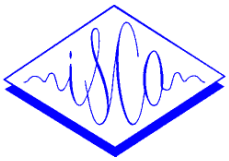
Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

IS PRIMITIVE AV COHERENCE AN AID TO SEGMENT THE SCENE?

J.P. Barker, F. Berthommier and J.L. Schwartz

Institut de la Communication Parlée, UPRESA CNRS No 5009, INPG,
46 Av. Félix-Viallet F38041 Grenoble Cedex 1, FRANCE
Tel no: +33-(0)4.76.57.48.38 Fax: +33-(0)4.76.57.47.10
(barker, bertho, schwartz)@icp.inpg.fr



ABSTRACT

In this paper we propose the existence of an audio-visual scene analysis (AVSA) module which is able to integrate primitive information from both auditory and visual modalities. This module forms correspondences between the auditory and visual streams based on primitive properties of either representation. Through these correspondences visual information is employed to aid the segregation of acoustic sources. This enhanced segregation may be partly responsible for the increased intelligibility of audio-visual speech.

The paper presents the initial results of a planned series of audio-visual speech experiments designed to test this account. Specifically, the experiments reported here address the question of whether visible movement of the speech articulators may protect speech from the effects of masking by noise. It is shown that a reduction in temporal uncertainty due to visual information may reduce the *detection* threshold for CVs in noise. The same performance increase was not observed in a parallel experiment testing consonant *identification*.

1. INTRODUCTION

It is well known that visual information may be employed to improve speech intelligibility in noisy conditions. It is generally considered that this intelligibility increase arises from the integration of the limited information separately contained in the auditory and visual signals. However, a recent study involving two simultaneous speakers has demonstrated that seeing the lip movements of speaker A can increase the intelligibility of an *unseen* speaker B (Driver, 1996 [2]). This result lies outside the scope of this simple model of audio-visual integration.

In this paper we propose the existence of a module that performs what may be termed audio-visual scene analysis (AVSA). In this module a correspondence is formed between the auditory and visual representations based on primitive properties of either representation i.e. spatial position, coherence of temporal modulations. This correspondence forms in a manner that is equivalent to the primitive grouping occuring between elements of the auditory representation in the traditional auditory scene analysis (ASA) account [1]. By interacting at this level the visual information may aid the segregation of the auditory sources. It is this enhancement of the auditory segregation that is at least partly responsible for the increased intelligibility of AV speech i.e. the speech is heard more clearly because the lip movements make it 'stand out' from the noise. This is shown schematically in figure 1 where visual information is integrated both at a primitive level in the AVSA module and at a phonetic level during AV speech recognition.

This paper reports preliminary results from two experiments in a series of studies planned to test this account. These two experiments examine the ability of visual cues arising from the speech articulators to protect the acoustic signal from

This work was supported by European Commission Network TMR ERBFMRXCT970150 (SPHEAR)

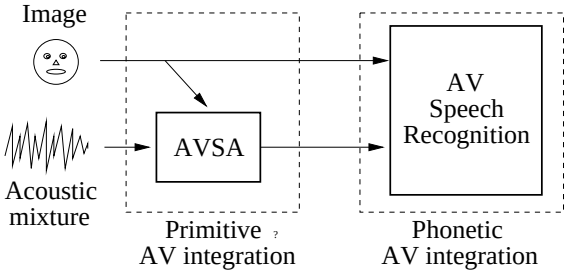


Figure 1: Primitive and phonetic AV integration.

masking. The first experiment examines whether these cues can lower the speech detection threshold. The second asks whether visually induced lowering of the detection threshold also leads to improved identification.

2. EXPERIMENT 1: DETECTION TASK

Repp et al. (1991) were the first to examine the potential influence of the visual modality on the detection of acoustic speech signals [4]. Their results failed to show a change in detection sensitivity thresholds. However, in this study the masking noise was itself modulated with the speech envelope and was therefore equally coherent with the visual signal. A more recent study of Grant and Seitz (1998) using an unmodulated noise masker and spoken sentences found a significant detection threshold reduction in the AV condition [3].

The experiment reported here aims to replicate Grant and Seitz's result employing simple tokens of /ga/ and /da/. Using these simple tokens rather than the sentences employed by Grant and Seitz's makes it possible to carefully control the visual information supplied. The control of the visual information is essential to the *identification* task described in section 3.

The stimuli for the *detection* experiment are designed such that there exists a large temporal uncertainty in the location of the target token relative to a long period of stationary masking noise. Is it possible to exploit the cue in the visual stream to help locate the token in the *acoustic* stream? If this audio-visual correspondence can be made then the temporal uncertainty of the stimuli is reduced and the tokens can be detected with greater reliability. This can be measured by comparing performances in audio-visual and audio-only conditions.

2.1. Method

Subjects

Eight volunteers were recruited from staff at the Institut de la Communication Parlée served to make up two groups of subjects. All subjects were adults (ages 24–30 years) who reported normal hearing in both ears, and who had normal speech and language abilities. All subjects were native speakers of French.

Stimuli

A single French speaker recorded a sequence consisting of 24 instances of the token /ga/ and 24 instances of the token /da/ occurring in a randomly permuted order. Each /ga/ and /da/ was preceded by an unvoiced initial vowel /a/ with a variable duration uniformly distributed between one and five seconds (see figure 2). This unvoiced vowel served to give the consonant and voiced vowel an unpredictable position within the noise band, and to strengthen the lip movement cues. Being unvoiced the initial vowel was below the level of the masking noise and inaudible to the subjects (i.e the articulation is that of a VCV, but the stimulus is heard as a CV). A computer driven 'auto-cue' was employed to provide the speaker with two prompts for each token: i) the first prompt signalled the speaker to begin the initial vowel, ii) the second prompt - occurring after a random delay - signalled the moment at which to produce the consonant, and the identity of the consonant which was to be produced. Using this technique anticipatory cues to the consonant identity are minimised.

The material was recorded in a sound proof chamber directly onto Betacam video tape. The speaker wore blue lip make-up to enable Chroma-key processing of the video image - this processing was not necessary for the detection experiment but is important for the identification experiment described later.

The recorded speech was lowpass filtered and digi-

tised using a 16 bit D/A converter and a sampling rate of 16 kHz. The level of each token was normalised so that the RMS energy measured over the voiced region was at a fixed value. A pink noise masker (i.e. equal energy per octave) was added at -12 dB SNR (as measured over the voiced region of the token). The noise began one second prior to the initial vowel onset, and ended two seconds after the latest possible time of consonant onset. Prior to adding the noise, a randomly selected set of half the tokens (i.e. 24) were was replaced by silence. The subjects' task is to detect the stimuli containing the /ga/ and /da/ items.

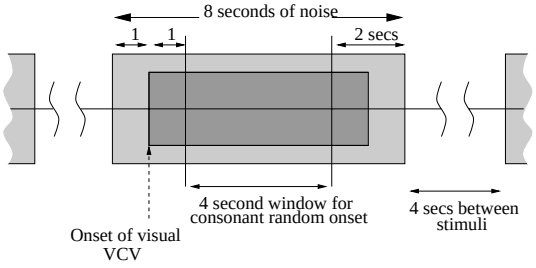


Figure 2: Construction of the VCV stimuli. The consonant may occur any time within the temporal window starting at two seconds and ending at five seconds after the onset of the masking noise.

Stimuli were created in both an audiovisual (AV) and audio-only (A) condition. Both the AV and A conditions were constructed from the same recorded block of 48 tokens and both had the same set of 24 tokens removed. However, in each condition the 48 stimuli were presented in a different random order.

For the AV condition, the video image was centred on the speaker's lips and included the speaker's neck and lower face, cutting off just below the level of the eyes and just above the speaker's shoulders.

Procedure

All subjects heard the stimuli in both the A and AV conditions. One subject group was presented with the AV condition followed by A (AV→A); for the other group the presentation order was reversed (A→AV).

The tests were performed in a sound proof chamber. The acoustic signal was delivered binaurally using headphones and an absolute sound level of approximately 60 dB SPL. The video was presented on a colour television set with the screen positioned approximately 3 metres in front of the subject.

Listeners were informed that half of the stimuli were tokens of /ga/ or /da/ masked by pink noise and that for the remaining 50% no speech was present. For each stimulus listeners had to make a forced-choice reporting whether or not they believed the speech token to be present. They were also required to supply a confidence rating ranging from 1 (unsure) to 3 (confident).

2.2. Results

Figure 3 shows the evolution of performance of each group of listeners when responding to the 1st of the two conditions (A or AV) that made up the experiment. Those who were first presented with the stimuli in an audio-only condition score little above chance levels throughout the session and show no significant learning effect. This contrasts with the steadily increasing performance of the listeners who were presented with the AV condition first. By the final block of 16 stimuli there is a large difference between the two groups with average performance in the AV task at 90% while that in the audio task is around 60%. A comparison of percentages test shows this difference to be significant at the $p < 0.01$ level. For none of the three blocks is the performance in the audio condition significantly above the 50% chance level. It is concluded that while subjects cannot detect the targets in the audio-alone condition, they may learn to detect them in the AV condition.

Figure 4 shows results for the final block of 16 stimuli for both groups in both conditions. The group who were presented with the audio condition first and failed to learn the task were able to perform the task well in the AV condition scoring over 90%. Interestingly, the group who were initially presented with the AV condition in which they performed well, appeared able to carry some of this high performance over to the subsequent audio condition where they scored over 80% (significantly higher than the audio-alone score of 58% achieved by the group who were presented with the audio-alone condition first).

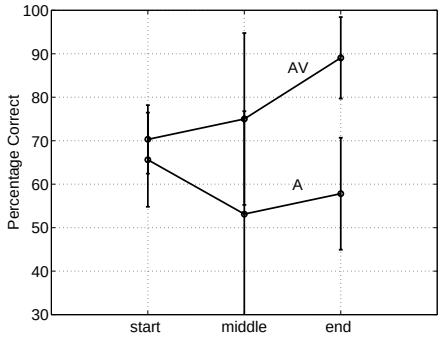


Figure 3: Detection scores averaged across listeners for A and AV when presented as the first condition. Scores are shown for the first, second and third block of 16 stimuli which compose the 48 stimuli sequence. Error bars show one standard deviation of the distribution of individual scores..

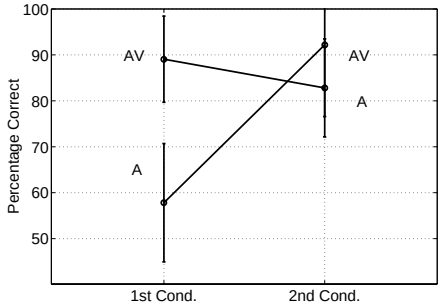


Figure 4: Detection scores averaged across listeners for the final stimuli block of the A and AV conditions for both the (A→AV) and (AV→A) orders of presentation.

3. EXPERIMENT 2: IDENTIFICATION TASK

The previous experiment demonstrated that a visual lip-movement cue can significantly increase the detectability of the tokens /ga/ and /da/ when there temporal location is uncertain. Subjects attested to the compellingness of the effect reporting that the tokens were perceived to be more audible in the audio-visual condition. However, does this apparent increase in audibility actually improve subject's identification performance? The visual cue may make the speech signal appear to 'stand out' but is it really heard with greater fidelity?

The second experiment addresses this question through an AV /ga/ and /da/ identification task. These tokens can be distinguished acoustically by detection of the difference in F2 transition. However, under the conditions of the experiment they are visually indiscriminable. The experiment compares listeners performance in an audio-alone and an audio-visual condition. Even though the visual information is of no direct benefit, the AVSA hypothesis suggests that subjects may be able to use the visual cues indirectly to facilitate the processing of the acoustic information - e.g. by enhancing the segregation of the speech tokens and the background noise. If this is the case then identification performance may be significantly higher in the audio-visual condition.

3.1. Method

Subjects

Six volunteers recruited from staff at the Institut de la Communication Parlée served as subjects. All subjects were adults (ages 25–30 years) who reported normal hearing in both ears, and who had normal speech and language abilities. All subjects were native speakers of French.

Stimuli

The experiment employed three sequences of 48 /ga/ and /da/ tokens which had been recorded in an identical manner and at the same time as the material employed in the previous experiment (see section 2.1 for details). A Chroma-key system was employed to process the video image so that only the lips remained visible. Hence, any visemic information arising from the face or from the inside the mouth (most importantly, the tongue) was removed.

Each token was randomly assigned an SNR of either -6 dB, -7 dB, or -8 dB such that there were 16 tokens at each noise level (note that the noise levels employed are all significantly lower than in the detection experiment). As previously, a pink noise masker was employed which began one second before the onset of each speech token and ended two seconds after the latest possible occurrence of the consonant.

Procedure

Subjects were presented with the three sequences of stimuli with the first sequence in a full audio-visual condition, the second with the visual in-

formation removed (i.e. audio-only) and the final sequence with the audio information removed (i.e. video only). This final sequence serves to validate that the tokens are indeed visually indiscriminable.

Listening conditions were the same as those described in the previous experiment. The video was displayed in black and white to reduce the distracting effect of the blue lip colouring.

For each stimulus listeners were asked to report whether they heard the consonant to be /g/ or /d/. They were also asked to supply a confidence rating ranging from 1 (unsure) to 3 (confident). In cases where they failed to hear the token they were told to respond with '?'. The experiment was scored by calculating the percentage of tokens correctly identified. In the audio-alone condition the '?' response was scored as being half correct so that chance performance remained at 50%.

3.2. Results

Figure 5 shows the identification scores averaged across listeners for each of the three conditions, and for each condition separate scores are shown for the first and last 24 stimuli of each 48 stimuli set. As expected, results for the visual-only condition are not significantly above the chance level of 50% and the scores for the second half of the block of stimuli are no greater that those for the first. This suggests that the measures taken to remove the visual cues for the /ga/ and /da/ tokens were successful.

For the audio-alone condition the mean identification score is around 60% and a textitt-test shows this result to be significantly above chance level ($p < 0.05$). The 'not heard' response was employed for only 4% of the tokens (and was never used in the AV condition). This confirms that the SNRs employed rested somewhere around the listeners identification threshold.

However, no systematic improvement is seen in the mean scores for the AV condition for the range of SNRs employed. No evidence of an effect is seen in the overall results (figure 5), neither can evidence be found at any of the individual noise levels (see figure 6).

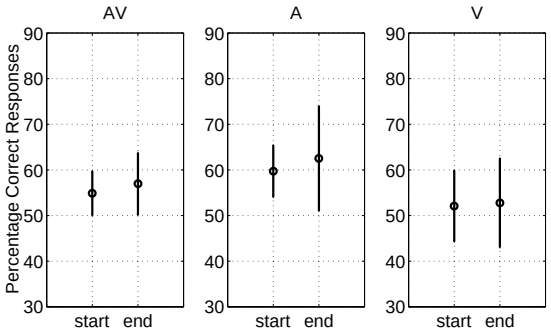


Figure 5: Identification scores averaged across subjects for the three conditions AV, A, and V. For each condition scores are calculated separately for the first and second half of the set of stimuli presented. Error bars show one standard deviation

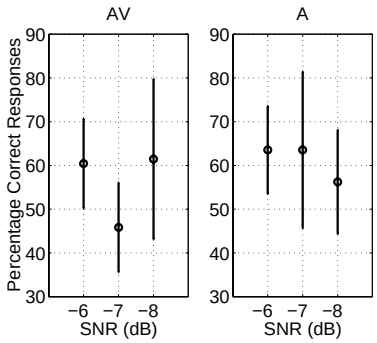


Figure 6: Identification scores averaged across subjects for conditions AV and A calculated for each of the three noise levels. Error bars show one standard deviation.

An examination of individual performances reveals large intra-subject variation (figure 7). There is possibly a pattern in that subjects who are not significantly above chance performance in the audio-alone condition may benefit from the visual information. Listeners who hear the tokens well enough to score above chance (the majority) have worse scores in the AV condition. A larger group of subjects would be needed to establish whether this pattern is genuine.

4. DISCUSSION AND CONCLUSION

Two experiments have been reported. The first confirms the result of Grant and Seitz demonstrating that a visual cue can improve speech detectability by reducing temporal uncertainty. The second experiment failed to find evidence that a reduction in temporal uncertainty can benefit

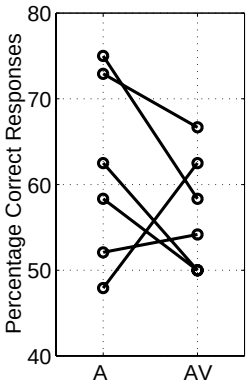


Figure 7: Overall identification scores for individuals for conditions A and AV.

phonetic identification.

It is worth considering the second experiment a little more closely before concluding that primitive audio-visual integration does not aid phonetic identification. There are features of the experimental design which may account for the failure to find a positive effect. First, most subjects unexpectedly reported that they found the AV task difficult to perform. This perhaps indicates that subjects 'over-attend' the visual stream in a futile attempt to locate non-existent discriminative cues. The large attentional resource expended may limit that which is available to the more informative auditory stream. This effect of visual distraction opposes the benefit that the temporal visual cue may have provided. The visual distraction may also be a symptom of the unnaturalness of the task - i.e. that of associating disembodied lips with acoustic speech signals - and if so, it could perhaps be reduced by more extensive training.

A second problem with the experiment arises due to the inter-subject variation in detection and identification thresholds. The visual cue is likely to have most impact over a narrow range of SNRs at a level relative to the thresholds of each individual. It is therefore likely that the fixed SNRs employed in the experiment were inappropriate for some of the subjects. Some of the high identification scores achieved in the audio-alone condition suggest that an effect may have been achieved at lower SNRs closer to detection thresholds. In any case, it appears that if visual masking protection is to benefit identification of these tokens, then the effect is small and a more sophisticated adaptive experimental procedure is required.

Although these preliminary results remain largely inconclusive, the work highlights problems that must be addressed in future experiments:

- *Sensitivity*: Designing a task that is highly sensitive to the effects of primitive AV integration.
- *Specificity*: Isolating the effects due to primitive integration from those of later occurring AV phonetic integration.
- *Attention*: The need to control the effects of attention across the conditions employed.

Future experiments are planned to determine whether visual information related to the masker can provide masking release (i.e. an AV analog of comodulation masking release). Synthetic AV stimuli will be employed and it is hoped that the greater control afforded may help to overcome the problems that have been discussed.

5. REFERENCES

[1] A.S. Bregman. *Auditory scene analysis.* MIT Press, Cambridge, Mass., 1990.

[2] J. Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381:66–68, 1996.

[3] K.W. Grant and P.F. Seitz. The use of visible speech cues (speechreading) for directing auditory attention. Presented at the *135th Meeting of the Acoustical Society of America*, Seattle, WA, June 1998.

[4] B.H. Repp, R. Frost, and E. Zsiga. Lexical mediation between sight and sound in speechreading. *Haskins laboratories status reports on speech research*, SR-107/108:243–254, 1991.