Auditory-Visual Speech Processing
(AVSP'98)
Terrigal - Sydney, Australia
December 4-6, 1998

Evaluation of a Visual-FM system to enhance speechreading

Jean-Pierre Gagné[1], *Kim Le Monday*[1], *Christine Desbiens*[1], *Marie Lapalme*[2] *& Luc Ducas*[2]

[1]École d'orthophonie et d'audiologie, Université de Montréal, Montréal, Québec, Canada, H3C 3J7

[2]Audisoft Technologie, Boucherville, Québec, Canada

ABSTRACT

An experiment was conducted to investigate the potential benefits of a visual-FM system for speechreading. Two speakers took part in the study. They were recorded while they spoke 36 sentences at a distance of 1.83, 3.66, and 7.32 m from a stationary Hi8 video camera. Under each experimental condition, the signal available from the camera of a visual-FM system was recorded simultaneously. A speechreading test consisting of a randomization of 432 recorded sentences was administered to a group of 16 subjects with normal hearing and normal (or corrected normal) visual acuity. The results revealed that for the recordings obtained with the Hi8 mm camera speechreading performance decreased as a function of distance. Specifically, there were no differences between the recordings made at 1.83 and 3.66 m. However, there was a significant difference between those two recordings distances and the recordings made at 7.32 m. For the recordings obtained with the visual-FM camera, speechreading performance did not vary significantly as a function of distance. Those findings indicate that the visual-FM system constitutes an effective method of providing visual-speech cues in environmental conditions where the distance between the speaker and the speechreader is not optimal for live-speechreading.

1. INTRODUCTION

There is ample evidence that visual-speech cues improve speech-perception performance among persons with normal-hearing [1] as well as among those with a hearing loss [2,3]. MacLeod and Summerfireld [4] reported that the improvements obtained by the provision of visual cues were equivalent to an effective increase in Signal-to-noise ratio (SNR) of more than 6 dB.

Several environmental factors are known to have a deleterious effect on speechreading performance. Clinical reports and experimental data [5] indicate that the proper illumination of the speaker's face is an important factor for speechreading. Specifically, speechreading performance is optimized when the predominant source of light is in front of the speaker and directed towards the person's oral cavity. Wozniak-Kaelin and Jackson [6] reported no difference in performance when normal-hearing subjects speechread a speaker from 0^{o} and 90^{o} on the horizontal plane. However, Erber [5] found that, relative to a viewing angle of 0^{o}, the performances of adolescents with a hearing loss were poorer by as much as 22% when speechreading was completed at a 90^{o} angle (re: the speaker's face). Erber concluded that speechreading performance was optimal when the viewing angle was between 0^{o} and 45^{o} on the horizontal plane. Speechreading performance decreases as a function of the distance between the speaker and the speechreader. According to Erber [5], speechreading performance decreases at a rate of 0.8% per foot for speaker-speechreader distances between 5 - 70 feet.

Many communicative environments are not optimal for speech perception, especially among persons with a permanent hearing impairment. Classrooms and lecture halls are good examples of settings in which the acoustic and optical characteristics are often not favorable for speech-perception. Often, the noise level in classrooms interferes with auditory-speech perception. Also, the distance and the viewing angle between a student and the teacher may not be optimal for speechreading. Moreover, teachers tend to move from one section of the room to another while they lecture and they must turn their backs to the students when they write on the chalkboard. Those conditions are not conducive to visual-speech perception.

Several audio-amplification devices (i.e., auditory-FM systems, infrared amplification devices, free-field amplification systems) have been designed to improve speech perception in settings with poor environmental characteristics [7]. However, few technical aids have been designed to optimize visual-speech cues which may also be beneficial for speech perception in those settings. Recently, *Audisoft Technologie*, a company based in Montréal, developed a prototype of a visual-FM system intended to optimize visual-speech perception impoverished speechreading conditions. The purpose of the present study was to evaluate the potential benefits of the visual-FM system. Specifically, the use of the visual-FM system to overcome difficulties related to visual-speech perception of a speaker located beyond the optimal distance for speechreading was investigated.

1.1 Description of the Visual-FM system

The underlying principles of the visual-FM system are similar to those applied to the better known auditory-FM sound transmission systems. The difference is that in the visual-FM system, the signal being transmitted is a visual signal, specifically the speaker's face and especially that person's lips.

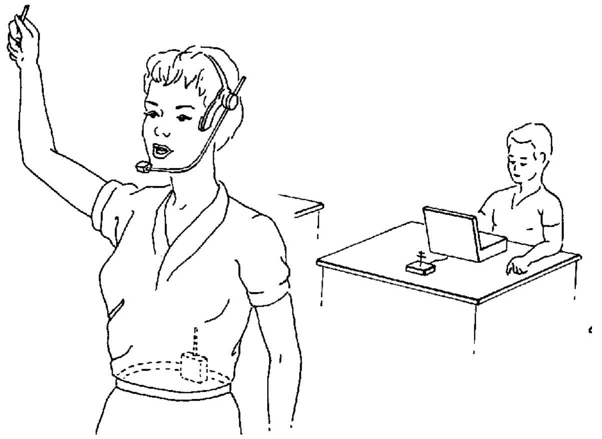

Figure 1: Illustration of a visual-FM system used to transmit speechreading information.

The prototype model of the visual-FM system consists of five basic components. First, a miniature camera (approx. 4 cm^3) is mounted on an adjustable headset worn by the speaker (see Figure 1). Typically, the camera is set at a distance of approximately 17 cm from the speaker's lips. The signal from the camera is routed (via a cable) to a decoder located on a belt worn by the speaker. The belt is also equipped with an fm-transmitter and a 12-volt rechargeable battery-pack used to power the system. In addition, the system includes an fm-receiver unit that can be placed anywhere in the room where the system is being used (see Figure 1). The signal from the fm-receiver is connected to a television monitor that is placed in a location that is easily visible to the speechreader. The signal consists of a portion of the speaker's face, extending from just below the eyes to just below the chin. The signal remains constant and stable regardless of the distance and the viewing angle between the speaker and the speechreader (i.e., the speaker's face is visible even when the person is positioned at an angle of 180^o relative to the speechreader's face) .

2. METHOD

2.1 Participants

Two adult female-speakers participated in the investigation. The primary language of communication of both speakers was French ("franco-québécois"). Both were completing their graduate degree in Audiology and neither of them displayed any atypical speech characteristics nor any noticeable oral-facial abnormalities, as judged by the primary investigator (J-PG). Sixteen (n=16) undergraduate students registered in the Speech-Language and Hearing program at the Université de Montréal completed the visual speech-perception task. They had normal hearing and normal (or corrected normal) visual acuity.

2.2 Test stimuli

The stimulus set consisted of sentences. Each sentence had the same syntactic structure and contained three critical elements (Subject + Verb + Complement) which served as the keywords in the visual-speech recognition task. Five interchangeable alternatives were selected for each critical element, generating a total of 125 different sentences. Within each critical element, the foils had the same number of syllables but the words were distinct from each other visually and acoustically (see Table 1).

2.3 Procedures

Recording procedures: First, the headset of the visual-fm system was positioned on the speaker. Adjustments were made to focus the recorded image on the speaker's mouth. A viewing television monitor (Sony PVM1350) was used to accomplish this task. Then, a conventional Hi8 mm video camera (Sony CCD-V5000), set on a tripod, was positioned at a distance of 1.83 m directly in front of the speaker (0^o on the horizontal plane and 0^o on the vertical plane). The lens of the camera was focused on the speaker's face and the zoom option of the lens was adjusted so that the image displayed on a television monitor (Sony PVM1350) approximated the size of the live image typically seen by a person from that distance. Once adjusted, the lens of the conventional camera remained unchanged for the remainder of the experiment. Recordings were obtained simultaneously from the camera of the visual-FM system and from the conventional 8-mm video camera. The recordings were made at three distances (between the conventional camera and the speaker): 1.83, 3.66, and 7.32 m. At each distance, the speaker spoke a set of 12 different sentences selected randomly from the set of 125 possible sentences (see Table 1). The recordings were obtained twice at each distance. For each speaker, the order in which the recordings were obtained (i.e., the distances) was randomized.

Subject	Verb	Complement
Les ***parents***	***marchent***	dans le ***jardin***.
Les ***personnes***	***dorment***	dans le ***canal***.
Les ***grands-pères***	***parlent***	dans le ***chalet***.
Les ***garçons***	***entrent***	dans le s***ous-sol.***
Les ***messieurs***	***rient***	dans le ***grenier***.

Table 1. List of keywords used in the sentence recognition task.

Editing procedure: A Sony Hi8 editing unit (EVO-9700) was used to edit the master recordings. A speechreading test consisting of 432 items (2 speakers x 12 sentences x 3 iterations x 3 distances x 2 types of cameras) was prepared. The test items were completely randomized across sentences, iterations, recording distances and speakers. Each test item consisted of a 3s test-item identification number, the test sentence, and a 6 s written message that appeared on the TV monitor and prompted the subjects to provide a response.

Test procedure: Five groups of 3 or 4 subjects completed the visual speech-perception task during one test session (approximately 90 minutes). The subjects were seated in a semi-circle approximately 1.25 m from a 33 cm color television monitor (Sony Trinitron model KV20510), at an angle that permitted an undistorted view of the video image. To reduce learning and familiarization effects, the starting position of the test tape was staggered across each group of subjects. The task consisted of close-set recognition: Each subject was given a response form that indicated the possible keywords for each test item. The task was completed in a visual-only mode. The subjects were instructed to attend to each test item and to indicate their response at the appropriate place on the response form by circling the three keywords that were presented in each sentence. Also the subjects were instructed to supply an answer for each keyword and to guess if they were not certain of the correct response. The subjects completed 10 practice trials (not included in the present investigation) before they began the experiment.

Scoring: The subjects were credited with a correct response only if they indicated the correct keyword at the appropriate place on the response form. A speaker's visual-speech intelligibility was determined by the mean performance obtained from the group of subjects in a given experimental condition. The two iterations produced by a given speaker for a given condition were averaged.

In order to produce a scale in which the size of the variance would be unrelated to the mean performance, a 'rationalized' arcsine transform (RAU) was applied to the raw data [8]. A visual-speech intelligibility score was calculated for each speaker in each of the six experimental conditions (3 distances x 2 types of camera).

3. RESULTS

The visual-speech intelligibility scores obtained as a function of distance, for each type of camera, are displayed in Figure 2. The results obtained for Speakers 1 and 2 are shown on the top and bottom panels, respectively. A 3-way ANOVA for repeated measures revealed that there was a significant main effect for speakers ($F = 32.791$, $df = 1$, $p < .0001$) and distance ($F = 9.062$, $df = 2$, $p <=.0002$). The effect of camera-type was not significant ($F = 2.615$, $df = 1$, $p = .1076$). Also, there was a significant Speaker x Camera-type interaction ($F = 19.464$, $df = 1$, $p < .0001$) as well as a significant Camera-type x Distance interaction ($F = 6.897$, $df = 2$, $p =.0013$). The 3-way Distance x Speaker x Camera-type interaction was not significant ($F = .297$, $df = 2$, $p = .7432$). Post-hoc analyses revealed that for the recordings obtained with the visual-FM camera the visual-speech intelligibility scores did not vary significantly as a function of distance. This result was observed for both speakers. For the recordings obtained with the conventional 8 mm camera the speech intelligibility scores differed significantly as a function of distance. Specifically, there were no difference between the recordings made at 1.83 and 3.66 m. However, there was a significant difference between those two recording distances and the recordings made at 7.32 m. This results was observed for both speakers.

4. CONCLUSION

In summary, the results of the investigation revealed that the visual-FM system constitutes an effective method of providing visual-speech cues in environmental conditions where the distance between a speaker and a speechreader is not optimal for live speechreading.

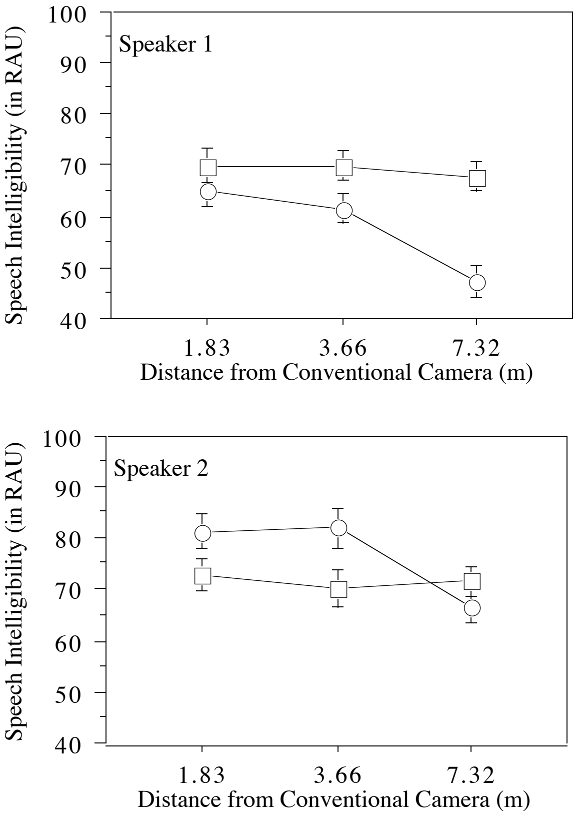

Figure 2. Mean visual-speech intelligibility of Speaker 1 (top panel) and Speaker 2 (bottom panel) as a function of distance between the speaker and the conventional camera (in meters). In each panel speech intelligibility scores obtained with the conventional camera (circles) and the visual-FM camera (squares) are shown separately. Error bars indicate one standard error.

The experimental data revealed that the visual-FM system was effective for speechreading a speaker up to a distance of 7.32 meters. However, informal tests conducted in our laboratory indicate that the optical signal available for the camera is of the fm-system is clearly visible to a distance of up to 30 m.

The prototype version of the visual-Fm system used for the present investigation was effective but not very convivial for the speaker. In its present form the system is heavy and the headset that must be worn by the speaker is not comfortable. However, based in part on the results of our preliminary investigations, the company that developed the first prototype of the visual-FM system is currently designing a second generation model that will be less cumbersome and more aesthetically pleasing.

5. ACKNOWLEDGMENTS

This project was supported in part by a research grant from the National Sciences and Engineering Research Council of Canada (NSERC) awarded to the first author (J-PG). Gratitude is also expressed to Marie Lapalme and Luc Ducas of Audisoft Technologie, for allowing the researchers to evaluate the visual-FM system developed by their firm. Finally, the authors would like to express their thanks to Monique Charest for her help with the statistical analyses and the preparation of the manuscript.

6. REFERENCES

1. Sumby, W.H., and Pollack, I. "Visual contribution to speech intelligibility in noise*," J. Acoust. Soc. Am., Vol. 26(2): 212-215,1954.*

2. Erber, N.P., "Interaction of audition and vision in the recognition of oral speech stimuli," *J. Speech Hear. Res., 12: 423-425, 1969.*

3. Erber, N.P., "Auditory-visual perception of speech with reduced optical clarity," *J. Speech Hear. Res., 22: 212-223, 1979.*

4. MacLeod, A., and Summerfireld, Q., "A procedure for measuring auditory and audio-visual speech-perception thresholds in noise: Rationale, evaluation, and recommendations for use," *Br. J. Audiol., 24: 29-43, 1990.*

5. Wozniak-Kaelin, V.D., and Jackson, P.,L., "Visual vowel and dipthong perception from two horizontal viewing angles*," J. Speech Hear. Res., 22: 354-365, 1979.*

6. Erber, N.P., "Effects of angles, distance, and illumination on visual reception of speech by profoundly deaf children*," J. Speech Hear. Res., 17: 99-112, 1974.*

7. Ross, M. (Ed.). (1994). *Communication access for persons with hearing loss: Compliance with the Americans with Disabilities Act.* Baltimore: York Press, Inc.

8. Studebaker, G.,A., "A 'rationalized' arcsine transform," *J. Speech Hear. Res., 28: 455-462, 1985.*