AUDITORY SYLLABIC IDENTIFICATION ENHANCED BY NON-INFORMATIVE VISIBLE SPEECH

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 show that seeing lip movements may improve the *detection* of speech sounds embedded in noise. 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 dubbing the same lip gesture on a number of visually similar but auditorily different configurations, e.g. [y u ty tu ky ku dy du gy gu] in French. The visual stimulus does not enable to identify the syllable, but it provides a temporal cue improving the audio identification of these stimuli embedded in a large level of cocktail-party noise, and particularly the identification of plosive voicing. Replacing the visual speech cue (the lip rounding gesture) by a non-speech one with the same temporal pattern (a red bar on a black background, increasing and decreasing in synchrony with the lips) removes the benefit.

1. INTRODUCTION

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. 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 [2,3] 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 [4]).

Nevertheless, detection and identification are very different tasks. Hence the question: 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 [5]) 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 [6] involving two simultaneous speakers, and showing that seeing speaker 1's lip movements could increase the intelligibility of the *unseen* speaker 2. In this case, seeing speaker 1 provides no phonetic information on speaker 2'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 an additional level of audiovisual interaction, preliminary to early or late fusion for phonetic identification, able to provide a "very early" contribution of vision to speech intelligibility, through audibility enhancement.

2. SETTING THE PARADIGM

We look for a situation in which the visual (V) contribution to audio (A) *detection* of speech cues would enhance speech *intelligibility*. However, in most situations, the visual input would also directly contribute to intelligibility improvement thanks to lipreading. Hence, our aim is to reduce lipreading contribution to zero. This provides the trick of our experimental setting: study the AV identification of speech gestures *similar on the lips*, embedded in acoustical noise. In this situation, the visual stimulus would contain no information about the phonetic content of the sound: it would just provide a series of cues about when and possibly where (in frequency) the auditory system should search for useful information. Therefore, if vision improves intelligibility, this will not be due to visual info *per se*, since utterances are visually similar – but because of visually-guided audio detection of acoustic cues.

We selected the ten French syllables: [y u ty tu ky ku dy du gy gu]. All these stimuli are associated to basically the same lip gesture towards a rounded vowel. They involve a mode contrast between voiced or unvoiced plosive or no consonant; a plosive place contrast between dentals [t d] and velars [k g]; and a vowel place contrast between front [y] and back [u]. We prepared a first experiment exploiting this corpus, embedded in a high level of cocktail-party noise [7]. The results were quite clear. While for vowel and plosive place both A, V and AV identification scores were almost 0, for voicing, the AV intelligibility was significantly higher than the A one, in spite of very weak V intelligibility, as expected. However, since we used natural stimuli in this preliminary experiment, we could not rule out the possibility that there remained small V voicing cues that could have provided some phonetic input improving A identification, in spite of the poor V scores. The objective of the present experiment is to exploit the same paradigm, but with a careful control of the V input, discarding any possibility that vision could enhance intelligibility through direct lipreading. For this aim, we prepared stimuli in which the sounds are dubbed on a fixed lip gesture. This is the focus of Exp. 1.

3. EXPERIMENT 1

3.1. Methodology

3.1.1. Original stimuli and audio/video analyses

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 1.2 and 3.9 s, with a mean of 2.3 s and a standard deviation of 0.7 s, for a total duration of 70 s. The recording was made audio-visually with a 3-CCD camera, in a sound-proof room, the video image centred on the speaker's face. Stimuli were recorded by the classical ICP experimental setup, including fixed head, excellent light conditions and blue lips allowing automatic detection of lip contours and area [8].

Then we performed a number of analyses on the stimuli, to extract the temporal coordinations between basic audible or visible events (Fig. 1). On the acoustic signal, we detected the beginning of consonantal prevoicing for voiced plosives (Consonantal Voicing Onset CVO), the burst onset for the voiced or unvoiced plosives (Consonantal Frication Onset CFO) and the beginning of the vowel (Vocalic Voicing Onset VVO: terminology from [9]). On the video signal, we noticed that the lip area was quite stable around 1.1 cm^2 during the preparation phase, and decreased towards a target value around 0.3 cm^2 for all stimuli. We detected the onset of lip movement from the preparation phase towards the target (Lip Onset LO) and the time of arrival at the target (Lip Target LT).

In Table 1, we report the major statistics of the temporal relationships between these events. The coordination between auditory events (columns 1, 2, 3) is rather strict (low standard deviations, around 20 ms) and classical: consonantal voicing for voiced plosives begins in average 157 ms before the vowel, while the burst precedes the vowel by 83 ms for unvoiced plosives and 26 ms for voiced ones; this corresponds to Voice Onset Time (VOT) values respectively around 80 ms for unvoiced plosives and –130 ms for voiced ones (these values are larger than classically reported for French; probably because of the rather slow rate used by the speaker). The visual gesture between LO and LT (column 4) spans over roughly 120 ms. The AV coordination values (columns 5, 6) are crucial for our enterprise. They are less stable (standard deviations around 40 ms), but the important point is the following: the initiation of the visible gesture always precedes sound. Indeed, LO happens at least 40 ms before CVO, with a mean around 100 ms, while the mean LO-VVO distance is 266 ms. This provides an essential temporal cue for enabling audition to focus on the spectro-temporal content of the sound for identification.

3.1.2. Modified stimuli with a fixed lip gesture

The further step consisted in replacing the natural video stimulus by a controlled one, in order to remove any visible information apart from temporal cues. In this procedure, the control of the AV synchrony is crucial. We selected an utterance with a rounding gesture beginning at a lip area value typical of the preparation phase, ending at a lip area value typical of the target, and changing from the first to the second value in exactly 120 ms. Since the video sampling frequency is 25 Hz[1], this provides altogether 5 images: I1, the basis, and I2-I3-I4-I5 the 120 ms-gesture *per se*, I5 being the target. We also selected an offset gesture made of 5 images allowing to come back from I5 to I1: let us call it I_back. Then we replaced the whole video stimulus by a baseline consisting of the repetition of I1, on which we applied for each of the 30 stimuli:

(1) The I2-I3-I4-I5 sequence beginning at a value, that we call LO* (corrected Lip Onset), as close as possible to 240 ms before the vowel onset VVO (it is not possible to perfectly control this, because of the 25 Hz video sampling frequency); in the case of voiced plosives, we imposed that LO* should always happen at least 80 ms before the beginning of prevoicing CVO; the end of the sequence, I5, provides a corrected Lip Target event LT*;

(2) A sequence of 10 repetitions of the target image I5 covering the whole vowel duration for all stimuli;

(3) The I-back sequence towards the basis I1.

Altogether, the whole duration of the dubbed video stimulus from I1 back to I1 is set at 19 images, or 760 ms (see Fig. 1). Columns 7-10 in Table 1 show that the "corrected" lip onset and target events are very close in average to the natural ones, and provide an AV pattern of coordination quite similar to the original one, though with much reduced standard deviations.

Finally, a cocktail-party crowd noise (http://citoyens.pays–allier.com.fr/Marmotte3D/Page_extra/Pagebruitages/ambiance/AMB10!!s.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), letting all stimuli audible.

3.1.3. Procedure

Twelve French subjects with no hearing trouble participated to the task. Firstly, they were shown a video tape, with the audio stimuli without noise, plus the dubbed video stimuli, just to become familiar with the material used. 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 (respectively conditions AV and A in the following).

To be able to study possible learning effects, we prepared two sequences of three tests. Half subjects (group 1 in the following) successively passed conditions A, AV and A, while the other half (group 2) successively passed conditions AV, A and AV. The first two tests (A then AV for group 1, AV then A for group 2) provided a balanced set of comparisons between conditions A and AV, taking into account a potential order effect. The third test aimed at further studying potential learning effects, respectively in the A (group 1) and AV (group 2) conditions. In each condition, subjects were asked to identify each utterance of the sequence, with no instruction about the response set (open choice paradigm).

[1] Let us recall that the video analysis may be done at 50 Hz, exploiting the separation of images in two frames 20 ms apart, but for dubbing, we are stuck to the 25-Hz frequency.

3.2. Results

3.2.1. Global identification scores

We began by studying the first two tests passed by the subjects, that is conditions A then AV for group 1, and AV then A for group 2. In this first stage, we did not consider order effects, hence we summed results over the two groups, and we obtained confusion matrices in the A and AV conditions, computed on 12 subjects. We also 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). Confusion matrices showed that both the two place features were very poorly identified, while the performance was much higher for plosive mode. Analysis of the confusion matrix for mode in the AV condition (Table 2) shows that, while there is no clear perceptual difference between the "no-plosive" and "unvoiced plosive" conditions, the voicing contrast seems well perceived in both conditions. This replicates a pattern of results already obtained in [7].

3.2.2. More on voicing

Therefore, we focussed on the voiced vs. unvoiced contrast. For this aim, we considered two categories of stimuli: a "voiced" category with stimuli containing a voiced plosive, and an "unvoiced" category with the others, that is stimuli containing an unvoiced plosive or no plosive at all. On the perceptual side, we defined two types of perceptual responses: either "voiced" when the subject's response included a voiced plosive, or a global "unvoiced" category grouping all other responses (including an unvoiced plosive, no plosive at all, or no response). The corresponding voiced-unvoiced confusion matrices are reported in Table 3. From these matrices, we can make the following analyses. Firstly, it is obvious that both in the A and AV conditions, the voiced-unvoiced distinction as we defined it is quite efficient. Secondly, voicing discrimination seems higher in the AV condition compared with the A condition. This is confirmed by chi2 analysis (0.63 *vs.* 0.52, $chi2(1)= 3.34$, $p<0.07$). Interestingly, while the scores don't seem to change much in the unvoiced category (82.8% correct A responses, vs. 81.9% correct AV responses), the increase is quite large in the voiced category, from 66% in A to 81% in AV, which is a quite large and significant gain (15%, $chi2(1) = 8.7$, $p<0.005$). Finally, we compared the A and AV conditions for each individual subject. On Table 4, we display the values of the difference of correct voicing responses between the two conditions. The mean gain is 1.67 correct response (30 being the total number of responses per subject), which is significantly higher than 0 in a t-test with paired samples ($t(11)= 1.95$, $p<0.04$). Interestingly enough, it appears that for all subjects but one, the gain is zero (1 subject) or larger than 0 (10 subjects), while for one subject there is a very large decrease in the number of correct responses from the A to the AV condition. It might be suspected that this subject focused too much on the video stimulus, desperately searching visual phonetic cues (actually lacking) to perform the task.

3.2.3. More on learning

Having at our disposal a good performance index, that is the number of correct identification of voicing, we applied it to estimate the role of learning. Indeed, the rather atypical nature of the stimuli – isolated syllables embedded in a very large amount of cocktail party noise – could lead to suspect that learning would be important, and play a potentially confusing role. Therefore, we searched for possible increase of performance from the first to the second to the third test. We performed two kinds of tests. Firstly, we compared the (AV-A) gain for the six subjects of group 1 for which the A condition preceded the AV one, and for the six subjects of group 2 with the inverse pattern: if performance increases with learning, the gain should be higher in group 1. The difference is small: the mean gain between A and AV is respectively 1.83 for group 1 (A then AV), 1.50 for group 2 (AV then A), the 0.33 difference being not significant ($t=0.19$, $p>0.8$). Secondly, we compared the scores in the first and second A session (group 1), and in the first and second AV sessions (group 2). No difference is significant: in the first case there is a mean gain from the first to the second A condition of 1 ($t=0.85$, $p>0.4$); in the second case, there is a mean *decrease* from the first to the second AV condition of 0.25 ($t=0.52$, $p>0.6$). Considered globally and comparing the first and third conditions for both groups (i.e. first and second A sessions in group 1, and first and second AV session in group 2), once again the difference is not significant (mean gain 0.44, $t=0.65$, $p>0.4$). Therefore, we did not find any significant learning effect in Experiment 1.

3.3. Summary

Experiment 1 appears as successful. Vision does increase the intelligibility of acoustic stimuli in spite of conveying no phonetic information *per se*, just because of temporal cueing provided by the lip rounding gesture. Experiment 2 aimed at testing whether this effect was speech specific.

4. EXPERIMENT 2

4.1. Methodology

4.1.1. Stimuli

In this experiment, we replaced the visual speech input of the previous experiment by a visual non-speech cue consisting in a red bar appearing and disappearing on a black 720x576 pixels background. The bar was a rectangle with a width set at 155 pixels and a height equal to 0 in the basis condition I'1 (no bar, just the black background) increasing to 320 pixels by 80-pixels steps from image I'2 to I'5, and decreasing back to I'1 in 3 steps in the I'-back sequence. The audio tape was the same as in Experiment 1, and the dubbing was also conserved from this experiment. Hence experiment 2 literally consisted in replacing a lip rounding-unrounding gesture by a bar increasing-decreasing sequence with exactly the same time course.

4.1.2. Procedure

The procedure was the same as in Experiment 1, with another set of 12 French subjects with no hearing trouble, passing the experiment in two groups: A, AV and A conditions for the six subjects of group 1, AV, A and AV conditions for the six subjects of group 2.

4.2. Results

4.2.1. Voicing

We focussed on the same criterion as in Experiment 1: the scores of correct identification of voicing, defined as

previously. The data for the 12 subjects of groups 1 and 2, computed on the first two series they passed (A then AV in group 1, AV then A in group 2) are displayed in Table 5. It appears that there is still a small gain from condition A to condition AV (4% in corrected scores) but it is not significant ($chi2(1) = 0.36$, $p=0.5$). The study of individual gains from A to AV, shows that there is in fact a decrease in 4 subjects, a zero gain in 1, and an increase in 7 subjects ($t(11)= 0.96$, $p > 0.1$).

4.2.2. Learning

The visual stimuli in the present experiment were not very classical for the subjects (though not completely atypical: they looked like a volume index on a sound system). Therefore, learning could be expected, even more than in Experiment 1. However, the same tests as previously were once again negative. The (AV-A) gain was larger in group 2 (with AV then A: mean gain 0.83) than in group 1 (A then AV: mean gain 0.33), with a not significant difference equal to -0.5 ($t=0.40$, $p>0.6$). The gain increase in both group 1 (mean gain from first to second A session equal to 1, $t=1.37$, $p>0.2$), group 2 (mean gain from first to second AV session equal to 0.5, $t=0.47$, $p>0.6$) and combined groups (mean gain from first to second A or AV session equal to 0.75, $t=1.22$, $p>0.2$) is not significant. Therefore, there is no demonstrated learning effect in Experiment 2 either.

4.3. Summary

Experiment 2 seems unsuccessful. Subjects do not seem to exploit the temporal cueing provided by vision to better identify the A stimuli in noise, though the cue is exactly the same as in Exp. 1, but presented in a non-speech mode. The two repetitions of the AV condition that were offered to subjects in group 2 did not seem enough to provide any significant learning benefit.

5. DISCUSSION

5.1. The lip gesture contributes to the audibility and intelligibility of audio speech cues

The results of Experiment 1 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 voicing. Our understanding is that 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 130 ms). During both Experiments 1 and 2, 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. Then, the visual input provides a very clear temporal cue for listening to the possible prevoicing phase, which begins 95 ms after the initiation of the lip rounding gesture. This fits well with the psychoacoustical case of audio detection without temporal uncertainty, which is known to provide a threshold by 2 to 3 dBs lower than detection with temporal uncertainty [10], and it is coherent with the strong decrease of “unvoiced” responses to voiced plosives from the A to the AV condition (almost 50%, see Table 4).

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] responses in both the A and AV responses, which is likely to be due to the auditory lack of this cue. However, we display on Fig. 2 the mean noise spectrum, superimposed on the spectrum of an [y] vowel nucleus, and of a typical prevoicing spectrum, in our corpus. It appears that the cocktail-party noise is very severe above 500 Hz. This explains why the detection of prevoicing stays possible (particularly with a temporal cue provided by the lips), while 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.

5.2. Possible mechanisms underlying these AV interactions

Neither late nor early integration can explain the present set of results. Late fusion models attempt to fuse an audio decision with a video decision fixed for all stimuli, hence AV intelligibility cannot be improved. For example, an FLMP [11] fit of the data in Table 3 with a fixed visual probability for all stimuli would be quite poor. Early fusion models add a fixed visual cue to a set of audio cues, and there is no reason to believe that AV intelligibility would be different from A intelligibility. In this respect, our results, though reminiscent of the AV VOT effect [12] through the common role played by voicing, are clearly different from all previous data displaying early effects on AV fusion (e.g. [12], [13]).

Extraction of auditory cues thanks to visual movements seems basic here, and it 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. If temporal cueing is indeed the explanation of the positive results in Experiment 1, the further step is to understand how it might proceed in an intermodal way in this paradigm, and why it does not function efficiently with non-speech visual cues, according to Experiment 2. The appeal to the concept of “Bimodal Coherence Masking Protection” [1] adapted from the audio “Coherence Masking Protection” paradigm [14] 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 suggests that Auditory Scene Analysis [15], that enable the auditory system to separate sounds into streams, could be extended towards Audio-Visual Scene Analysis (AVSA).

The fact that the effect does not seem to resist to the replacement of a speech cueing stimulus by a non-speech one in Experiment 2 is reminiscent of a famous pioneer study by Summerfield [16] showing that the ecological “speech nature” of the visual input could be necessary for V enhancement of noisy speech. In that study, the replacement of the lips by a simple Lissajous curve varying with the audio amplitude provided no intelligibility enhancement for speech in noise. AVSA, in this respect, should rather by understood as “Audio-Visual *Speech* Analysis”, implying that there is a set of sensory algorithms processing AV speech in a specific way (not to say “special”!): see e.g. [17]; and a recent prolongation of the AV “sine-wave speech” paradigm confirming the existence of speech-specific effects [18].

6. REFERENCES

[1] Grant, K.W., & Seitz, P. (2000). *JASA, 108*, 1197-1208.
[2] Grant, K.W. (2001). *JASA, 109*, 2272-2275.
[3] Kim, J., & Davis, C. (2001). *Proc. AVSP'2001*, 127-131.
[4] Barker, J., & Berthommier, F. (1999). *ICPhS '99*, 199-202.
[5] Sumby, W.H., & Pollack, I. (1954). *JASA, 26*, 212-215.
[6] Driver, J. (1996). *Nature, 381*, 66-68.
[7] Schwartz, J.L. et al. (2002). *Proc ICSLP'02*, 1937-1940.
[8] Lallouache, M.T. (1990). *Proc. XVIII JEPs*, 282-286.
[9] Abry, C. et al. (1990). *European Bull. of Cogn. Psych., 10*, 269-288.
[10] Watson, C.S., & Nichols, T.L. (1976). *JASA, 59*, 655-668.
[11] Massaro, D.W. (1987). *Speech perception by ear and eye.* London: Laurence Erlbaum Associates.
[12] Breeuwer, M., & Plomp, R. (1986). *JASA, 79*, 481-499.
[13] Green, K.P. (1998). In Campbell *et al.* (eds.) *Hearing by eye, II* (pp. 3-25). Hove (UK): Psychology Press.
[14] Gordon, P.C. (1997). *JASA, 102*, 2277-2283.
[15] Bregman, A.S. (1990). *Auditory Scene Analysis.* Cambridge, Mass, MIT Press.
[16] Summerfield, Q. (1979). *Phonetica*, 36, 314-331.
[17] Remez, R.E. (1996). *Proc. ICSLP'96*, 1660-1663.
[18] Tuomainen, J., et al. (2002). *1st Pan-Amer./Iberian Conf. Acoust.*, Mexico.

Acknowledgements

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

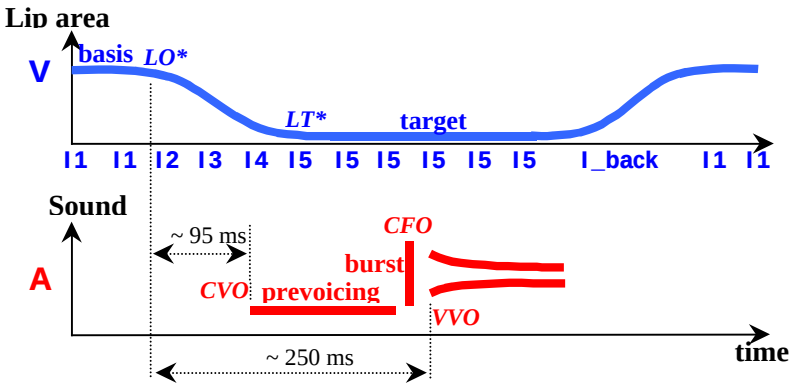

Fig. 1 – Schematic structure of the AV stimuli (see text)

	1	2	3	4	5	6	7	8	9	10
temporal differences	VVO–CVO	VVO–CFO (unvoiced)	VVO–CFO (voiced)	LT–LO	CVO–LO	VVO–LO	LO* – LO	LT* – LT	CVO–LO*	VVO–LO*
min (ms)	120	50	10	80	40	330	-60	-80	80	230
max (ms)	190	110	40	200	160	180	80	80	120	290
mean (ms)	157	83	26	123	100	266	15	12	95	250
std(ms)	25	18	10	23	40	41	42	37	14	14

Table 1 – Temporal differences between pairs of audio and/or video events

Utterance	No plosive	Unvoiced plosive	Voiced plosive
No plosive perceived	0.36	0.21	0.10
Perceived unvoiced	0.36	0.59	0.06
Perceived voiced	0.24	0.15	0.81
No response	0.04	0.05	0.03

Table 2 – Mode confusion matrix in the AV condition, Exp. 1
Percentage of each perceptual response (lines) for each occurrence mode (columns)

A Condition: cor. score=0.52

Uttered stimulus	Unvoiced	Voiced
Perceived unvoiced	179	49
Perceived voiced	37	95

AV Condition: cor score=0.63

Uttered stimulus	Unvoiced	Voiced
Perceived unvoiced	177	27
Perceived voiced	39	117

Table 3 – Voicing identification in Exp. 1
Number of "voiced" and "unvoiced" responses (grouping, in the second case, "no plosive", "no response" and "unvoiced" choices) to stimuli either containing a voiced plosive, or containing an unvoiced plosive or no plosive at all.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
(AV-A) Gain	-6	0	+1	+1	+1	+1	+2	+3	+3	+4	+4	+6

Table 4 – (AV-A) gain for each subject in Experiment 1

A Condition: cor. score=0.48

Uttered stimulus	Unvoiced	Voiced
Perceived unvoiced	183	61
Perceived voiced	33	83

AV Condition: cor score=0.52

Uttered stimulus	Unvoiced	Voiced
Perceived unvoiced	175	46
Perceived voiced	41	98

Table 5 – Voicing identification in Exp. 2
Number of "voiced" and "unvoiced" responses (grouping, in the second case, "no plosive", "no response" and "unvoiced" choices) to stimuli either containing a voiced plosive, or containing an unvoiced plosive or no plosive at all.

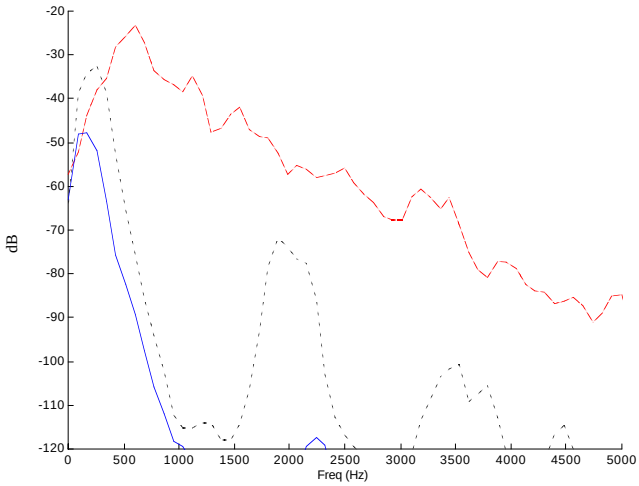

Fig. 2 – Noise and speech acoustic spectra
Long-term noise spectrum in dashed, short-term spectrum of the prevoicing component (solid) and of the vowel nucleus [y] (dotted).