VIRTUAL TALK: A MODEL-BASED VIRTUAL PHONE USING A LAYERED AUDIO-VISUAL INTEGRATION

Yao-Jen Chang, Chih-Chung Chen, Jen-Chung Chou, and Yung-Chang Chen

Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 30055, R.O.C.
Email: ycchen@ee.nthu.edu.tw

ABSTRACT

In this paper, an implementation of a one-to-one model-based virtual phone system, the Virtual Talk, is presented. In addition to the conventional ability of Internet phone applications to transmit speech messages between two users, the facial expressions of each user are real-time analyzed, transmitted, and synthesized in a model-based approach. Thereby, a joint audio-visual communication can be achieved in less than 10 kbps. To resolve the problem of audio-visual synchronization, a layered architecture is also proposed for audio-visual integration.

1. INTRODUCTION

After the telephone was invented in 1875, the phone service has become one fundamental means for personal communications. However, the traditional plain-old telephone service (POTS) is facing severe competitions from two newly developed technologies. One is the mobile phone service based on wireless communication. And the other is the Internet phone application based on various low-bit rate speech compression technologies.

The main advantages of Internet phone application lie in its economical efficiency and service variability. Single fee is accounted to everywhere the Internet is distributed. And besides the essential voice services, many Internet phone applications provide additional services such as text communications, electronic whiteboard, and file transfer. Recently, several noted software such as Microsoft’s NetMeeting [1], and VocalTec’s Internet Phone [2] also incorporate video service with their applications. However, when high network bandwidth is not available, the visual performance will be degraded either by sacrificing video quality or by reducing the video frame rate. Therefore, the model-based approach [3-5] has been received wide attentions as an alternative solution. The users’ facial expressions are analyzed and converted to some predefined Facial Action Units [6] or Facial Animation Parameters (FAP) defined in MPEG-4 standard [7]. These parameters are transmitted and synthesized on a computer graphics generated 3-D head wireframe model. The 3-D model can be a generic model or a user head adapted model textured with user’s face.

Since the audio and visual information may be transmitted separately, a lip-synch mechanism is necessary to prevent the “McGurk Effect” [8]. Therefore, both in [9] and [10], each audio frame is analyzed to derive the LPC cepstral coefficients, and the corresponding mouth shapes are generated from a trained neural network as proposed in [10] or the joint audio-visual Gaussian mixtures as presented in [9]. However, the visual analysis is not addressed in [9] and [10] since the avatar’s facial expression is fully voice driven.

Hence, in this work, a virtual phone application is implemented to perform audio and visual analysis contemporarily. And a layered mechanism is proposed for the integration of audio and visual information. In the next section, the proposed layered audio-visual integration is introduced. The model-based visual analysis and the audio analysis are described in Section III and IV. The network architecture and performance analysis is presented in Section V. And finally, Section VI concludes the paper.

2. THE LAYERED AUDIO-VISUAL INTEGRATION

In our scheme, the layered audio-visual integration is inspired by the “subsumption architecture” proposed by Rodney A. Brooks for robust mobile robot control [11]. In contrast to traditional robot control system, where each function module is executed in a successive manner, the “subsumption architecture” decomposes the control system based on task achieving behaviors. All modules are executed in parallel, which form levels of competence. Higher levels are permitted to inject data into lower levels and thereby suppressing the normal data flow [11]. With the same concept, the paralleled audio and visual processes in our virtual phone system can also be integrated as two layers of competence. Each layer has its own sensors (microphone, camera) and actuators (speaker, synthesized head). The audio is acted as the higher level such that it can alter the visual synthesis procedure to achieve lip-synch.

The conceptual diagram of the system is depicted in Figure 1. At the encoding side, the visual analysis and audio compression are performed in parallel using multi-thread programming. The objective of the visual analysis is to examine the user’s head motion and the non-rigid facial expression. The analysis result is converted to the MPEG-4 FAF values using several predefined rules. The audio compression can be achieved by any low bit rate speech compression standard. The ITU-T G.723.1 standard [12] is adopted in our system. At the decoding side, the visual synthesis and audio decompression are also performed at the same time. In addition, the energy of each audio frame is transmitted with the audio data. The energy information is utilized to distinguish the voiced audio frame and unvoiced ones. The voiced audio frame is analyzed to derive the corresponding mouth shapes represented in FAP values. The audio generated
FAPs are used to overwrite the lip-related FAPs while synthesizing the facial expression.

3. THE MODEL-BASED VISUAL ANALYSIS

3.1 Facial Model Adaptation

In the Virtual Talk system, a 3-D wireframe head model is utilized for synthesis of user's facial expressions. The adopted facial model is developed by the Instituto Superior Técnico for the complete human head [13]. It is described by 1038 points and 1702 polygons, which can be divided into different important regions such as hair, eyebrows, eyes, nose, mouth, teeth, and ears. The generic model is smoothly rendered in predefined color according to different region of the head. To enhance the visual perception, a user head adapted model textured with user's face image would be much better. Therefore, a semi-automatic facial model adaptation is developed for this task.

The facial model adaptation process is done before each communication session. Firstly, user's face images from frontal and right side views are acquired and some salient points are automatically extracted to roughly adapt the model's 3-D size. Secondly, the 3-D facial model is orthographically projected to the two images. And more feature points corresponding to the facial definition parameters (FDP) defined in MPEG-4 standard [7] are placed on the two images for manual adjustment. The 3-D positions of these feature points are calculated, and the positions of other vertices on the wireframe model are interpolated by using the reference displacement vector measure originally proposed in [14], where the displacement of each vertex is determined by

\[
d_j = \frac{\sum_{i=1}^{N} \frac{1}{s_{ij}} v_i}{\sum_{i=1}^{N} \frac{1}{s_{ij}}}
\]

where the \( s_{ij} \) is the distance between the \( j \)-th vertex and the \( i \)-th control point, \( v_i \) is the displacement of the control point from its original position and \( N \) is the number of total control points.

Once the model is adapted, the texture is extracted for each triangular polygon, and a texture map is reconstructed on a virtual cylinder enclosing the 3-D face model [15]. The texture map is then blended from the extracted textures from two images; thereby the mismatch lighting conditions from the two orthogonal images can be compensated. Figure 2 depicts the results from the texture mapping procedure.

![Figure 2](image-url)

Figure 2. (a) and (b) The adapted wireframe model projected on the frontal view image and the right-side view image after user refinement, (c) the texture map on the virtual cylindrical coordinate, (d) the result texture mapped facial model.

3.2 Facial Expression Analysis

For analysis of user's facial expression, the user's face is firstly located using a skin-color-based face detector performed in normalized r-g space [16,17] and the head orientation is coarsely estimated using the skin-color classification map for the roll angle and the estimated eye line and face center line for the yaw and pitch angles. The estimated head angles are used to rectify
4. THE AUDIO-TO-VISUAL INTERPRETATION

For audio-to-visual interpretation, the LPC cepstral coefficients are widely used in the literature [9,10,19]. In the proposed scheme, the line spectrum pairs (LSP) parameters [20] are adopted for its better quantization property and the computations of LSP parameters is one built-in procedure in G.723.1 speech codec [12]. Therefore, these parameters can be derived directly from the audio decoding process.

In G.723.1, ten LSP parameters are calculated for each frame with 240 samples. It is then linearly interpolated for each subframe composed with 60 samples. These interpolated LSP parameters are converted to three mouth shape parameters: the mouth width, the mouth height, and the upper mouth height to lower mouth height ratio using a GMM based method proposed in [9]. Compared to other methods such like classification-based conversion and neural networks, the GMM based method provides more accurate and continuous estimation [21]. In the Virtual Talk system, nine joint audio-visual Gaussian mixtures are utilized for audio-to-visual interpretation. Three mouth shape parameters are converted to FAP values for voiced audio frame to overwrite the lip-related FAP values from the visual process. Thus, the audio-visual synchronization can be achieved.

5. THE NETWORK ARCHITECTURE AND PERFORMANCE ANALYSIS

The network communication of the Virtual Talk forms a simple client-and-server architecture as shown in Figure 5. Since it is one-to-one communication system, the server is only used for keeping users properties including the IP addresses, names, head wireframe models, and texture maps. These properties cannot tolerate errors; therefore the server and clients are connected with TCP channels. Each client connects to the server for acquiring another client's information, and then these two clients are directly connected using separate audio and visual channels. These channels are UDP channels for the audio and visual data are sensitive to delays.

![Network Architecture Diagram](image)

**Figure 5.** The network architecture of the Virtual Talk. (SND: Send Port, RCV: Receive Port)
For the network traffic analysis, the audio data is compressed with a 3.723.1 codec, which consumes bandwidth with 6.3 kbps. The FAP values are quantized and entropy coded with an arithmetic coder as defined in MPEG-4 standard [7]. Depending on the frame rate of the visual analysis, the bandwidth requirement ranges from 2.5 kbps to 3.5 kbps for visual frame rate from 10 fps to 15 fps. Thus good audio-visual quality can be achieved with bandwidth less than 10 kbps. Figure 6 illustrates the user interface of Virtual Talk.

6. CONCLUSIONS AND FUTURE WORKS

The Virtual Talk system demonstrates the Internet virtual phone in a model-based approach for audio-visual services. The layered audio-visual integration also provides a good mechanism for audio-visual synchronization. Currently, the system is only implemented for communication between two users. Future works will apply the layered audio-visual integration to our previous work [18] for providing multi-user audio-visual communications.

7. REFERENCES

[1] Microsoft NetMeeting 3, online Internet conferencing software.
[2] VocalTec Internet Phone 5, online Internet telephony software.
[3] K. Aizawa, H. Harashima and T. Saito, "Model-based analysis synthesis image coding (MBASIC) system for a person's face," Signal Processing: Image Communication, vol. 1, Oct. 1995, pages 587-590.
[4] M. Kampmann and J. Ostermann, "Automatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer," Signal Processing: Image Communication, vol. 9, Mar. 1997, pages 201-220.
[5] L. Zhang "Automatic Adaptation of a face model using action units for semantic coding of videophone sequences" IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 6, Oct. 1998, pages 781-795.
[6] P. Ekman and W. V. Friesen, "The Facial Action Coding System," Palo Alto, CA: Consulting Psychologist Press, 1977.
[7] "Text of ISO/IEC FDIS 14496-2: Visual, " ISO/IEC JTC1/SC29/WG11 N2502, Atlantic City MPEG Meeting, Oct. 1998.
[8] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, vol. 264, Dec. 1976, pages 746-748.
[9] Wing Ho Leung and Tsuhan Chen, "Networked Collaborative Environment with Animated 3D Avatars," Proceedings of MMSP'98, Los Angeles, California, Dec. 1998.
[10] S. Morishima and T. Yotsukura, "Face-to-face communicative Avatar Driven by Voice," Proceedings of ICIP '99, Kobe, Japan, Oct. 1999.
[11] Brooks, R. A. "A Robust Layered Control System for a Mobile Robot," IEEE Journal of Robotics and Automation, Vol. 2, No. 1, March 1986, pages 14-23.
[12] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," International Telecommunication Union, 1996.
[13] Gabriel Antunes Abrantes and Fernando Pereira, "An MPEG-4 SNHC Compatible Implementation of a 3D Facial Animation System," Proceedings of IWNSHC3D1'97, Rhodes, Greece, 1997.
[14] Markus Kampmann and Ridha Farhoud, "Precise face model adaptation for semantic coding of video sequences," PCS'97, 1997.
[15] Frederic Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin, "Synthesizing Realistic Facial Expressions from Photographs," Proceedings of SIGGRAPH'98, Orlando, Florida, July 1998.
[16] Nuria Oliver, Sandy Pentland and Francois Berard "LAFTER: A Real-time Lips and Face Tracker with Facial Expression Recognition," Proceedings of CVPR'97, S.Juan, Puerto Rico, June 1997.
[17] Jie Yang and Alex Waibel, "A Real-Time Face Tracker," Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, Sarasota, Florida, 1996.
[18] Yao-Jen Chang, Chih-Chung Chen, Jen-Chung Chou, and Yung-Chang Chen, "Implementation of a Virtual Chat Room for Multimedia Communications," Proceedings of MMSP'99, Copenhagen, Denmark, 1999.
[19] KyoungHo Choi and Jenq-Neng Hwang, "Baum-Welch Hidden Markov Model Inversion For Reliable Audio-To-Visual Conversion," Proceedings of MMSP'99, Copenhagen, Denmark, 1999.
[20] J. R. Deller, J. G. Proakis, J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan Publishing Company, 1993.
[21] Ram R. Rao, Tsuhan Chen, and Russell M. Mersereau, "Audio-to-Visual Conversion for Multimedia Communication," IEEE Transactions on Industrial Electronics, vol. 45, No. 1, Feb. 1998, pages 15-22.