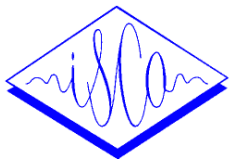

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Enhanced Auditory Detection with AV Speech: Perceptual Evidence for Speech and Non-Speech Mechanisms

Lynne E. Bernstein, Sumiko Takayanagi, Edward T. Auer, Jr.

Department of Communication Neuroscience
House Ear Institute, Los Angeles, CA
lbernstein@hei.org

Abstract

Speech in a noisy or reverberant environment is more detectable and more intelligible when the listener can see the talker. How to explain these perceptual phenomena is a fundamental problem for AV speech research. We have undertaken a series of behavioral and electrophysiological experiments to investigate the perceptual and neural bases for enhanced auditory speech detection in noise with AV stimuli. We hypothesize that the enhancement effect arises due to at least two neurophysiologically distinct mechanisms, one in no way specialized for speech and the other specific to speech stimuli. Here we report results of a perceptual experiment in which an auditory /ba/ token was presented adaptively to obtain its 71% detection threshold [1] in white noise. Participants were tested in three conditions, auditory-only speech, audiovisual speech, and auditory speech with a visual dynamic Lissajous figure. The Lissajous figure was a control for many of the complex visual features of speech. Evidence was obtained for two separate sources of AV detection enhancement: Detection thresholds were highest for the auditory-only speech, lower for the auditory speech with the Lissajous figure, and lowest for the audiovisual speech. Our Discussion section outlines the implications and limitations of the current results for explaining the AV speech detection enhancement effect.

1. Audiovisual Speech Enhancement

Speech in a noisy or reverberant environment becomes first more detectable and then more intelligible when the listener can see the talker [2-4]. Under noisy or otherwise difficult listening conditions, the detection of speech signals could be expected to lead the listener to make various adaptations to improve the communication channel. For example, she might try to move closer, ask the talker to slow his speech or speak louder, or she might initiate manual or facial gestures to augment the information. That is, detection of speech information typically would seem to precede information exchange under difficult conditions.

Grant conducted two experiments on the detectability of spoken sentences in noise [2, 3]. The perceiver's task was to detect the speech in noise under two conditions, either when the talker's face was also presented or when only the audio stimuli were presented. An estimated 2-3 dB reduction in the signal-to-noise ratio (SNR) at the detection level was obtained when the talker could be seen. That is, 2-3 dB of increased noise could be applied when the talker was visible. Although the magnitude of this effect appears small, 3 dB SNR is equivalent to a 50% increase in sound pressure level, and at suprathreshold levels, a 3 dB signal change corresponds to approximately 30% improvement in speech intelligibility.

In order to explain the audiovisual detection enhancement effect, Grant correlated the sound amplitude and mouth area opening of his stimuli. The positive correlation between the two led to the hypothesis that, 'Essentially, watching the variations in the movement of the mouth during speech production informs listeners both when in time and where in the spectrum to expect signal energy. By focusing the attention of the listener to specific spectro-temporal locations in the speech waveform, the ability to hear speech in noisy backgrounds is improved' (2001, p. 2274-2275). This explanation, involving attentional focusing, implies that higher level, top-down, processing is responsible for the enhancement effect.

In other words, one interpretation of this hypothesis is that a perceiver sees the speech movements first and uses what has been seen to focus attention to the peak amplitudes in the acoustic speech signal. But neurophysiological research in humans briefly described below has shown that information from the ear reaches the cortex before information from the eye. Thus, the visual speech information that is hypothesized to be available to focus on the speech spectrum is apparently processed by the brain *after* the sought-for acoustic events have passed.

1.1. Neurophysiological Constraints on AV Speech Processing

Processing times through the auditory and visual pathways set constraints on when and where auditory and visual speech information could possibly interact neurophysiologically [5]. The first volley of stimulus driven activity into the auditory core cortex occurs around 11-20 ms post stimulus onset [6, 7], and the auditory speech percept appears to develop within 150-200 ms post-stimulus onset [8]. 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 [9, 10]. Given the early discrepancy in latencies across auditory and visual primary cortices, and given that trans-cortical processing also requires time [5], interactions between auditory and visual speech information processed via the bottom-up cortical pathways should occur beyond 100 or 150 ms post-stimulus onset. These latencies are at least equivalent to half the duration of the acoustic signal of syllable length. So by the time the visual information is being processed by the cortex at a level specific to speech, the peak amplitude in the syllable has likely passed. This suggests that if the enhancement effect is specific to speech, its explanation is likely different than the one offered by Grant [2, 3].

The finding that acoustic speech is more detectable when the talker can also be seen is not entirely surprising. Research in the human and animal literature has been reported showing that weak acoustic, optical, and vibrational signals are more easily detected when they occur together [11, 12]. One possible explanation for enhanced detectability is that multisensory neurons in the superior colliculus (a sub-cortical structure concerned with stereotyped detection and co-localization of events in extra-personal space) respond more vigorously in the presence of multisensory stimulation than in the presence of single-sensory stimulation [12]. The multisensory superior colliculus neurons are not specialized for speech. If that sub-cortical system contributes to the AV speech enhancement effect, then it is not linguistic information *per se* that is likely relevant. Instead, merely the synchronization of auditory and visual signals of any kind should cause activation at the low-level of the superior colliculus, a generic detection system available to humans and other mammals such as cats and monkeys. Another alternative is that relatively short latency pathways to multisensory cortical convergence sites exist [5], which might account for heightened activity in unimodal cortices through feedback mechanisms [13]. But again, this mechanism need not be specific to speech.

1.2. The Current Study

The current study was undertaken as the first in a series to investigate the perceptual and neural basis for the AV speech enhancement effect. In order to investigate the hypothesis that auditory speech detection can be enhanced with only a non-speech visual stimulus, a dynamic Lissajous figure was used as a control stimulus for the talking face. The Lissajous figure was an oval that changed in its vertical extent as a function of the acoustic speech amplitude. Thus, it presented an audio-to-visual correlation but by itself was not a speech stimulus. If it is the case that only low-level acoustic and optical stimulus information is needed to produce enhanced detection thresholds, the Lissajous figure should be sufficient. If there is, however, a special role for visual speech information given by a talking face, there should be a difference between the condition with the Lissajous figure versus the natural face. Finally, both AV conditions were compared with an auditory-only (AO) condition.

2. METHOD

2.1. Participants

Seven native speakers of American English (2 males and 5 females, ages 18 to 30 years, with mean age of 21 years) participated in this experiment. All had normal hearing with hearing thresholds $\leq$ 15 dB HL [14] at audiometric test frequencies from 250 to 8000 Hz. Their speech reception thresholds in noise were tested using the HINT [15]. Their composite HINT scores (comprising measures of noise front, noise right, noise left) ranged from 69 to 94 percentile of the normal-hearing range, with a mean of 78 percentile. Participants were screened for normal vision, and they were screened to be average or better lipreaders. They were paid $10 per hour for their service.

2.2. Stimuli

The stimuli were an acoustic /ba/ token, its corresponding video /ba/ token, and a synthesized dynamic Lissajous figure whose vertical extent changed as a function of the speech amplitude in the acoustic /ba/ token. The Lissajous figure was synthesized based on the amplitude envelope of the acoustic /ba/ at the field rate of the video speech, that is, 59.94 frames per second.

The speech was recorded with a UVW-1800 SONY Betacam SP recorder and a SONY production camera. The component signal from the Betacam SP video 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 740 x 486.

To prevent perceivers from relying on the timing of the two intervals of each adaptive trial, each condition had a set of trial types that varied in terms of the onset of each of the stimulus intervals. The onset delays spanned 11 steps, with each step equivalent to one video frame (at 33.37 ms/frame).

For each trial type, the sequence of uncompressed frames for the video /ba/ or the Lissajous figure was built into an AVI (Audio Video Interleave) file for software MPEG compression, along with the appropriate frames needed to jitter the onset times. All of the AVI files for the different trial types were transferred to 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 x 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/sec. 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 AVI 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. The audio associated with the different trial types in video were stored in a separate file uncompressed with a sample rate of 48 kHz. As with the video, all of 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 files is performed using custom software that ensures frame locked audio of 8008 audio samples/ 5 video frames. Figure 1 shows the video and Lissajous stimuli. The hyperlinks demonstrate the AV stimuli.

Speech signal level was set following a pilot experiment with five participants. In the pilot study, the speech was presented at 65 dB SPL. The mean detection thresholds in dB SPL were: (1) AO, -15.8; (2) AVL, -16.2; and (3) AVS, -17.1. But 65 dB SPL for the speech signal, which is comfortable for speech alone, requires unacceptably high noise levels to obtain detection thresholds when the video is presented. Therefore, the speech level for this experiment was reduced to 60 dB SPL. The noise masker was white noise. Noise was stored in a long audio file. Each noise interval was selected at random from the large file and output through a sound card. The onset trigger for the noise was a signal recorded on the second audio track of the DVD.

2.3. Procedure

There were three different conditions in the experiment. In each condition, the experimental task used a two-interval forced-choice adaptive threshold paradigm. In this task, there were two observation intervals for each trial, and the perceiver indicated whether the acoustic speech token occurred in the first or second interval. The adaptive rule was as follows: Two correct responses and the noise was increased, and one incorrect response and the noise was decreased. The rule converges on the 71% detection threshold [1]. At the beginning of testing, the acoustic signal was -4 dB below the level of the noise. After the first ten adaptive reversals at decreasing step sizes (2, 1, .5, .2, dB), the noise levels then changed in .1-dB increments. For each trial, the interval with the signal and the temporal jitter in stimulus onset were randomly selected. Twelve reversals were used to estimate each threshold.

Thresholds were obtained in three conditions, auditory-only speech (AO), audiovisual speech (AVS), and audio /ba/ with the visual Lissajous figure (AVL). During the AO testing, only one interval contained the audio /ba/. During the AVS and AVL testing, only one interval contained the audio /ba/, and both intervals contained the visual stimulus (visual speech—AVS or the Lissajous figure—AVL) (see Figure 2). Each participant received testing in all of the conditions. The order of conditions was randomized within sets of the three types of conditions, so that each participant completed four randomized sessions. Participants were asked to respond as quickly and accurately as possible. They responded 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. Prior to testing, participants received a set of practice trials to learn the procedure.

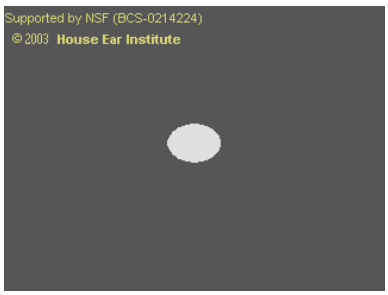

Figure 1. Video speech to (AV ba.avi) and dynamic Lissajous figure bottom (AV Lissajous ba.avi).

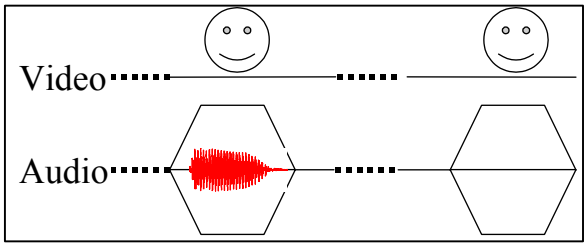

Figure 2. Two-interval stimulus presentation format. The audio speech occurred randomly in one interval only. The noise masker occurred in both intervals. A variable time interval occurred at the onset of each observation intervals. For AV testing, each visual stimulus was placed within the interval in what would be the correct synchrony with the speech. The video was either a talking face or a Lissajous figure.

3. Results

Figure 3 shows the mean results for the seven participants. Each participant contributed 24 threshold estimates across three conditions. A systematic effect occurred such that AO speech thresholds were higher than were AVL thresholds, which in turn were higher than AVS thresholds.

Multivariate analysis of variance showed that the main effect of the stimulus type was significant [F (2, 5) = 49.50, p = .001]. The mean threshold for AO was –15.23, for AVL was –17.28, and for AVS was –18.60. The analysis of the simple contrasts between stimulus types showed that the difference between AO and AVL was significant [F (1, 6) = 55.92, p < .001], and the difference between AO and AVS was significant [F (1, 6) = 82.22, p < .001]. The difference between AVS and AVL was also significant [F (1, 6) = 10.06, p = .019]. Interestingly, as Figure 3 shows, all of the participants improved their thresholds with the AVL stimuli, but not all of them obtained an additional improvement with the AVS stimuli.

Figure 4 shows the mean response times across all of the SNRs that were presented during the adaptive runs in the fourth session. The means in the figure were computed only on responses that were correct. These response time results were not statistically reliable, likely due to the anomalous increase in some response times with the AVL stimuli in the –13.1 to –16 dB SNR range. Nevertheless, the results are suggestive: The AVS response times were longer than the AO response times, consistent with additional processing involved in using the visual information. Also, in the region of the AVL and AVS thresholds, the AVL response times were shorter than the AVS response times, consistent with the Lissajous figure involving less processing. At supra-threshold levels (-13 dB SNR or higher), response times were equal across conditions, suggesting perhaps that the visual stimuli did not contribute much at these levels.

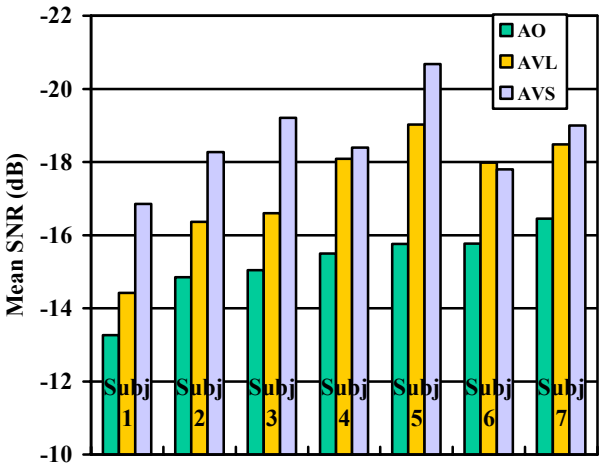

Figure 3. Mean thresholds for three conditions obtained with five subjects.

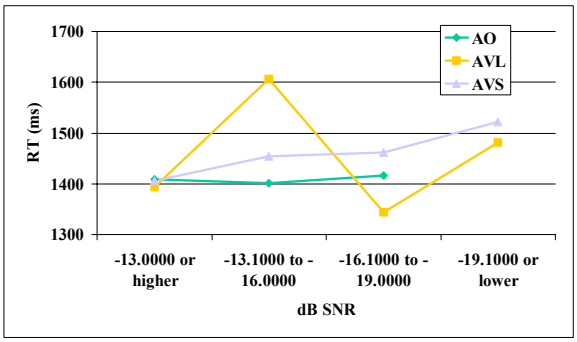

Figure 4. Mean response times for correct responses in the first stimulus interval. Times obtained during the fourth test session.

4. Discussion

The perceptual results reported here are consistent with the hypothesis that there are at least two neurophysiological mechanisms that might lead to the enhancement of speech detection in noise with AV stimuli. The results showed that a simple dynamic Lissajous figure decreased thresholds, but the talker's face further decreased thresholds. Response time measures were consistent with the view that the combining of the AV speech stimulus information involves some later processing than that for AO stimulus information.

The effect of the Lissajous figure could be accounted for by a sub-cortical neural process such as multisensory neurons in the superior colliculus [12]. It might also be accounted for by a cortical interaction that is not specialized for speech, such as feedback from the visual system to the auditory system [13, 16]. The additional increment in the AV speech thresholds could be accounted for by a process specific to speech but not specific to processing phonetic speech features. For example, it could be related to a higher level of attention, when the stimulus is the more complex face image than the Lissajous figure.

Electrophysiological studies of AV effects with simple non-speech stimuli have demonstrated early AV effects attributable to attention or expectancy and sensitive to the type of stimulus [17-19]. Alternatively, the possibility remains that a genuinely phonetic process is responsible for the threshold shift with AVS.

Bernstein, Auer, and Moore [20] have observed that the brain mechanisms responsible for speech processing likely are complex and non-linear, and the expectation that the translation between perceptual theory and neural implementation could be simple is likely to be overly optimistic (see also, [21, 22]). That is, perceptual evidence is insufficient by itself to determine processing mechanisms. Therefore, we intend to study the AV speech detection enhancement effect using recordings of event-related potentials. These electrophysiological measures can be used to obtain evidence concerning the latency and brain locations at which perceptual effects arise.

At the same time, we believe that neural explanations cannot be developed in isolation from additional investigations of perceptual effects. For example, an alternative possibility to the two-process hypothesis needs consideration: Perhaps, the Lissajous figure is sufficiently similar to a talker's mouth opening movements that it receives speech processing. Although this possibility must be considered in light of the fact that (1) by itself the Lissajous figure does not give the impression of speech, and (2) in combination with the just-detectable auditory speech stimulus, the Lissajous figure is not identifiable as /ba/; some further control conditions are needed.

In an ongoing experiment, we are presenting a static oval for the same duration as the movement by the Lissajous stimuli. This control for the speech-driven dynamics of the visual stimulus. A second control condition is a rectangle that expands in the horizontal direction as a function of speech amplitude, to control for the mouth-like appearance of the dynamic Lissajous figure.

5. Conclusions

Evidence was obtained consistent with the existence of two processes responsible for the enhancement of auditory speech detection by a dynamic visual stimulus. A Lissajous figure driven by the speech envelope, but otherwise without visual speech features, resulted in a significant decrease in auditory speech detection thresholds. An additional decrement in the SNR at threshold was obtained when the talker's face was presented. This pattern of results could be due to sub-cortical detection processing, perhaps, at the level of the superior colliculus. The additional decrement in the threshold with visible speech could be due to a higher level process specific to speech, such as a higher level of attention to a talking face. Or it could be due to the specific processing of visual phonetic information. Additional perceptual experiments are needed along with experiments using event-related potentials to refine our understanding.

6. Acknowledgements

This research was supported by the National Science Foundation (BCS-0214224. We thank Brian Chaney and John Jordan for their engineering assistance.

7. References

[1] Levitt, H., "Transformed up-down method in psychoacoustics", *J. Acoust. Soc. Am.*, *49*(2 Suppl. 2): 467-477, 1971.

[2] Grant, K.W., "The effect of speechreading on masked detection thresholds for filtered speech", *Journal of the Acoustical Society of America*, *109*(5): 2272-2275, 2001.
[3] Grant, K.W. and P.F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences", *J. Acoust. Soc. Am.*, *108*(3 Pt 1): 1197-208, 2000.
[4] Sumby, W.H. and I. Pollack, "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, *26*: 212-215, 1954.
[5] Schroeder, C.E. and J.J. Foxe, "The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex", *Cognit. Brain Res.*, *14*: 187-198, 2002.
[6] Steinschneider, M., I.O. Volkov, M.D. Noh, P.C. Garell, and M.A. Howard, 3rd, "Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex", *J. Neurophysiol.*, *82*(5): 2346-57, 1999.
[7] Yvert, B., A. Crouzeix, O. Bertrand, A. Seither-Preisler, and C. Pantev, "Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans", *Cereb. Cortex*, *11*: 411-423, 2001.
[8] Naatanen, R., "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, 2001.
[9] Foxe, J.J. and G.V. Simpson, "Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing", *Exp. Brain Res.*, *142*: 139-150, 2002.
[10] Krolak-Salmon, P., M.A. Henaff, C. Tallon-Baudry, B. Yvert, C. Fischer, A. Vighetto, O. Betrand, and F. Mauguiere, "How fast can the human lateral geniculate nucleus and visual striate cortex see?" *Soc. Neurosci. Abstracts*, *27*: 913., 2001.
[11] Welch, R.B. and D.H. Warren, *Intersensory interactions*, in *Handbook of perception and human performance, volume I: Sensory processes and perception*, J.P. Thomas, Editor. Wiley, NY, p. 25-1-25-36, 1986.
[12] Stein, B.E. and M.A. Meredith, *The Merging of the Senses*, Cambridge, MA, MIT, 1993.
[13] Calvert, G.A., M.J. Brammer, E.T. Bullmore, R. Campbell, S.D. Iversen, and A.S. David, "Response amplification in sensory-specific cortices during crossmodal binding", *Neuroreport*, *10*(12): 2619-23, 1999.
[14] American National Standards Institute, *Specification for audiometers (ANSI S3.6-1989)*, New York, Author, 1989.
[15] Nilsson, M., S.D. Soli, and J.A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise", *J. Acoust. Soc. Am.*, *95*(1085-1099), 1994.
[16] Molholm, S., W. Ritter, M.M. Murray, D.C. Javitt, C.E. Schroeder, and J.J. Foxe, "Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study", *Cognit. Brain Res.*, *14*: 115-129, 2002.
[17] Fort, A., C. Delpuech, J. Pernier, and M.H. Giard, "Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans", *Cereb. Cortex*, *12*(10): 1031-1039, 2002.
[18] Giard, M.H. and F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study", *J. Cogn. Neurosci.*, *11*: 473-490, 1999.
[19] Teder-Salejarvi, W.A., J.J. McDonald, F. Di Russo, and S.A. Hillyard, "An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings", *Brain Res. Cogn. Brain Res.*, *14*(1): 106-114, 2002.
[20] Bernstein, L.E., E.T. Auer, and J.K. Moore, *Audiovisual Speech Binding: Convergence or Association?*, in *Handbook of Multisensory Processing*, B.E. Stein, Editor. MIT, Cambridge, MA, in press.
[21] Friston, K.J., C.J. Price, P. Fletcher, C. Moore, R.S. Frackowiak, and R.J. Dolan, "The trouble with cognitive subtraction", *Neuroimage*, *4*(2): 97-104, 1996.
[22] Picton, T.W., S. Bentin, P. Berg, E. Donchin, S.A. Hillyard, R. Johnson, Jr., G.A. Miller, W. Ritter, D.S. Ruchkin, M.D. Rugg, and M.J. Taylor, "Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria", *Psychophysiology*, *37*(2): 127-152, 2000.