Is it featural or holistic? Which type of visual input facilitates speech perception best?

Grozdana Erjavec and Denis Legros
Laboratory CHART
University of Paris VIII
Saint-Denis, France
grozdana.erjavec@etud.univ-paris8.fr, legrosdenis@yahoo.fr

***Abstract*—While speech perception is well-known to be multimodal, resulting from a fusion between visual and acoustic information, the goal of the present study was to explore the contribution of holistic (whole-face) and featural (mouth-only) visual input to speech perception. Sixteen adult participants were asked to repeat mildly and strongly acoustically degraded syllables presented in audio-only (AO), audio-visual whole face (AVF), and two audio-visual mouth conditions within which one contained high contrast regions (AVM-W) and the other did not (AVM-E). Participant's correct repetitions and fixations duration in the talker's oral area were analyzed to find that the featural AVM-E format was the most facilitative of perception of phonological information. This was also the format that yielded the longest fixations durations in the talker's oral region. The results are interpreted in line with a possible association between speech and face processing.**

Keywords—audio-visual speech perception, noise degradation paradigm, eyetracking, featural processing, holistic processing, adults

I. Introduction

Over the last seven decades, research on speech processing has well established that visual information plays an important role in speech perception. Firstly, visual information has a facilitative effect on speech perception under conditions in which the auditory input is degraded (e.g., [1], [2], [3]). Secondly, speech perception appears to be multimodal in its nature, resulting from the fusion between auditory and its corresponding visual input [4].

One of the problematics in multimodal speech perception research is to establish the nature of visual input (featural mouth-only *vs* holistic full-face displays) that facilitates best speech perception in noise. On one hand, some studies provide evidence for a possible holistic nature of visual speech cues processing either by finding a facial inversion effect [5], [6] or by observing an activation of fusiform gyrus in multimodal speech perception in noise [7]. On the other hand however, studies comparing subjects' performance in multimodal speech perception under whole-face and mouth-only visual conditions yield divergent results. Indeed, if some of these studies found the whole-face displays as being the most facilitating of speech perception [8], [9], others found no differences in subjects' performance between holistic and featural visual input conditions [10], [11], [12].

According to Thomas and Jordan [12], the reasons for the discrepancies in these results are due to the differences in the featural visual information format which firstly, contained areas not strictly confined to the mouth of the talker in some studies and, secondly, was often obtained by covering irrelevant regions of talker's face with high-contrast mask. Yet high contrast regions present in visual stimuli could have affected attentional processing either by diverging viewer's gaze from the exposed mouth area to the occluded areas or, in the contrary, by attracting viewer's attention to the exposed mouth area. The goal of our study was to address the exposed issue by taking into account both subjects' verbal performance and eye movement behavior in conditions containing holistic visual input and featural visual input with and without high contrast areas.

II. Method

A. Participants

Sixteen adults (age range: 18-40; mean age: 25.13; *SD:* 6.57) with no reported psychiatric and/or neurological disorders and either normal or corrected to normal vision, as well as normal hearing capacities participated in the study. They were all native speakers of French.

B. Stimuli

Critical trial stimuli consisted of 16 consonant-vowel syllables (/ba/, /da/, /fa/, /ga/, /ka/, /la/, /ma/, /na/, /pa/, /ʁa/, /sa/, /ʃa/, /ta/, /va/, /za/, /ʒa/). They were presented under one audio-only (AO) and three audio-visual conditions which differed in the format of visual input presentation: audio-visual face format (AVF), audio-visual mouth "extraction" format (AVM-E) (with no high-contrast areas) and audio-visual mouth "window" format (AVM-W) (with high contrast areas between the talker's mouth and the rest of facial context which was occluded by a high-contrast mask). Two demonstration trial stimuli, (/oua/ and /ŋa/), were also included in the study. They were presented under the same conditions as the critical trial stimuli.

In addition to different visual presentation conditions, the stimuli were presented in two auditory conditions which differed in the level of signal's degradation by noise: In the signal to noise ratio -6 (SNR-6) the purple noise covering the signal was of 6 dB stronger than the signal and in the SNR-12 condition the difference in decibels between the noise and the signal was of 12.

All the videos were of 2 seconds length. Within each, the target syllable of approximately 1 second length was centered.

C. Procedure

The participants' task was to repeat, as it was perceived, each critical trial stimulus. For this purpose, an inter-stimuli interval of 2 seconds was included in the stimuli presentation.

The stimuli were presented in blocks, each corresponding to one of 4 overall presentation conditions. Each block contained two sub-blocks which corresponded to auditory degradation conditions. For half of the participants each block started with low degradation sub-block and, for the other half, with high degradation sub-block. The order of stimuli presentation within each sub-block was random. The order of the four blocks was partially counterbalanced in such a way that each condition was preceded and followed at least once by every other condition, which resulted into four different condition orders. Four participants were assigned to each order of block presentation. In the experiment, the eye movement behavior of participants was recorder with the Tobii 1750 eye-tracker.

In addition to the critical trials, the experiment started with a demonstration of experimental material: the syllables /oua/ and /ŋa/ were presented in all eight conditions (audio-visual presentation X level of auditory degradation).

III. Results

For the verbal data, the differences in the correct repetitions between the audio-only and each audio-visual condition, also known as audio-visual (AV) gain, and for eye movement data, the fixations duration in the mouth area of the talker were taken into account. Both types of data were submitted to a Signal to Noise Ratio x Visual Information Format (2x3) within-subject analysis of variance.

A. Mean percentage of AV gain

The main effect was significant for both factors, SNR ($F(1,15)=9.741$; $p<0.070$; $\eta_p^2=0.394$) and Visual Information Format ($F(2,30)=4.577$; $p<0.020$; $\eta_p^2=0.234$). The AV gain was greater in the SNR-12 condition, and lower in the AVF format condition as opposed to the to the AVM-E format ($t(31)=-3.572$; $p<0.001$) and the AVM-W format ($t(31)=-2.784$; $p<0.009$). The interaction effect was significant as well ($F(2,30)=8.316$; $p<0.002$; $\eta_p^2=0.357$). At the SNR-6, the AVM-E format yield the best performances with the other two formats being somehow comparable (AVM-E *vs* AVM-W ($t(15)=2.449$; $p<0.027$) and AVM-E *vs* AVF ($t(15)=3.000$; $p<0.009$)). At the SNR-12 condition subjects did better with the AVM-W than the AVF format ($t(15)=4.044$; $p<0.001$). No difference was found at this level between both featural formats. (See Figure 1 for a graphical representation of results.)

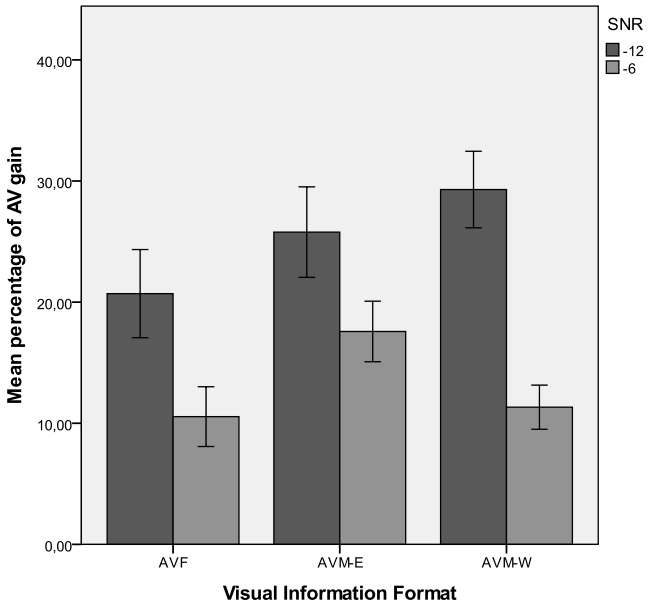

Figure 1. *Mean percentage of AV gain with the AVF, the AVM-E and the AVM-W formats under the two SNR conditions. (The bars are representing standard error.)*

B. Fixations duration in the talker's mouth area

Only the main effect of Visual Information Format was found to be significant ($F(2,30)=7.193$; $p<0.009$; $\eta_p^2=0.643$). Overall, subjects' fixations in the talker's mouth area were longer in the AVM-E format than in the AVF ($t(31)=-3.774$; $p<0.001$) and the AVM-W format ($t(31)=5.284$; $p<0.001$) conditions. The difference between the AVF and AVM-W was also significant with the benefit for the AVF condition ($t(31)=2.304$; $p<0.028$). (See Figure 2 for a graphical representation of results.)

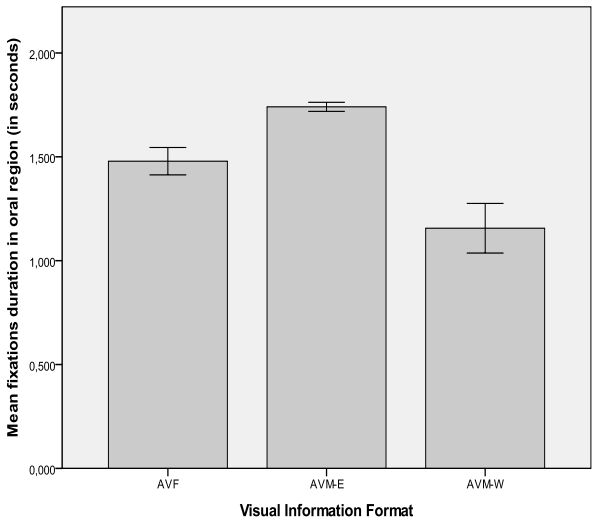

Figure 2. *Mean fixations duration in the talker's oral area under the three Visual Information Format conditions. (The bars are representing standard error.)*

IV. Discussion

While expecting to find that holistic facial information would be the most facilitative of audio-visual speech perception in noise, our results surprisingly suggest that this hypothesis is to be rejected. Such results are however somehow in line with those obtained by Thomas and Jordan [12]. These authors also tested the efficiency of whole face and featural mouth only format containing no high contrast areas and found a tendency, although not significant, towards the featural format being more advantageous for speech perception. Since Thomas and Jordan [12] used words as their experimental stimuli participants' lexical knowledge could have constituted a factor that masked a possibly significant effect of information presentation format. In our study, on the contrary, the stimuli used contained minimal lexical information and participants' task was mainly of phonological nature – detecting the first consonant of a consonant-vowel syllable for the vowel was maintained fixed across all stimuli. Thus, it seems that when it comes to audio-visual processing of speech production in its phonological aspect alone, featural mouth only format of visual information presentation containing no high contrast areas (AVM-E condition) allows the viewer to focus his/hers visual attention on the oral region and facilitates best speech perception. While featural mouth-only window (AVM-W) format attracted visual attention focus less on the mouth than the AVM-E format, it allowed a successful encoding of phonological speech information especially in a condition with a high cognitive load on such processing (SNR-12 condition). Finally, while visual attentional focus on oral region was comparable to that in the AVM-W condition, the full face (AVF) holistic format yield the lowest AV gain overall. In line with previous studies which found an association between audio-visual speech perception and face processing, one might suggest that face processing, in which we are highly specialized as extremely social beings, could have interfered with processing of phonological information relative cues. As such, the results of present study offer numerous possibilities for further research on problematics in line with normal development (children are known to be less efficient in holistic face processing than adults) as well as pathological developmental conditions (particularities in face processing are a well-known feature in autism). The interpretation of the results would also greatly benefit from a study of patterns of neural activity underlying the processing induced by each visual format with a special attention of the activity in the region highly specialized for face processing, the right fusiform gyrus.

References

[1] Binnie, C. A., Montgomery, A. A. and Jackson, P. L., "Auditory and visual contributions to the perception of consonants", J. Speech Hear. Disord., vol. 17, pp. 619-630, December 1974.

[2] Eramudugolla, R., Henderson, R. and Mattingley, J. B., "Effects of audio-visual integration on the detection of masked speech and non-speech sounds", Brain Cognition, vol. 75, pp. 60–66, February 2011.

[3] Neely, K. K., "Effects of visual Factors on the intelligibility of speech", J. Acoust. Soc. Am., vol. 28, pp. 1275-1277, November 1956.

[4] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", Nature, vol. 264, pp. 746-748, April 1976.

[5] Rosenblum, L. D., Yakel, D. A. and Green, K. P., "Face and mouth inversion effects on visual and audiovisual speech perception", J. Exp. Psychol.-Hum. Percept. Perform., vol. 26, pp. 806-819, April 2000.

[6] Thomas, S. M. and Jordan, T. R., "Determining the influence of gaussian blurring on inversion effects with talking faces", Percept. Psychophys., vol. 64, pp. 932-944, August 2002.

[7] Kawase, T., Yamaguchi, K., Ogawa, T., Suzuki, K., Suzuki, M., Itoh, M., Kobayashi, T. and Fujii, T., "Recruitment of fusiform face area associated with listening to degraded speech sounds in auditory–visual speech perception: A PET study", Neurosci. Lett., vol. 382, pp. 254-258, July 2005.

[8] Greenberg, H. J., and Bode, D. L., "Visual discrimination of consonants", J. Speech Hear. Res., vol. 11, pp. 869-874, December 1968.

[9] IJsseldijk, F. J., "Speechreading performance under different conditions of video image, repetition, and speech rate", J. Speech Hear. Res., vol. 35, pp. 466-471, April 1992.

[10] Montgomery, A. A. and Jackson, P. L., "Physical characteristics of the lips underlying vowel lipreading performance", J. Acoust. Soc. Am., vol. 73, pp. 2134-2144, June 1983.

[11] Stone, L., "Facial clues of context in lip reading", Los Angeles: John Tracy Clinic, 1957.

[12] Thomas, S. M. and Jordan, T. R., "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception", J. Exp. Psychol.-Hum. Percept. Perform., vol. 30, pp. 873-888, October 2004.