Effects of Mouth-Only and Whole-Face Displays on Audio-Visual Speech Perception in Noise: Is the Vision of a Talker's Full Face Truly the Most Efficient Solution?

Grozdana Erjavec [1], *Denis Legros* [1]

[1] Department of Cognition, Language and Interaction, University of Paris VIII, France

grozdana.erjavec@etud.univ-paris8.fr, legrosdenis@yahoo.fr

10.21437/Interspeech.2013-423

Abstract

The goal of the present study was to establish the nature of visual input (featural *vs* holistic) and the mode of its presentation that facilitates best audio-visual speech perception. Sixteen participants were asked to repeat acoustically strongly and mildly degraded syllables, presented in auditory and three audio-visual conditions, within which one contained holistic and two contained featural visual information. The featural audio-visual conditions differed in characteristics of talker's mouth presentation. Data on correct repetitions and participants fixations duration in talker's mouth area were collected. The results showed that the facilitative effect of visual information on speech perception depended upon both auditory input degradation level and the visual presentation format, while eye-movement behavior was only affected by the visual input format. Featural information, when presented in a format containing no high contrast elements, was overall the most efficient visual aid for speech perception. It was also in this format that the fixations duration on talker's mouth was the longest. The results are interpreted with a stress on differences in attentional and perceptual processes that the different visual input formats most likely induced.

Keywords: Audio-visual speech perception, featural processing, holistic processing, noise degradation paradigm, eye-movement

1. Introduction

In order to achieve an optimal speed and accuracy, human perception is often based upon multisensory integration; that is, an integration of information on events and/or objects provided by different sensory modalities. Perceiving natural speech when hearing a talker's voice and seeing her/his face is a perfect example of multimodal perception. Indeed, the beneficial nature of a multimodal, audio-visual, input is well established for speech perception. Not only does adding visual information on articulatory movement to a congruent auditory speech input lower auditory signal detection threshold in noisy settings [1], it also facilitates speech perception and comprehension either with degraded [2]-[5] or intact auditory input [6]. A further evidence for the implication of visual information in speech perception comes from the McGurk effect which demonstrates that there is an audio-visual fusion in multimodal speech processing, leading to a unified percept which combines the characteristics of both auditory and visual modalities [7].

An important issue in multimodal speech perception is to identify the visual cues that are needed for optimal accuracy in perception of phonetic information and, in extension, the nature of visual speech processing (featural *vs* holistic). On one hand it has been well established that the lips of a talker, and even more so her/his entire mouth, are the most crucial facial components of visual speech [8]-[10]. On the other hand however, there is much less clarity about the possible additional facilitative role of a holistic facial input compared to the featural, mouth-only information. Indeed, in the latter vein of research, the results are somewhat conflicting. A few authors found some evidence about the holistic nature of visual and audio-visual speech perception by observing either facial inversion effects in speech perception [11], [12], or an important activity in the right fusiform gyrus during multimodal speech perception when auditory signal was degraded by noise [13]. In the first case, the fact that visual facilitation of speech perception is smaller when seeing inverted faces is attributed to a disruption that the face inversion causes in processing of spatial relationships between facial features, thus primarily affecting holistic type of face processing [14]. In the second case, the right fusiform gyrus seems to be implicated in establishing representations of facial identity which is thought to require featural, and even to a greater extent holistic processes [15].

If visual speech perception is indeed holistic in its nature, one would expect that a featural, mouth-restricted information should have a lesser facilitative effect on audio-visual speech perception than a holistic, whole-face-related information. Nevertheless, in this vein of reasoning, the results of previous studies do not unanimously confirm this hypothesis. Indeed, while using the noise-degradation paradigm, some authors found poorer speech perception with mouth-only displays [16], [17] but others observed no difference between mouth-only and full-face displays [9], [18], [19].

The question of why such differences in the results of previous studies have appeared was addressed by Thomas and Jordan [19]. The authors pointed out methodological issues related to the mode of presentation of featural, mouth-only visual input. Firstly, mouth-only displays were sometimes not restricted to oral area alone but were extended to mandible and larynx [16], [18]. Secondly, in some previous studies, a window technique was used in editing mouth-only displays in which facial context was covered with an opaque masque, thus creating high visual contrast between the region that was left exposed and the rest of the display. However, it was suggested that such a format of presentation could lead to ambiguity in the interpretation of visual information [20]. Moreover, the use of window technique may also affect attentional and perceptual processing either by diverting a viewer's gaze from the region left exposed to the occluded area [8], or, in the contrary, by encouraging the focus of the highlighted unoccluded area. When studying a possible difference in the role of holistic *vs* fetural visual information in multimodal speech perception, each of these methodological issues seems thus clearly problematic for the validity of the results.

Keeping in mind these methodological concerns, Thomas and Jordan [19] addressed the question by simply extracting the talker's mouth and placing it on a uniform background of

facial color. When degrading the auditory input with noise, the authors found a slight but insignificant tendency towards the superiority of featural presentation format which led them to conclude the vision of a talker's mouth is sufficient for an optimal processing of phonetic cues carried by the visual information.

However, since the study did not include a window mouth-only format, nothing can be said about whether or not the differences in the results of previous studies were indeed related to the mode of visual information presentation. Furthermore, another important methodological question, notably the noise level used to degrade the auditory input, was not taken into account in Thomas and Jordan's study [19], although it is thought to affect the size of facilitative effect of visual information in multimodal perception conditions [20], [21]. The influence that the noise level has on multimodal speech perception might also come as a consequence of its potential effect on attentional and perceptual processes. Thinking intuitively, it seems possible that higher noise intensity would encourage the viewer to fixate in a greater extent the mouth of the talker, whereby creating a condition in which a possible advantage of a holistic visual input is diminished.

The goal of the present study was to contribute to the debate on the nature of visual information needed for optimal extraction of phonetic information from visual speech by taking into account both window and extraction format of fetural, mouth-only input presentation, as well as the level of degradation of auditory input. Moreover, in order to further explore attentional processes underlying multimodal perception in this context, we considered not only subjects' verbal responses to multimodal speech stimuli, but also their eye-movement behavior during visual information processing. Our hypotheses were as follows: i) In line with the results of Ross et al. study [21], [22], we predicted that the overall contribution of visual information to speech perception would be more important in the conditions with poorer intelligibility of auditory signal. ii) Following Thomas and Jordan's reasoning [19], we expected that the contribution of visual information to speech perception would be greater and similar for the holistic full-face and the featural mouth-extracted formats of presentation as compared to the featural mouth-window format. iii) Taking into account our intuitive proposition that the higher level of auditory signal degradation would affect the viewer's attentional processes in such a way that she/he would be mostly focused on the speaker's mouth, we thought that the difference in the facilitative effect of visual information between the full-face and the mouth-extracted formats on one hand and the mouth-window format on the other hand should be diminished in high noise condition. As for the eye-movement behavior, we hypothesized that i) the observers would fixate the speaker's mouth longer when the visual information would be presented with the holistic and the mouth-extracted formats; ii) the observers would fixate the speaker's mouth longer in the condition with higher noise level of auditory signal degradation; iii) the difference in the fixations duration of the talker's mouth between the full-face and the mouth-extracted format on one hand, and the mouth-window format on the other hand would decrease with increasing level of auditory signal degradation.

2. Method

2.1. Participants

Sixteen adults participated in the study (age range: 18-40; mean age: 25,13; *SD:* 6,57). The criteria for subjects' inclusion in the study were i) being a native speaker of French; ii) having no neurological and/or psychiatric condition; iii) having normal or corrected vision and normal hearing capacities. These were all reported data collected through an interview prior to the testing phase of the experiment.

2.2. Stimuli

Critical trials in the experiment consisted of 16 consonant-vowel syllables (/ba/, /da/, /fa/, /ga/, /ka/, /la/, /ma/, /na/, /pa/, /ʁ a/, /sa/, /ʃ a/, /ta/, /va/, /za/, /ʒ a/) which were presented audio-visually under two signal-to-noise ratio levels (SNR-6 and SNR-12) and within four visual formats. One of the visual formats contained only a static play icon and was used for audio-only (AO) condition (see Fa_AO_6.avi and Fa_AO_12.avi for an example of this type of stimuli in the two noise level conditions); while the other three formats each comprised visual speech elements. Notably, in the audio-visual face (AVF) format the speaker's entire face and neck were visible (see Fa_AVF_6.avi and Fa_AVF_12.avi for an example of this type of stimuli in the two noise level conditions), in the audio-visual mouth-extracted (AVM-E) format only the talker's mouth was present, placed on the facially-colored background (see Fa_AVM-E_6.avi and Fa_AVM-E_12.avi for an example of this type of stimuli in the two noise level conditions), finally, in the audio-video mouth-window (AVM-W) format, only the talker's oral area was visible through a static rectangular window with the rest of the face occluded by obscure black regions (see Fa_AVM-W_6.avi and Fa_AVM-W_12.avi for an example of this type of stimuli in the two noise level conditions). The rectangular area in the AVM-W format was as large as to allow a clear scene of the talker's mouth at the maximum mouth opening.

In addition to the critical trials items, two other syllables (/oua/ and /ŋa/) were recorded for the same noise and presentation format conditions. They were used as demonstration items.

The talker was a 30 year old female who was a native speaker of French. She was recorded upon a light background under normal lightning conditions and was given the instruction to pronounce the items clearly but with no exaggeration in the articulation and with a constant, monotone intonation.

The videos were digitally remastered with the Adobe Premier Pro CS5.5. Their final length was of 2 seconds within which the pronounced syllable was centered. The audio recordings were processed with the Adobe Audition CS5.5 software. The syllables were kept at -32dBFS (RMS) and the added pink noise was kept at -20dBFS (RMS) for SNR-12 and at -26dBFS (RMS) for SNR-6. The noise was presented in synchrony with the videos. Participants were wearing headphones in the experiment and were asked to adjust the volume level by themselves.

2.3. Task

The participants were asked to pay attention to the videos and to repeat orally, as clearly as possible, the sounds they have

perceived. They were encouraged to try even if they were not completely sure of what they have perceived.

2.4. Procedure

The experiment consisted of two phases: i) demonstration of testing material; ii) testing phase. During the first phase, the participants were shown examples of different types of experimental stimuli. Syllables /oua/ and /ŋa/ were shown in all eight modes of presentation. During the testing phase, the stimuli were administered by blocks, each corresponding to one of four modes of visual information presentation (AO, AVF, AVM-E and AVM-W). Each block was further divided into two sub-blocks, each corresponding to one of the two SNR levels. The order of the four audio-visual presentation format 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. Out of the 16 participants, 4 were assigned to each condition order, for half of them the blocks started with a sub-block of SNR-12 and ended with the sub-block of SNR-6, and for the other half of participants the order of the sub-blocks was inversed. Inside of each sub-block stimuli were presented in a random order. The inter-stimuli interval was of 2 seconds. During the testing phase the participants' eye-movements were recorded with the use of Tobii 1750 eye-tracker.

3. Results

A two-way repeated measures ANOVA was conducted on the collected verbal and eye-movement data. The factor SNR consisted of two modalities, SNR-6 for medium to weak auditory signal degradation (which corresponded to 55% of correctly identified stimuli in the AO condition) and SNR-12 for strong auditory signal degradation (which corresponded to 31% of correctly identified stimuli in the AO condition). The factor Visual Information Format consisted of three modalities corresponding to the AVF, the AVM-E and the AVM-W format. As usual, the AO mode of stimuli presentation was used as a comparison condition.

The verbal data consisted of the differences in the number of correct repetitions between each AV and the AO condition. These differences, expressed in terms of percentage in the present study, are commonly referred to as the AV gain. As for the eye-tracking measures, the total fixations duration in the talker's oral-region (as defined in the AVM-W format and transposed to the AVF and the AVM-E formats) was taken into account for ANOVA analysis.

3.1. Verbal data (AV gain data)

The ANOVA analysis showed a significant main effect of both factors, SNR ($F(1,15)=9,741$; $p<0,007$; $\eta_p^2=0,394$) and Visual Information Format ($F(2,30)=4,577$; $p<0,020$; $\eta_p^2=0,234$). Post-hoc paired *t*-test analysis showed that a mean AV gain was significantly lower with the AVF format as opposed to the AVM-E format ($t(31)=-3,572$; $p<0,001$) and also as opposed to the AVM-W format ($t(31)=-2,784$; $p<0,009$). The effect of Noise Level x Visual Information Format turned out to be significant as well ($F(2,30)=8,316$; $p<0,002$; $\eta_p^2=0,357$). Paired *t*-test post-hoc analysis revealed the following significant differences in mean AV gain: i) the AV gain in the AVM-E was superior to the one in the AVF condition at the SNR-6 ($t(15)=3,000$; $p<0,009$); ii) at the SNR-12, the AV gain in the AVM-W was also superior to the one obtained with the AVF format ($t(15)=4,044$; $p<0,001$); iii) finally, the AV gain for the AVM-E format was superior to the one found in the AVM-W condition at the SNR-6 ($t(15)=2,449$; $p<0,027$). In addition, at the SNR-12, the difference in the AV gain between the AVB-E and the AVF conditions was very close to being statistically significant ($t(15)=2,030$; $p<0,060$) with the gain being higher in the AVB-E condition. (See Figure 1 for a graphical representation of results.)

3.2. Eye-movement data

For the eye-movement data, only the main effect of factor Visual Information Format was significant ($F(2,30)=7,193$; $p<0,009$; $\eta_p^2=0,643$). Paired *t*-test post-hoc analysis revealed that all mean differences were significant (AVF *vs* AVM-E condition ($t(31)=-3,774$; $p<0,001$); AVF *vs* AVM-W ($t(31)=2,304$; $p<0,028$); AVM-E *vs* AVM-W condition ($t(31)=5,284$; $p<0,001$). The longest fixations duration of the talker's oral area was found in the AVM-E condition, while in the AVM-W condition the fixations duration of the area of interest was the shortest in the experiment. (See Figure 2 for a graphical representation of results.)

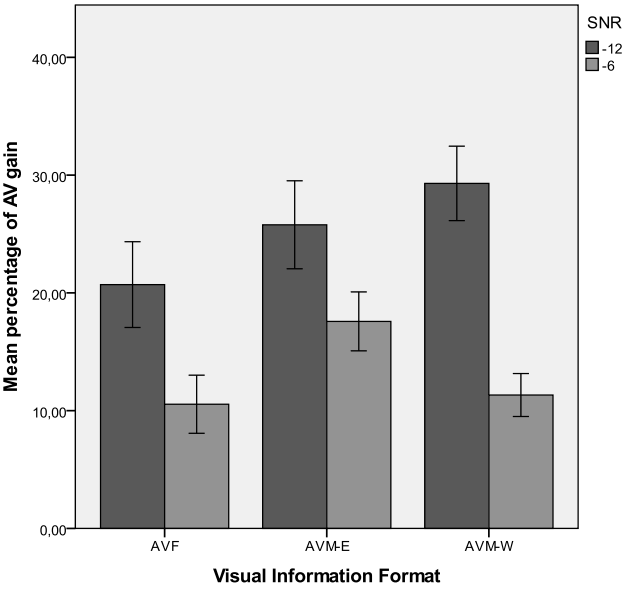

Figure 1. *Mean percentage of AV gain with the AVF, the AVM-E and the AVM-W formats under the two SNR conditions. (The bars are representing standard error.)*

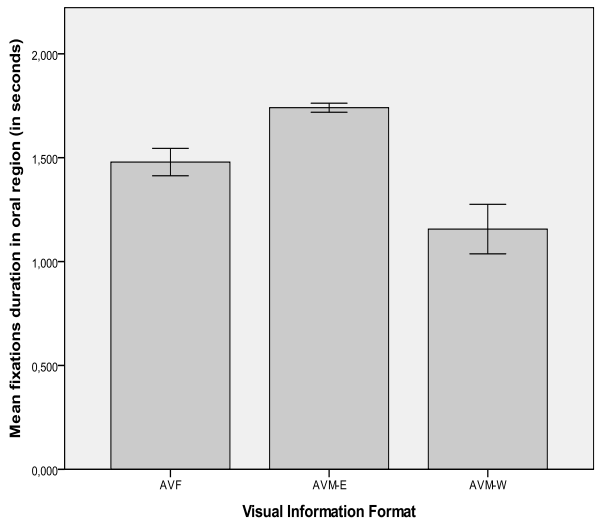

Figure 2. *Mean fixations duration in the talker's oral area under the three Visual Information Format conditions. (The bars are representing standard error.)*

4. Discussion

4.1. Results from verbal data

In the exception of significant effect of factor SNR, the results from verbal data do not confirm our hypotheses. As expected, a greater AV gain was found in the lower SNR condition (SNR-12) which shows that the visual speech input is especially helpful for speech perception when the unintelligibility, and thus unreliability of the auditory input is high.

The rest of the results might appear as somewhat surprising. In point of fact, both featural information presentation formats (AVM-E and AVM-W) turned out to be more efficient for multimodal speech perception than the holistic information presentation format (AVF). While differences between AVM-E and AVF formats were either significant or very close to being statistically significant at both SNR levels, the difference between AVM-W and AVF format was only (but highly) significant at the lowest SNR. On one hand, such results suggest that, when presented without any high-contrast regions, featural mouth-centered information is more facilitative of multimodal speech perception than holistic information, which might either not or only slightly depend upon the quality of auditory signal. On the other hand, even when featural information is presented within obscure regions, it is more helpful for speech perception than holistic information when the auditory signal is highly unreliable. It thus seems that, in adults, featural visual input is more successfully processed in respect to phonetic information that it carries than the holistic input.

4.2. Eye-movement results

The eye-movement results confirm one of our initial hypotheses. We found that the viewer does indeed fixate the talker's oral area longer when it is presented in a holistic AVF or in a featural AVM-E format than in a featural AVM-W format. This result implies that the use of high-contrast regions when aiming to expose a featural information does indeed affect eye-movement behavior in the way that the viewer's gaze is drawn from the exposed area towards the regions that were surrounding it. From our initial rationale, such a result indicates that the mode of visual input presentation affects attentionnal mechanisms in viewer. Nevertheless, the viewer's visual attention does not seem to depend upon the degree of auditory input intelligibility.

4.3. Results from both sets of data combined

Taking the results of both sets of data into account, the fact that a viewer fixates longer on the talker's oral area seems only partially assimilated to a greater multimodal speech perception accuracy. As a matter of fact, while the AVM-E format in which the speaker's oral area was fixated the longer was also the format which was the most facilitative for multimodal speech perception, the AVF and the AVM-W formats represent a difference in sense of relationship between fixations duration in the speaker's oral area (high in the AVF and low in the AVM-W format) and the importance of facilitation of speech perception (low in the AVF and high in the AVM-W format). This is an important aspect in the comparison of the results of both sets of data that might mark a difference between different types and levels of processing. It seems possible that, in the AVM-W condition, the viewer's gaze was drawn towards obscure regions which did not really trigger profound processing (indeed, there was nothing there to process). The viewer could thus successfully process featurally-presented visual speech that was encoded from peripheral vision. It is known that the peripheral vision is particularly sensitive to movement which is the type of information that is very important in multimodal speech perception. On the contrary, in the AVF condition, while viewer's gaze stayed well fixated on the speaker's oral area, the information from peripheral vision could have engaged the viewer in somewhat more profound processing of facial-context-related holistic information which was possibly aimed at the establishing of some level of representations about talker's identity. The fact that i) adults are experts in holistic face processing [23], [24]; ii) the implication of the right fusiform gyrus in acoustically-degraded multimodal speech perception was found [13]; and iii) audio-visual speech perception is sensitive to face inversion effect [11], [12] do indeed point towards an implication of holistic processing of visual speech perception in conditions where the talker's face is visible. Yet, none of these elements indicates that holistic processing of facial information is facilitative of extracting acoustic cues from visual input; it might as well be the opposite.

5. Conclusions

The results of the present study suggest that multimodal speech perception, as well as visual attention mechanisms, depend greatly both on the type of visual input (featural *vs* holistic), and on the characteristics of visual input presentation. Adult subjects seem to benefit the most from featural visual information, possibly because it allows the viewer to engage exclusively (or almost exclusively) in processing of visually-conducted phonetic information. On the contrary, holistic, face-related visual information might induce some level of face identity processing, thus allowing a less profound processing of speech-related visual information. Further research is needed on the subject.

6. References

[1] Grant, K. W. and Seitz, P, "The Use of Visible Speech Cues for Improving Auditory Detection of Spoken Sentences", J. Acoust. Soc. Am., 108(3): 1197-1208, 2000.

[2] Binnie, C. A., Montgomery, A. A. and Jackson, P. L.,. "Auditory and Visual Contributions to the Perception of Consonants", J. Speech Hear. Disord., 17(4): 619-630, 1974.

[3] Eramudugolla, R., Henderson, R. and Mattingley, "Effects of Audio-Visual Integration on the Detection of Masked Speech and Non-Speech Sounds", Brain Cognition, 75(1): 60–66, 2011.

[4] Neely, K. K., "Effects of Visual Factors on the Intelligibility of Speech", J. Acoust. Soc. Am., 28(6): 1275-1277, 1956.

[5] Schwartz, J.-L., Berthommier, F. and Savariaux, C., "Seeing to Hear Better: Evidence for Early Audio-Visual Interactions in Speech Identification", Cognition, 93(2): 69-78, 2004.

[6] Arnold, P. and Hill, F., "Bisensory Augmentation: A Speechreading Advantage when Speech is Clearly Audible and Intact", Brit. J. Psychol., 92(2): 339-355, 2001.

[7] McGurk, H. and MacDonald, J., "Hearing Lips and Seeing Voices", Nature, 264(5588), 746-748, 1976.

[8] Marassa, L. K. and Lansing, C. R., "Visual Word Recognition in 2 Facial Motion Conditions: Full Face versus Lips-plus-Mandible", J. Speech Hear. Res., 38(6): 1387-1394, 1995.

[9] Montgomery, A. A. and Jackson, P. L., "Physical Characteristics of the Lips Underlying Vowel Lipreading Performance", J. Acoust. Soc. Am., 73(6): 2134-2144, 1983.

[10] Summerfield, A. Q., "Use of Visual Information for Phonetic Perception", Phonetica, 36(4/5): 314-331, 1979.

[11] 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., 26(3): 806-819, 2000.

[12] Thomas, S. M. and Jordan, T. R., "Determining the Influence of Gaussian Blurring on Inversion Effects with Talking Faces", Percept. Psychophys., 64(6): 932-944, 2002.

[13] 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., 382(3): 254-258, 2005.

[14] Goffaux, V., Rossion, B., Sorger, B., Schiltz, C. and Goebel, R., "Face Inversion Disrupts the Perception of Vertical Relations between Features in the Right Human Occipito-Temporal Cortex", Journal of Neuropsychology, 3(1): 45-67, 2009.

[15] Sergent, J., Ohta, S. and MacDonald, B., "Functional Neuroanatomy of Face and Object Processing. A Positron Emission Tomography Study", Brain, 115(1): 15-36, 1992.

[16] Greenberg, H. J., and Bode, D. L., "Visual Discrimination of Consonants", J. Speech Hear. Res., 11(4): 869-874, 1968.

[17] IJsseldijk, F. J., "Speechreading Performance under Different Conditions of Video Image, Repetition, and Speech Rate", J. Speech Hear. Res., 35(2): 466-471, 1992.

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

[19] 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., 30(5): 873-888, 2004.

[20] Nelson, R. and Palmer, S. E., "Of Holes and Wholes: The Perception of Surrounded Regions", Perception, 30(10): 1213-1226, 2001.

[21] Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D. and Foxe, J. J., "The Development of Multisensory Speech Perception Continues into the Late Childhood Years", Eur. J. Neurosci, 33(12): 2329-2337, 2011.

[22] Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C. and Foxe, J.J., "Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments", Cereb. Cortex, 18(6): 1147-1153, 2007.

[23] Bahrick H. P., Bahrick P. O. and Wittlinger R. P., "Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach", J. Exp. Psychol.: Gen., 104(1): 54-75, 1975.

[24] Carey S. and Diamond R., "Are Faces Perceived as Configurations More by Adults than by Children?", Vis. Cogn., 1(2/3): 253-274, 1994.