SPEECH INTELLIGIBILITY OF SYNTHETIC LIPS AND JAW

T. Guiard-Marigny (1), C. Benoît (1) and D.J. Ostry (2)
(1) Institut de la Communication Parlée, INPG-Université Stendhal, Grenoble, France
(2) Psychology Department, McGill University, Montreal, Canada.

ABSTRACT

Audio-visual speech intelligibility was tested using high-quality 3D models of the lips and jaw. The models were animated on the basis of six parameters obtained from the actual movements of a speaker's face and synchronized with the original audio utterances. Eighteen French nonsense utterances were presented to twenty subjects at five levels of added noise. Intelligibility was best when the lip and jaw animations were presented along with the acoustic speech signal.

INTRODUCTION

Even though the auditory modality is dominant in speech perception, it has been shown that seeing the speaker's face increases intelligibility, especially in a background noise [12, 3, 4, 13, 2]. Synthetic faces are thus expected to enhance the intelligibility of speech synthesizers which is still far lower than that of humans.

It has been shown in English [9], and then in French [8] that the human lips carry more than a half of the visual information provided by the whole natural face. Moreover, vision of the teeth increases the intelligibility of a message: the teeth help disambiguate sounds differing in jaw position like "bib" versus. "bab" [9].

In this paper we evaluate through a perception test the contribution to speech intelligibility of a lip model alone and of the same lip model superimposed upon a synthetic jaw and upper skull.

THE 3D LIP MODEL

The 3D model of the lips used in this study was developed on the basis of a geometrical analysis of the natural lip movements of a French speaker [5]. The model is controlled with five parameters which can be measured directly from the recorded lip movements of a real speaker's face. A specially designed workstation [7] is used to obtain accurate measures of the parameters from a videotape. The measurement procedure produces an output file which contains the five parameters measured at 20 ms intervals; this file is used as a command file to our model. The digitized voice of the natural speaker is synchronized with the visual display.

THE JAW MODEL

Apart from the lips, the most visible articulator is the jaw and with it, the chin and the teeth. Since the jaw is a rigid skeletal structure, the animation process is easier than with the lips. Like all rigid objects, jaw motions have six degrees of freedom. Thus, its position relative to the skull can be defined with three orientation angles (yaw, pitch, roll) and three positions (horizontal, vertical, lateral).

The synthetic jaw which was used for our model was developed at McGill University [6] in order to visualize jaw motion kinematics that are recorded with an optoelectronic measurement system. It comprises a 3D digitized upper skull and jaw along with their corresponding teeth. The jaw model is animated using empirically recorded jaw orientation angles and jaw positions [6]. The visual display of the synthetic upper skull and jaw was synchronized with the corresponding natural audio signal.

ANIMATION OF THE MODELS

The lip and jaw models were integrated in a single display. The lip model was directly superimposed on the 3D skull and jaw. For tests of the model, lip movements were obtained using the video analysis technique described above. Jaw movements were obtained in a similar manner from the motion of the chin of the speaker using image processing techniques like those developed for lip movement. It should be noted that while it would have been desirable to use the optoelectronic measurement system at McGill to obtain jaw motions, this technique requires the use of an acrylic and metal dental

appliance which makes it difficult to measure lip movement.

Jaw motions in speech are controlled primarily in three degrees of freedom [10], namely the pitch angle, the vertical position and the horizontal position. The positions of two points on the jaw are sufficient to reconstruct these three motions in the sagittal plane. However, since the jaw is not directly visible and the overlying skin moves relative to the jaw, the points needed to reconstruct sagittal plane jaw motion cannot be obtained with non-intrusive methods. Nevertheless, it can be seen from the data reported in [11] that the basic parameters of jaw motion are often strongly correlated in running speech. To a first approximation, the three basic jaw motions can thus be predicted from the displacement of a single point on the jaw. Since the teeth are not always visible, we have decided to obtain this single point by tracking a dot on the chin. Of course, in so doing, a discrepancy cannot be avoided between the actual jaw motion and that of the reference point on the chin.

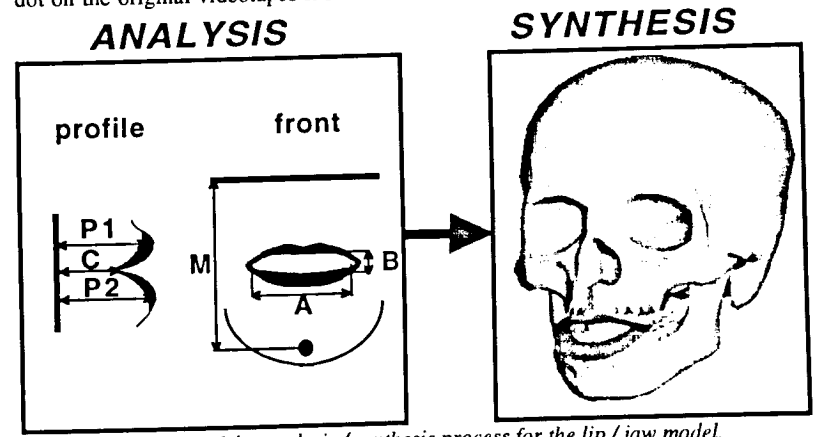
For purposes of our first tests of the lip / jaw synthesizer we have used an audio-visual corpus which has already been used extensively at ICP in order to make geometric measurements [1, 5] and to evaluate the contribution of vision to speech intelligibility [2, 8]. Since the speaker's chin was made up with a single dot on the original videotapes it seemed sufficient for the initial evaluation. A schematic of the analysis and synthesis process used to obtain the animation is presented in Figure 1.

INTELLIGIBILITY OF THE MODELS

Following two previous experiments [2, 8], the audio-visual intelligibility of the lip model and of the superimposed models of the lips and jaw (called the lip / jaw model) have been tested at five levels of acoustic degradation.

Preparation of the Stimuli

The speech material consisted of the natural acoustic utterances of a French speaker synchronized with three kinds of display: no video, synthetic lips, synthetic lips and jaw. The corpus consisted of VCVCV nonsense utterances. The vowels tested were /a/, /i/ and /y/. The consonants were /b/, /v/, /z/, /ʒ/, /ʁ/ or /l/. The test words were embedded in a carrier sentence of the form "C'est pas VCVCVz ?". Eighteen different sentences were first digitized and then acoustically degraded by addition of white noise, at five signal to noise levels, in 6 dB steps. Thus overall, there were 90 different acoustic stimuli. A pseudo-random order was used for presentation of the stimuli to subjects. Ten additional stimuli preceded the actual test in order to help subjects adapt to the test conditions.



Figue 1. Schematic of the analysis / synthesis process for the lip / jaw model.

The same sequence of acoustic stimuli was used in all three experimental conditions. The visual stimuli for the two synthetic models were recorded frame by frame on a videotape. The models were synthesized at a 25 ips rate with the virtual camera located at a 25° angle from the sagittal plane. The audio stimuli were subsequently synchronized with the visual display.

Twenty normal French listeners took part in the experiment. They were seated at a 1m distance from a 15" color monitor equiped with a loudspeaker. The order of presentation of the three sub-tests was balanced across the subjects. The subjects were required to identify both the vowel and the consonant in each utterance.

Global intelligibility

A test word was considered correct only if both the vowel and the consonant were correctly identified. The intelligibility scores obtained with the audio alone and with the lip model in this experiment were comparable to those reported in [8]. The data showed that the lip model restored approximately a third of the missing information when the acoustic signal was degraded. Moreover, we obtained a noticeable gain in speech intelligibility when the synthetic jaw was added to the synthetic lips, as shown on Figure 2.

Confusions

When the visual display of lip movement was added, the identification of /b/ was improved. However, /b/ was often given as the response to /v/ stimuli. The identification of the other consonants was also improved except for /ʒ/. For the vowels, /y/ was almost always correctly identified but /i/ and /a/ were still confused.

When synchronized with the lip model, the jaw model generally enhances intelligibility. The number of cases in which subjects were unable to respond at all was reduced by a factor of two. The vowel /i/ was confused less with the vowel /a/, mostly in consonantal contexts that close or tend to close the lips (/b/ or /v/). In addition, there were less confusions between /ʒ/ and /ʁ/, regardless of the vocalic context. Moreover, /b/ was no longer confused with /v/, especially in the context of the vowel /i/. The visibility of the teeth presumably accounts for this disambiguation.

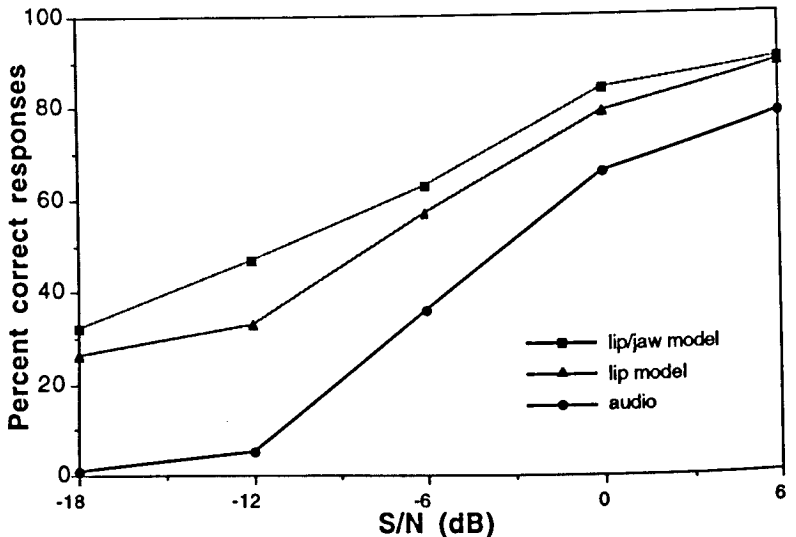


Figure 2. Audio-visual intelligibility of the lip model and of the lip / jaw model compared with the intelligibility of the auditory signal alone. The data are shown for various levels of noise.

On the other hand, with the combined lip and jaw models, /l/ and /v/ were more often confused with /ʒ/. However, this only occurred in rounded vocalic contexts , such as, /y/. Vision of the jaw also led to a greater number of confusions between /i/ and /a/ in a /z/ context. This was mostly due to the individual utterances /izizi/ and /azaza/ selected as stimuli for this experiment.

CONCLUSION

We obtained a noticeable gain in speech intelligibility when a synthetic jaw was displayed along with synthetic lips. However, compared to the intelligibility scores obtained in [8] with a synthetic face, the gain is small. This is likely due to the unnatural display of the lips superimposed on the skeletal skull and the jaw. Nevertheless, we can speculate that a semi-transparent skin overlaid on a display of the intrinsic articulators of the vocal tract may further improve intelligibility. The intelligibility scores in such an "augmented reality" of visible speech should be tested in the near future.

ACKNOWLEDGEMENT

This research was supported by the CNRS and by a grant from the ESPRIT-BRA progamme ("MIAMI" project No 8579), and by NIH Grant DC-00594 from the National Institute on Deafness and Other Communication Disorders.

REFERENCES

[1] Benoît, C., Lallouache, M.T., Mohamadi, T. & Abry, C. (1990), *A set of French visemes for visual speech synthesis*, Talking Machines, Bailly & Benoît, Eds, Elsevier B.V, Amsterdam, pp. 485-504.

[2] Benoît, C., Mohamadi, T. & Kandell, S.D. (1994), *Effects of phonetic context on audio-visual intelligibility in French*, Journal of Speech & Hearing Research, vol. 37, pp. 1195-1203.

[3] Erber, N.P. (1969), *Interaction of audition and vision in the recognition of oral speech stimulii*, Journal of Speech & Hearing Research, vol. 12, pp. 423-425.

[4] Erber, N.P. (1975), *Auditory-visual perception of speech*, Journal of Speech & Hearing Disorders, vol.40, pp. 481-492.

[5] Guiard-Marigny, T., Adjoudani, A. & Benoît, C.(1994), *A 3D model of the lips for visual speech synthesis*, Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA, pp. 49-52.

[6] Guiard-Marigny, T. & Ostry, D.J (1995), *Three-dimensional visualization of human jaw motion in speech*, 129th Meeting of the Acoustical Society of America, Washington, USA, to appear.

[7] Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*, Unpublished Thesis Dissertation INP, Grenoble, France, 214 pp.

[8] Le Goff, B, Guiard-Marigny, T., & Benoît, C. (1995), *Real-time analysis-synthesis and intelligibility of talking faces*, Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA, pp. 53-56.

[9] McGrath M. (1985), *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces*, Ph.D. Thesis, Univ. of Nottingham, UK.

[10] Ostry, D.J. & Bateson, E.V. (1994), *Jaw motions in speech are controlled in (at least) three degrees of freedom*, Proceedings of the International Conference on Spoken Language Processing, vol. 1, pp.41-44.

[11] Ostry, D.J. & Munhall, K.G. (1994), *Control of jaw orientation and position in mastication and speech*, Journal of Neurophysiology, vol 71, No 4.

[12] Sumby, W.H. & Pollack, I. (1954), *Visual contribution to speech intelligibility in noise*, Journal of the Acoustical Society of America, vol. 26, pp. 212-215.

[13] Summerfield, Q. (1979), *Use of visual information for phonetic perception*, Phonetica, vol. 36, pp. 314-331.