EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

A MULTIMEDIA PLATFORM FOR AUDIO-VISUAL SPEECH PROCESSING

A. Adjoudani, T. Guiard-Marigny, B. Le Goff, L. Reveret & C. Benoît
Institut de la Communication Parlée
UPRESA CNRS n° 5009
INPG-ENSERG / Université Stendhal, Grenoble, France
Tel. +33 4 82 43 36, FAX: +33 4 82 43 35, E-mail: benoit@icp.grenet.fr

ABSTRACT

In the framework of the European ESPRIT Project MIAMI ("Multimodal Integration for Advanced Multimedia Interfaces"), a platform has been developed at the ICP to study the various combinations of audio-visual speech processing, including real-time lip motion analysis, real-time synthesis of models of the lips and of the face, audiovisual speech recognition of isolated words, and text-to-audio-visual speech synthesis in French. All these facilities are implemented on a network of three SGI computers. Not only this platform is a usefull research tool to study the production and the perception of visible speech as well as audio-visual integration by humans and by the machines, but it is also a nice testbed to study man-machine multimodal interaction and very low bit rate audio-visual speech communication between humans.

1. INTRODUCTION

The user interface of the platform (see Figure 1) mostly consists of a light head-mounted device to which are attached a microphone, an earphone, and a microcamera focusing on the mouth area of the user whose lips are made-up in blue.
The AV platform (see Figure 2) allows a user to audio-visually communicate with other humans or with the machine through different means of bimodal communication, including analysis/synthesis, automatic recognition, and text-to-speech synthesis of facial gestures. These three main applications use a set of compatible modules installed on a network of SGI computers communicating via PVM (Parallel Virtual Machine). The modules. are presented below. They are also evaluated through an intelligibility assessment of the three above mentioned applications.

Figure 1. Photograph of the user's head mounted device showing the microcamera and the microphone

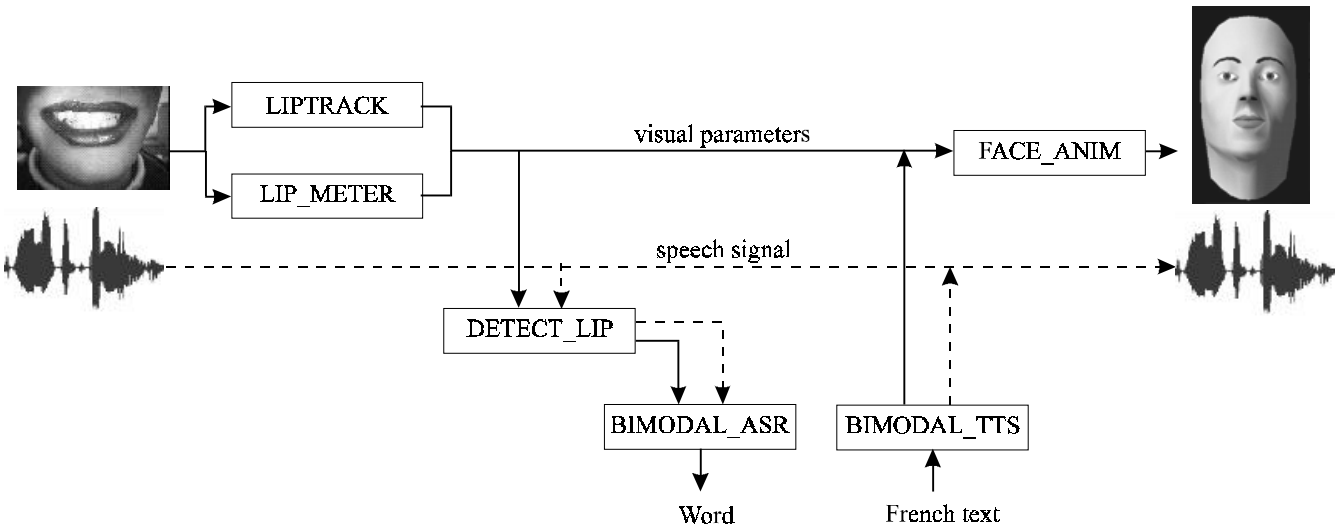

Figure 2. Overview of the AV platform showing its main components.

2. LIP GESTURES TRANSMISSION

We have implemented an analysis-synthesis of the facial gestures associated with speech production in order to allow real-time tele-lipreading at a very low bit rate through the network. Thus far, six parameters are measured on the speaker's face and simultaneously used to control 3D models of the face. Analysis id achived by the LIPMETER software as well as by the LIPTRACK dedicated hardware. Synthesis is performed by the FACE_ANIM software. Control parameters are sent from the analysis device (SGI Indy or LIPTRACK) to the synthesis workstation (an SGI or a PC equiped with a graphic accelerator).

2.1. The LIPMETER software

The LIPMETER program includes a digital chroma-key and measures anatomical parameters characteristic of lip motion 50 times per second, or at 60 Hz when an NTSC microcamera is used [7][11]. The algorithm is mostly based on that originally developped by Lallouache [9] for high accurate measurement of lip gestures in speech production.

2.2. The LIPTRACK hardware

The LIPTRACK system ("*Labiomètre Indépendant Programmable Temps Réel Avec Chroma-Key*") is a real-time independent device based on a Motorola 56002 Digital Signal Processor which detects the lip contours and measures lip geometry from the binary vector that comes out of a LookUp Table directly connected to a video digitizer (BrookTree BT812 chip). Lip parameters are then accessible to any computer (or to any transmission line) through an RS232 interface [1]. Again, the same lip detection algorithm has been implemented in assembler onto the DSP. The analysis rate is 50 or 60 parameters per second, depending on the video signal format of the input.

2.3. The FACE_ANIM software

The FACEANIM program allows real-time animation of a lip model, a jaw model, a face model, or combination thereof [7][8][11]. It takes corresponding articulatory and audio files as the input (or live streams in the real-time analysis/synthesis version) and synchronizes the visual animation with the acoustic display.

2.4. Evaluation of the Analysis-Synthesis

Intelligibility tests have been run under the paradigm widely used since the Fifties for the assessment of natural speech presented auditorily and audio-visually to normal hearers under various conditions of background noise [4][5][6][16][17]. The same corpus as that used by Benoît et al. [4] with natural stimuli was systematically used in all our experiments with synthetic stimuli. It consists of 18 non-sense words of the form VCVCV with V = [a, i, y] and C = [b, v, z, J, R, l]. Percent correct identification of these words utterd by the synthetic lips, jaw, face, or combination thereof, presented or not in synchrony with the original audio waveform have been evaluated in [8][10][11]. Results with the synthetic face are presented on Figure 3. We observe a dramatic improvement of speech intelligibility when a synthetic face (or even part thereof) is available to the subjects, even though the facial models only require a few hundred bits/s to be animated [10].

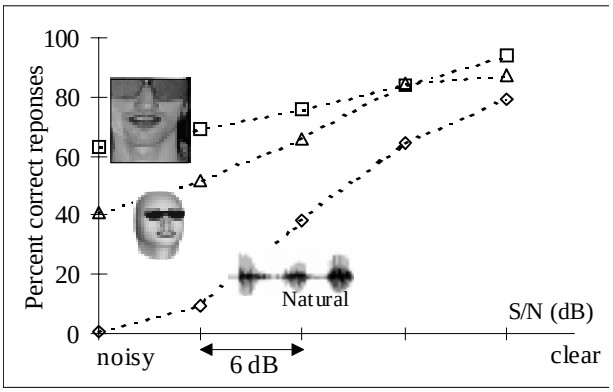

Figure 3: Audio-visual intelligibility of a 3D model of a talking head controlled from an analysis of the lips and chin gestures of a real speaker, under five levels of degradation of the acoustic signal. The bottom curve shows the audio alone scores.

3. AUTOMATIC AV RECOGNITION

From the remarkable ability of humans to benefit from lipreading in the speech decoding process, it is expected that ASR machines also improve their recognition scores if optical information is provided together with the acoustic information, especially when acoustics is degraded. The challenge here is to integrate the outputs of the two channels so that AV scores are always higher than A and than V scores alone. In this perspective, we developed an AV recognizer based on conventional HMM architectures. Beside the acoustic regognizer, it includes the LIPMETER above presented, a video-based silence-speech detector robust to background noise, and an integration module.

3.1. The DETECT_LIP software

The DETECTLIP program detects lip motion in order to segment isolated speech utterances based on visual cues, thus making it robust to background noise. Speech vs. non-speech is detected according to parameters adjustable to the speaker and to the speech material used. DETECTLIP also stores on disk the synchronized audio and lip parameter files. The acoustic and articulatory files are then available for transmission to animate local or distant talking heads (see above § 2.) as well as for automatic recognition.

3.2. The BIMODAL_ASR software

The BIMODAL_ASR program allows audio, visual or audio-visual recognition of isolated words. The whole system consists of two independent HMM-based automatic recognizers, for resp. auditory and visual speech recognition. Input data comes from the DETECLIP program. The output probabilities are then processed through an integration module where they are weighted depending on the reliability of each channel, estimated from the dispersion of the first four candidates in each modality [1] [2]. The schematic of the recognition architecture is presented on Figure 4.

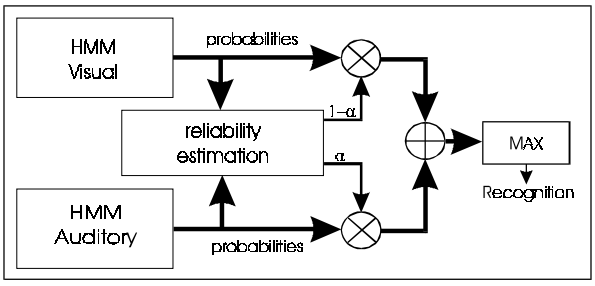

Figure 4. Sschematic of the Late Integration Model. The reliability is estimated from the standard deviation of the first candidates in each modality: The more dispersed the probabilities, the more reliable the modality.

3.3. Evaluation of the AV Recognition

The Bimodal_ASR module has been evaluated in terms of recognition scores along the same paradigm as above, under six conditions of S/N [1][2]. Figure 5 only shows the scores corresponding to one combination of 7 training + 2 test stimuli repetitions (that leading to the lowest V score) among a vocabulary of 54 different non-sense words of the same form as above, except that the central vowel could differ from the extreme vowels in CV1CV2CV1. With the architecture described on Figure 4, we see that AVrecognition scores are higher than A scores and than V scores whatever the acoustic degradation. Even the high score obtained in pure lipreading allows us consider potential applications in oral man-machine communication with confidentiality.

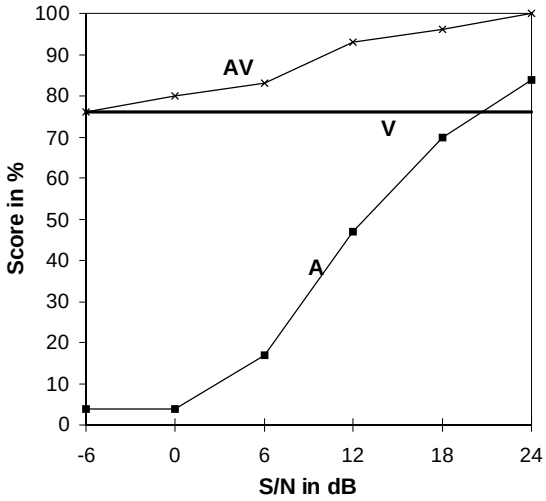

Figure 5. AV scores obtained with weighted output probabilities.

4. TEXT-TO-AV-SPEECH SYNTHESIS

The Bimodal_TTS consists of a phoneme-to-visual speech synthesizer, developed by [12], and synchronized to a text-to-acoustic speech synthesizer. The animation of the face is achieved by the FACE_ANIM program. The facial model is controled by eight articulatory parameters (five for the lips, one for the chin and two for the tongue) whose time varying trajectories are predicted by rules (see [13] for more details).

4.1. The BIMODAL_TTS software

In order to predict the eight parameters of the 3D face model, a coarticulation model based on spline-like functions has been implemented. The numerous coefficients of the coarticulation model have been identified through a automatic data-driven approach [12] [13].

4.2. Evaluation of the Text to AV speech synthesis

The system has been evaluated through the same intelligibility test and corpus as the ones described in paragraph 2.4. The global intelligibility scores obtained with the synthetic face are summerized on Figure 6.

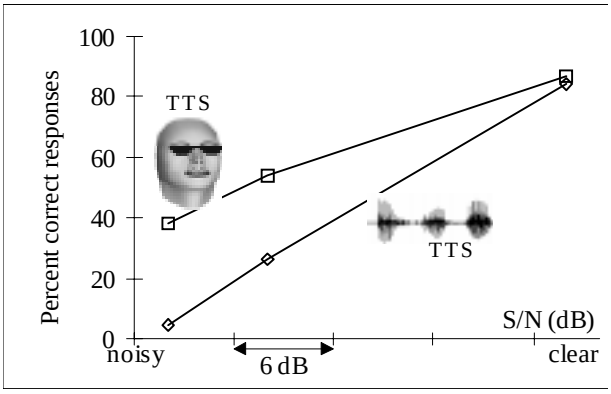

Figure 6. Intelligibility scores obtained with the text-to-audiovisual speech synthesizer.

5. PERSPECTIVES AND CONCLUSION

We have presented the architecture of a platform dedicated to audio-visual speech processing. All its components run in real-time on a network of three SGI computers. Some pieces of software also run on a PC. One piece of hardware has been specially designed.ll the modules of the platform have been evaluated in terms of recognition scores, whether by humans or by machine. Results show that the platform is well suited to basic studies on how speech is produced and perceived by humans through its two main modalities. It also serves as a functional testbed to evaluate bimodal speech communication between humans and machine. It has been largely used in this purpose within the ESPRIT-BRA "MIAMI" project ('Multimodal Integration fo Advanced Multimedia Interfaces'), where it was possible to communicate with the platform through automatic recognition of handwriting or of speech (whether aloud or wispered), and where feedback was sent to the user via facial animation either driven by prestored parameters or from unmimited French text. The platform has been nicely advertised in the short movie "Innovating Tomorrow" commissioned in 1997 by Edith Cresson, European Commisioner for Research.

In the close future, it is expected that visual input will not be limited to users with lips specially made up in blue: [14] [15] are investigating new ways to detect lip contours on natural lips.

6. REFERENCES

[1] A. Adjoudani., «*Reconnaissance de la parole audiovisuelle : stratégies d'intégration et réalisation du LIPTRACK*», PhD thesis, Signal Image Parole, Institut National Polytechnique, Grenoble, France, 1997.

[2] A. Adjoudani and C. Benoît., «*Audio-visual speech recognition compared across two architectures*», Proceedings of the 4th EUROSPEECH Conference, volume 3, pages 1563-1566, Madrid, Spain, 1995.

[3] G. Bailly and M. Guerti, *"Synthesis-by-rule for French*", Proceedings of the 12th International Congress of Phonetic Sciences, Aix-en-Provence, France, Volume 2, PP 506-511, 1991.

[4] C. Benoît, T. Mohamadi and S. Kandel, «Audio-visual intelligibility of French speech in noise», *Journal of Speech & Hearing Research*, Vol. 37, pp. 1195-1203, 1994.

[5] N.P. Erber, «Interaction of audition and vision in the recognition of oral speech stimuli», *Journal of Speech & Hearing Research*, Vol. 12, pp. 423-425, 1969.

[6] N.P. Erber, «Auditory-visual perception of speech», *Journal of Speech & Hearing Research*, Vol. 40, pp. 481-492, 1975.

[7] T. Guiard-Marigny, «*Modélisation tridimensionnelle des articulateurs de la parole : implémentation temps réel et mesures d'intelligibilité bimodale*», PhD thesis, Signal Image Parole, Institut National Polytechnique, Grenoble, France, 1996.

[8] T. Guiard-Marigny, A. Adjoudani and C. Benoît, «3D models of the lips and jaw for visual speech synthesis», *Progress in speech synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive & J. Hirschberg Editors, Springer Verlag New York, pp. 247-258, 1996.

[9] T. Lallouache, «*Un poste visage parole couleur. Acquisition et traitement automatique des contours des lèvres*», PhD thesis, Signal Image Parole, Institut National Polytechnique, Grenoble, France, 1991.

[10] B. Le Goff, T. Guiard-Marigny and C. Benoît, «*Read my lips ... and my jaw ! How intelligible are the components of a speaker's face ?*», Proceedings of the 4th EUROSPEECH Conference, Madrid, Spain, Vol. 1, pp. 291-294, 1995.

[11] B. Le Goff, T. Guiard-Marigny and C. Benoît, «Analysis-Synthesis and Intelligibility of a Talking Face», *Progress in speech synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive & J. Hirschberg Editors, Springer Verlag New York, pp. 235-246, 1996.

[12] B. Le Goff and C. Benoît, «*A text-to-audiovisual-speech synthesizer for French*», Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, Vol. 4, pp. 2163-2166, 1996.

[13] B. Le Goff and C. Benoît, «*Automatic modeling of coarticulation in text-to-audiovisual speech synthesis*», Proceedings of the 5th EUROSPEECH Conference, Rhodes, Greece, in volume, 1997.

[14] L. Reveret, «*From raw images to articulatory parameters: viseme-based prediction*», Proceedings of the 5th EUROSPEECH Conference, Rhodes, Greece, in volume, 1997.

[15] L. Reveret, F. Garcia, C. Benoit, and E. Vatikiotis-Bateson, «*An hybrid image processing approach to liptracking independent of head orientation*», Proceedings of the 5th EUROSPEECH Conference, Rhodes, Greece, in volume, 1997.

[16] W.H. Sumby and I. Pollack, «Visual contribution to speech intelligibility in noise», *Journal of the Acoustical Society of America*, Vol. 26, pp. 212-215, 1954.

[17] Q. Summerfield, A. MacLeod, M. McGrath and M. Brooke, «Lips, teeth, and the benefits of lipreading», *Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis Editors, Elsevier Science Publishers, pp. 223-233, 1989.