2nd International Conference on Speech Synthesis, Newark (NY), September 1994

Real-Time Analysis-Synthesis and Intelligibility of Talking Faces

B. LE GOFF [(1)], T. GUIARD-MARIGNY [(1)], M. COHEN [(2)], & C. BENOIT [(1)]

(1) **Institut de la communication parlée**
U.A. CNRS N°368,
INPG/ENSERG - Université STENDHAL,
BP 25X - 38040 GRENOBLE, FRANCE

(2) **Program in Experimental Psychology**
University of California
UCSC, Kerr Hall,
Santa Cruz, CA 95064, USA

Abstract

Analytic measurement of visual parameters relevant to the labial production of speech as well as real-time 3D computer animated models of the lips and of the face have been implemented on two coupled computers, so that synthetic lips alone or a whole facial model can mimic on line (or play back) the actual gestures of a natural speaker. The geometric measurements performed on the speaker's lips and jaw are made through image processing of the front and profile view of the speaker's face. Data are transmitted to a display computer through a control interface which delivers the proper parameters to control the animation of the 3D models. The lip model uses five control parameters and the facial model uses one extra one: jaw lowering. At present, the tongue is not controlled. We here present the real-time techniques used for analysis, animation of the 3D models, and synchronization of the two processes. Finally, we evaluate the bimodal intelligibility of speech under five levels of acoustic degradation by added noise. We compare the intelligibility of the speech signal presented alone, with the lip model, with the facial model, and with the original speaker's face. Our results confirm the importance of visual information in the perception of speech: The whole natural face restores two thirds of the missing auditory intelligibility when the acoustic transmission is degraded or missing; the facial model (tongue movements excluded) restores half of it; and the lip model alone restores a third of it.

1. INTRODUCTION

Parametric models of the face aim at providing the viewer with the maximum quantity of information carried on by the speaker's face, through a small number of commands . We here present a system for analysis-synthesis of speaking faces in which a small number of parameters allow lip and jaw gestures to be accurately reconstructed. First, we present models of the lips and face that we used as well as their control parameters. We then describe our parameter acquisition technique based on a video processing of a speaker's face. Finally, we discuss the performance of our analysis-synthesis system in terms of the intelligibility of the visual information and we compare it as a function of the kind of visual display used.

2. THE PARAMETRIC MODELS

The high resolution model of the lips we used is presented in a related paper published in these proceedings (Guiard-Marigny et al., 1994). This 3D model is controlled through only five easy to measure parameters on a speaker's face: internal mouse width (A), internal mouse height (B), lip contact protrusion (C), upper lip protrusion (P1) and lower lip protrusion (P2).

The model of the whole face was first designed by Parke (1974). It was then implemented on a Silicon Graphics and improved for speech production by Cohen & Massaro (1993, 1994).

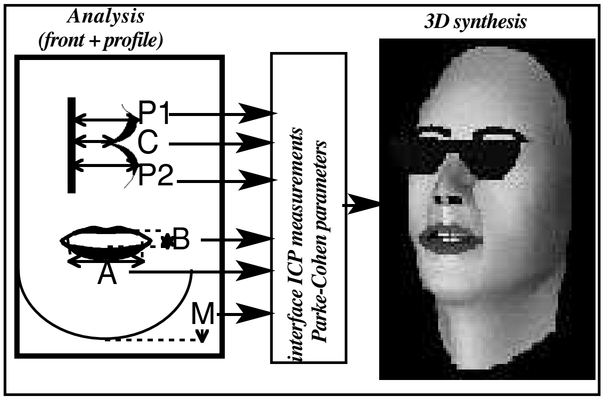

Figure 1□: Schematic of the analysis-synthesis process

In order to animate this face using the above parameters and chin lowering (M), Le Goff (1993) developed a control interface that allows the internal parameters of this face model to be predicted from six parameters directly measurable on a speaker's face. He then superimposed the lip model onto the face model. Figure 1 displays the basic principle of the whole analysis-synthesis process.

3. VIDEO ANALYSIS

The speaker is filmed from front and profile by two video cameras. To ease the video processing, which is based on chromanance, the speaker's lips and a mark on his chin had chroma-key blue makeup applied. A chroma-keyer transforms this blue color into black so that the lip vermillion area and the chin dot are the darkest parts of the screen, which allowed easier automatic analysis. After digitizing the video frames on a grey scale, the software developed by Lallouache (1991) allows the contours, of the lips, chin dot and reference rulers to be extracted on each field. Then, geometric measurements are made on each contour. A highly accurate version of this automatic labiometer is installed on a PC-based workstation connected to a VCR. A real-time version of this software has also recently been installed on a SGI Indigo computer directly connected to the cameras (Angola et al., 1994).

4. REAL-TIME ANALYSIS-SYNTHESIS

The real-time analysis-synthesis system presented on Figure 2 is implemented today on two SGI computers connected via ethernet. One is used for video analysis and the other for image synthesis. The video signal comes either from the cameras through the preprocessing devices or from a VCR. The command parameters are transmitted at a 12.5 ips rate (due to hardware limitations of the image digitizer. The graphics computer thus receives a new set of parameters every 80 ms. The parameters are lineraly interpolated between two consecutive images so that the models can be synthesized at a 25 ips rate. In parallel, the acoustic signal is directly digitized on the graphics computer where it is delayed in order to resynchronize the audio and video signals. The actual delay between audio and video is today 200 ms.

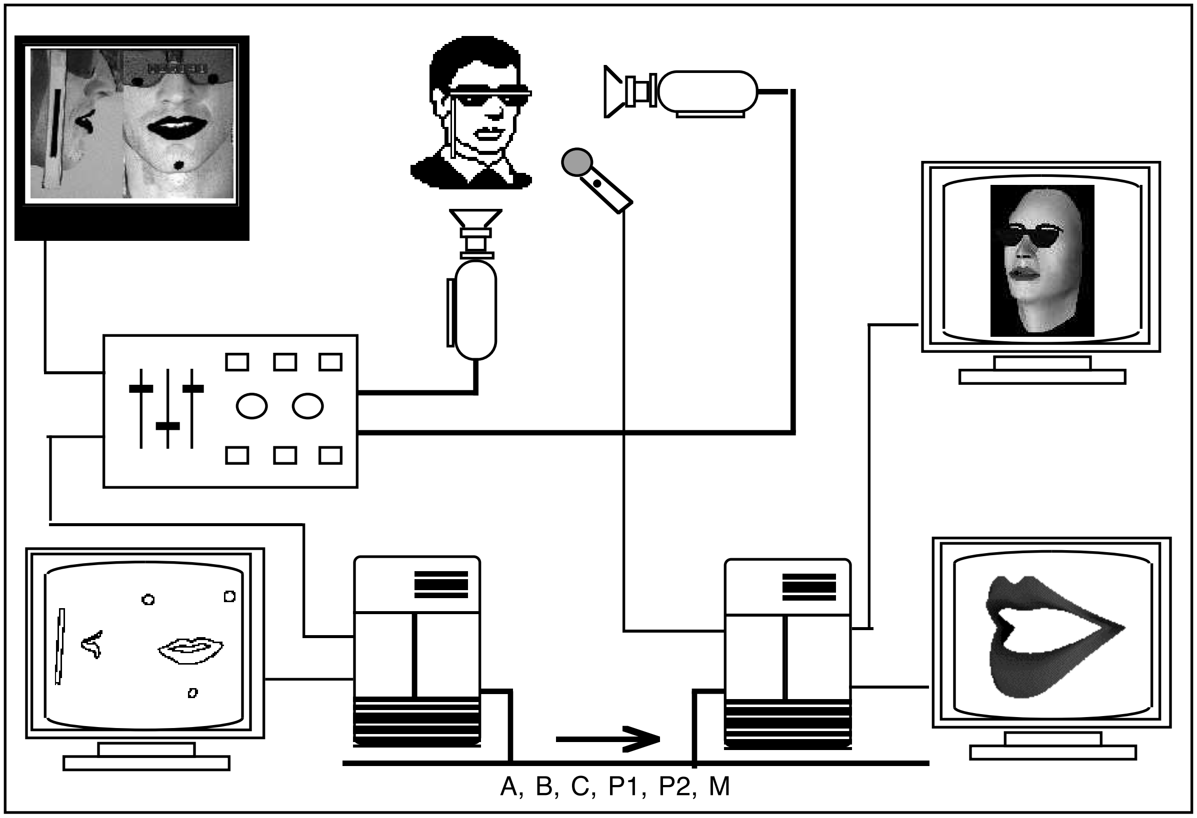

Figure 2☐: Schematic of the real-time analysis-synthesis process for the animation of a speaking face.

5. INTELLIGIBILITY TESTS

Reproducing an experiment by Benoit et al. (1994), Le Goff (1993) quantified how the intelligibility of bimodal speech is affected by the contribution of the visual information from the speaker's face under several conditions of acoustic degradation. He extended this test by adding two additional conditions of visual presentation where either the lip model or the facial model were presented to the subjects in synchrony with the natural audio utterances. The lip model was analyzed (off-line) and synthesized every 20 ms. The face model -- on which the lip model had been superimposed -- was synthesized every 40 ms. In each condition of acoustic degradation (S/N ranging from –18 to +6 dB) and of visual presentation (no, lip model, face model, and natural face), 14 subjects identified the vowel (V = /a/, /i/, or /y/) and the consonant (C□= /b/, /v/, /z/, /J/, /R/, /l/) in a phonetic string VCVCV. Le Goff's results are in agreement with those of Benoit et al. (1994). Figure 3 shows the intelligibility scores obtained depending on the mode of presentation. In all cases, the (natural) auditory information was the same at a given S/N ratio. The bottom curve figures the audio alone condition. The top curve shows the scores obtained bimodally with the natural voice and face of the speaker. In between are the curves for the whole face and the lips alone models.

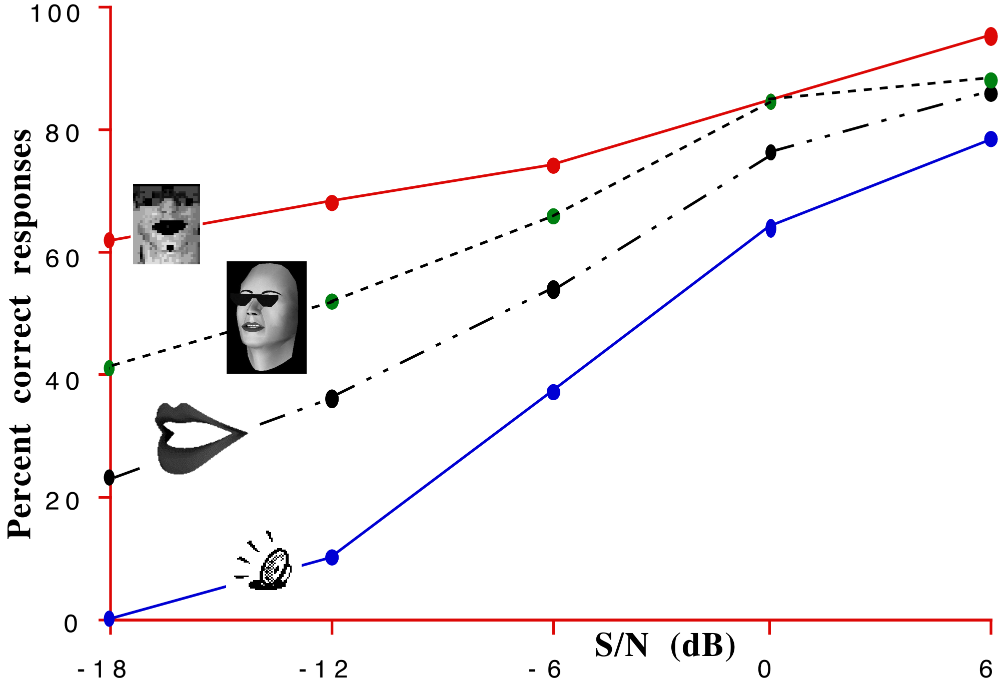

Figure 4□: Intelligibility scores obtained by 18 subjects in the identification of 18 stimuli, as a function of acoustic degradation, depending on the mode of presentation: Audio alone, audio plus the lip model, audio plus the face model, audio plus the whole natural face (from bottom to top).

Sumby and Pollack (1954) proposed an index of the visual contribution to the missing auditory information: (I[AV] – I[A]) / (1 – I[A]) where I[AV] and I[A] are the AudioVisual and Visual intelligibility scores in a given S/N condition. Figure 5 shows the evolution of this index along the acoustic degradation at the three S/N conditions where all differences in intelligibility are significant (between -18 dB and -6 dB), and in the three conditions of visual information added to the acoustic information. The index is remarkably constant over the acoustic conditions of degradation.

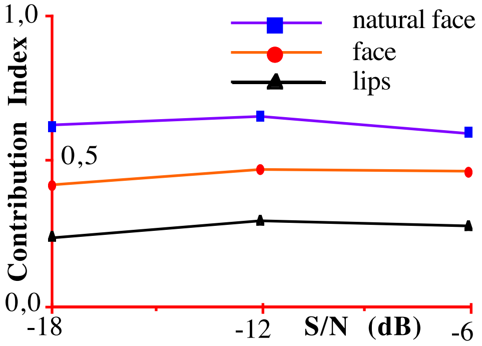

Figure 5: Contribution index of the visual information to missing acoustic information

Overall, the whole natural face restores the two thirds of the missing auditory intelligibility when the acoustic transmission is degraded or missing; the facial model (tongue movements excluded) restores half of it; and the lip model restores a third of it. This is a strong evidence that a very low bit rate of information (five or six parameters 25 times per second!) is sufficient to restore a great deal of the visual information carried on by the speaker's natural face even though the tongue gestures were not here controlled.

5. CONCLUSION

A first prototype of analysis-synthesis of speaking face runs in real-time at the ICP. The synthesis module only uses five anatomical parameters to animate the lip model developed at the ICP, or six to animate the Parke's model of the face adapted to speech at the UCSC on which the ICP model of the lips was superimposed. Those parameters are easily measured on the front and profile views of a speaker's face through video analysis. The intelligibility carried on by these parameters remarkably complements the intelligibility of the acoustic signal. The quantity of information to be transmitted is very low, in the order of a few hundred bits per second. Therefore, many applications can be forseen in the area of multimodal telecommunications as well as in the the automatic animation of synthetic actors.

Furthermore, this analysis-synthesis process allows the influence of a given parameter to be tested in order to better understand the psychophysics of speech perception by the eye, since each control parameter can be individually modified, either automatically or by hand, before being applied to one model or the other. When used off-line, the analysis module allows temporal modifications to be processed, for instance, so that the influence of rate can be tested in the bimodal perception of speech.

References

Adjoudani, A. (1993), *Élaboration d'un modèle de lèvres 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Angola, O. et al. (1994), Analyse-synthèse devisages parlants*, 20èmes Journées d'Etude sur la Parolee*, , Société Française d'Acoustique, Trégastel, France.

Benoît, C., **Mohamadi**, T. & **Kandel**, S. (1994), Audio-Visual Intelligibility of French speech in noise, *J. Speech & Hearing Res.*., (to appear in October).

Cohen, M.M. & **Massaro**, D.W. (1993), Modelling coarticulation in synthetic visual speech, *Proceedings of Computer Animation'93,* Magnenat-Thalmann & Thalmann Eds, Genève, Suisse.

Cohen, M.M. & **Massaro**, D.W. (1994), Development and experimentation with synthetic visible speech, *Behavioral Research Methods, Instrumentation, & Computers, 260-265.*

Guiard-Marigny, T., **Adjoudani**, A., & **Benoit**, C. (1994), A 3D model of the lips, *Proceedings of the 2nd ETRW on Speech Synthesis*, New Platz, USA.

Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat de l'Institut National Polytechnique de Grenoble, 214 pp.

Le Goff, B. (1993), *Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Parke, F.I. (1974), *A parametric model for human faces*, PhD Dissertation, University of Utah, Department of Computer Sciences.

Sumby, W.H., & **Pollack**, I. (1954), Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am., 26, 212-215.