ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HYPER-ARTICULATED SPEECH:
AUDITORY AND VISUAL INTELLIGIBILITY

Denis Beautemps, Pascal Borel, Sébastien Manolios
Institut de la Communication Parlée
CNRS ESA 5009, INPG/Univ Stendhal
46 Avenue Félix Viallet
38031 Grenoble Cedex 1, France

10.21437/Eurospeech.1999-31

ABSTRACT

In the field of speech adaptability to environmental conditions, the intelligibility of hyper auditory and visual speech versus normal speech was investigated. The « Lombard » reflex was used to obtain auditory sequences produced with vocal effort for a set of French /b, d, g, p, t, k/ plosives. Analysis of plosive identifications showed an improvement with the auditory hyper condition in the plosive place of articulation for dentals and velars. Instruction to produce hyper-visual speech provided hyper-articulated labial movements for a set of French /a, i, y/ vowels and /b, v, z, ʒ, l, r/ consonants. Analysis of visual identification scores showed a gain for the hyper-visual condition on the /i/ versus /a/ contrast, but to the detriment of identification of lip movement for consonants — results which were moreover confirmed by the lip shape analysis.

Keywords: intelligibility, hyper audio-visual speech, Lombard speech.

1. INTRODUCTION

This study focuses on auditory and visual speech adaptability to environmental conditions. The feasibility of varying and increasing the speech intelligibility was addressed through the hyper-articulated speech paradigm [1]. The Lombard reflex (e.g. see [2]) was used to obtain auditory data produced with vocal effort. Instruction to produce whispered hyper-visual speech provided hyper-articulated labial movements. The impact on perceptual factors has been analysed in relation to lip contours of facial views for a set of French vowels and consonants.

2. HYPER AUDITORY SPEECH

2.1 The data

A French speaker uttering a set of 28 /VCVsV/ sequences was recorded in quiet environmental conditions and with a 80 dB SPL white noise presented at both ears for a corpus made of the French /a, i, u, y/ vowels, /b, d, g, p, t, k/ plosives. The noisy environmental condition was used to encourage the speaker to produce an auditory vocal effort (the so-called « Lombard reflex ») with a natural hyper-articulation. For each of the two environmental conditions the speaker PB was asked to pronounce the sequence and to repeat it with an emphasis on the first consonant. This ensemble was pronounced three times thus leading to 288 acoustic stimuli, digitized at 16000 Hz.

2.2 The auditory identification test

The contribution of the hyper-articulated auditory speech to the intelligibility of the consonants was evaluated using a procedure of comparative identification under noise degradation of the plosives produced in the quiet condition and in the repeated Lombard speech condition (i.e. with an emphasis on the consonant in the noisy environment). Thus, a subset of 2 x 24 stimuli x(t) composed of the first realisation of the quiet condition and of the first realisation of the repeated Lombard condition were degraded with white noise *b(t)* at seven *t* signal to noise ratio levels and presented to the left ear of 10 native French subjects with no known hearing troubles:

$y(t) = x(t) + \alpha .\ b(t)$ with $\alpha = 10^{-t/20}$ where the power of *b(t)* is identical to that of *x(t)*.

The mean energy of the stimuli *x(t)* of the quiet condition was in average 5.8 dB lower than the mean energy of the repeated Lombard condition. Thus to cancel out this «sound volume effect » the stimuli *x(t)* of the quiet condition were first amplified by 5.8 dB before being mixed with the white noise *b(t)*.

The subjects were asked to identify the plosive from the auditory degraded /VCVsV/ sequences among the proposed /p, t, k, b, d, g/ consonants. They could choose a response /$/ when unable to identify a plosive from the proposed choices.

The analysis of plosive identification scores from the auditory degraded sequences revealed a decrease in the global score with the signal to

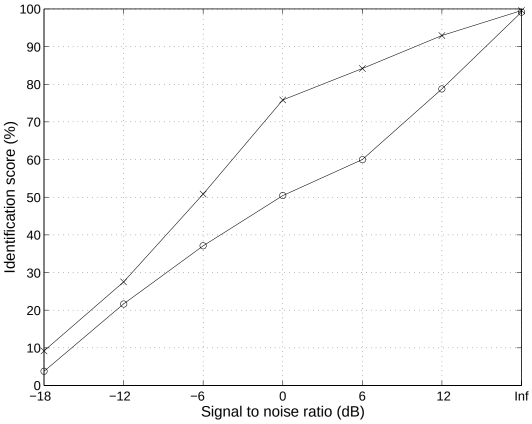

Figure 1: Scores of correct identification (all plosives taken into account) in function of the signal to noise ratio. The repeated Lombard condition (lines with crosses) and the quiet condition (lines with circles).

noise ratio (see *figure 1*) for both speech conditions but with better robustness for the repeated Lombard one: a noticeable identification score of more than 75 % at t=0 dB was obtained, which was nearly identical to the score of the quiet speech condition at t=12 dB.

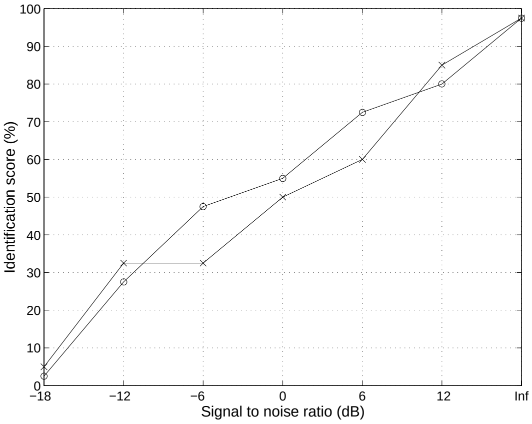

Figure 2: Identification scores for /p/.

More precisely, the analysis of the scores showed an improvement in the velar (see *figures 6 and 7*) and dental place (see *figures 4 and 5*) of articulation in the case of the auditory hyper-articulated speech condition whereas no significant gain was noticed for the bilabials /p/ and /b/ plosives (see *figures 2 and 3*). This later fact is certainly due to the particular lip articulation feature minimising the interaction with the internal part of the vocal tract. Finally, the comparison of the results for each place of articulation showed that in degraded environmental conditions the voicing was less robust than the place of articulation for both speech conditions.

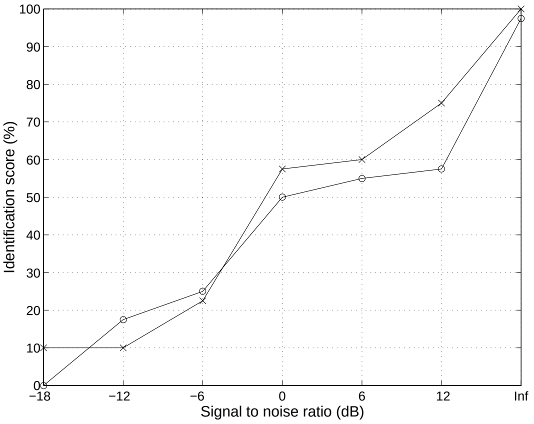

Figure 3: Identification scores for /b/.

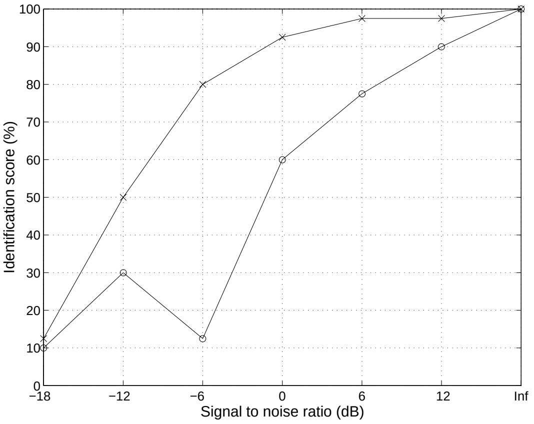

Figure 4: Identification scores for /t/.

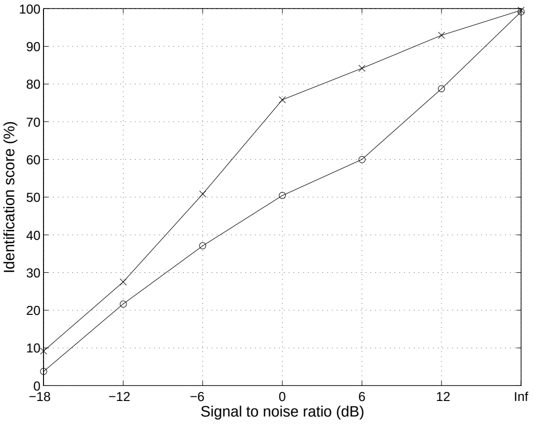

Figure 5: Identification scores for /d/.

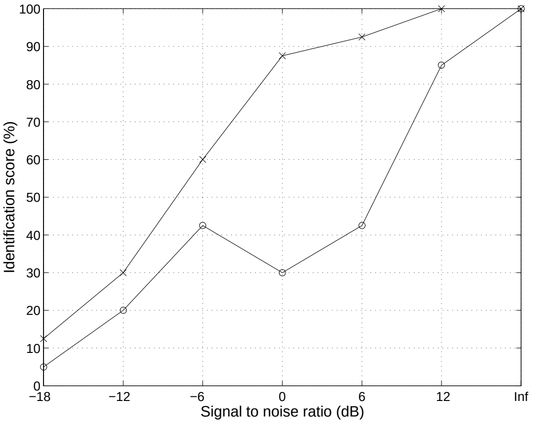

Figure 6: Identification scores for /k/.

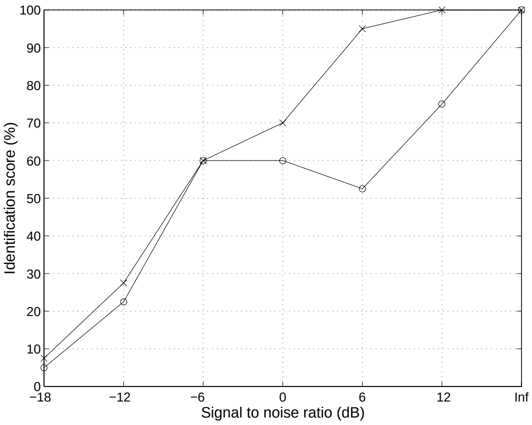

Figure 7: Identification scores for /g/.

3. HYPER VISUAL SPEECH

3.1 The data

A French speaker uttering a set of 63 $/V_1CV_2CV_1z/$ items made of /a, i, y/ vowels, /b, v, z, J, l, r, #/ consonants (absence of consonant noted /#/) was recorded using the ICP audio-visual acquisition system ([3] and [4]). The front and profile views of the face with the acoustic signal were recorded. An audio-visual reference signal was recorded before each pronunciation of the sequences to maintain the synchronisation between the audio and visual parts. The lips of the speaker were painted in blue to allow image processing using the LipSink software (www.ganymedia.com).

The speaker JLS was asked to pronounce the phrase « C’est pas $/V_1CV_2CV_1z/$ » ([5]) with no vocal-tract effort at first (so-called normal condition) followed by three repetitions of the phrase at the same speaking rate but in whispered hyper-visual speech style to obtain hyper-articulated movements of the lips (so-called hyper-visual condition).

The LipSink software was used to extract a set of lip dimensions from the front views of the face at a rate of 50 Hz: the inner lip height (B), the outer lip height (B’), the inner lip width (A) and the outer lip width (A’). The position of the vowels and of the consonants were manually located on the resulting lip trajectories.

This resulted in an audio-visual database for one speaker containing 50 Hz video frames of the front and profile views of the face (between the larynx and the eyes), 50 Hz lip dimensions in synchrony with the acoustic signal for 63 $/V_1CV_2CV_1z/$ sequences in normal condition and 3 x 63 $/V_1CV_2CV_1z/$ with labial hyper-articulated movements. The increase in speaking rate between the normal and the hyper-visual condition is on average 3.5 % (27 ms) and was neglected.

3.2 The visual identification test

Following the approach of Benoît et al. [5], the intelligibility of the visual sequences was quantified for the two speech conditions. The 2 x 63 video facial views made of the normal condition and of the second repetition of the hyper-visual condition of the «C’est pas $/V_1CV_2CV_1z/$» sequence were presented without the corresponding sound to 21 native French subjects. They were asked to identify the $/V_1CV_2/$ part among the /a, i, y/ vowels, /b, v, z, ʒ, l, r, #/ consonants. The subjects used the /?/ marginally for 1.6 % of the responses in the case of no phoneme identification among the proposed choices. The *Figure 8* shows a statistically significant increase (at the 1% risk level) of the identification score for the vowels produced under the hyper-visual speech condition. The analysis of the confusions showed a gain for the /i/ versus /a/ contrast but to the detriment of the identification of lip movement for the consonants, clearly noticed in the increase in the confusion between /b/ and /v/. The effect of coarticulation of /z/ and /ʒ/ with the following vowel, which was better identified in the hyper-visual condition deeply affected the identification rate. With no labial articulation the /l/ and /r/ consonants were not well identified in both speech conditions. However a tendency towards improvement of identification was noted mainly due to a better view of the tongue movement inside the vocal-tract and of visible larynx effort at the neck for the hyper-visual condition.

3.3 Analysis of the lip parameters

From the whole set of data, the mean values and the 2 standard deviation limit of the A and B lip parameters were plotted for the vowels which position was noted V_1, V_2 and V_3 in reference to their position in the $/V_1CV_2CV_3z/$ sequences (see *figures 9 to 11*). A high correlation of 0.84 and 0.97 above the whole set of data was also noticed between the non-null values of A and B with the outer lip parameters A’ and B’ respectively. For V_1 and V_3 positions, the lip width A dimension is the determinant factor in separating the rounding /y/ vowel from the spread lips vowel /i/, whereas the lip aperture B dimension is the determinant factor of the /i/ versus /a/ contrast. The effect of the hyper-visual speech condition is noticeable on the coarticulated V_2 vowel for which the discrimination between the three vowels is consequently improved. In particular the vowel effect in conjunction with the speech condition effect revealed an improvement of the /i/ — /a/ contrast in hyper-visual speech, statistically

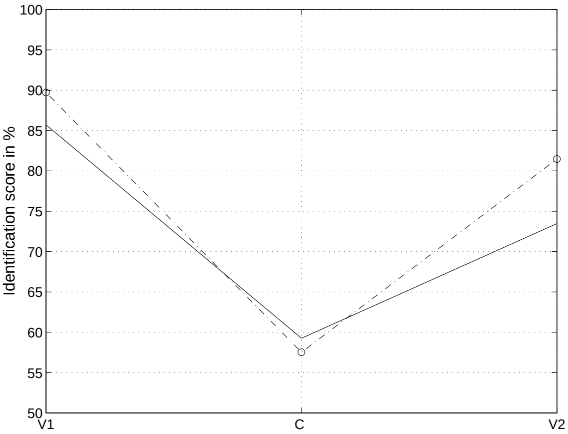

Figure 8: Identification of the /V_1CV_2/ sequences. Hyper-visual speech condition (dotted line with circles) and normal speech condition (thin line).

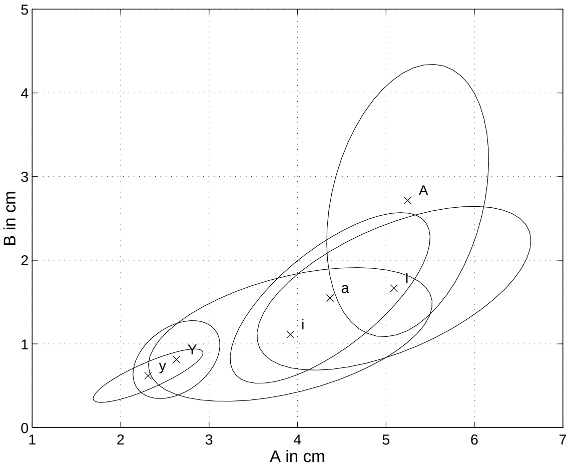

Figure 9: Mean (cross) and 2σ deviation of the vowel V_1 in the (A, B) plan. Hyper-visual condition in upper case.

significant at the 2.5% risk level for each of the three repetitions of the hyper-visual speech condition. This is consistent with the consequent gain in the identification score observed on the perceptual results of the vowel V_2.

4. CONCLUSION

The effect of the auditory Lombard reflex on plosive consonants showed a gain in speech intelligibility for the dentals and the velars but not for the bilabials. The place of articulation is more robust to white noise environmental degradation than the voicing feature. The hyper visual speech on /VCVCVz/ sequences showed that the improvement of the intelligibility was real for the vowels but to the detriment of the consonants.

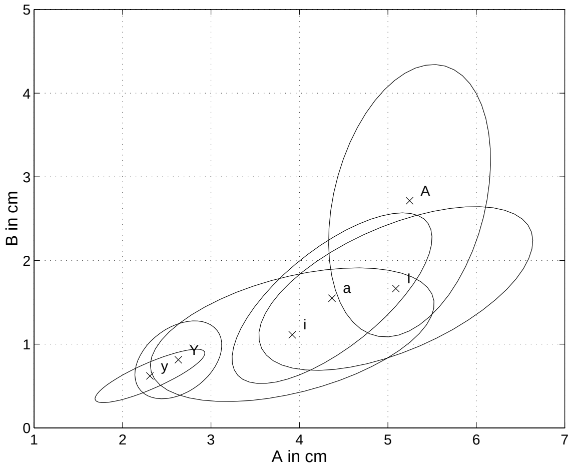

Figure 10: Mean (cross) and 2σ deviation of the vowel V_2 in the (A, B) plan. Hyper-visual condition in upper case.

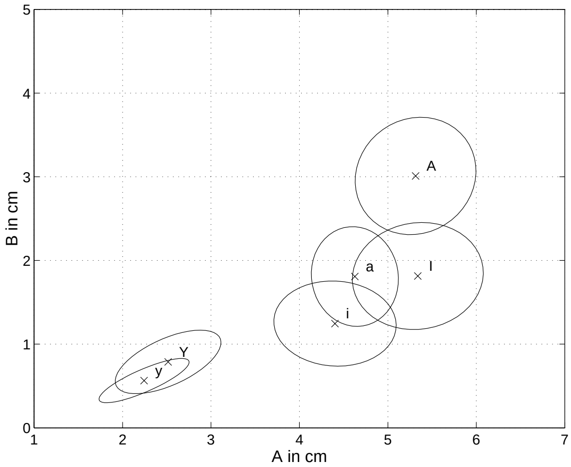

Figure 11: Mean (cross) and 2σ deviation of the vowel V_3 in the (A, B) plan. Hyper-visual condition in upper case.

5. REFERENCES

[1] Lindblom B. (1986-1987). Adaptative variability and absolute constancy in speech signals : two themes in the quest for phonetic invariance. *Perilus*, 5, pp. 2-20.

[2] Summers W., Pisoni D., Bernacki R., Pedlow R. & Stokes M. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.*, 84 (3), pp. 917-928.

[3] Badin P., Motoki K., Miki N. (1994). Some geometric and acoustic properties of the lip horn. *J. Acoust. Soc. Jpn*, 15, pp. 243-253.

[4] Lallouache M. T. (1990). Un poste visage-parole; Acquisition et traitement de contours labiaux. *In proceedings of the 18th JEP*, Montréal, pp. 27-28.

[5] Benoît C., Mohamadi T., Kandel S. (1994). Audio-visual intelligibility of French speech in noise. *Journal of Speech and Hearing Research*, 37, pp. 1195-1203.