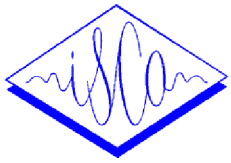
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 – September 3, 2007

Auditory-Visual Perception of Acoustically Degraded Prosodic Contrastive Focus in French

Marion Dohen, Hélène Lœvenbruck

ICP, Speech & Cognition Department, GIPSA-lab, UMR CNRS 5216, INPG, Univ. Stendhal, UJF, Grenoble, France
{Marion.Dohen,Helene.Loevenbruck}@icp.inpg.fr

Abstract

Previous studies have shown that visual only perception of prosodic contrastive focus in French is possible. The aim of this study was to determine whether the visual modality could be combined to the auditory one and lead to a perceptual enhancement of prosodic focus. In order to examine this question, we carried out auditory only, audiovisual and visual only perception tests. In order to avoid a ceiling effect, auditory only perception of prosodic focus being very high, we used whispered speech for which the acoustic prosodic information is degraded. The productions of two different speakers were used. This test showed that adding the visual modality enhances auditory perception of prosodic focus when the acoustic prosodic cues are degraded.

Index Terms: audiovisual, prosody, perception, whispered speech, contrastive focus

1. Introduction

Adding the visual modality to the auditory one enhances speech perception. Several studies have shown that audiovisual perception is better than auditory alone perception of speech in noise for example ([1-9]). An audiovisual advantage was also observed in silent environment for the perception of a non-native language or of semantically complex utterances ([10]). These studies dealt with segmental perception of speech i.e. identifying what is being said. It seems reasonable to think that the visual modality could also be useful for the perception of prosody i.e. for the perception of supra-segmental speech features. In previous studies ([11]), we showed that it is possible to detect prosodic contrastive focus in French from the visual modality alone. This showed that there are potential visual cues to prosodic focus perception. It is however unsure whether the audio and visual modalities can interact to enhance audiovisual perception of prosody.

Swerts & Krahmer examined the possible interactions between the auditory and visual modalities during the perception of focus in Dutch [12,13]. Both studies showed that there seemed to be visual information to prosodic focus for Dutch, just as we observed for French ([11]). The authors observed that the focus perception process appears to result from a combination of the information coming from the auditory and the visual modalities leading to a perceptual decision. When the auditory and visual modalities are conflicting, perception appears to be more difficult and slower. The authors also found that the upper and left parts of the speakers' faces seemed to provide more visual information. We suggest that this may be partly due to the unnaturalness of the task used for recording the material from the speakers (especially the one used to trigger focus) which could have resulted in exaggerated facial gestures or unnatural facial motion. These two studies provide very interesting preliminary information concerning the interactions between the auditory and visual modalities and their relative importance in perception.

The aim of the present study was to test the bimodal perception of French prosodic contrastive focus and to see whether the visual modality, added to the auditory one, could contribute to enhance perception.

2. Measuring enhancement: the "ceiling effect" problem

It has already been shown that auditory only perception of prosodic contrastive focus in French reaches near to perfect identification scores (e.g., [14]). Therefore, the probability of measuring any significant improvement when adding the visual modality is very low. The challenge was thus for us to design auditorily degraded prosodic stimuli in order to lower the auditory only perception scores and make improvement possible. The classical speech perception paradigm used to measure the audiovisual advantage is the speech in noise perception paradigm. However, even though adding noise to a signal reduces its lexical intelligibility, it does not alter the global fundamental frequency (F0) contour. Actually, [2] showed that voicing was the most robust speech feature to noise. This is why we designed a paradigm using whispered speech for which there is no F0 information at all since there is no vocal fold vibration. Moreover, whispering is used when one wants to be understood by the person he/she is speaking to but not overheard by others. The task given to the speakers during the audiovisual recordings described hereafter, was to make themselves understood by someone located at a certain distance from them (and not to whisper in someone's ear). It is therefore possible that the speakers might compensate for the lack of auditory cues by emphasizing visual cues.

3. Experimental Methods

3.1. Materials

3.1.1. Corpus

The corpus used consisted of four sentences with a Subject-Verb-Object (SVO) structure and with CV syllables. An example of a sentence used is given in (1).

(1) Romain ranima la jolie maman.
'Romain revived the good-looking mother.'

3.1.2. Audiovisual recording

We recorded two male native speakers (A and B) of French as they were whispering the sentences from the corpus under

four focus conditions: neutral, subject focus (SF), verb focus (VF) and object focus (OF). The two speakers were the same as those recorded for the studies described in [11]. A total of 16 utterances were thus recorded for each speaker. The recordings were done in a sound attenuated room at the Institut de la Communication Parlée (ICP). The speakers were not directly asked to produce focus. A correction task was used instead in order to trigger focus in the most natural way possible. The speakers listened to a prompt in which two speakers (S1 and S2) were talking. S1 first pronounced a sentence from the corpus which S2 then repeated in a question mode because he was not sure to have understood correctly one of the constituents from the sentence (S, V or O). The recorded speaker then had to correct S2 and thus produced contrastive focus on the mispronounced constituent. The recording therefore went as follows (capital letters signal focus):

Audio prompt: S1: Romain ranima la jolie maman.
S2: S1 a dit : Denis ranima la jolie maman?
'S1 said: Denis revived the good-looking mother?'

Speaker uttered: ROMAIN ranima la jolie maman.

No indication was given to the speakers on how to produce focus (e.g., which syllable(s) was(were) to be focused). When S2 had correctly understood (he produced the correct sentence in a question mode), the recorded speaker was instructed to produce a neutral version (broad focus) of the sentence i.e. without focusing any particular constituent.

The speakers movements were monitored with front and profile cameras. An example of the recorded images is given in Figure 1. The speakers wore blue markers on the lips and chin (see Figure 1) in order to make it possible to use an automatic lip-tracking device designed at ICP [15] to extract articulatory features.

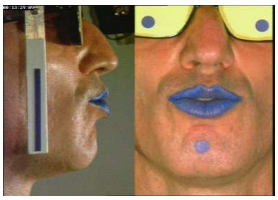


Figure 1: *Image from the video signal recorded.*

3.1.3. Stimuli manipulation

A preliminary informal auditory only perception test was conducted to examine whether the detection of focus using the whispered audio signals was low enough to be possibly improved. This test showed that the perceptual performances were quite high. An acoustic analysis actually showed that the intensity cues seemed to be boosted in whispered speech as if to compensate for the lack of F0 cues. We therefore decided to weigh the intensities of all the utterances recorded in order to bring each constituent of the utterance (S, V and O) to the same level as in the neutral version. A second informal auditory only perception test showed that perception scores were lower using the modified auditory files.

3.2. Experimental paradigm

The tests took place in a quiet room at ICP in which the participants were isolated both from outside noise and from the experimenters. The videos were shown on a video monitor placed approximately one meter away from the participants. The speaker's head on the screen was approximately real size. The participants were told that they would be following part of a conversation between two people (S1 and S2). S1 would first utter a SVO sentence. Not having heard the sentence very well, S2 would question S1 by repeating the sentence the way he had understood it, in a question mode. S1 would then repeat the first sentence he had uttered correcting the constituent (S, V or O) that S2 had misunderstood. He would therefore insist on this particular constituent (i.e. focus it).

The participants were told that they would neither hear nor see S1's first utterance as well as S2's. They would either see only (V), hear only (A) or hear and see (AV) S1's correction.

Participants were told that, in some cases, no correction would be performed by S1 because S2 would have correctly understood. In that case, they would just hear or see or see and hear S1 repeating the initial sentence without performing any correction (i.e. neutral version of the sentence). The task was for the participants to identify which constituent (S, V, O or none) had been misunderstood by S2 and thus corrected by S1. They were asked to highlight the constituent they had identified as being corrected on an answer sheet such as the one presented below:

Romain	ranima	la jolie maman.	

If they thought that S1 had performed no correction, they were asked to highlight the empty column on the right.

The participants were thus indirectly asked to identify whether a constituent had been focused and which one. They were never told about "focus" or about the experiment's aim.

Three movie clips were elaborated combining the videos recorded and the degraded auditory signals (one clip for each condition: AV, A and V). Each movie consisted of two sequences each corresponding to a random combination of the 32 stimuli. One of the sequences was to be seen with the front view by the participants and the other with the profile view.

Two separate tests (a and b) were designed using the same three movie clips. Each participant went through one of the two tests. For test a, the presentation modality order was: AV, A, V and for test b: A, AV, V. The aim being to analyze the contribution of the visual modality, this paradigm allowed comparisons between the performances corresponding to audio only and audiovisual conditions. For test a, for example, there could indeed be a training effect during the audiovisual session which would affect the performances during the audio only session and vice versa. This is why two tests were necessary. The visual only perception condition represented a control.

A total of 32 stimuli were thus evaluated by the subjects under three conditions (A, V and AV) and two different views (front and profile). This represents a total of 192 stimuli.

A total of 13 native speakers of French (8 men and 5 women) aged 19 to 57 participated in this experiment.

4. Results

4.1. General results

Figure 2 provides the percentages of correct answers for each participant and for each condition (A, V and AV). The means over all the participants for each condition correspond to the

thick dark lines. As could have been expected, the perceptual performances are better for the AV condition than for the other conditions. It therefore appears that when the acoustic prosodic cues are lacking, visual information can help recover at least part of the information. For all the conditions, the results are significantly above chance (25%): AV: t=31.478, p<0.001; A: t=13.369, p<0.001; V: t=13.374, p<0.001. The fact that audio only performances turned out to be better than chance was predictable since the durational information was still available for the audio detection. It is also possible that the intensity cues were not entirely "erased".

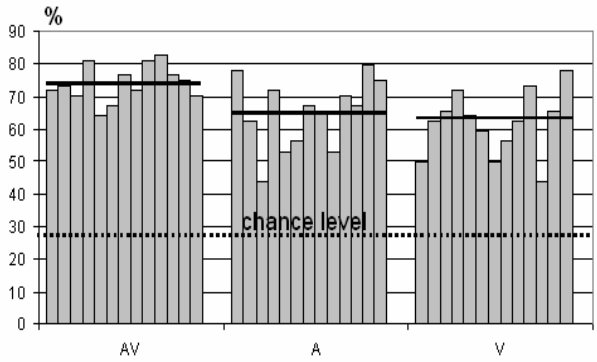


Figure 2: *Percentages of correct answers for each condition (AV, A and V) and each participant (1 bar = 1 participant). The thick dark lines represent the means of these percentages over all the participants.*

Figure 3.a provides the means of the percentages of correct answers over all the participants for each focus condition and for each speaker. It appears that the performances are better for speaker A which reflects what had already been observed in previous visual only perception tests (see [11]). It also appears that the audiovisual advantage is smaller for speaker A than for speaker B (+3.9% for speaker A and +14.4% for speaker B). The performances corresponding to the audio only condition for speaker A are very good (80.8% correct answers) and much better than those for speaker B (63.5%). There must therefore be a ceiling effect for speaker A: the performances are too good in the audio only condition to get improved to a significant extent. This is probably due to the fact that speaker A is a trained speaker. The acoustic cues he produced were very strong and we had great difficulties bringing back the intensity to an average level without distorting the signals. In addition, hyper-articulation has visual as well as acoustic consequences: formant patterns are less reduced for instance, which may be an additional acoustic cue (see [16] for the difference in auditory perception between hypo and hyper-articulated /iai/ sequences). This may have slightly biased the results and probably explains the small visual supplement measured for speaker A. Since speaker B was a naive speaker, it is possible that the results corresponding to the perception of his productions better reflect what would happen in natural communication.

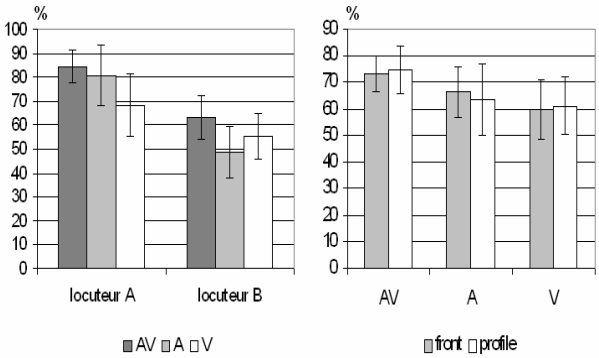


Figure 3: *Means of the percentages of corrects answers a.(left) for each speaker depending on the condition (AV, A and V) and b.(right) for each view point depending on the condition.*

Figure 3.b provides the means of the percentages of correct answers across the participants for each view (front and profile) and for each condition (AV, A and V). It shows that there are no differences from one view to another which confirms what had been observed for previous visual only perception tests ([11]).

A four intra-subject factor ANOVA was conducted on the results. The four factors were: modality (three levels: AV, A and V), speaker (two levels: speakers A and B), view point (two levels, front and profile) and focus condition (four levels: neutral, SF, VF and OF).

There appears to be a significant modality effect: $F(2,24)=7.232$, $p=.003$. The analysis of the ANOVA contrasts shows that the AV performances are significantly higher than those corresponding to the other modalities ($p<.001$). The results corresponding to the audio and visual only modalities are not significantly different ($p=.451$). The perceptual performances are thus significantly better when both modalities are available.

There is also a significant effect of the speaker: $F(1,12)=121.384$, $p<.001$. As explained before, the general perceptual results are better for speaker A.

Both the view and focus condition factors do not have a significant effect on the perceptual performances (view point: $F(1,12)=0.244$, $p=.63$; focus condition: $F(1.716,36)=11.231$, $p<.001$).

4.2. Further analysis

The mean percentages of correct answers over all the participants were calculated for each stimulus in each condition. The stimuli were then classified into three categories for which explanations were put forward as to what improved or degraded perception:

- **"AV≥A,V"** category (AV performances are the best): The majority of stimuli (46.9%) belong to this category. In this case, there were not many acoustic cues to contrastive focus left after degradation except for the durational ones. Auditory perception was thus not very good. Some visual cues were present and contributed to enhancing perception: when both modalities were combined, perceptual performances got better.
- **"A≥AV>V"** category (auditory only performances are the best and AV perception is better than V perception): 28.1% of the stimuli belong to this category. In this case, it seems quite clear that adding the visual modality impaired perception. Two explanations can be put forward. They correspond to two different cases. The

first case (A~AV) corresponds to one for which the AV and A performances are approximately the same. In this case, there must only be few visual correlates and adding the visual modality does not bring anything. Half of the stimuli from this category correspond to this case. In the second case, the A performances are better than the AV ones (A>AV). The visual cues must not only be absent but they seem to mislead perception. The other half of the stimuli from this category correspond to this case.

– **"V>AV,A"** category (visual only performances are the best): 25% of the stimuli belong to this category. In this case, the visual cues must be present since visual only perception is good but it seems that, when the auditory information is added, performances get lower. If the AV performances are better than the audio only ones (AV>A), it is possible that the auditory signal was poorly manipulated and that it tended to confuse perception and resulted in V>AV. This is the case for half of the stimuli from this category. If AV=A, the auditory signal may have been modified too much resulting for example in an unnatural production. In this case, the visual information was not sufficient to help the participants. This is the case for 25% of the stimuli from this category. Only two stimuli correspond to cases for which AV<A, they are isolated cases for which no explanation could be found.

5. Conclusions and Discussion

This audiovisual perception test of prosodic contrastive focus in French, conducted with whispered speech for two speakers, showed that visual cues can help detect focus when the auditory information is deteriorated. Perceptual performances are enhanced when the visual modality is added. Performances were better for speaker A than for speaker B which was predictable from the results of the visual only tests performed before ([11]). The audiovisual advantage is however much stronger for speaker B. It is actually possible that speaker A's auditory performances reached a ceiling that could not be improved.

6. Acknowledgments

The authors would like to thank Coriandre Vilain, Christophe Savariaux and Alain Arnal for their technical help. We are grateful to our two speakers as well as to all the participants to the test. We also thank Jean-Luc Schwartz and Marie-Agnès Cathiard for their help and comments.

7. References

[1] Sumby, W.H. and Pollack, I., "Visual contribution to Speech Intelligibility in Noise", J. Acoust. Soc. Amer., 26(2): 212-215, 1954.

[2] Miller, G. A. and Nicely, P., "An Analysis of Perceptual Confusions among some English Consonants", J. Acoust. Soc. Amer., 27(2): 338-352, 1955.

[3] Neely, K. K., "Effects of visual factors on the intelligibility of speech", J. Acoust. Soc. Amer., 28(6): 1275-1277, 1956.

[4] Erber, N. P., "Auditory-visual perception of speech", J. Spe. Hear. Dis., 40(4): 481-492, 1975.

[5] Binnie, C. A., Montgomery, A. A., and Jacskon, P. L., "Auditory and visual contributions to the perception of consonants", J. Spe. Hear. Res., 17(4): 619-630, 1974.

[6] Summerfield, A. Q., "Use of visual information for phonetic perception", Phonetica, 36: 314-331, 1979.

[7] MacLeod, A. and Summerfield, A. Q., "Quantifying the contribution of vision to speech perception in noise", Brit. J. Audiol., 21: p. 131-141, 1987.

[8] Grant, K. W. and Braida L. D., "Evaluating the Articulation Index for audiovisual input", J. Acoust. Soc. Amer., 89: 2952-2960, 1991.

[9] Benoît, C., Mohamadi, T., and Kandel. S., "Effects of Phonetic Context on Audio-Visual Intelligibility of French", J. Spe. Hear. Res., 37: 1195-1203, 1994.

[10] Reisberg, D., McLean, J., and Goldfield, A., "Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli", In Dodd, B. and Campbell, R. (Eds.), Hearing by eye: The psychology of lip-reading, Lawrence Erlbaum Associates, Hillsdale (USA), p. 97-114, 1987.

[11] Dohen, M. and Lœvenbruck H., "Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study", Interspeech/Eurospeech 2005, Portugal, 2005, p 2413-2416.

[12] Swerts, M., and Krahmer, E., "Congruent and Incongruent Audiovisual Cues to Prominence", Proceedings of Speech Prosody 2004, Japan, 2004, p 69-72.

[13] Swerts, M., and Krahmer, E., "Cognitive processing of audiovisual cues to prominence", Proceedings of AVSP 2005, Canada, 2005, p. 29-30.

[14] Dahan, D. and Bernard, J.-M., "Interspeaker Variability in Emphatic Accent Production in French", Language & Speech, 39(4): 341-374, 1996.

[15] Lallouache, M.-T., Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours de lèvres, PhD Thesis, Institut National Polytechnique de Grenoble, France, 1991.

[16] Lœvenbruck, H., Pistes pour le contrôle d'un robot parlant capable de réduction vocalique, PhD Thesis, Institut National Polytechnique de Grenoble, France, 1996.