AUDIO-VISUAL SCENE ANALYSIS
Evidence for a “very-early” integration process in audio-visual speech perception

Jean-Luc Schwartz, Frédéric Berthommier, Christophe Savariaux
Institut de la Communication Parlée (ICP) (Speech Communication Institute)
CNRS UMR 5009, INPG / Université Stendhal, Grenoble France
schwartz, bertho, savario@icp.inpg.fr

ABSTRACT
Recent experiments suggest that audio-visual interaction in speech perception could begin at a very early level, in which the visual input could improve the detection of speech sounds embedded in noise [1]. We show here that the “speech detection” benefit may result in a “speech identification” benefit different from lipreading per se. The experimental trick consists in using a series of lip gestures compatible with a number of different audio configurations, e.g. [y u t y k y k y d u g y g u] in French. We show that the visual identification of this corpus is random, but, when added to the sound merged in a large amount of cocktail-party noise, vision happens to improve the identification of one phonetic feature, i.e. plosive voicing. We discuss this result in terms of audio-visual scene analysis.

1. AUDIO-VISUAL SPEECH FUSION: LATE, EARLY OR ... VERY EARLY?
The literature on audiovisual (AV) fusion in speech perception is largely organized around the question of the fusion level: late or early whether it follows or precedes phonetic identification. While late-integration models account for a large body of experimental evidence and provide the basis to most developments in the framework of AV speech recognition, a number of experimental data appear incompatible with late integration. Let us mention the AV-VOT problem [2]. AV interaction in the processing of rate and voicing [3] and even difficulties with the McGurk effect [4,5].

Late- and early-integration models share a common assumption of independence of the primitive monosensorial processing. That is, information would be first extracted separately in each sensorial channel before fusion. However, a number of recent studies have raised serious doubts about this assumption. The first study by Grant and Seitz [1] showed that visible movements of the speech articulators allowed to improve the detection of speech embedded in acoustical white noise, with a gain of about 2 dBs. Further experiments [6,7] confirmed this result, and showed that the correlation between energy in the F2-F3 region and the variation of inter-lip separation was the main determinant of the detection improvement (see also [8]). Extraction of auditory cues thanks to visual movements can be understood as some kind of “very early” fusion process, which should occur prior to the fusion/identification stages considered by early- or late-integration models; we shall come back to this question later.

Does this process add anything to the intelligibility of speech in noise? The role of lipreading in understanding noisy speech is quite well-known (since [9]) but the question here is different. Independently of what can be understood per se on the speaker’s lips, and since the seen speaker’s gestures seem to improve the audibility of the sound, does this provide some additional gain to intelligibility? An indirect positive evidence is found in the study by Driver [10] involving two simultaneous speakers, and showing that seeing speaker A’s lip movements could increase the intelligibility of the unseen speaker B. In this case, seeing speaker A provides no phonetic information on speaker B’s speech. However, this experiment is rather global, and it could also be interpreted as some top-down attentional mechanism allowing to group phonetic audio information thanks to phonetic video information, hence its results must be interpreted cautiously. Our aim in the present study is to try to provide some clear new evidence in favour of this “very early” contribution of vision to speech intelligibility, through audibility enhancement.

2. EXPERIMENT
2.1. Principle
We look for a situation in which the visual contribution to audio detection of speech cues would enhance speech intelligibility. However, in most situations, the visual input would also contribute directly to intelligibility improvement thanks to lipreading. Hence, our aim is to reduce this contribution to zero. This provides the trick of our experimental setting, using a series of lip gestures compatible with a number of different audio configurations. If the visual input alone provides no information on the sound, and when added to the noisy sound it happens to improve its identification, this will be interpreted as no direct lipreading benefit, but as a gain due to enhanced audio detection of speech cues thanks to vision.

In a previous experiment based on this reasoning, we studied the audibility and intelligibility of visually similar [da] and [ga] sequences embedded in a large amount of white noise [11]. This experiment first confirmed that vision improves audio detection. Then, in an identification experiment, only the lips were provided thanks to a chroma-key process, hence the visual-only stimuli (V) were almost indistinguishable. Unfortunately, we observed no benefit of the audiovisual (AV) condition over the audio-only (A) condition. However, in addition to some caveats raised in the original paper, we noticed a posteriori two factors that could have blurred a possible effect in the original setup. Firstly, only the place contrast [d] vs. [g] had been studied, while the experiment could also have involved other invisible contrasts such as voicing (e.g. [d] vs. [t]) or vowel place (e.g. [y] vs. [u] in French). Secondly, the choice of white noise could be arguable, and stronger effects could be expected with more “unpredictable” noise such as...
cocktail party noise made of merged speakers, which is by the way more in line with the two-speakers experiment by Driver [10]. Hence the setting of the present experiment.

2.2. Method

2.2.1. Corpus

We used ten French syllables: [y u t y k y k u d y d y g u], involving a mode contrast between no consonant, or a voiced or unvoiced plosive; a plosive place contrast between dentals [t d] and velars [k g]; and a vowel place contrast between front [y] and back [u]. All these stimuli are associated to basically the same lip gesture towards a rounded close vowel.

2.2.2. Procedure

We prepared an experimental design with large temporal stimulus uncertainty, to increase the chance that the visual contribution to auditory detection would be significant. A French male speaker (the first author of this study) recorded three times each stimulus in a random order, and with a variable amount of silence between two consecutive utterances. This resulted in a set of inter-stimuli silences varying between 0.9 and 4.2 s, with a mean of 2.2 s and a standard deviation of 0.9 s. 

A posteriori examination of the stimuli showed that the last two utterances (which happened to be [du] and [g u]) were technically incorrect, hence the target sequence was finally made of 28 stimuli (3 repetitions of 10 sequences, minus 2 stimuli) and lasted roughly 80 s. The recording was made audio-visually, with a 3-CCD camera and Betacam SP storage, in a sound-proof room, the visual image centred on the speaker’s face. Then a cocktail-party noise (http://citoyens.pays-after.com/Marmotte3D/Pageextra/Pagebruitsages/ambiance/AMBI107 .mp3) was added to the sound, with a mean SNR around −9dB (measured as the ratio of the mean noise power to the mean power of the vocalic portions of the target sequence).

Eight subjects participated to the task. Firstly, they were shown the original AV tape without noise, just to become familiar with the stimuli it contained, and they were asked to identify each stimulus and repeat it to allow the experimenter know how it had been identified. Then they had to do the same task with the noisy tape, presented with or without the visual input: half subjects passed the AV condition before the A one (group 1 in the following) and half passed the A condition before the AV one (group 2). Finally, all subjects passed a visual-only condition (V) with no audio input. In each condition, they were asked to identify each utterance of the target sequence, within a close set [y u t y k y k u d y d y g u]. The task being done in real time hence with some time pressure, some stimuli were not identified at all, or identified as another category (such as [a], [bu] or [sl]). The experimenter, present in the room, noted all the answers. We checked that all stimuli were correctly detected in all three conditions and particularly in the A one, hence the SNR was above the detection threshold.

2.3. Results

All scores in the following are computed thanks to formula (1) (corrected scores to account for random answers):

\[
\text{corrected\_score} = \frac{\text{number\_correct\_answers}}{\text{number\_stimuli} - \frac{1}{\text{number\_categories}}} - 1
\]

Corrected scores vary between 0 for random responses (or possibly less than 0) to 1 for all correct answers. No-responses were replaced in the computations by equal probabilities of all categories. It appeared that all scores were similar between groups 1 and 2 (no significant learning effect in the task), hence the two groups were merged in the following scores.

2.3.1. Global identification scores

All global identification scores are close to zero for all conditions (Table 1).

2.3.2. Identification scores for each phonetic contrast

Then we analyzed the responses in terms of the identified mode (no plosive vs. voiced plosive vs. unvoiced plosive), plosive place (dental vs. velar) and vowel place (front vs. back). The scores are reported in Table 1. They are all close to zero (random responses), except mode identification in the AV conditions. Analysis of the corresponding confusion matrix (Table 2) displays the reason of this rather high score. It appears that, while there is no clear perceptual difference between the “no-plosive” and “unvoiced plosive” conditions, the voicing contrast seems well perceived in the AV condition.

2.3.3. More on voicing

Therefore, we focussed on the voiced vs. unvoiced contrast, dropping the six vowel utterances. Since there are 12 voiced plosives and only 10 unvoiced plosives for technical reasons (2 incorrect stimuli, see above), we dropped the last two voiced utterances to define a restricted even set with 10 voiced and 10 unvoiced plosives. Then we considered two groups of perceptual responses: either “voiced”, or a global “unvoiced” category grouping all other responses (including “unvoiced”, “no plosive” or “no response”). The corresponding voiced-unvoiced confusion matrices are reported in Table 3. From these matrices, we can make the following analyses:

- The voicing identification score, defined as in (1), is very low (0.1) for the V condition, and not significantly different from 0 (chi2(1)=1.60, NS).

- It is significantly higher than 0 in both the A (score=0.34, chi2(1)=18.2, p<0.001) and AV (score=0.60, chi2(1)=57.6, p<0.001) conditions; but significantly higher in the AV than in the A condition (0.60 vs. 0.34, chi2(1)=7.1, p<0.01).

- We can use these matrices to compute d’ scores as in the Signal Detection Theory [12] evaluating a psychophysical normalized distance between the two categories. This provides d’ values respectively equal to 0.25 (V), 1.00 (A), and 1.68 (AV). Since d’ values close to zero correspond to no perceptual contrast while values higher than 1 signal an important contrast, this confirms the previous portrait.

3. DISCUSSION

3.1. Visual contribution to the audibility and intelligibility of audio speech cues

Our results display a clear effect of vision, enhancing A intelligibility without directly providing any significant visual cue differentiating vowel place, or plosive place or mode. The effect is restricted to plosive mode. The negative finding on the other two studied contrasts agrees well with the negative finding on the place contrast [d] vs. [g] in the previously cited study [11] with the same paradigm. How can this positive result on voicing be explained? The gain could be due to improved
prevoicing detection. Remember that in French there is an important prevoicing phase for voiced plosives. In our corpus, the mean duration of this phase is about 110 ms. During our experiment, we observed that all stimuli were detected – if not correctly identified – in the A condition. This means that vowel nuclei were always audible. But the low-frequency prevoicing component is much weaker than the vowel nucleus. Hence, we can assume that this typical and important speech audio cue was often not detected in the A condition. This pattern may be made clearer by inspection of the mean noise spectrum, superimposed on the spectrum of an [y] vowel nucleus, and of a typical prevoicing spectrum, in our corpus (Fig. 1).

Could vision have improved its detection? Looking at our visual stimuli, it appears that the initiation of the rounding gesture begins about 150 ms before the rounding climax, which is reached quite precisely at the beginning of the vowel (see [13]). In consequence, the visual input provides a very clear temporal cue for listening to the possible prevoicing phase, which begins roughly 30-50 ms after the initiation of the lip rounding gesture. This fits well with the psychacoustical case of audio detection without temporal uncertainty, which is known to provide a threshold 2 to 3 dBs lower than detection with temporal uncertainty [14]. This effect is quite likely the basis of our phenomenon. It is coherent with the strong increase of “voiced” responses from the A to the AV condition (more than 50%, see Table 3). Notice that, in this vein, an effect could also be expected on the detection of the high frequency F2 component around 2 kHz for [y]. Indeed, there were about twice as many [u] than [y] in both the A and AV responses, which is likely to be due to the auditory lack of this cue. However, inspection of Fig. 1 shows that the detection of the [y] second formant is quite unlikely. Future experiments should aim at studying what could happen with a noise masker having less energy in the medium frequency range around 2000 Hz.

3.2. AV scene analysis and AV speech perception
Can the effect displayed by the present experiment be called “AV Scene Analysis” (AVSA) and where would it intervene in the course of AV speech perception? Firstly, let us quote Bregman himself in his famous “ASA” book [15]. He announced very clearly that he expected the kind of data we present here: “The correction of one sense by another has great utility. The senses operate in different ways and are good at different things. (...) The best strategy then is to combine the information from the two (ear and eye)” (p. 184). “Although there is no formal research on the role of auditory-visual temporal correspondences in primitive scene analysis there are some informal examples that suggest that they may be used.” (p. 291).

The kind of effect we display here is likely to be an example of multisensorial physical coherence, that could be described as an “audiovisual common fate”, at least in temporal terms. The phonetic extraction/interpretation of audio cues would be guided by the visual information, even though this one is ambiguous and thus phonetically uninterpretable per se.

The fact that the present effect is likely to be “primitive” (or “very early”) does not mean that it could not be speech specific. The appeal to the concept of “Bimodal Coherence Masking Protection” [1] adapted from the audio “Coherence Masking Protection” [16] does not solve the problem. Indeed one position could be that any kind of audio-visual “comodulation” reducing the spectro-temporal uncertainty could improve audio-visual intelligibility. But the ecological “speech nature” of the visual input could be necessary (see the negative finding by Summerfield [17], when he replaced the lips by a simple Lissajous curve varying with the audio amplitude, with no intelligibility enhancement). This suggests that the speech nature of audiovisual coherence could be crucial for audiovisual fusion [18]. This could also be compatible with the “audiovisual” input recently introduced by Massaro in his FMLP [19], though this point deserves more thinking and debate. Whatever the position adopted, it remains that the “very early” aspect of this mechanism fits well with recent electrophysiological data on AV speech perception [20, 21].

3.3. Future experiments
Our major experimental finding is clear, but the experiment could suffer from one criticism. Though the visual information is a priori very weak on all the phonetic contrasts studied, and provides a posteriori identification scores not significantly different from random, we cannot completely eliminate the hypothesis that very weak visual cues could be present, such as small lip area differences between [y] and [u] or small movements in the lower part of the face signalling plosive voicing. Such differences could be too weak to provide significant V intelligibility, but enough to increase the A one. There is a way to precisely discard this hypothesis. It would consist in dubbing on all the audio stimuli the same visual gesture, for example the lip trajectory (extracted by chroma-key) for a typical rounding gesture towards [y] or [u]. This should not be very difficult, since all the lip trajectories are quite similar in our corpus. Demonstration of an (AV>A) effect in this condition would discard the “V cue” hypothesis. Another control experiment could involve replacing the visual speech gesture by a visual non-speech cue providing the same temporal information (e.g. a visual flash synchronous with the lip movement initiation, or with the whole trajectory), to test the role of the “speech” nature of the visual input in AVSA.

Last but not least, we can imagine a “technological counterpart” of this work, exploiting joint audiovisual processing techniques in order to better process and enhance speech before identification or transmission. Since a few years, we attempt to develop algorithms to enhance noisy speech thanks to the visual input [22], and we are presently developing an “AV source separation” algorithm, from “blind source separation” techniques plus an additional term quantifying the audiovisual coherence between sound spectrum and lip characteristics [23]. This and other techniques should be developed in the future.

Acknowledgements
This work was partly funded by the EC program TMR-SPHEAR and is a part of the ESPRIT-BR project RESPITE.

REFERENCES
Table 1 – Global scores and identification scores for individual phonetic features in the three conditions
Corrected scores computed according to (1)

<table>
<thead>
<tr>
<th></th>
<th>Mode</th>
<th>Plosive place</th>
<th>Vowel place</th>
<th>Global score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Condition</td>
<td>0.07</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>AV Condition</td>
<td>0.30</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>V Condition</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 2 – Mode confusion matrix in the AV condition
Percentage of each perceptual response (lines) for each occurrence mode (columns)

A Condition: corrected score=0.34, chi2=18.2

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>68</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>12</td>
<td>39</td>
<td></td>
</tr>
</tbody>
</table>

AV Condition: corrected score=0.60, chi2=57.6

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>63</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>17</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

V Condition: corrected score=0.10, chi2=1.6

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>47</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>33</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 – Voicing identification
Number of “voiced” and “unvoiced” responses (grouping, in the second case, “no plosive”, “no response” and “unvoiced” choices) to sequences containing voiced and unvoiced plosives, in the three conditions

A Condition: corrected score=0.34, chi2=18.2

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>68</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>12</td>
<td>39</td>
<td></td>
</tr>
</tbody>
</table>

AV Condition: corrected score=0.60, chi2=57.6

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>63</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>17</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

V Condition: corrected score=0.10, chi2=1.6

<table>
<thead>
<tr>
<th></th>
<th>Uttered plosive</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived unvoiced</td>
<td>47</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Perceived voiced</td>
<td>33</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 – Noise and speech acoustic spectra
Long-term noise spectrum in dashed, short-term spectrum of the prevoicing component (solid) and of the vowel nucleus [y] (dotted).