noise; Speech in noise; Audiovisual speech perception; Speech processing; Lipreading; Speechreading

☆ A subset of the results in this paper was presented at AVSP03., St. Jorioz, France, September 4–7, 2003. This research was supported by the National Science Foundation (BCS 0214224). This article was written with support of the National Science Foundation. The views expressed here are those of the authors and do not necessarily represent those of the National Science Foundation.

* Corresponding author. Address: Department of Communication Neuroscience, House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA. Tel.: +1 213 353 7044; fax: +1 213 413 0950.
E-mail addresses: lbernstein@hei.org (L.E. Bernstein), auer@ku.edu (E.T. Auer Jr.), stakayanagi@hei.org (S. Takayanagi).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2004.10.011

1. Introduction

Audiovisual speech stimuli produce a diverse set of perceptual phenomena. For example, under low noise and good audibility conditions, phenomena such as the McGurk effect (McGurk and MacDonald, 1976) and the ventriloquist effect (De Gelder and Bertelson, 2003) show that vision influences auditory perception. Under good listening conditions, being able both to hear and see the talker can also enhance comprehension of the message (Arnold and Hill, 2001; Reisberg et al., 1987). Under noisy conditions as well, intelligibility is enhanced when the talker can be seen: Seeing the talker can be functionally equivalent to an increase in the acoustic signal-to-noise ratio (Sumby and Pollack, 1954). Recently, speech *detection* in noise has been shown to be enhanced under audiovisual conditions: Grant (2001) and Grant and Seitz (2000) showed that a spoken sentence masked by acoustic white noise is detectable at a lower signal-to-noise ratio (SNR) when the talker's speech movements can be seen.

In Grant's experiments, sentences were presented in white acoustic noise, in a two-interval forced-choice adaptive paradigm. Participants were asked to listen during both intervals and detect the acoustic stimulus sentence, which was presented in only one of the intervals. The sentences were presented under auditory-only (AO) and audiovisual (AV) conditions. The mean improvement (AV − AO threshold) in the detection threshold was 1.6 dB SNR (0.8–2.2 dB SNR) (Grant and Seitz, 2000). Similar results were obtained in (Grant, 2001). In the former study, a control experiment examined whether reading the text of each sentence prior to a detection trial also enhanced the detection threshold. A mean improvement of 0.5 dB SNR (0.33–0.78), which was statistically significant, was obtained when the sentences were read in advance. However, visual speech was significantly more effective for enhancing the threshold than was reading. The reading effect was attributed to a reduction in stimulus uncertainty.

In order to explain the AV detection enhancement effect, Grant calculated correlations between speech amplitude and the area of the mouth opening. The rationale for undertaking these correlations came from studies showing systematic relationships between the cross-sectional area of the front cavity of the vocal tract and the acoustic speech amplitude (Stevens, 1998). Pearson correlations for RMS energy of stimulus sentences versus area of mouth opening were in the range of 0.35–0.52 (Grant and Seitz, 2000), or 12–27% variance accounted for. Somewhat higher, but not statistically significant changes in improvements in the correlations were obtained when the analysis used speech that was bandpass filtered in the region of the second formant.

Grant (2001) reported higher average local correlations of 0.82, or 67% variance accounted for, when the analysis was focused on restricted portions of the stimulus with the highest amplitudes and the acoustic signal was restricted to the region of the second formant. These local high correlations were forwarded as the driver for the audiovisual speech detection enhancement effect. Grant (2001) and Grant and Seitz (2000) suggested that the primary mechanism of the AV detection enhancement depends on perception of brief high positive correlations between lip area and acoustic amplitude peaks. By observing the visual stimulus, the perceiver was theorized to become "alerted to temporal, and possibly spectral, locations of the acoustic noise-plus-speech stimulus where the S/N [SNR] is most favorable for detecting the speech. By this account, visually congruent speech information may serve to direct auditory attention, thereby reducing temporal and spectral uncertainty" (Grant and Seitz, 2000, p. 1206).

A problem with this explanation is that the time course of neural processing varies across stimulus attributes within sensory-perceptual systems and across sensory-perceptual systems, even leading under certain conditions to stimulus features being erroneously bound together (Moutoussis and Zeki, 1997; Treisman, 1996; von der Malsburg, 1995; Zeki, 1998). Yet, as Grant points out (Grant, 2001; Grant and Seitz, 2000), the detection of speech in noise would have to rely on brief portions of speech whose amplitude briefly exceeds the noise background. If visual processing of mouth gestures were to direct auditory attention,

visual processing would have to be fast enough to extract the gesture while the auditory system still has access to the brief acoustic peak.

Many temporal constraints of the central nervous system are known. Neural processing times through the auditory and visual pathways from the periphery to the first several levels of the cortex (Mesulam, 1998) set constraints on when and where auditory and visual speech information could possibly interact neurophysiologically (Schroeder and Foxe, 2002).

The first volley of stimulus driven activity into the auditory core cortex (the entry point to cortex for auditory information) occurs around 11–20 ms post-stimulus onset (Steinschneider et al., 1999; Yvert et al., 2001), and the conscious auditory speech percept appears to develop within 150–200 ms post-stimulus onset (Naatanen, 2001). In comparison, intra-cortical recordings in V1/V2 (the entry point to cortex for visual information) have shown the earliest stimulus-driven response to be at approximately 56–60 ms (Foxe and Simpson, 2002; Krolak-Salmon et al., 2001). Trans-cortical processing—processing required to extract stimulus features—requires time (Schroeder and Foxe, 2002). Evidence suggests that the latency for combining visual form and motion at the level of the cortex is at least 100 ms, and face motion processing might require latencies closer to 170 ms (Puce and Perrett, 2003). These estimates of processing times suggest that by the time that a visual mouth gesture has been processed, the acoustic stimulus is likely buried in the noise background again. At a cortical level, the temporal dynamics of auditory and visual perceptual stimulus processing do not seem well suited to using brief fine-grained correlations for detection.

An alternate explanation for the AV speech detection enhancement effect—one that is not dependent on perceiving complex visual speech features, such as mouth area, but would require merely the co-presentation of an auditory and visual stimulus—is excitatory–excitatory convergence, such as the type demonstrated by multisensory neurons in the superior colliculus (Meredith, 2002; Stein and Meredith, 1993). The superior colliculus is a sub-cortical structure in the bottom-up pathway, prior to the higher cortical levels of stimulus feature analysis, and is concerned with the detection of events in extra-personal space (Meredith, 2002; Stein and Meredith, 1993). Superior colliculus neurons can respond weakly to AO or visual-only stimulation but very strongly to their combination, frequently super-additively at threshold levels. Their responses are sensitive to the temporal relationship of multisensory stimuli, with responses greatest when the stimuli occur within 100 ms of each other (Meredith et al., 1987). Rather than relying on speech feature processing, the AV speech detection effect could rely on early sub-cortical processing that is not specialized for speech (Bernstein et al., 2004) and does not require top-down attention. Other AV phenomena listed earlier in this introduction are also not all specific to speech, and some effects might engage more than one perceptual mechanism. For example, the ventriloquist effect, which involves mislocation of an auditory stimulus to that of a visual stimulus, can be demonstrated with both speech and non-speech stimuli (De Gelder and Bertelson, 2003) and has been attributed to early bottom up processing (Colin et al., 2002).

1.1. The current study

With the above considerations in mind, the current perceptual study was undertaken to investigate the stimulus conditions that lead to the AV speech enhancement effect. The study was designed to test whether enhanced auditory speech detection depends on seeing a speech stimulus, or whether a simple, non-speech visual stimulus is sufficient. The study was designed also to test whether the effect relies on processing a fine-grained correlation between the area of a dynamic visual stimulus and the acoustic amplitude envelope, or whether merely presenting a constant visual stimulus during a speech token is sufficient to achieve enhanced detection.

The study used an adaptive two-interval forced-choice paradigm (Levitt, 1971) to obtain detection thresholds for an acoustic speech token /ba/ whose level was fixed across trials. The syllable was presented in an adaptively adjusted white noise masker, where the noise sample was randomly selected from trial to trial. The acoustic token

was presented in only one of the two intervals of each trial, and participants were instructed to select the interval in which the syllable occurred.

In Experiment 1, in addition to the AO speech token, four different types of visual stimuli were used in separate adaptive threshold runs. One of the visual stimuli was the face of the talker, recorded at the same time as the recording of the audio stimulus. To test whether merely presenting a simple visual stimulus during the speech syllable could enhance the detection threshold for the /ba/ stimulus, a static filled rectangular shape was synthesized, whose presentation duration was equal to that of the audio /ba/, and therefore, had zero correlation with the acoustic envelope during the course of the /ba/ stimulus. A significant effect of this stimulus would be consistent with a low-level excitatory–excitatory interaction mechanism (Meredith, 2002; Stein and Meredith, 1993).

To test whether the effective stimulus required a fine-grained correlation between the amplitude envelope of the speech and the area of the visual stimulus, related to the mechanism hypothesized by Grant, a dynamic rectangle whose horizontal extent was correlated with the speech amplitude envelope of the /ba/ was generated. The dynamic rectangle expanded horizontally, so that it would not have the appearance of a mouth opening, although its area was correlated with speech energy. A fourth visual stimulus was generated to capture the same dynamics as the rectangle but did have the potential to appear mouth-like. It was a filled dynamic Lissajous figure, that is, a filled oval shape whose vertical extent was correlated with the acoustic speech amplitude. Thus, it presented an audio-to-visual correlation that might have a mouth-like appearance to participants; however, if the Lissajous figure were presented by itself, it would not convey phonetic information. The Lissajous figure tested the possibility that a very schematic mouth-like gesture, in combination with the audio /ba/, could create a speech impression that might result in a speech-specific effect. That is, a possible outcome would be that a similar enhancement would be achieved with the Lissajous figure as with the natural video token.

During the AO speech detection threshold runs, a fixation cross was presented during each observation interval, so as to reduce uncertainty within the context of each trial. That is, the fixation cross indicated the time periods during which participants should attend for the auditory stimulus.

In summary, if it were the case that merely presenting a simple non-speech visual stimulus with an acoustic speech syllable is sufficient to enhance detection thresholds, the static rectangle should significantly reduce audio /ba/ detection thresholds. This result would be consistent with a bottom-up, excitatory–excitatory mechanism that did not require higher-level perception. If, however, dynamic properties are required, then the dynamic rectangle should be significantly more effective than the static rectangle, implicating higher-level processing of visual stimulus properties. If the effect were specific to visual speech, then the natural video token should result in the lowest detection thresholds. If the dynamic Lissajous figure produced thresholds similar to the natural video token and lower than the other video stimuli, the implication would be that stimuli need only be grossly speech-like and need not convey specific phonetic information. Following on the finding that there did seem to be a special advantage to only the natural visual speech token, a second experiment was run to investigate that effect further.

2. Experiment 1 methods

2.1. Participants

Eleven participants were recruited, and 10 completed the experiment. All were native speakers of American English (three males and seven females, ages 19–40 years, mean age 26.4 years), with normal hearing (hearing thresholds $\leqslant$15 dB HL at audiometric test frequencies from 250 to 8000 Hz) (American National Standards Institute, 1989). Their speech reception thresholds in noise were tested using the Hearing in Noise Test (HINT) (Nilsson et al., 1994). Their composite HINT scores (comprising measures of noise front, noise right, noise left) were normal. Participants

were screened for normal vision, and they were screened to be average or better lipreaders, as referenced to the distribution of performance of a larger group of hearing lipreaders (Bernstein et al., 2000). Screening was used to assure that participants were individuals who were likely to be users of visual speech information. They gave their informed consent, and they were paid $10 per hour for their participation.

2.2. Stimuli

The stimuli for this study were based on a single videorecorded /ba/ token. The token was produced by an experienced female talker as part of a much larger database of syllables. A UVW-1800 SONY Betacam SP recorder and a SONY production camera were used to make the recording. The natural audiovisual speech (AVS) token was used in one of the conditions.

Synthesized video stimuli were generated following computation of the amplitude envelope of the acoustic /ba/ signal. Fig. 1 shows the amplitude envelope curve of the acoustic /ba/ token. Fig. 1 also shows the area of the mouth opening, which was obtained by manually selecting the pixels within the mouth opening for each video field and computing the total for each field. Fig. 1 demonstrates that the amplitude of the acoustic signal rose sharply and slightly prior to the full mouth opening. The amplitude peak was extremely brief. Fig. 1 shows that in the natural video token, the talker also slightly opened and closed her lips in advance of the acoustic bilabial release of the /b/. The dynamic synthesized video stimuli were correlated with the amplitude curve rather than the mouth opening area function. The full amplitude of the synthesized dynamic stimuli was achieved approximately two video frames (67 ms) ahead of the open mouth position. The correlation between synthesized motion stimuli and the acoustic envelope was 0.996. The correlation between the natural mouth opening and the acoustic envelope was 0.76, and between the natural mouth opening and the synthetic stimuli was 0.77.

The dynamic Lissajous figure was synthesized at the field rate of the video speech, that is, 59.94 frames/s. Its vertical extent was correlated with the speech amplitude envelope (see Figs. 1 and 2). A dynamic rectangle was synthesized in a similar manner, but its horizontal extent was correlated with the speech amplitude envelope (see Fig. 2). The areas of the dynamic Lissajous and rectangle stimuli were equated so that neither had an energy advantage. A static rectangle corresponding to the largest rectangular video frame

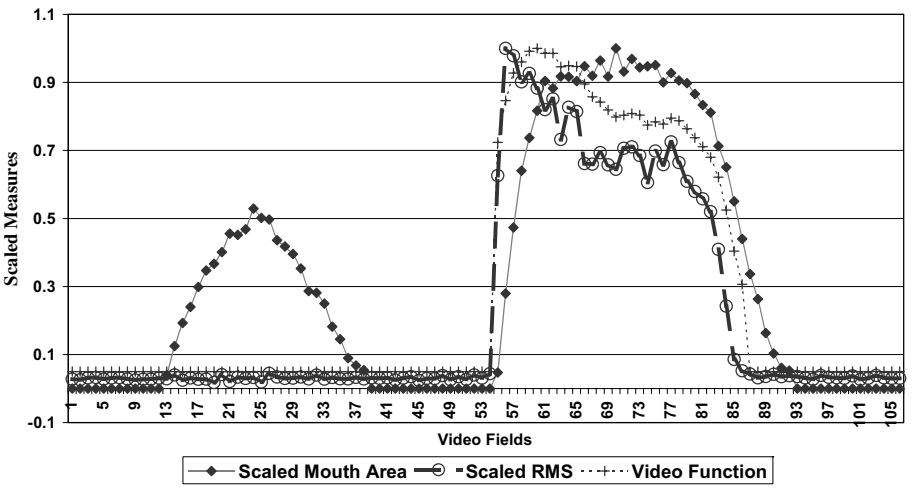

Fig. 1. Normalized speech amplitude (rms), normalized mouth opening area (pixels), and drive signal for dynamic synthesized video (AVL and AVR).

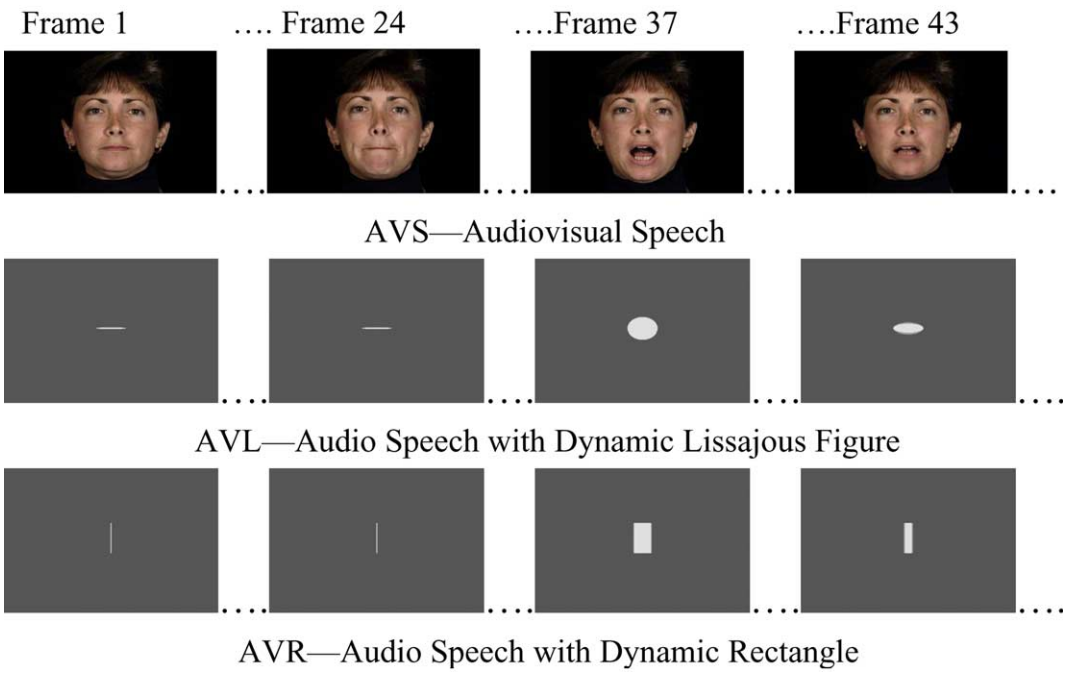

Fig. 2. Four video frames from each of the conditions with dynamic video stimuli. The top row is from the natural AVS condition. The second row is from the dynamic Lissajous condition (AVL), and the third row is from the dynamic rectangle condition (AVR).

was also synthesized with temporal duration equivalent to the acoustic syllable duration. During the AO condition, a fixation cross appeared on the video monitor during each stimulus observation interval in the two-interval forced-choice trial. All video stimuli were presented on a SONY Trinitron monitor.

The presentation amplitude used for the auditory /ba/ did not vary throughout the experiment. It was set following a pilot experiment with five participants. In the pilot study, the speech was presented at 65 dB SPL. But this resulted in unacceptably high noise levels whenever the speech was presented under audiovisual conditions. Therefore, the speech level for the current experiment was set at 60 dB SPL.

Computer-generated white masking noise was stored in a long audio file. Each interval of masking noise was selected at random from the long noise file and output through a sound card and a calibrated programmable attenuator. The speech and noise were mixed in real time and presented binaurally over TDH 49 headphones.

All of the AV stimuli (natural video, synthesized video, and audio) were transferred to a DVD for use during the experiment. The component signal from the original Betacam SP video of the /ba/ stimulus was digitized on an ACCOM 2XTREME real-time digital disk recorder. Uncompressed video frames were transferred to a PC as individual frame files with a spatial resolution of 720×486.

For every AV stimulus, a sequence of uncompressed frames for the video was built into an AVI (audio video interleave) file for software MPEG compression. All of the MPEG files for the different conditions were transferred to the DVD. MPEG Level 2 compression was accomplished using the LIGOS LSX MPEG-Compressor (Version 3.5). We have obtained excellent results in direct comparisons between DVD and laserdisc, suggesting that this level of compression does not compromise video quality. The video input format was 720×480, interlaced with the top field first and frame rate of 29.97. For compression, the frame sequence was all I frames, with a constant

bitrate specified at 7700 Kbits/s. The bitrate was selected so as not to exceed the peak rate allowed by DVD when including uncompressed 48-kHz locked audio with the video. The MPEG files were authored to DVD using the ReelDVD (Version 2.5.1) software package from SONIC. The resulting DVD contained a single sequential program chain, which is required by the Panasonic V7400 player to allow frame-based searching and access. By this method, random access of the stimuli for each trial was made possible. The audio /ba/ associated with the different visual stimuli was stored in a separate file, uncompressed with a sample rate of 48 kHz. As with the video, the audio associated with the different trial types was concatenated into a single long file for production of the DVD. The concatenation of the audio was performed using custom software that ensures frame-locked audio of 8008 audio samples/5 video frames.

2.3. Procedure

There were five different conditions in the experiment: (1) auditory-only speech (AO); (2) audiovisual speech (AVS); (3) audio /ba/ with the dynamic visual Lissajous figure (AVL); (4) audio /ba/ with a dynamic rectangle (AVR); and (5) audio /ba/ with a static rectangle (AVSR). Across all conditions in the experiment, the task was based on a two-interval, forced-choice adaptive threshold paradigm. In this paradigm, there were two observation intervals for each trial, and the participant indicated whether the acoustic speech token occurred in the first or second interval (see Fig. 3). For each trial, the observation interval in which the signal was presented was randomly selected. The noise masker began before the first interval in which a stimulus could occur and ended after the second such interval. The onset trigger for the noise masker was a signal recorded on the second audio track of the DVD that stored the stimuli.

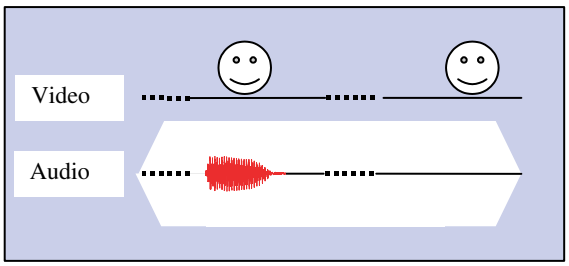

Fig. 3. Trial structure. Each trial comprised two intervals. The acoustic /ba/ stimulus could occur in only one of the intervals. For trials with visual stimuli, the visual stimulus appeared in both intervals. The dots in the figure correspond to the temporal jitter in the timing of the two intervals (see text).

The adaptive rule that was used to adjust the masker noise was as follows: Three correct responses and the noise was increased, and one incorrect response and the noise was decreased. The rule converges on the 79.4% detection threshold (Levitt, 1971). At the beginning of testing, the acoustic signal was −6 dB below the level of the noise. The step sizes during the adaptive testing were adjusted so that at the beginning, step changes following the adaptive rule were 3 dB. Then the step changes were reduced to 2 dB until the second and third reversals occurred, as specified by the adaptive rules. Changes in the noise step sizes then followed the schedule of 1 dB until the fifth reversal started; 0.5 dB until the seventh reversal started; 0.2 dB until the tenth reversal started; and 0.1 dB for the final two reversals. The threshold was the mean calculated using all of the 12 SNR levels at reversal points.

To prevent participants from relying on the timing relationships within the two stimulus observation intervals in each of the adaptive trials, a set of trials was generated for each condition that varied in terms of the onset of the stimulus within the observation interval (see Fig. 3). (Only the timing of the stimulus presentation was jittered, not the relationship between audio and video for AV stimuli.) The total duration of the observation interval remained fixed at 64 frames. The stimulus onset jitter spanned six steps (0–5 frames, represented by the dots in Fig. 3), with each step equivalent to one video frame (at 33.37 ms/frame). The duration of onset jitter was randomly selected. In order to hold observation intervals constant across trials, whatever number of jitter steps prior to the stimulus onset was subtracted from 6, and the reminder was added following the stimulus.

The method that was used to obtain the required timing relationships for each trial involved creating each in advance on the DVD. For each

trial type, the sequence of uncompressed frames for the video /ba/ or the synthesized visual stimulus was built into an AVI (Audio Video Interleave) file for software MPEG compression, along with the appropriate frames needed to jitter the onset times. During the AO trials, a fixation cross was presented on the video monitor during each observation interval.

Participants were asked to respond as quickly and accurately as possible using a button box for which the first interval corresponded to the left button and the second interval corresponded to the right button. Response times were recorded by an external response time clock. On the first day of the experiment, participants received a set of practice trials in each stimulus condition to learn the procedure. During the practice, the same adaptive rules were used as during the testing, but the step size was maintained at 3 dB. Also, during practice no more than 10 trials were presented in each condition. Each participant was tested in each of the conditions a total of eight times. The order of conditions was randomized within sets of the five conditions, so that each participant completed eight randomized sets. The number of days that participants required to complete the experiment varied between 2 and 6 days. The interval of time in which the sessions took place varied between 2 and 32 days. Ample rest times were given when the period of testing was reduced to 2 days. Testing was administered in a double-walled IAC booth.

3. Results

3.1. Speech detection thresholds

Fig. 4 shows the group mean thresholds across the five conditions and 10 participants. Examination of the figure suggests that thresholds were highest in the AO condition, lowest in the AVS condition, and intermediate for the other AV conditions (AVSR, AVL, AVR). Fig. 5 shows the mean results for each participant and condition. This figure shows that the group pattern held generally across participants but with some individual variations.

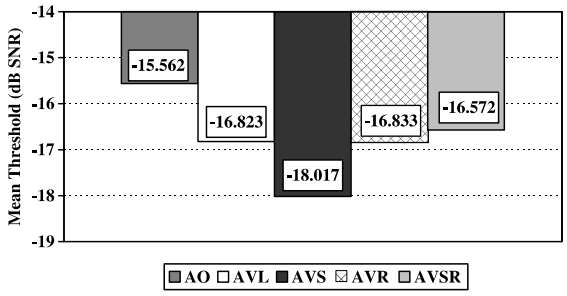

Fig. 4. Group means for thresholds across all sessions and participants. Means are dB signal-to-noise ratio at the estimated 79.4% threshold.

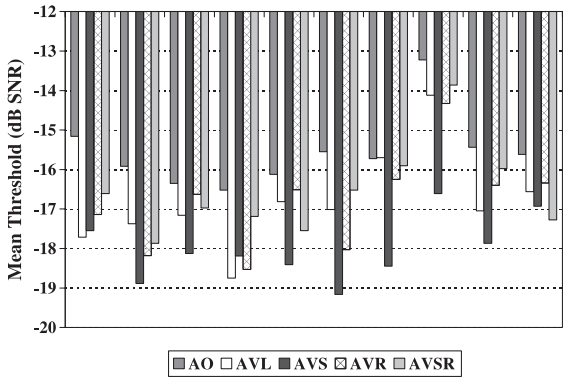

Fig. 5. Individual means for thresholds in each condition and for each participant. Means are dB signal-to-noise ratio at the estimated 79.4% threshold.

Analysis of variance was applied to the results in a repeated measures design. The repeated factors were condition (AVS, AVL, AVR, AO, and AVSR), session (four), and run (two per session). There was a significant main effect of condition [$F(4,6) = 50.49$, $p < 0.001$]. The threshold means (in dB SNR) for the five conditions were AVS = −18.017; AO = −15.562; AVL = −16.823; AVR = −16.833, AVSR = −16.572. There were no other significant main effects or interactions.

Contrast analyses were used to test the specific hypotheses of the study. The results of the contrast analyses are shown in Table 1. AO thresholds were significantly higher than the thresholds in the four other conditions (AVS, AVL, AVR, and AVSR). AVS thresholds were significantly lower than in each of the four other conditions. There were no

Table 1
Results of the contrast analyses for thresholds in the five conditions of Experiment 1

	AO	AVL	AV	AVR	AVSR
AVL	$F = 27.506$ $p = 0.001$	–	–	–	–
AV	$F = 108.791$ $p < 0.001$	$F = 11.594$ $p = 0.008$	–	–	–
AVR	$F = 23.314$ $p = 0.001$	$F = 0.002$ $p = 0.961$, *NS	$F = 19.677$ $p = 0.002$	–	–
AVSR	$F = 30.925$ $p < 0.001$	$F = 0.969$ $p = 0.351$, *NS	$F = 21.341$ $p = 0.001$	$F = 0.962$ $p = 0.352$, *NS	–

* NS—Not significant.

significant differences among AVL, AVR, and AVSR conditions.

3.2. Response time measures

Response times collected during the experiment were analyzed to determine whether there were patterns that could provide additional insight into the participants' performance. Several different analyses were performed, using responses to correct trials only. If the participant responded before the audio began, even if the response was correct, that response time was not included in any latency analysis. Two measures of central tendency were computed, the arithmetic mean and the harmonic mean, because response times are known to have non-normal distributions and to be sensitive to outliers (Ratcliff, 1993). The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the scores. By using two measures, a better estimate of the stability of the results could be achieved. Both measures are reported only when the analyses produced different results.

Response times were entered into a repeated measures ANOVA, with stimulus interval (first or second), session (4), run (2) and condition (5) as the repeated factors. This analysis used data from eight of the 10 participants, because there were a few missing cells for the other two participants. For those participants, data for both intervals were not available for some sessions and/or conditions. Interval was the only significant effect for the mean response time measure [$F(1,7) = 7.49$, $p = 0.029$], with the second interval faster (1593 ms) than the first interval (1662 ms). There were no other significant main effects or interactions. Mean response times for the five conditions were AO, 1677 ms; AVL, 1619 ms; AVS, 1606 ms; AVR, 1605 ms; and AVSR, 1632 ms. (Note that these long response times were an artifact of measuring latency from the point corresponding to the onset of the visual speech stimulus.) The advantage to the second interval can be attributed to participants having quickly determined, particularly during suprathreshold trials, that the auditory stimulus was not in the first interval and could therefore be expected in the second.

The analysis above did not take into account the effect of the varying SNRs across the individual trials of the adaptive runs. Additional analyses were undertaken to determine whether response times varied across conditions in the neighborhood of the detection threshold. For each participant, the response times for trials that were within ±0.5 dB SNR of the final threshold estimate were extracted from the data. Although stimulus observation interval was a significant factor in the previous analysis, it was not used in this analysis, because of empty cells for several of the participants who had many errors whenever the stimuli were in the first interval.

A repeated measures ANOVA was carried out with condition (5) as the repeated factor. The analysis of the mean response time measure failed to produce any significant effects. A significant main effect for condition was obtained with the harmonic means [$F(4,6) = 5.161$, $p = 0.038$]. Table 2 shows the contrast analyses on the harmonic means. The significant main effect of condition

Table 2
Results of the contrast analyses for response times estimated with the harmonic means in the five conditions in the experiment

	AO	AVL	AVS	AVR	AVSR
AVL	$F = 6.939$ $p = 0.027$	–	–	–	–
AVS	$F = 1.322$ $p = 0.280$, *NS	$F = 0.144$ $p = 0.713$, *NS	–	–	–
AVR	$F = 15.931$ $p = 0.003$	$F = 0.000$ $p = 0.986$, *NS	$F = 0.140$ $p = 0.717$, *NS	–	–
AVSR	$F = 2.133$ $p = 0.178$, *NS	$F = 1.149$ $p = 0.312$, *NS	$F = 0.247$ $p = 0.631$, *NS	$F = 1.958$ $p = 0.195$, *NS	–

These contrast analyses used harmonic mean RTs obtained across both stimulus observation intervals for responses within ±0.5 dB SNR of the estimated threshold for the run ($N = 10$).
* NS—Not significant.

was due to the AO condition being slowest and significantly different from the AVL and AVR conditions. The AVS condition was not different in latency from the AO condition.

Lastly, the response time measure was analyzed to determine whether the participants used different strategies across the two intervals as a function of condition. The response times in the first versus second interval for each condition and participant were entered into a repeated measures ANOVA. This analysis used an expanded range of SNR values (within plus and minus 1.5 dB of the threshold for each run). All 10 participants contributed data. The main effect of interval was significant [$F(1,9) = 55.352$, $p < 0.001$] (first interval 13% correct responses; second interval 40% correct responses). But the main effect of condition was not reliable, nor was the interaction between interval and condition. Overall, response time measures suggest that participants did not change their response strategies as a function of condition.

3.3. Discussion

Experiment 1 was undertaken to investigate whether the AV speech detection enhancement effect reported by Grant (2001) and Grant and Seitz (2000) is specific to visual speech stimuli, and whether it relies on perceptual analysis and attention to the fine-grained correlated dynamics of an audiovisual speech stimulus. The threshold estimates obtained in Experiment 1 showed an unqualified advantage for audiovisual stimuli. But the results with the static rectangle showed that the static visual stimulus was sufficient to enhance the detection thresholds of an acoustic speech syllable. Animating the visual stimuli as a function of the acoustic amplitude envelope of the /ba/ syllable did not result in further improvements to the threshold levels over the static rectangle. Participants appeared not to benefit from the correlation between the dynamics of the speech envelope and the dynamics of the Lissajous and rectangle shapes. The foregoing results are compatible with the hypothesis that the AV speech detection enhancement effect does not require high-level analysis of visual mouth features.

But significantly lower thresholds were obtained with the natural AVS stimulus. This result raised the questions whether visual speech engages additional or different mechanisms than does non-speech visual stimuli, or whether the visual speech token provided additional useful stimulus information in the threshold task. Fig. 1 shows that the natural visual speech movement began with a small lip opening followed by lip closure, prior to the acoustic bilabial release. The preliminary gesture was not in the synthesized visual stimuli, and the synthesized visual stimuli were shorter in duration, coinciding with the acoustic syllable amplitude envelope. The preliminary visible gesture of the natural stimulus was at a fixed duration from the amplitude peak in the acoustic stimulus and could have functioned as a cue to the peak's location. The relationship between the preliminary lip gesture and the acoustic syllable was fixed. The

preliminary information could have functioned as a pre-cue and provided the advantage obtained in the AVS condition.

Experiment 2 was undertaken to determine whether the advantage in the AVS condition was due to the preliminary mouth gesture in the natural token. The visual token in the AVS condition was edited to remove the frames comprising the preliminary mouth opening and closing gesture. Following the same general methods as in Experiment 1, four conditions were compared in Experiment 2: AO, AVS, AVR, and AVSE (audiovisual speech edited).

4. Experiment 2 methods

4.1. Participants

Four participants were recruited. Two had participated in Experiment 1, and two had participated in pilot experiments. Thus, all had normal pure tone averages, normal or corrected-to-normal vision, passing or better scores on the lipreading screening, and normal HINT scores. Three were female. Their ages ranged between 20 and 28 years (mean age 25 years). They gave their informed consent, and they were paid $10 per hour for their participation.

4.2. Stimuli

The AVS, AVR, and AO were the same stimuli as in Experiment 1. The AVSE stimulus was the same as the AVS stimulus, except that the visual token began with the 25th frame (see Fig. 1, and note that the data are presented at the field rate). The duration of the AVS and AVSE visual stimuli were the same, however, because the 25th frame was repeated 25 times initially.

4.3. Procedure

The same procedure was followed as in Experiment 1, except that this experiment had four conditions (AVS, AVR, AO, and AVSE) rather than five.

5. Results

Fig. 6 shows the individual mean thresholds in Experiment 2 and the individual mean AVS threshold in a previous experiment (AVS-P). Mean thresholds from Experiment 2 are shown on the right of the figure. Examination of the figure suggests that thresholds were highest in the AO condition and lower but similar in all the audiovisual conditions. Most striking is the high similarity between AVS and AVSE thresholds within individual participants.

A repeated measures analysis of variance was applied to the results with the factors of condition (AVS, AVSE, AVR, and AO), session (four), and run (two per session). There was a significant main effect of condition [$F(3,9) = 22.07$, $p < 0.001$]. The threshold means (in dB SNR) for the four conditions were AVS = −17.149; AVSE = −17.091; AVR = −16.336; and AO = −15.516. There were no other significant main effects or interactions.

Contrast analyses showed that AO thresholds were significantly higher than the thresholds in the three other conditions (AVS, AVSE, and AVR): AO versus AVR [$F(1,3) = 11.559$, $p = 0.042$]; AO versus AVS [$F(1,3) = 68.568$, $p = 0.004$]; and AO versus AVSE [$F(1,3) = 69.739$,

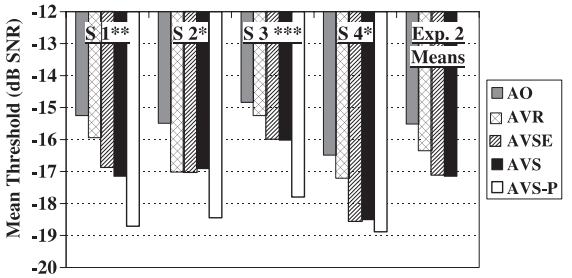

Fig. 6. Individual means for thresholds in each condition of Experiment 2 with the participants' previous AVS (AVS-P) thresholds. Participants are designated S1–S4. Experiment 2 means shown on the right are across all threshold estimates for the four participants. Means are dB signal-to-noise ratio at the estimated 79.4% threshold. Note: "*" Designates participant was in Experiment 1. "**" Designates participant was in a pilot study in which the 71% threshold (a more difficult level) was tested to obtain the AVS-P threshold. "***" Designates participant was in a pilot study in which the 50% threshold was obtained, and the speech was presented at 65 dB SPL (yet more difficult).

$p = 0.004$]. But there were not any other significant differences among audiovisual conditions.

The results in Fig. 6 suggest that not only was there no difference between AVS and AVSE in Experiment 2, but in addition, the removal of the preliminary mouth gesture resulted in higher thresholds than these participants had obtained with the AVS stimulus previously. When these participants had been tested previously, they obtained lower AVS thresholds, even under more difficult testing conditions. That is, although all of the participants in Experiment 2 received the same adaptive threshold rule, and the previous results for Participants 2 and 4 used the same threshold rule, the previous results shown in Fig. 6 for Participants 1 and 3 were obtained under more difficult conditions: Participant 1 had been tested in a pilot experiment in which the 71% adaptive threshold rule was used. The previous results for Participant 3 were from a pilot experiment in which the 50% adaptive threshold rule was used, and the speech acoustic stimulus was presented at 65 dB SPL.

5.1. Discussion

The results of Experiment 2 can be interpreted as evidence that the audiovisual effect of Experiment 1 is replicable. But the additional advantage for visual speech appears to be fragile. Participants in Experiment 1 apparently took advantage of the temporal relationship between the visible preparatory mouth motions and the consonant gestures, which was a reliable temporal cue to the location of the upcoming acoustic amplitude peak. This advantage seems to have been defeated in Experiment 2, in which participants' AVS and AVSE thresholds were not different from the AVR thresholds, and in which their AVS thresholds rose relative to previous AVS thresholds.

During Experiment 2, eight thresholds were obtained in the AVS and eight in the AVSE conditions. The order of conditions was randomized within sets of the four conditions (i.e., AO, AVR, AVS, and AVSE), so that each participant completed eight randomized sets. Thus, during a set of four thresholds, with randomly ordered conditions, the preliminary mouth gesture was present for only one of the threshold runs. A possible explanation for the results in Experiment 2 is that when the preliminary gesture was no longer reliably available across the context of the experiment, the participants no longer took advantage of it when it was present. That is, they no longer attended to the preliminary mouth gesture. On the other hand, the advantage of AV stimuli relative to the AO condition remained a robust, statistically significant difference. This result is compatible with the hypothesized bottom-up excitatory–excitatory detection mechanism, which to be effective should operate independent of stimulus identity.

6. General discussion

In Experiment 1, the AO detection thresholds were significantly higher than the thresholds obtained with AV stimuli. But the natural AV speech token produced significantly lower thresholds than those obtained with the synthesized video stimuli (AVR, AVSR, and AVL). Importantly, there was not any difference among the synthesized video stimuli, suggesting that the dynamics of the AVR and AVL stimuli did not contribute additional advantage beyond the static rectangle. Experiment 2 compared the AVS stimulus to the AVSE stimulus for which the preliminary mouth gesture was removed, but the total stimulus was equal in duration. When the preliminary speech gesture was removed, both the original AVS and the AVSE stimuli produced similar thresholds, but ones that were higher than those obtained earlier by these participants. This result suggests that the advantage of the preliminary gesture depended in part on the experimental context.

Taken together, the two experiments do not support the theory forwarded by Grant (2001) and Grant and Seitz (2000), that the primary mechanism of the audiovisual speech detection enhancement effect is perception of the fine-grained correlation between lip area and acoustic amplitude peaks, and that this correlation is used by top-down attentional mechanisms. First, the data with the static rectangle in Experiment 1 show that fine-grained correlations between the sound

and the lips are not required for the effect. Second, the additional speech advantage found in Experiment 1 can be interpreted as a reduction in stimulus uncertainty by providing a pre-cue to the location of an upcoming acoustic peak, rather than an effect of audiovisual correlation.

Previously, Schwartz et al. (2002) conducted a syllable identification experiment in −9 dB SNR, using babble noise. Performance was extremely inaccurate except for the identification of consonant voicing. This was attributed to making use of the temporal cue afforded by preliminary mouth gestures. A top-down attentional strategy was hypothesized to be the mechanism responsible. In (Schwartz et al., 2003), the initial lip gesture was specifically investigated and found to improve audiovisual speech identification, even though the visual information specific to the consonant identity was replaced by a video sequence that was fixed across syllables. When, in their second experiment, they replaced the lip gestures with a red bar that increased and then decreased in height, no audiovisual gain was obtained. Given that theirs was an identification experiment and was conducted at a more favorable SNR, generalization across theirs and our experiments is hazardous. Nevertheless, both supported some role for pre-cuing in enhancing performance, and both showed that pre-cuing is a relatively fragile effect. Given the longer visual (Foxe and Simpson, 2002; Krolak-Salmon et al., 2001) than auditory (Naatanen, 2001; Steinschneider et al., 1999; Yvert et al., 2001) system processing latencies described earlier, a mechanism that helps to initiate auditory attention in advance of the acoustic stimulus could provide a strategic advantage.

In fact, none of the results obtained in the current study are support for speech-specific mechanisms in enhancing the auditory speech detection thresholds. If the preliminary mouth gesture were a cue that automatically engaged speech-specific mechanisms, it might be expected that it would function whenever present. But that is not what was found in Experiment 2.

All of the visual stimuli (speech and non-speech) could have participated in early (possibly sub-cortical, superior colliculus) bottom-up, excitatory–excitatory neural mechanisms that lead to response gain under multisensory stimulus conditions, particularly when a stimulus is at threshold level (Meredith, 2002). This type of response apparently does not require visual stimulus feature analysis beyond locating audiovisual correspondence in time and/or in space (Stein and Meredith, 1993).

6.1. Conclusions

This study examined whether fine-grained audiovisual correlations are responsible for the AV detection enhancement effect reported by Grant (2001) and Grant and Seitz (2000). Here, comparisons between speech and non-speech stimuli were used to investigate what is special about audiovisual speech processing, and what is more likely attributable to more general audiovisual processing capacities. Our results with the simple static visual non-speech (non-phonetic) stimulus suggest that fine-grained correlations are not the basis for the effect. Overall, the results did not support the hypothesis that the primary mechanism of the AV detection enhancement depends on perception of brief high positive correlations between lip area and acoustic amplitude peaks. The results across Experiments 1 and 2 support the conclusion that speech can provide a pre-cue that enhances acoustic speech detection, but that the cue use is relatively fragile. The comparison across experiments raises a cautionary note for attributing effects to hypothesized mechanisms. Because the only visual speech stimulus in Experiment 1 produced a significant advantage, the possibility was raised that visual speech was somehow special. Results from Experiment 2 were not consistent with that possibility. Consideration of known neurophysiological processing constraints informed the design of the current study. The results were seen to be consistent with the possibility that a bottom-up excitatory–excitatory mechanism is responsible for the AV speech detection enhancement effect.

References

American National Standards Institute, 1989. Specification for audiometers (ANSI S3.6-1989), New York: Author.

Arnold, P., Hill, F., 2001. Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. Br. J. Psychol. 92, 339–355.

Bernstein, L.E., Auer Jr., E.T., Moore, J.K., 2004. Audiovisual speech binding: convergence or association?. In: Calvert, C., Spence, C., Stein, B.E. (Eds.), Handbook of Multisensory Processing. MIT, Cambridge, MA.

Bernstein, L.E., Demorest, M.E., Tucker, P.E., 2000. Speech perception without hearing. Percept. Psychophys. 62, 233–252.

Colin, C., Radeau, M., Soquet, A., Dachy, B., Deltenre, P., 2002. Electrophysiology of spatial scene analysis: the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. Clin. Neurophysiol. 113, 507–518.

De Gelder, B., Bertelson, P., 2003. Multisensory integration, perception and ecological validity. Trends Cogn. Sci. 7, 460–467.

Foxe, J.J., Simpson, G.V., 2002. Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. Exp. Brain Res. 142, 139–150.

Grant, K.W., 2001. The effect of speechreading on masked detection thresholds for filtered speech. J. Acoust. Soc. Amer. 109, 2272–2275.

Grant, K.W., Seitz, P.F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Amer. 108, 1197–1208.

Krolak-Salmon, P., Henaff, M.A., Tallon-Baudry, C., Yvert, B., Fischer, C., Vighetto, A., Betrand, O., Mauguiere, F., 2001. How fast can the human lateral geniculate nucleus and visual striate cortex see?. Soc. Neurosci. Abstracts 27, 913.

Levitt, H., 1971. Transformed up-down method in psychoacoustics. J. Acoust. Soc. Amer. 49, 467–477.

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

Meredith, M.A., 2002. On the neuronal basis for multisensory convergence: a brief overview. Cognit. Brain Res. 14, 31–40.

Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J. Neurosci. 7, 3215–3229.

Mesulam, M.M., 1998. From sensation to cognition. Brain 121 (Pt 6), 1013–1052.

Moutoussis, K., Zeki, S., 1997. A direct demonstration of perceptual asynchrony in vision. Proc. Roy. Soc. Lond. B, Biol. Sci. 264, 393–399.

Naatanen, R., 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38, 1–21.

Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Amer. 95.

Puce, A., Perrett, D., 2003. Electrophysiology and brain imaging of biological motion. Philos. Trans. Roy. Soc. Lond. B 358, 435–445.

Ratcliff, R., 1993. Methods for dealing with reaction time outliers. Psychol. Bull. 114, 510–532.

Reisberg, D., McLean, J., Goldfield, A., 1987. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. Lawrence Erlbaum, London, pp. 97–113.

Schroeder, C.E., Foxe, J.J., 2002. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Cognit. Brain Res. 14, 187–198.

Schwartz, J.-L., Berthommier, F., Savariaux, C., 2002. Audio–visual scene analysis: evidence for a "very-early" integration process in audio–visual speech perception. In: Proc. 7th Internat. Conf. on Spoken Language Processing, Denver, CO.

Schwartz, J.-L., Berthommier, F., Savariaux, C., 2003. Auditory syllabic identification enhanced by non-informative visible speech. In: Proc. 7th AVSP (Audiovisual Speech Processing) 2003, St. Jorioz, France.

Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT, Cambridge, MA.

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

Stevens, K.N., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Amer. 26, 212–215.

Treisman, A., 1996. The binding problem. Curr. Opin. Neurobiol. 6, 171–178.

von der Malsburg, C., 1995. Binding in models of perception and brain function. Curr. Opin. Neurobiol. 5, 520–526.

Yvert, B., Crouzeix, A., Bertrand, O., Seither-Preisler, A., Pantev, C., 2001. Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans. Cereb. Cortex 11, 411–423.

Zeki, S., 1998. Parallel processing, asynchronous perception, and a distributed system of consciousness in vision. The Neuroscientist 4, 365–372.