

Seeing to hear better: Evidence for early audio-visual interactions in speech identification

Jean-Luc Schwartz, Frédéric Berthommier, Christophe Savariaux

Institut de la Communication Parlée,

CNRS-INPG-Université Stendhal, Grenoble, France

Running Heading: Seeing to hear better

Total number of words (including abstract / non including bibliography and figures): 2907

Corresponding author:

Dr. J.L. Schwartz

Institut de la Communication Parlée,

INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex 1

France

Email: schwartz@icp.inpg.fr

Abstract

Lip reading is the ability to partially understand speech by looking at the speaker's lips. It improves the intelligibility of speech in noise when audio-visual perception is compared with audio-only perception. A recent set of experiments showed that seeing the speaker's lips also enhances *sensitivity* to acoustic information, decreasing the auditory detection threshold of speech embedded in noise (Grant & Seitz, 2000; Grant, 2001). However, detection is different from comprehension, and it remains to be seen whether improved sensitivity also results in an *intelligibility* gain in audio-visual speech perception. In this work, we use an original paradigm to show that seeing the speaker's lips enables the listener to hear better *and hence to understand better*. The audio-visual stimuli used here could not be differentiated by lip reading *per se* since they contained exactly the same lip gesture matched with different compatible speech sounds. Nevertheless, the noise-masked stimuli were more intelligible in the audio-visual condition than in the audio-only condition due to the contribution of visual information to the extraction of acoustic cues. Replacing the lip gesture by a non-speech visual input with exactly the same time course, providing the same temporal cues for extraction, removed the intelligibility benefit. This early contribution to audio-visual speech identification is discussed in relationships with recent neurophysiological data on audio-visual perception.

Keywords: Lipreading; audio-visual interactions; speech detection; speech identification; temporal cues.

Introduction

Lip reading, that is the visual identification of speech gestures from the moving face, improves the intelligibility of speech in noise when audio-visual perception is compared with audio-only perception (Sumbly & Pollack, 1954; Erber, 1975). Another part of our experience as listeners is that speech sounds “louder” when we look at the speaker, as if audio cues were visually enhanced. This has been demonstrated in auditory detection experiments, where the auditory threshold was decreased by 1 to 2 dBs when the sound was accompanied by a lip gesture (Grant & Seitz, 2000). Analyses of correlations between area of mouth opening and energy fluctuations in different spectral bands showed that for utterances with larger correlation, the visually-driven detection gain was larger: hence, the auditory-visual temporal coherence was basic there (Grant, 2001; Kim & Davis, 2003a).

The question that arises is whether this “speech enhancement” effect due to visual information also improves the intelligibility of speech in noise. In other words, does the gain in *detection* result in a gain in *intelligibility*? The answer to this question is not trivial, since these tasks are quite different, and probably involve different mechanisms in the human brain. Searching experimental evidence raises a serious problem. Indeed, it is very difficult in a speech comprehension task to disentangle the contribution of lip reading *per se* from the potential contribution of an additional mechanism, that is visually-guided sensitivity enhancement of audio cues. The primary objective of this work is to propose and test an original paradigm in order to demonstrate the existence of such an additional component to audio-visual speech intelligibility, different from lip-reading (Experiments 1 and 2). In addition, the “speech-specific” nature of this component is tested in Experiment 3.

The potential contribution of visually-guided sensitivity enhancement of audio cues would be displayed if we could show an intelligibility gain due to visual information using a paradigm that completely eliminated the contribution of lip reading. Our experiments were designed to do just that: we examined the audio-visual identification of speech stimuli *with identical lip gestures*, embedded in noise. The addition of noise to the stimuli is intended to at least partially simulate real listening conditions—conversations do not typically take place in sound booths and natural settings are often quite noisy. The ability to focus one’s attention and understand speech in such conditions has been called the “cocktail party effect” (Cherry, 1953). In our design, the visual stimulus contains no information about the phonetic content of the sound: it just provides potential cues about when and possibly where (in frequency) the auditory system should expect useful information. We selected the ten French syllables: [y], [u], [ty], [tu], [ky], [ku], [dy], [du], [gy], [gu]. All these stimuli are associated with basically the same lip gesture towards a rounded vowel (either [y], the vowel in *tu* ‘you’ or [u], the vowel in *tout* ‘all’). They involve a “mode” contrast between a voiced or unvoiced plosive consonant (e.g., [ty] vs. [dy]) or no plosive consonant at all ([y]); a “consonant place of articulation” contrast between dentals [t d] and velars [k g]; and a “vowel place of articulation” contrast between front [y] and back [u]. If visual information improves intelligibility, this improvement is not likely to be due to visual information *per se*, since the stimuli are visually similar. Rather, the improvement should be mainly due to visually enhanced detection of acoustic cues.

Experiment 1: Displaying a visual enhancement with visually ambiguous stimuli

Method

Experiment 1 tested the potential contribution of visual information using natural stimuli. A French male speaker recorded each of the 10 stimuli [y u ty tu ky ku dy du gy gu] three times in a random order, with variable inter-stimulus intervals (between 1 and 4 s). This variable temporal rhythm ensured that the time of presentation of each syllable was quite unpredictable. A cocktail-party crowd noise was added to the sound signal, with a signal-to-noise ratio of approximately -9dB (measured as the ratio of the mean power of the vocalic portions of the stimuli to the mean noise power). This manipulation severely decreased the audio intelligibility of the stimuli, while preserving the listener's ability to detect the presence of each stimulus syllable because of the high energy in the vowel nucleus. A panel of eight French subjects with no reported hearing problems then completed an identification task in three conditions: audio-only (A), video-only (V) or audio-visual (AV) presentation (Fig. 1a). Half completed the audio-visual (AV) condition, then the audio-only (A) one, half did the reverse order. All subjects completed the visual (V) condition last. They were asked to identify each stimulus and repeat it aloud.

Results

The responses of the subjects were grouped into confusion matrices, from which we computed global A, V or AV identification scores thanks to formula (1) (corrected scores to account for random answers, see Robert-Ribes et al., 1998):

$$corrected_score = \frac{\frac{number_correct_answers}{number_stimuli} - \frac{1}{number_categories}}{1 - \frac{1}{number_categories}} \quad (1)$$

Corrected scores vary between 0 for random responses (or possibly less than 0) to 1 for all correct answers. We also analysed 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). All the scores are reported in Table 1. It appears that they are all quite low, except for mode in the AV condition. Further analysis of the mode confusion matrices showed that the only distinction that was partly identified was the presence vs. absence of a voiced plosive, that is, [gy gu dy du] vs. [ty tu ky ku y u]. We therefore focussed all further analyses on this distinction. The percentage of identification of the presence vs. absence of a voiced plosive for the three conditions is presented on Fig. 1b. The results show that AV intelligibility was significantly higher than A intelligibility ($\chi^2(1)=6.0, p<0.05$), while V intelligibility was, not surprisingly, very poor.

Discussion

The likely explanation of this effect is displayed in Fig. 2. The typical acoustic structure of a consonant-vowel syllable is illustrated in Fig. 2*a*. In French, voiced plosives such as [d] or [g] begin with a low frequency prevoicing bar: before the tongue leaves the palate to produce the consonant, the glottal source has already begun its action, and is audible by the listener as a major voicing cue (Lisker & Abramson, 1964). Next comes an acoustic burst, and a spectral transition towards the vowel target (vowel nucleus). In Fig. 2*b*, we see that the speaker begins to move his lips towards the rounded vowel [y] or [u] about 100 ms before the beginning of the prevoicing bar, if it exists, and 240ms before the vowel nucleus. This provides a temporal cue likely to improve the detection of upcoming acoustic cues, and particularly the prevoicing bar indicating the presence or absence of a voiced plosive.

Experiment 2: Removing any residual lip reading cues

Method

In the previous experiment dealing with natural audio-visual stimuli, we could not yet rule out the possibility that there remained small visual voicing cues that could have provided the increase in intelligibility from the A to the AV condition. The purpose of Experiment 2 was to discard any possibility that vision could enhance intelligibility through direct lip reading, using stimuli in which different sounds were dubbed onto a fixed lip gesture. For this purpose, a new recording of 30 stimuli was made with the same speaker. The speaker's lips were coloured using blue make-up, to allow precise video analyses using a chroma-key process (Lallouache, 1990). We first performed various analyses of the lip profiles and acoustic content of all the stimuli to verify that they were globally compatible with the portrait in Fig. 2. We selected an utterance with a rounding gesture beginning at a lip area value of 1.1 cm² typical of the "basis" period, ending at a lip area value of 0.3 cm² typical of the rounded target, and changing from the first to the second value in exactly 120 ms. We added a 400 ms rounded plateau and a 240-ms unrounding gesture coming back to the 1.1 cm² basis. This fixed 760 ms video stimulus was dubbed onto all 30 acoustic stimuli, in the precise configuration displayed in Fig. 2. We ensured that the lip onset phase occurred at least 100 ms before the acoustic prevoicing phase (if there was one) and roughly 240 ms before the acoustic initiation of the vowel (beginning of the formant transition). As in Experiment 1 stimuli were merged with high-level cocktail party noise, and a new panel of twelve French subjects with no reported hearing problems completed an identification task in two conditions: audio-only (A) or audio-visual (AV) presentation. Half of the subjects completed the AV condition followed by the A condition, whereas half did the reverse order. The video-only condition was not used here, since it would have involved 30 repetitions of exactly the same video stimulus.

Results

The results are displayed in Fig. 3, still focussing on the [gy gu dy du] vs. [ty tu ky ku y u] contrast. Percentages of correct identification of the presence vs. absence of a voiced plosive

computed from the 30 stimuli are displayed. Identification is significantly higher in the AV condition compared to the A condition: $\chi^2(1)=3.3$, $p<0.1$. Since the average AV-A difference is rather small, we also compared the number of correct responses per subject, and then evaluated the AV-A difference by a paired t-test. The mean gain is 1.7 correct responses (30 being the total number of responses per subject), which is significantly higher than 0 in a t-test with paired samples ($t(11)=1.95$, $p<0.04$).

Discussion

The significant gain in intelligibility in the AV condition compared to the A condition replicates the result in Experiment 1, but now the interpretation is straightforward. The gain due to visual information cannot be based on lip reading *per se*; it must be due to the enhanced sensitivity to the voicing cues (basically, the prevoicing bar) contributed by the temporal cue of the lip gesture onset. This fits well with psychoacoustic data showing that audio detection without temporal uncertainty provides a threshold by 2 to 3 dBs lower than detection with temporal uncertainty (Egan, Greenberg & Shulman, 1961); but it shows for the first time that the audio detection gain may be converted into a speech audio-visual intelligibility gain, providing, independently of lip reading *per se*, an additional contribution to the visual enhancement of auditory speech intelligibility.

Experiment 3: Testing the speech-specificity of the visual cuing effect

Method

We conducted a third experiment to attempt to test whether the effect displayed in Experiment 2 was speech specific. In Experiment 3, we replaced lip movements by a visual non-speech cue consisting of a red bar appearing and disappearing on a black 720x576 pixel background in synchrony with each stimulus syllable (Fig. 4a). The bar was a rectangle with a width set at 155 pixels and a height equal to 0 in the “basis” period (no bar, just the black background) increasing to 320 pixels by 80-pixel steps in the 120-ms sequence towards the target, stable at 320 pixels during the 400-ms plateau and decreasing back to 0 in 240 ms. The audio tape and the dubbing were the same as in Experiment 2. Hence Experiment 3 literally consisted of replacing a lip rounding-unrounding gesture by a bar increasing and decreasing in height in exactly the same time course. The temporal auditory-visual coherence was therefore preserved, while the nature of the visual input as a speech stimulus was disrupted. Another set of 12 French subjects with no reported hearing problems completed the AV and A conditions in the same way as in Experiment 2.

Results

The percentages of correct identification of the presence vs. absence of a voiced plosive computed from the 30 stimuli are displayed in Fig. 4b. Strikingly, the potential advantage due to visual information disappeared here. Indeed, intelligibility is not significantly higher in the AV

condition compared with the A condition: $\chi^2(1)=0.4$, $p>0.5$. A paired t-test of the AV-A difference in the number of correct responses per subject is also negative: mean AV-A difference 0.6, $t(11)=0.96$, $p>0.1$. Hence the temporal cueing gain provided by visual information seems to be, at least partly, speech specific.

General discussion

This work shows that seeing the speaker's lips enables the listener to better *extract useful acoustic information* embedded in cocktail party noise. This results in an additional increase in audio-visual speech intelligibility, different from lip reading *per se*. In Experiments 1 and 2, while listeners hear all vowel nuclei in the acoustic stimulus, seeing the speaker's lips enables them to extract an additional cue, that is the prevoicing bar, if it exists. The speech understanding system then integrates these features to identify the syllable. These results show that cross-modal interactions can occur early to enhance speech in noise and improve intelligibility.

Visual temporal cueing is likely to provide the explanation of the positive results in Experiments 1 and 2. In this context, the appeal to the concept of “Bimodal Coherence Masking Protection” (Grant & Seitz, 2000) adapted from the audio “Coherence Masking Protection” paradigm (Gordon, 1997) is logical. It would expand “co-modulation” between frequency bands in the audio spectrum to audio-visual co-modulation reducing the spectro-temporal uncertainty and thus improving audio-visual intelligibility. This mechanism, clearly different from lip-reading, provides an “early stage” for audio-visual interaction in speech identification, which could be part of a multisensorial selective listening process (Driver, 1996). It suggests that auditory scene analysis (Bregman, 1990), that is the set of mechanisms that enable the auditory system to separate sounds into streams, could be extended towards audio-visual scene analysis (Barker, Berthommier & Schwartz, 1998).

The negative finding with non-speech visual cues in Experiment 3 is in line with other findings in audio-visual speech detection experiments. Bernstein, Takayanagi & Auer (2003), replacing lip movements with a simple Lissajous curve varying with the audio amplitude, found an audio-visual detection threshold of speech in white noise lower than the audio one, but higher than for audiovisual speech. Kim & Davis (2003b) showed that the AV speech detection advantage was removed by replacing the original video speech stimulus by the synthetic output of a computer-generated talking face. The present results also fit well with a previous negative finding in which replacing lip movements with an audio amplitude driven Lissajous curve lead to no visual benefit in the cocktail party effect (Summerfield, 1979). Notice that the “speech specificity” of the visual cueing effect is not unambiguously demonstrated in the present work. Firstly, the synthetic visual display used in Experiment 3 may have failed to capture critical aspects of the more natural visual display used in Experiments 1 and 2, though the very precise coherence in time makes this assumption a bit unlikely. Second and more importantly, it could well be the case that this display

needs some training by the subjects before it can be used efficiently. Checking this assumption in more detail will need further studies.

The neural bases of early audio-visual interaction are of course not completely clear. A number of recent electrophysiological data suggest that audiovisual integration could indeed happen very early in the perceptual processes, both for speech (Colin, Radeau, Soquet, Demolin, Colin & Deltenre, 2002) and non-speech stimuli (Giard & Peronnet, 1999). Importantly, the fact that the anticipatory lip movement modulates auditory neural processing as early as 50-100 ms after the auditory onset (Lebib, Papo, de Bode & Baudonnière, 2003; van Wassenhove, Grant & Poeppel, 2003) suggests that the 100-ms advance of the visual lip onset phase on the acoustic prevoicing phase in Experiment 2 is suitable for audio-visual efficient interaction. These data are compatible with fMRI evidence that the heteromodal cortex could provide a good site for audio-visual binding and signal enhancement (Calvert, Campbell & Brammer, 2000).

Such early interactions in audio-visual speech perception call for the development of a new generation of models, incorporating a low-level processing stage in the general architecture (e.g. Berthommier, 2003). They provide the basis for new signal processing techniques for joint audio-visual speech analysis, exploiting audio-visual coherence to better process and enhance speech (Sodoyer, Schwartz, Girin, Klinkisch & Jutten, 2002) before transmission or recognition in telecommunication systems or hearing aids.

Acknowledgements.

We thank Jon Barker for participation in a preliminary set of experiments, and Pauline Welby for her help in the preparation of the paper. This work was supported by CNRS, INPG and Université Stendhal.

References

- Barker, J., Berthommier, F., & Schwartz, J.L. (1998). Is primitive coherence an aid to segment the scene? Proceedings of AVSP'98, pp. 103-108. Terrigal, Australia.
- Bernstein, L.E., Takayanagi, S., & Auer, E.T.Jr. (2003). Enhanced auditory detection with AV Speech: Perceptual evidence for speech and non-speech mechanisms. Proceedings of AVSP'2003, pp. 13-18. St Jorioz, France.
- Berthommier, F. (2003). A phonetically neutral model of the low-level audiovisual interaction. Proceedings of AVSP'2003, pp. 89-94. St Jorioz, France.
- Bregman, A.S. (1990). *Auditory Scene Analysis: the perceptual organization of sound*. Cambridge, Mass, Bradford Books, MIT Press.
- Calvert, G.A., Campbell, R., & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, **10**, 649-657.
- Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.
- Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, **113**, 495-506.
- Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, **381**, 66-68.
- Egan, J.P., Greenberg, G.Z., & Shulman, A.I (1961). Interval of time uncertainty in auditory detection. *Journal of the Acoustical Society of America*, **33**, 771-778.
- Erber, N.P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, **40**, 481-492.
- Giard M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, **11**, 473-90.
- Gordon, P.C. (1997). Coherence masking protection in speech sounds: The role of formant synchrony. *Perception & Psychophysics*, **59**, 232-242.
- Grant, K.W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, **108**, 1197-1208.

Grant, K.W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*, **109**, 2272-2275.

Kim, J., & Davis, C. (2003a). Hearing foreign voices: does knowing what is said affect masked visual speech detection? *Perception*, **32**, 111-120.

Kim, J., & Davis, C. (2003b). Testing the cuing hypothesis for the AV speech detection. Proceedings of AVSP'2003, pp. 9-12. St Jorioz, France.

Lallouache, M.T. (1990). Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. Proceedings of the XVIII Journées d'Études sur la Parole, pp. 282-286. Montréal, Canada.

Lebib, R., Papo, D., de Bode, S., & Baudonniere, P.-M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the P50 event-related brain potential modulation. *Neuroscience Letters*, **341**, 185-188.

Lisker, L., & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, **20**, 384-422.

Robert-Ribes, J., Schwartz, J.L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual and audiovisual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, **103**, 3677-3689.

Sodoyer, D., Schwartz, J.-L., Girin, L., Klinkisch, J., & Jutten, C. (2002). Separation of audio-visual speech sources: A new approach exploiting the audiovisual coherence of speech stimuli. *Eurasip Journal on Applied Signal*, **11**, 1165-1173.

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

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

van Wassenhove, V., Grant, K.W., & Poeppel, D. (2003). Electrophysiology of auditory-visual speech integration. Proceedings of AVSP'2003, pp. 37-42. St Jorioz, France.

Vitae

Jean-Luc Schwartz is a member of the Centre National de la Recherche Scientifique. He led the Speech Perception Group at ICP for 10 years. He is now the head of ICP, a laboratory exploring all aspects of speech communication. His main areas of research involve auditory modelling, psychoacoustics, speech perception, auditory front-ends for speech recognition, bimodal integration in speech perception and source separation, perceptuo-motor interactions and speech robotics. He organised with Frédéric Berthommier, Marie-Agnès Cathiard and David Sodoyer the International Workshop “Audio-Visual Speech Processing” AVSP’2003.

Frédéric Berthommier received the M.D. degree from the University of Paris 7 (Lariboisière St-Louis) and his ph.D. in biomedical engineering from the University of Grenoble I in 1992. He is a CNRS researcher in the "Institut de la Communication Parlée" in Grenoble since 1993. His research interests include auditory scene analysis, speech perception, audio-visual speech processing and auditory modelling. He now leads the Speech Perception Group at ICP.

Christophe Savariaux received his Ph.D. in “Signal, Image and Speech Processing” from the "Institut National Polytechnique of Grenoble” in 1995. He is a CNRS Engineer and Researcher in the "Institut de la Communication Parlée" in Grenoble since 1999. His main research interests are focused on speech production, speech pathology and audio-visual speech processing. He is in charge of data acquisition (acoustic, video, electromagnetic, physiological, etc) at ICP.

Table 1. Corrected global scores and corrected identification scores for individual phonetic features in the three conditions of Experiment 1. All scores are computed according to Eq. (1).

Corrected scores	Mode	Plosive place	Vowel place	Global score
A Condition	0.07	-0.03	-0.03	0.02
AV Condition	0.30	0.06	-0.03	0.06
V Condition	0.01	0.05	-0.01	0.02

Figure 1. In Experiment 1, French subjects identified natural utterances of the 10 stimuli in three conditions: AV, A and V(a). Percentage of identification of the presence vs. absence of a voiced plosive (b).

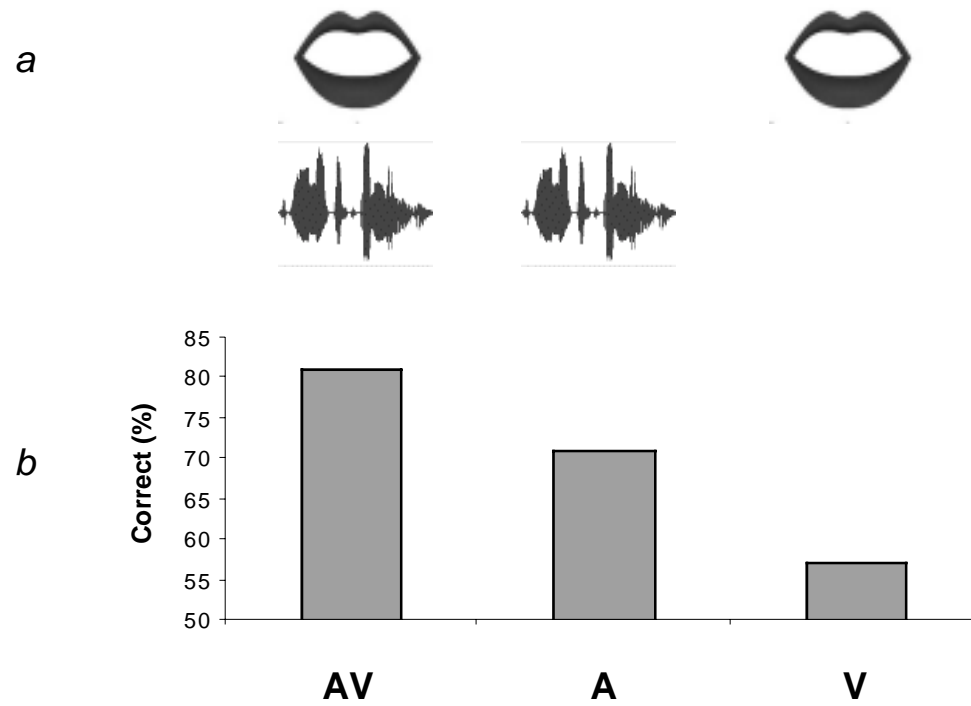


Figure 2. A sketch of the acoustic structure of a plosive – vowel syllable in French (*a*) and of the corresponding lip gesture (*b*).

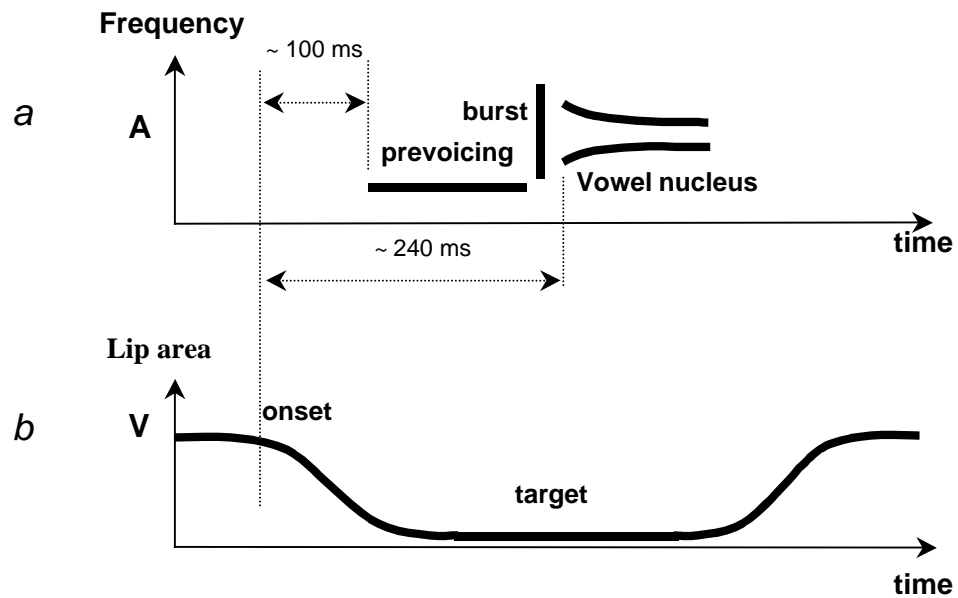


Figure 3. Percentage of identification of the presence vs. absence of a voiced plosive with the dubbed stimuli of Experiment 2.

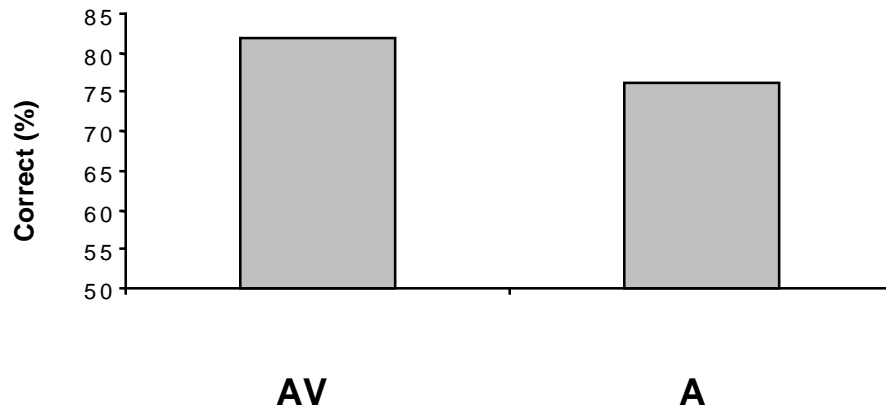


Figure 4. In Experiment 3, the speaker's lips were replaced by a red bar with varying height on a black background (*a*). Percentages of identification of the presence vs. absence of a voiced plosive computed from the 30 stimuli are displayed (*b*).

