A Speech Driven Talking Head System Based on a Single Face Image

I-Chen Lin, Cheng-Sheng Hung, Tzong-Jer Yang, Ming Ouhyoung
Communication and Multimedia Laboratory,
Dept. of Computer Science and Information Engineering,
National Taiwan University, Taipei 106, Taiwan

Abstract

In this paper, a lifelike talking head system is proposed. The talking head, which is driven by speaker independent speech recognition, requires only one single face image to synthesize lifelike facial expression.

The proposed system uses speech recognition engines to get utterances and corresponding time stamps in the speech data. Associated facial expressions can be fetched from an expression pool and the synthetic facial expression can then be synchronized with speech.

When applied to Internet, our web-enabled talking head system can be a vivid merchandise narrator, and only requires 50 K bytes/minute with an additional face image (about 40Kbytes in CIF format, 24 bit-color, JPEG compression). The system can synthesize facial animation more than 30 frames/sec on a Pentium II 266 MHz PC.

1. Introduction

Compared with the rapid growth of the Internet usage, the bandwidth of Internet does not grow as much, and so can easily get congested. Besides, users are not satisfied with static information such as homepages with only static images and text; hence communications with video and audio will be the future trend. Therefore, developing very-low bit-rate but high-resolution communication tools become more and more necessary. Using synthetic faces and talking heads instead of current frame-based videos appears to be a good way to reduce the bit-rate of communication dramatically, because a remote computer can reconstruct the animation by some key parameters. In the international standard MPEG-4 [1][2], synthetic heads are also included; the controls of synthetic face expressions are defined as a set of FAPs (Face Animation Parameters).

A video using synthetic face can be very entertaining. A user can choose not only the model of the user himself, but also whichever model he (or she) wants to use in the synthetic video, such as a movie star, a statesman, or even an animal. A user can also let the talking head make amusing facial expressions to entertain other people.

In previous researches, most approaches try to synthesize one's facial expressions with a 3D model. Waters [3] proposed to use physical and anatomical models such as bones and muscles to synthesize facial expressions. Most researches use a generic 3D model with texture mapping from a set of images. Using 3D model is suitable when the taking head acts with big motions and rotations, but it must take a lot of efforts to fit the 3D model for the set of images. Pighin et al. [4] proposed a delicate method to reconstruct one's 3D head model from a set of images, and developed a method to generate new facial expressions. However, the whole process is considered complex for general users. Furthermore, the hair is not considered.

2D image warping or morphing has proved to be a powerful tool for visual effects [5][6][7]. Synthetic taking heads with this technique can look quite real when the talking head moves with only small-scale translation and rotation. Some researches are proposed to synthesize talking heads by combining individual parts of face features extracted from video clips of a talking person [8][15].

In this paper, we propose to synthesize facial expressions using the 2D image warping technique. This system is the second generation of our Chinese text-to-speech talking head system, the Image Talk [9]. It first applies a generic Talking Mask to a given image. Then the character in the image can blink eyes, move its head, and even talk in Mandarin Chinese (see Figure 1). Facial expressions of this photo-realistic talking head can be synchronized with and driven by speech from a wave file, or from a microphone. Speech recognition techniques are used to conjecture the phonemes in the speech and then fetch the associated facial expression.

In Section 2, we describe the head model of our proposed talking head system. How to generate facial expressions are described in section 3. Section 4 shows how speech recognition techniques are used to synchronize

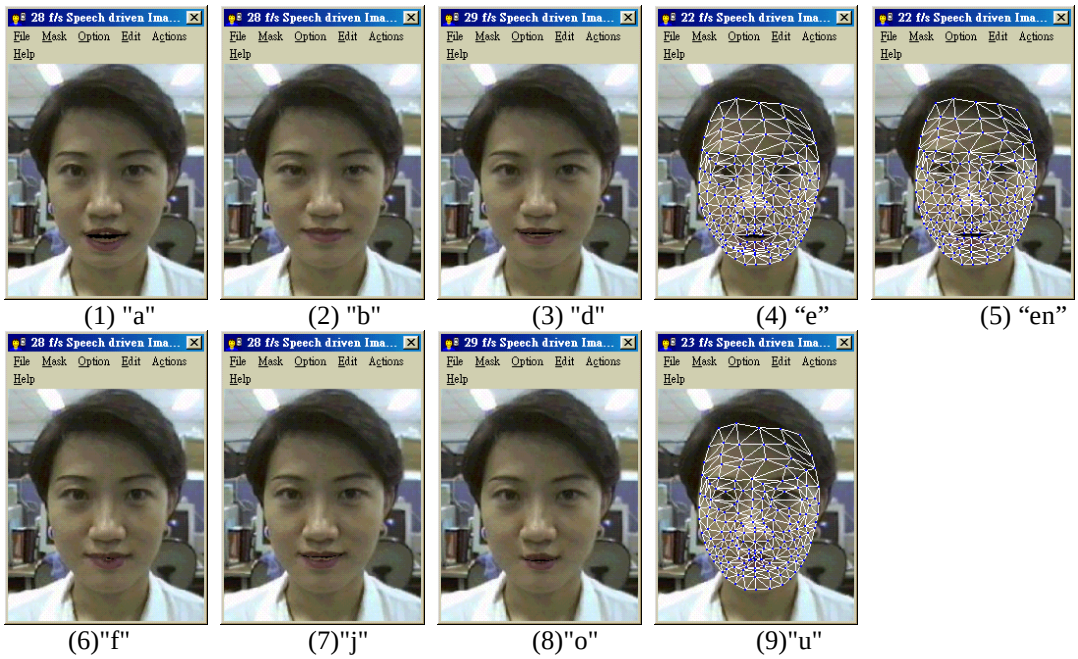


Figure 1**.** Basic expressions of a female reporter (1) ~ (9).

facial expressions with the speech. The result and applications of our system is in section 5 and 6.

2. Head model

Since our talking head is supposed not to have much head rotation and translation, our proposed talking head models a human head by applying a two-dimensional mesh model, and uses real time view dependent mesh warping for animations. Because it is simple but with good visual effect, it can run on common Pentium II PCs in real-time.

A neutral face is the frontal face image without specific facial expressions. The color information of an input neutral face provides a base image for the system. By warping the input image, we can morph the neutral face into various expressions. If the picture taken is not from a neutral face, the furrows caused by squeezing the muscles while smiling or frowning will sometimes make the synthesized expression unnatural.

How to warp the input image into various expressions? A set of interesting spots are marked as control vertices. These vertices were placed around the contour of specific features on the face such as eyes, nose, mouth, and chin. These control vertices were then connected into convex polygons, such as triangles. Then the problem of warping a whole image can be simplified to a set of polygon texture mapping problems.

For real time processing, the less the number of the vertices is, the better is the performance. In our work, less then 200 control vertices were marked on the facial area, and the final two-dimensional mesh model comprises less than 350 triangles; still the synthetic faces look natural.

2.1 Mouth shape generation

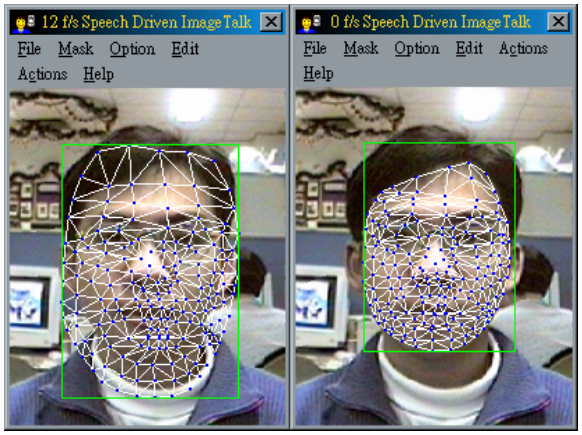
At this moment, our system is developed for Mandarin Chinese, and can be extended to other languages. In general, there are a total of 408 Chinese utterances without tone variation [10], and 1333 Chinese utterances with tone variation. Many mouth shapes of these utterances are quite similar to each other, and all mouth shapes of these utterances can be simulated by combining basic mouth shapes. In our system, 9 basic mouth shapes are adopted, as shown in Figure 1 [9].

2.2 Face mesh fitting

The first stage to use the talking head system is to fit a generic two-dimensional face mesh to a user’s face image. We provide user-friendly tools for users to go through this process. As shown in Figure 2, after a user inputs a new frontal facial image, a generic 2D mesh is applied to the face image. A boundary box is used to approximate the head size in the image, and a user can manually adjust control points to fit with feature points, for example, eyes, nose and lips on the image. Most efforts are on the adjustment of eyelids and lips, which takes about one to three minutes to adjust the mask for an experienced user already using the system for more than two times.

Figure 2. Adjustment of a mesh mask for a new face.

(a) The original image.



(b) A generic mesh mask (c) The modified mesh mask

3. Synthesizing facial expressions

3.1 Key frames specification

As mentioned above, various facial expressions can be synthesized by mapping parts of the original texture to certain polygons defined by control vertices. Putting the texture mapped mesh model and the background together, the scene now looks just like the original face with certain facial expressions.

The first step to animate facial expressions is to define the key frames. The neutral face without any facial expressions can be thought as a key frame that contains a neutral facial expression. Key frames in this proposed system are from our previous system: the Image Talk[9].

The actual content of key frames saved in our system library are the vector differences of each control vertex from the neutral face, and normalized according to the size of the generic mask. In summary, a synthetic face can be represented as weighted combinations of the neutral facial mask (from the generic one adjusted to match the input image) and the expression mask.

3.2 Generating intermediate expressions by interpolation

In order to generate the animation of a face giving a talk, our proposed system has to generate intermediate frames or expressions between two key frames. Since the key frames are saved as normalized vectors according to the size of the generic mask, the intermediate expressions can be generated by time-driven interpolation.

To make our talking head's mouth actions look more realistic, the problem of co-articulation, which means the current mouth shape does not only depend on the current pronunciation but the one coming before and after, must be taken into account.

There are 408 possible combinations of phonemes in a Mandarin Chinese word. In our previous system [9], we only use linear interpolation to generate intermediate expression; in the current system, we try to use an exponential decay model [21] to make the animation of mouth shape look more natural.

3.3 Head motion

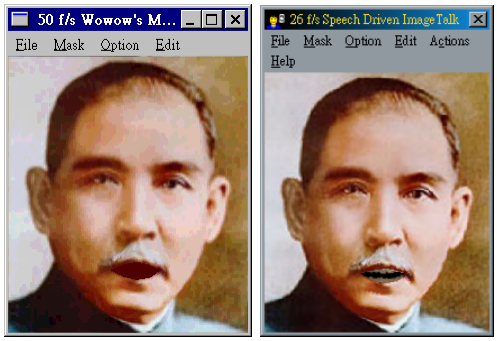
In addition to our previous algorithm developed for head motion [9], a 3D-projection method is also tried to make the head motion more realistic. In this way, the 2D mesh of the talking head is acquired from the projection of a 3D head model. To perform head motion, the 3D head model is rotated first, and then projected onto the 2D mesh. After moving the vertices of the 2D mesh to the new projection location, the image of the talking head is morphed to perform the head motion.

3.4 Generic teeth model

Because the input image is only a static neutral face, there is no image information inside the talking head's mouth. We propose a generic teeth model to simulate teeth inside the mouth. The teeth model is separated into two parts: the upper teeth and the lower ones. The upper teeth model is moved according to the control vertex at the philtrum, and the lower one is moved according to the control vertex at the chin. This generic teeth model can resize according to the mouth size in the neutral face.

In addition, there is a basic assumption that the larger one's mouth is opened, the more light his teeth can be illuminated. Our proposed system adjusts colors of the generic teeth model according to the longest distance between the upper lip and the lower one. The smaller the distance is, the darker the teeth are (see Figure 3).

4. Speech driven face synthesis



(a) without teeth (b) with teeth

Figure 3**.** A generic teeth model inside the mouth

After the 2D face mesh is adjusted, it can be used to animate facial expressions driven by speech.

To synthesize animations of facial expressions synchronized with speech data, we must know which utterances appear in the input wave data. In addition, the start and stop time of a certain utterance should be obtained to synchronize the mouth shapes with wave data.

For example, in Figure 4, it is the PCM data of a Chinese sentence "ni hau ma" spoken by the first author. After getting this wave file, our system invokes a speech recognition engine and finds that from *StartTime* to *TimeA* is silence; *TimeA* to *TimeB* should be "ni"; *TimeB* to *TimeC* should be "hau"; *TimeC* to *EndTime* should be "ma". Our system then translate these results into "neutral (from 0 to *TimeA*), ni (from *TimeA* to *TimeB*), hau (from *TimeB* to *TimeC*), ma (from *TimeC* to *EndTime*) and appropriate key frames are fetched from the expression pool.

Figure 5 is the flow diagram of our proposed system. First, wave data from a speech file or a microphone are fed to a speech recognition engine that helps us to conjecture the phonemes of the speech. The engine compares the input wave data with its own database; then reports the most possible utterance and the time stamps of each utterance in the sequence. A table of mapping from utterances to phonetic notations is used to get basic facial expressions. Thus, we can get a sequence of basic facial expressions according to the input speech data. With this information, facial animation synchronized with the input wave data by techniques mentioned above can be generated. For example, Mandarin Chinese word "good" pronounced as /hau/ is converted to be /h/+/au/, and the corresponding mouth shape is from "h" then gradually morphed to "au".

Since our purpose is to synthesize facial expressions according to speech data and many mouth shapes of Chinese utterances are quite similar to each other, the recognized results don't need to have high recognition rate. Currently an efficient speech recognition engine can be used to generate facial animation in near real time.

In general, the recognition rate for a speaker independent recognition engine is only around 60%. Since many different pronunciations have similar mouth shapes; that is, the real difference is inside the mouth that is not visible, the overall "recognition rate" for this talking head system is higher than 90%. Actually, the visual effect is so strong that most people can not see the difference. The details of how speech recognition engines are applied for two kinds of speech data sources are described in the following.

4.1 Generating facial animation driven by speech

Since speech signals saved as files can be played repetitively, this kind of wave sources is proper for preprocessing computation. We will to preprocess the input speech file and save the recognition results as an index file.

When users want to see facial expressions driven by a speech wave file, our system checks whether there is an index file associated with it first. If there is one, the system plays back the speech file and the talking head makes facial expressions according to the information in that index file. If there is no index file, a speech recognition engine is invoked to recognize the speech data, and an index file can then be generated.

At this moment, we can invoke two kinds of speech recognition engines. The first group are those speech recognition engines that support Microsoft DirectSpeechRecognition API from Microsoft SAPI 4.0[11] (Microsoft Speech Application Interface version 4.0); the other one is the speech recognition engine from Applied Speech Technologies [12].

For first-time users with MS DirectSpeechRecognition API, they are encouraged to have a training session to increase the recognition rate by reading some specified texts for 10 minutes. The main parts we need are the functions of wave indexing. Wave indexing is a way to index each segment of wave data from files with an appropriate utterance or a word by parsing wave data segments with a given CFG (context free grammar). We have defined a CFG describing the basic 408 Mandarin Chinese utter-

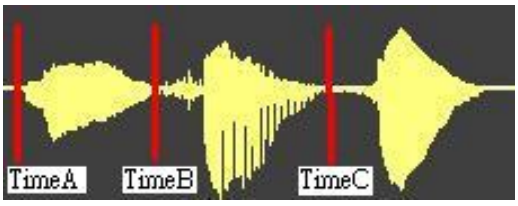


Figure 4**.** The speech data shape of a Chinese sentence pronounced as "ni hau ma" which means as "How are you?" in English. The vertical lines in the picture are marked to separate three different utterances (words) "ni", "hau", "ma".

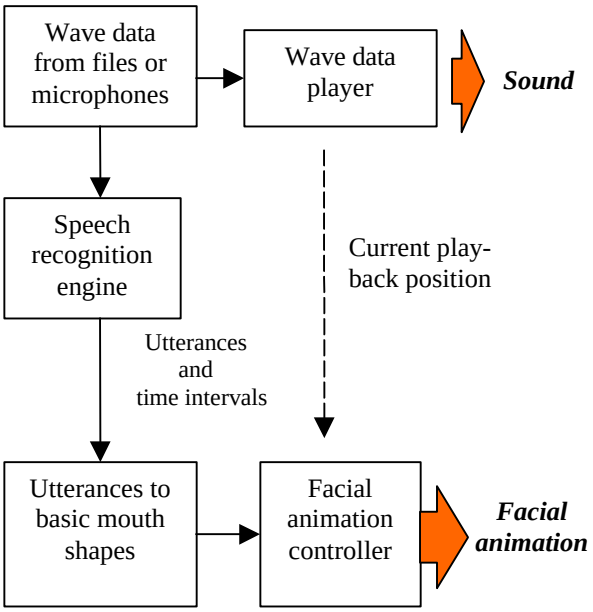


Figure 5. the flow diagram of the speech-driven talking head system.

ances [10] for the API, and the most matched utterances can therefore be reported.

The speech recognition engine we use from Applied Speech Technologies is a simplified version without parsing grammar and therefore can recognize Chinese utterances directly. It can index speech data from files or from input devices such as microphones.

4.2 Near real-time speech driven facial animation by dual recognition engines

To apply our speech driven talking head to visual communication, a real-time system is required. However, the required time for speech recognition techniques to look ahead is usually more than one second, and so we should reduce the response time as much as possible. The pre-processing actions for the input speech in previous sections are not practicable for real-time applications.

Since the speech recognition library from Applied Speech Technologies is a simplified version and focussed just on Chinese utterances, the recognition time of this engine can be less than the playback time of input on an Intel Pentium II 266 MHz PC. To display facial animation driven by speech from a microphone with the shortest response time, the concept of double buffers in graphics is applied in our system. Because of constraints of wave I/O control in the recognition library we used, a pipeline structure does not work. We use dual recognition engines to recognize input speech data blocks alternately.

We make 2 second long speech data a block, and so the input data sequences consist of a sequence of blocks. Two recognition engines are opened at the same time but work alternately. *Engine A* is in charge of odd blocks; *Engine B* is in charge of even ones. While *Engine A* is recognizing data in *block N*, *Engine B* receives the wave data of *block N+1* from a microphone. While *Engine B* is recognizing data in *block N+1*, *Engine A* receives the wave data of *block N+2* from the microphone. These processes are demonstrated in Figure 6.

4.3 Experiments

To verify our viseme recognition rate of speech, eleven subjects participated in our experiment. Eight of them are male; three of them are female.

Each subject gave an introduction of himself first, and then read two short paragraphs in newspaper. All these speech data are recorded in our system and then are used to generate corresponding viseme sequences. Except one male whose visemes are only 77% recognized correctly, others' viseme recognition rates range from 87.5% to 97.1%. The average recognition rate in this experiment is 89.7%.

In those 10.3% errors, 74% are cases where utterances and visemes were misjudged; 26% are obvious er-

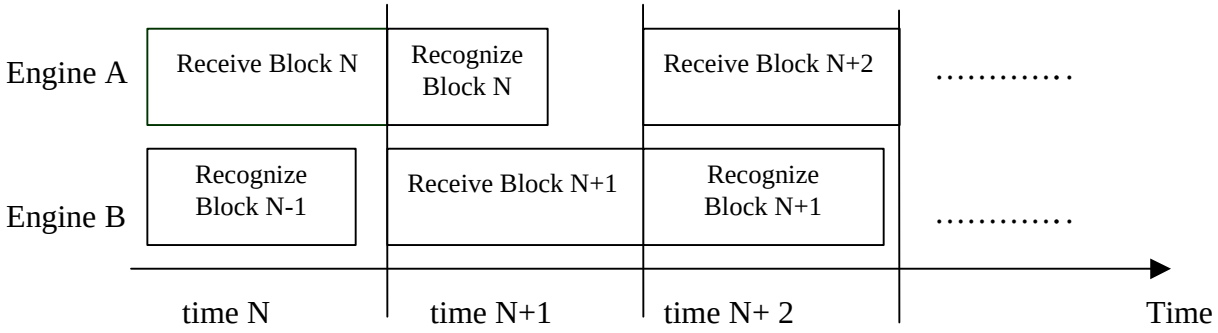


Figure 6. Dual Speech Recognition engines approach.

rors in utterance segmentation.

Since the viseme recognition rate is 90%, most of viseme animation in our system is correct. However, to be a practical application, a few errors still can be perceived by a sensitive observer. Thus, we developed an interactive tool for users to adjust the recognized results such as reported visemes and the time boundaries of visemes in speech.

5. Results

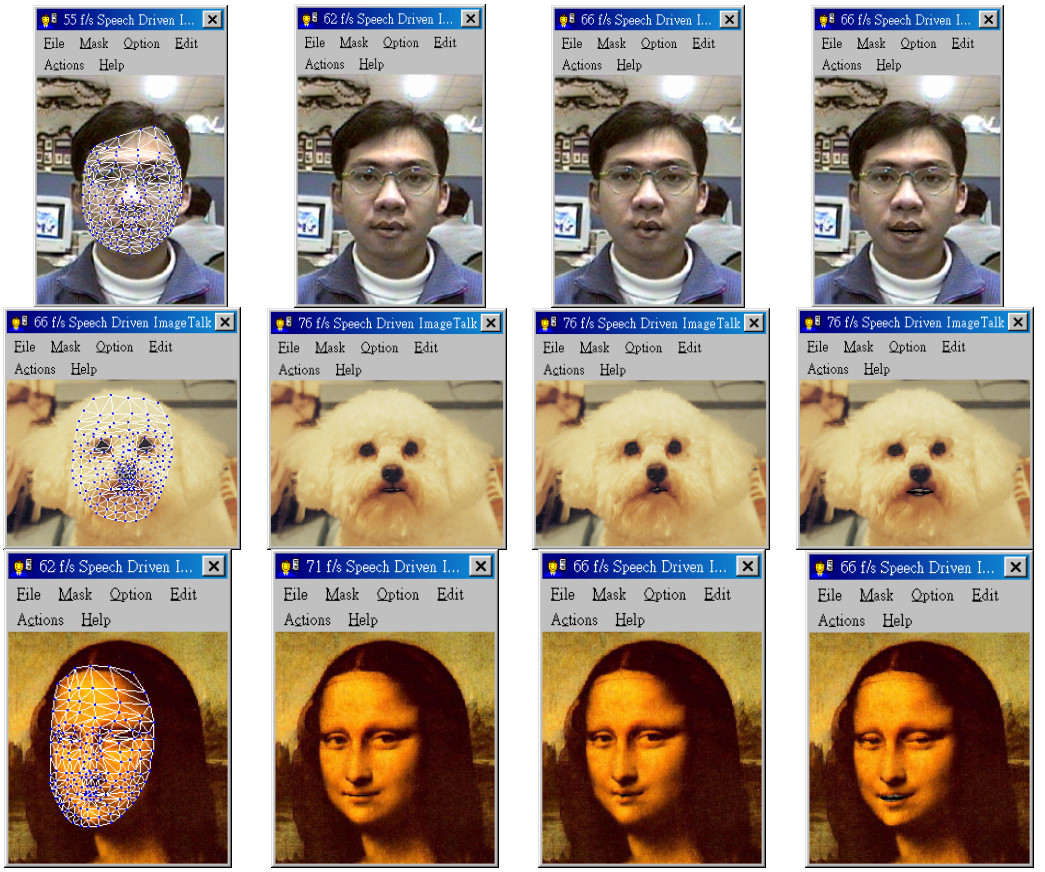
The current performance of the proposed web-enabled talking head system without real-time speech recognition is about 30 frame/sec in CIF format on a Pentium II 266MHz PC. The speech recognition is processed off-line using Microsoft's DirectSpeechRecognition API in Speech SDK 4.0 or the speech recognition library from Applied Speech Technologies.

Facial animations driven by live speech has limited response time, which is about 4 seconds, because the speech recognition module needs to accumulate a certain length of signals to begin the recognition process. However, there are still two practical problems. The first one is the possible discontinuity (less than 0.3 sec) when two recognition engines "swap". The second problem is that the display frame rate sometimes can not keep at 20 frames per second. On a Pentium III 500MHz PC, the performance is about 15-24 frames per second with image size 120x150. This is due to the fact that the recognition engine shares the CPU with the display engine.

To make our speech-driven talking head with full-scale head rotation, we also developed a talking head system based on 3D models [13]. An executable demo program is put in the web page http://www.cmlab.csie.ntu.edu.tw/~ichen.

6. Applications

One immediate application of the talking head system is a web-enabled merchandise narrator. Speech data about the products and the corresponding index file, which is generated by an off-line process, are put on the web-site, and the talking head animator is packed in active X control module. After installing the module, when users click



(a) 2D mesh mask (b) a neutral face (c) pronouncing "u" (d) pronouncing "en"

Figure 7. Three examples of pronouncing a Chinese syllable "wen" (u+en).

icons of our talking head on the homepage, speech data and the index file are sent to the client side, and then the talking head can give a vivid speech to introduce the product.

Compared to frame-based video such as H.263 etc., which requires about 400Kbytes to 4Mbytes per minute depending on the video quality, the proposed system will only need 50Kbytes/minute (voice is compressed using G.723.1) with an additional face image (about 40Kbytes compressed using JPEG). Another application is the idea of "VR-talk" that can be applied to Internet chat rooms. Since the procedures of adjusting our generic mask to fit to a new face image are simple, users can easily use new face images such as movie stars or even animals to represent themselves in chatrooms. This feature appears to satisfy people's intension of concealing and disguising themselves on the Internet for fun or for other purposes, and this can also make Internet chat rooms more fun.

7. Future Works

A speech-driven talking head with real-time response is our goal. As mentioned in subsection 4.2, the practical problems should be solved. Similar to our generic teeth model, an eyeball model can be developed to change the viewing direction of eyes. Speaker-independent speech recognition engines also need improvement.

Acknowledgement

This project is partly sponsored by the research grant from National Science Council (Taiwan), under the grant number NSC88-2622-E-002-002.

We would also like to thank Applied Speech Technologies, since it provides its speech recognition library and helps our research.

Reference:

[1] MPEG4 Systems Group, *"Text for ISO/IEC FCD 14498-1 Systems,"* ISO/IEC JTC1/SC29/WG11 N2201, 15 May 1998.

[2] J. Ostermann, *"Animation of Synthetic Faces in MPEG-4"*, Proc. of Computer Animation, pp.49-51, Philadelphia, Pennsylvania, USA, June 8-10, 1998.

[3] Demetri Terzopoulos, Keith Waters, *"Analysis and synthesis of Facial Image Sequences using Physical and Anatomical Models,"* IEEE Tran. On Pattern and Machine Intelligence,15(6), Jun.1993, pp.569-579.

[4] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Pichard Szeliski, David H. Salesin, *"Synthesizeng Realistic Facial Expressions from Photographs,"* Proceedings of ACM Computer Graphics (SIGGRAPH 98), pp. 75-84 Aug-1998.

[5] Thaddeus Beier, Shawn Neely *"Feature-Based Image Metamorphosis"*, Proc.of SIGGRAPH 92. In Computer Graphics, pp. 35- 42

[6] Steven M.Seitz, Charles R. Dyer, *"View Morphing"*, Proc. SIGGRAPH 96, pp. 21-30.

[7] Nur Arad, Nira Dyn, Daniel Resfeld, Yehezkel Yeshurun, *"Image Warping by Radial Basis Functions: Application to Facial Expressions"*, CVGIP: Graphical Models and Image Processing", Vol. 56, No.2, pp.161-172, 1994.

[8] Eric Cosatto, Hans Peter Graf, *"Sample-Based Synthesis of Photo-Realistic Talking Heads"*, Proc. of Computer Animation 98, pp. 103-110, Philadelphia, Pennsylvania, June 8-10, 1998.

[9] Woei-Luen Perng, Yungkang Wu, Ming Ouhyoung, *"Image Talk: A Real Time Synthetic Talking Head Using One Single Image with Chinese Text-To-Speech Capability"* Proc. of PacificGraphics 98, pp. 140-148, Singapore, Oct 1998.

[10] Lin-Shan Lee, Chiu-Yu Tseng, Ming Ouhyoung, *"The Synthesis Rules in a Chinese Text-to-Speech System"*, IEEE Trans. On Acoustics, Speech and Signal Processing. Pp.1309-1320. Vol.37, No.9, 1989.

[11] Microsoft Speech Technology SAPI 4.0 SDK, http://www.microsoft.com/iit/projects/sapisdk.htm

[12] Applied Speech Technologies Corporation. http://www.speech.com.tw

[13] Tzong-Jer Yang, I-Chen Lin, Cheng-Sheng Hung, Chien-Feng Huang and Ming Ouhyoung, *"Speech Driven Facial Animation"*, to appear in the Proceedings of EuroGraphics CAS'99, Milan, Italy, Sept. 1999.

[14] M.Esoher and N.M. Thalmann, *"Automatic 3D Cloning and Real-Time Animation of a Human Face"*, Proc. Computer Animation 97, pp.58-66, 1997.

[15] C. Bregler, M.Covell, M.Slaney, *"Video Rewrite: Driving Visual Speech with Audio"*, Proc. SIGGRAPH'97, pp.353-360, 1997.

[16] B. Guenter, c. grimm, D. Wood, H. Malvar, F. Pighin, *"Making Face"*, Proc. of Computer Graphics (SIGGRAPH '98), pp. 55-66, Aug. 1998.

[17] D. Decaolo, D. Metaxas, M. Stone, *"An Antropometric Face Model Using Variational Techniques"*, Proc. Computer Graphics (SIGGRAPH '98), pp. 67-74, Aug. 1998.

[18] T. DeRose, M. Kass, T. Truong, *"Subdivision Surfaces in Character Animation"*, Proc. of Computer Graphics (SIGGRAPH'98), pp85-94, Aug 1998.

[19] P.E Kmon, W.Fresen, *"Facial Action Coding System: A Technique for the Measurement of Facial Movement"*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[20] S. Morishima, H.Harashima, *"A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface"*, IEEE J. Selected Areas in communications, 9, pp. 594-600, 1991.

[21] M.M. Cohen and D.W. Massaro. *"Modeling co-articulation in synthetic visual speech"*. In N.M. Thalmann and D. Thalmann, editors, Models and Techniques in Computer Animation. Springer-Verlag, 1993.