Mixing faces and voices: a study of the influence of faces and voices on audiovisual intelligibility

Jérémy Miranda, Slim Ouni

Mixing Faces And Voices:

A Study of the Influence of Faces and Voices on Audiovisual Intelligibility

Jérémy Miranda and Slim Ouni

Université de Lorraine. LORIA. UMR7503. Villers-lès-Nancy. F-54600. FRANCE

Jeremy.miranda@loria.fr. Slim.ouni@loria.fr

Abstract

This study examined the influence of mixing faces and voices on the audiovisual intelligibility. The goal is to study the effect of combining two sources of information on the audiovisual intelligibility. Cross-talker dubbing was performed between faces and voices of 10 meaningful sentences pronounced by 10 talkers: 5 females and 5 males. Human subjects were asked to rate the articulation of the output videos. Comparisons were made between results of original and dubbed video. Almost across all the combinations, the audiovisual intelligibility was acceptable. The intelligibility of the speakers varied, however. We observed an influence of the audio/visual channel on the overall intelligibility that can increase or decrease depending the intelligibility results of this channel.

1. Introduction

Human communication is naturally based on audiovisual speech, in the majority of cases. Audiovisual speech can be considered as the combination of two channels: audio and visual. This bimodal signal allows communicating an intelligible message. Research in audiovisual speech intelligibility has shown the importance of the information provided by the face especially when audio is degraded [1]. This importance of audiovisual speech has been promoted thanks to research in audiovisual speech synthesis, i.e., the generation of facial animation together with the corresponding acoustic speech. Audiovisual speech synthesis is considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the facial animation [2,3]. However, achieving perfect synchronization between these two streams is not straightforward and presents several challenges related to audio-visual intelligibility. In fact, humans are acutely sensitive to any incoherence between audio and visual animation. This may occur as an asynchrony between audio and visual speech [4], or a small phonetic distortion compared to the natural relationship between the acoustic and the visual channels [5,6].

The McGurk effect [7] describes the case when the mismatch is more important: when an auditory stimulus “ba” is paired with a visual stimulus “ga”, and the perceiver reports that the talker said “da”. This is called a fusion effect. We can observe a combination effect when pairing an auditory “ga” with a visual “ba”, and the perceived result is a combined “bga”. Some perceptual studies may suggest that the acoustic and visual information is processed as a “whole unit” [7,8]. In the field of audiovisual synthesis, it has been shown that the degree of coherence between the auditory and visual modalities has an influence on the perceived quality of the synthetic visual speech [9]. The question about synchrony and coherence of two channels has its importance when dealing with synthetic talking heads. But, when dealing with humans, does the combination of two different channels has an effect on the coherence of the output message? In other words, what happens when mixing voices with faces of different speakers? Does the resulting bimodal signal still intelligible?

To try to answer these questions, we designed an experiment, where voices and faces of human speakers were mixed together and human subjects were asked to rate the articulation of the output video. This experiment should give insight on the robustness of dubbing, and the interaction of the intelligibility of the original audiovisual presentations with the dubbed presentations.

2. Method

2.1. Stimulus materials

Ten speakers participated in this experiment: 5 males and 5 females. All were native speakers of French and had no reported speech or hearing disabilities. 9 speakers were within an age range of 22 to 30, and one was 41 years old.

Ten sentences from the Fournier's sentence list [10] were used in this experiment. These sentences are based on familiar words and are widely used in phonetic tests and intelligibility measures in audiometry. They contain a high semantic influence and can be easily understood. Each speaker uttering the 10 sentences was recorded in a reasonable soundproofing room with a SONY DCR-SR90E digital camera, placed one meter from the speaker's face who was seated in front of a black background. The entire face of the speaker was visible, including the neck. One of the 10 speakers was chosen as a model and was asked to pronounce the 10 sentences in a natural manner. Every sentence has been displayed on a monitor placed in front of the speaker's face and behind the camera. The played videos were then treated using the Adobe Premiere CS6 software in order to extract the soundtrack for each sentence. Those soundtracks were then played to the 9 others speakers who were asked to repeat each heard sentence in exactly the same speaking rate in front of the camera in order to get synchronized articulation as the model speaker. This method allowed keeping the initial audio and video signals without any signal processing (as using, for instance, the TDPSOLA technique [11]) which may alter the acoustic signal and thus one can detect easily that the audio is not original. The videotaped utterances (audio and video) were post-processed, in such a way that there was one-second-silence before the start and after the end of the uttered sentences. The audio of each sentence and

each speaker were dubbed on the video of the same sentence of a same gender speaker. The asynchrony between each audio and video component was less than 20 *msec*, which is substantially less than the asynchrony of 80 *msec* that can be noticed by some subjects [12]. This technique of dubbing provided good quality of combination of faces and voices. In several cases, it is extremely hard to tell that video was dubbed, when not familiar with the speaker. The resulting stimuli consisted of 500 video sequences (2 genders x 5 faces x 5 voices x 10 sentences). This set of videos contains also the original audiovisual sentences (10 videos for each speaker), which allows evaluating the audiovisual intelligibility of each speaker.

2.2. Subjects

Thirteen native French subjects, aged 20 to 33 years, participated in the experiment. They all had normal or corrected-to-normal vision. None of them reported hearing disabilities. None of them were familiar with the speakers. Each subject participated in approximately one hour-length session.

2.3. Procedure

Participants were seated approximately 50 cm from a 19-inch computer-screen. They were instructed to watch *and* listen to a list of small videos, and then to answer the following question: "*in this video, is the sentence correctly articulated?*", i.e., whether both face and voice look natural in terms of articulation and pronunciation. The experimenter explained well the meaning of this question. Participants were instructed to pay careful attention to both audio and the face simultaneously, and to watch the video only twice. No information about the fact that some videos were dubbed has been provided. We used the mean opinion score (MOS) test to subjectively measure intelligibility as perceived by participants. Participants were asked to rate each video by answering the question above using a Likert scale from 1 (not at all) to 5 (well articulated). We decided to use a mean opinion score (MOS) instead of a perceptual recognition experiment, for practical reason. In fact, to use perceptual recognition design, the audio or visual component needs to be degraded (adding some noise). However it will be difficult to quantify the intelligibility in the case of mixing voices with faces as one of the components is biased. We will consider in our future work different experimental design where we can tackle recognition experiments without penalizing any components and interpreting correctly the results.

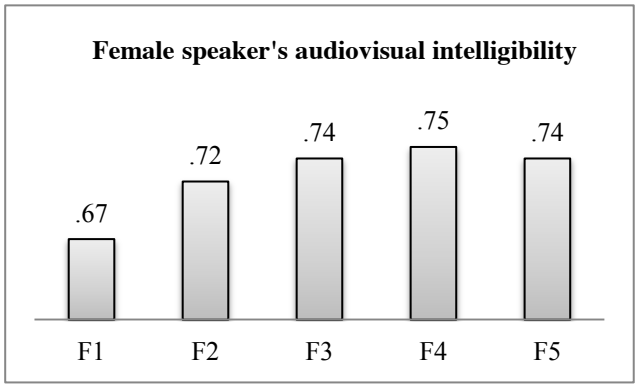

Figure 1: *Audiovisual intelligibility mean scores for the female speakers (face and audio from the same speaker).*

3. Results

In this section, we present the results of scoring the different videos by participants. Overall, 500 videos were presented where faces and voices were mixed. Among these videos, 50 were original videos (i.e., the face and the voice were of the same speaker). We tested the effects of the different presentation conditions (face and voice presentations) on the intelligibility using a Friedman test [14] for each sentence. This test revealed a significant effect for the presentation conditions ($Q = 215.611$. $p<.0001$ for female speakers and $Q = 195.198$. $p<.0001$ for male speakers). Post-Hoc tests have been done using pairwise comparisons (Nemenyi test [15]) between presentations in order to detect which presentations are significantly different from other. Differences were significant when comparisons were done between speakers who had the highest audiovisual intelligibility score and speakers who had the lowest audiovisual intelligibility score; this was the case for the original videos, but also for the mixed ones.

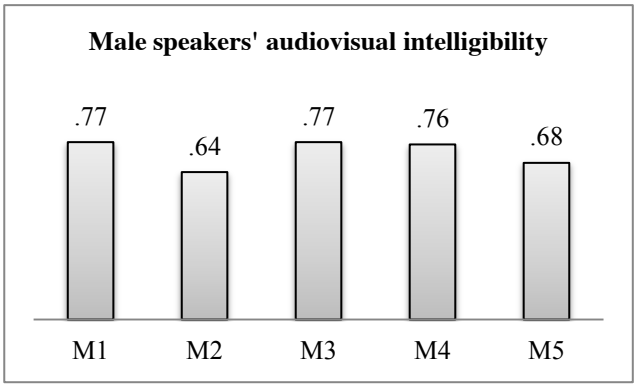

Figure 2: *Audiovisual intelligibility mean scores for the male speakers (face and audio from the same speaker).*

Normalized Audiovisual intelligibility mean scores of female and male speakers are presented respectively in Figure 1 and Figure 2. The audiovisual intelligibility means varied from .67 to .75 for female speakers and from .64 to .77 for the male speakers. Figure 1 suggests that the female speaker F1 presents the lowest audiovisual intelligibility score, while F4 has the highest audiovisual intelligibility score, closely followed by F3 and F5. F1's score is significantly different from F4's ($Q(F1,F4)=0.485$. $p=0.013$). However, there is no other significant difference between the other female speakers.

Figure 2 suggests that the male speaker M2 presents the lowest audiovisual intelligibility score, while M1 and M3 present the highest audiovisual intelligibility scores. The results showed that the score of M2 is significantly different from M1 ($Q(M2,M1)=0.888$. $p<0.0001$), M3 ($Q(M2,M3)=0.842$. $p<0.0001$) and M4 ($Q(M2,M4)=0.846$. $p<0.0001$) according to the Nemenyi test. The score of M2 is followed by one of M5, which is also significantly different from M1 ($Q(M5,M1)=0.658$. $p=0.001$), M3 ($Q(M5,M3)=0.612$. $p=0.002$) and M4 ($Q(M5,M4)=0.615$. $p=0.002$). But the differences between the

scores of speakers M2 and M5 are not significant. Tests showed that there is no significant difference between speakers M1, M3 and M4.

Table 1 and Table 2 present the normalized means of intelligibility scores and the standard deviation (STD) for the different dubbed and original conditions for the female and male speakers. The original video results are those of the diagonal. The rows present the scores when a given voice was used with different faces, and the columns the scores when a given face is used with different voices. The standard deviation did not exceed .25 for female speakers and .23 for male speakers. Across all the faces-voices combinations, no mean scores were lower than .56, which is good result *per se*. The mean values corresponding to the original presentations (face and voice congruent) were not systemically the highest values for every speaker. Some speakers, as F1 for instance, have the lowest scores when their voices are presented with the other faces. The speaker M1 has the best results when his voice is presented with the other faces and also when his face is presented with the other voices. This suggests that M1 has the highest audiovisual intelligibility.

Table 1. *Normalized means of intelligibility scores (M) and standard deviation (STD) for the female speakers. Each cell A_{ij} presents the results when the voice of the row i is dubbed with the face of the column j. The cells A_{ii} are the results of the original video.*

Female speakers		Face									
		F1		F2		F3		F4		F5	
		M	STD	M	STD	M	STD	M	STD	M	STD
Voice	F1	**.67**	.24	**.60**	.24	**.56**	.23	**.62**	.22	**.62**	.25
	F2	**.64**	.22	**.72**	.20	**.68**	.22	**.72**	.21	**.67**	.23
	F3	**.74**	.23	**.75**	.19	**.74**	.22	**.77**	.18	**.74**	.21
	F4	**.68**	.21	**.73**	.20	**.71**	.21	**.75**	.22	**.72**	.23
	F5	**.69**	.21	**.69**	.20	**.68**	.22	**.70**	.22	**.74**	.20

Table 2. *Normalized means of intelligibility scores (M) and standard deviation (STD) for the male speakers. Each cell A_{ij} presents the results when the voice of the row i is dubbed with the face of the column j. The cells A_{ii} are the results of the original video.*

Male speakers		Face									
		M1		M2		M3		M4		M5	
		M	STD	M	STD	M	STD	M	STD	M	STD
Voice	M1	**.77**	.20	**.67**	.22	**.73**	.21	**.76**	.21	**.70**	.20
	M2	**.72**	.21	**.64**	.22	**.69**	.21	**.69**	.22	**.67**	.20
	M3	**.80**	.19	**.69**	.22	**.77**	.21	**.76**	.22	**.76**	.20
	M4	**.76**	.19	**.64**	.22	**.71**	.22	**.76**	.20	**.68**	.19
	M5	**.72**	.22	**.63**	.23	**.70**	.21	**.71**	.22	**.68**	.22

The speaker M2 has the lowest intelligibility scores when his face is presented with the other speakers' voices, but we can observe that his audiovisual intelligibility score when his face is presented with his own voice is lower than the cases where his face is presented with the voices of the speakers M1 and M3.

To study the effect of the voice over the face, and the effect of the face over the voice, we pulled the means of each presentation where a voice (*resp*. a face) was dubbed with the different faces (*resp*. the different voices). The goal is to give a global intelligibility score for the voice independently of the face and a global score for the face independently of the voice, for each speaker. This should be considered as an approximation and should not be considered as a classical evaluation in the case of unimodal audio presentations. Based on this definition, the audio intelligibility scores and the visual intelligibility scores for female and male speakers are presented respectively in Figure 3 and Figure 4, in addition to the audiovisual intelligibility scores.

Figure 3 and Figure 4 give an explanation for the low audiovisual intelligibility of F1 and M5. In fact, it is very likely that the voice has a lower intelligibility than the face, if evaluated separately. This is the case for F1, M1, M4 and M5, but not for the other speakers. The speakers F2, F4, F5 and M1, M3, M4 have a better audiovisual intelligibility score than their audio and visual intelligibility scores when considered separately. The intelligibility of the face seems to be insufficient to improve the overall audiovisual intelligibility. As can be observed for F2, F3, F4, F5 and M3, when the voice intelligibility is higher than the face intelligibility, the audiovisual intelligibility seems to increase. This was not the case for M2, which is probably due to the lower overall intelligibility.

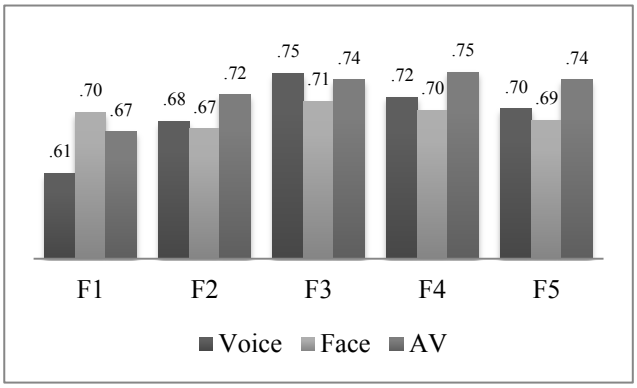

Figure 3: *Comparisons between means of global audio intelligibility (speaker's* Voice *across all the speakers' faces), global visual intelligibility (speaker's* Face *across all the speakers' voices), and global audiovisual intelligibility (speaker's face and voice:* AV*) for each female speaker.*

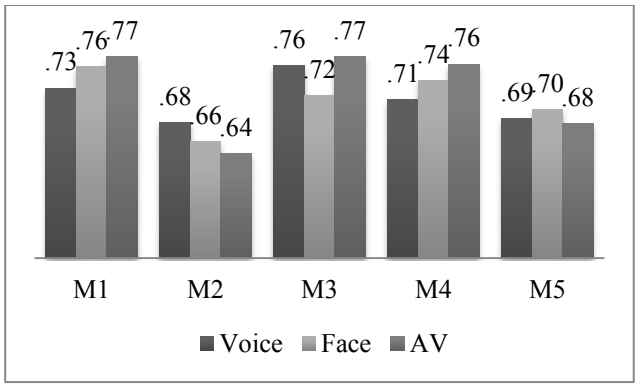

Figure 4: *Comparisons between means of global audio intelligibility (speaker's* Voice *across all the speakers' faces), global visual intelligibility (speaker's* Face *across all the speakers' voices), and global audiovisual intelligibility* (*speaker's face and voice:* AV) *for each male speaker*

4. Discussion

A very interesting observation of Table 1 and Table 2 is that when combining a highly intelligible face (that of the speaker with the highest audiovisual intelligibility score) with other voices, it provides higher intelligibility results than when the voice is used with its own face. It is also the case, when combining a highly intelligible voice (that of the speaker with the highest audiovisual intelligibility score) with the other faces; it provides higher intelligibility results when the face is used with its own face. The results of our study clearly suggest the influence of the voice and/or the face on audiovisual intelligibility. This influence can depend on the lack of information from the other component. The information given by the face seems to complete the deficiency of information from the voice and to increase the intelligibility, when the voice has a low intelligibility. This can also be the case when the visual information delivered by the speaker's face is not sufficient to fully understand the message (due to a bad articulation, for example). The voice can add auditory information to increase the intelligibility. However, this influence is not the same across all the speakers. This can be explained by the subjective representation of the speaker: when the face of the speaker is shown, the receiver can use the visual information to create the own representation of what the speaker may sound like. Thus, there are cases where the voice of the speaker is not compatible with this representation and the influence of the voice and/or the face on the intelligibility can be reduced. This can be a reason why the influence of the voice or the face is not the same in a specific combination than another. This hypothesis is compatible with the idea that listeners integrate the visual and auditory integration to understand a message.

The speaker with lowest audiovisual intelligibility scores tends to lower the intelligibility when using his/her own voice or own face with the other speakers. Nevertheless, and in this case, speakers with high intelligibility seem to be robust to this degradation and the result of this mixing provides higher intelligibility than the audiovisual intelligibility of the speaker with lowest intelligibility scores.

In the conditions of this experiment, when dubbing with the same speech rate voice, the impact on the intelligibility seems to be very limited. Almost in all the combinations the intelligibility scores were acceptable and above average (0.5). We should notice also that the highest intelligibility scores did not exceed (0.8). Depending on the audiovisual intelligibility of the speakers, the dubbed video may provide higher intelligibility results than the original video, particularly when the original video present low intelligibility results.

Based on the outcome of this study, we can speculate that synthesizing talking heads by combining audio and face originating from different sources can still provide good result and the audiovisual intelligibility can be very good. However, the important condition, as implicitly suggested by this study, is that both audio and visual channels should have the same speaking rate, i.e., "well articulated": the articulatory trajectory patterns for both acoustics and facial deformation should be very close. For visual channel, this can be modeled by a coarticulation model. For the acoustic channel, we suggest to keep a tight link with the visual channel to conserve some coarticulation features, which at least reflect speaking rate.

We should mention that these conclusions are based on a subjective perceptual evaluation. The evaluation gives a global score for the articulation of a given utterance. It is not possible to conclude on what part of the utterance there might be intelligibility issues. We plan to extend this work with an objective perceptual evaluation (a recognition experiment), where we can develop more accurate analysis at the phoneme level. This may, however, increase the complexity of the experimental design.

5. References

[1] Sumby, W. H. and Pollack, I., "Visual contribution to speech intelligibility in noise". J. Acoust. Soc. Am. 26. 212–215. 1954.

[2] Bailly, G., Bérar, M., Elisei, F. and Odisio, M., "Audiovisual speech synthesis". International Journal of Speech Technology 2003, 6(4):331–346, 2003.

[3] Theobald, B. J., "Audiovisual speech synthesis". In ICPhS, Saarbrucken, Germany, 2007.

[4] Dixon, N. F. and Spitz, L., "The detection of audiovisual desynchrony". Perception, 9:719–721, 1980.

[5] Green, K. P. and Kuhl, P. K., "Integral processing of visual place and auditory voicing information during phonetic perception". Journal of Experimental Psychology: Human Perception and Performance, 17:278–288, 1991.

[6] Jiang, J., Alwan, A., Keating, P. A., Auer, E. T. and Bernstein, L. E., "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech". EURASIP Journal on Applied Signal Processing, 11:1174–1188, 2002.

[7] McGurk, H. and MacDonald. J., "Hearing lips and seeing voices". Nature 264, 746–748, 1976.

[8] Green, K. P. and Kuhl, P. K., "The role of visual information in the processing of place and manner features in speech perception". Perception and Psychophysics, 45:34–42, 1989.

[9] Mattheyses, W., Latacz, L. and Verhelst, W., "On the importance of audiovisual coherence for the perceived quality of synthesized visual

speech". EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[10] Fournier, J. E., "Audiométrie vocale," les épreuves d'intelligibilité et leurs application au diagnostic. à l'expertise et à la correction prothétique des surdités",Paris VI, édition Maloine, 1951.

[11] Moulines, E. and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". Speech Communication, 9 (5 - 6):453 – 467, 1990.

[12] McGrath, M. and Summerfield, Q., "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults". Journal of the Acoustical Society of America. 77(2), 678–685, 1985.

[14] Friedman, M., "A comparison of alternative tests of significance for the problem of m rankings". The Annals of Mathematical Statistics 11 (1): 86–92, 1940.

[15] Nemenyi. P.B., "Distribution-free Multiple Comparisons". PhD thesis. Princeton University, 1963.