Lifelike Speech Driven Talking Head from a Single Face Image

I-Chen Lin, 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 speech recognition, requires only one single face image to synthesize lifelike facial expression animation. When applied to video e-mail, our synthetic video e-mail requires 100 K bytes/Min with an additional face image (about 40Kbytes in CIF format, 24 bit-color JPEG compression). Our system can synthesize facial animation more than 30 frames/sec on a Pentium II 233 MHz PC.

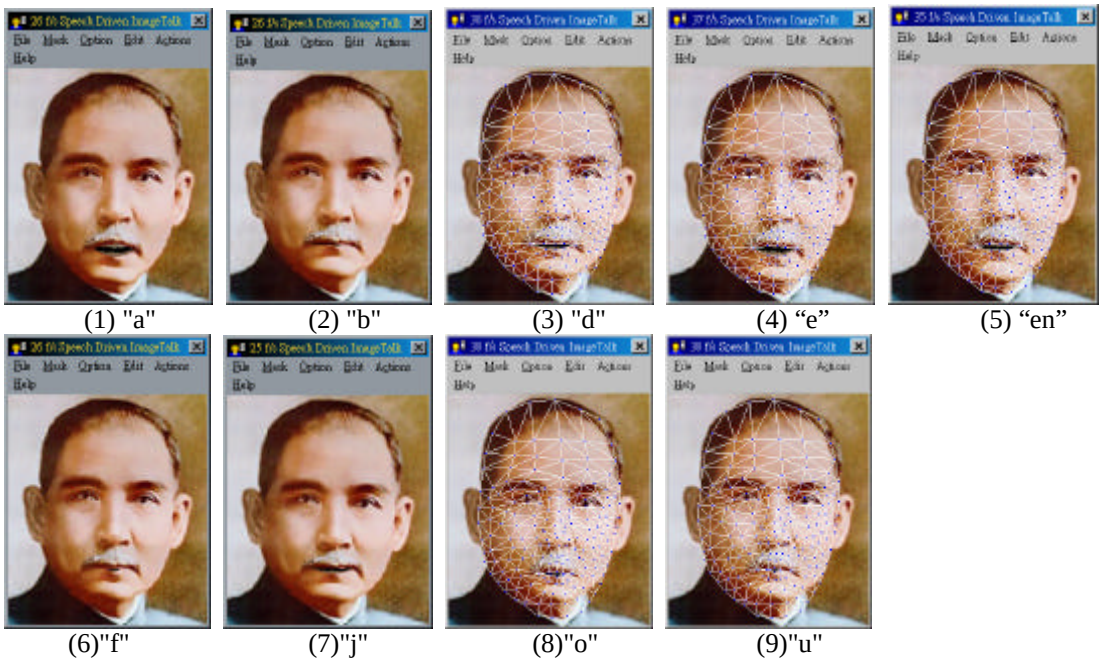
1. Introduction

Because of the emergence of MPEG-4 [1], face synthesis gets more and more attentions. In previous researches, most approaches try to synthesize one's facial expressions with a 3D model. Waters [2] proposed to use physical and anatomical models to synthesize facial expressions; however, the results don't yet look realistic. Pighin et al. [3] proposed a delicate method to reconstruct one's 3D head model from 5 images, and developed a method to generate new facial expressions. However, the whole process is still too complex for general users. Furthermore, the hair is not considered. In this paper, we propose to synthesize facial expressions using 2D images. This system is the second phase of a Chinese text-to-speech talking head system, Image Talk [4].

2. Proposed Talking Head System

2.1 Mouth shape generation

At this moment, our system is developed for Mandarin Chinese, and can be extended to other languages. There



(1) "a" (2) "b" (3) "d" (4) “e” (5) “en”

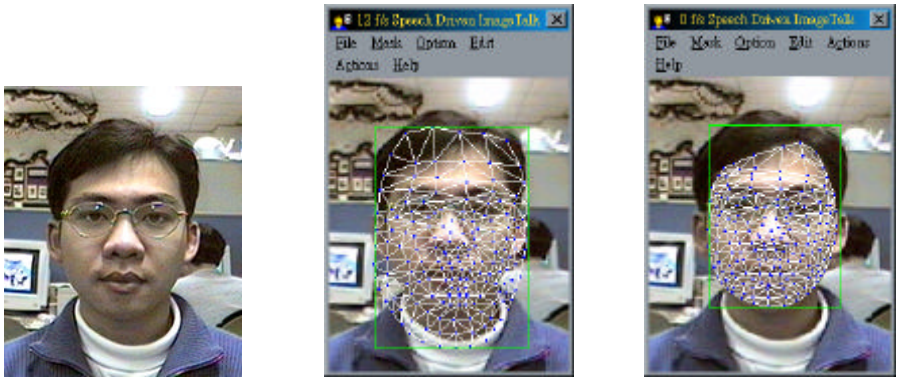
(6)"f" (7)"j" (8)"o" (9)"u"

Figure 1. Basic expression (1) ~ (9). The original image is the founding father of modern China (R.O.C), Dr. Sun Yet-Sen.

are 408 Chinese utterances without tone variations [6], 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 [4].

2.2 Face Mesh Fitting

The first stage to use the talking head system is to fit a generic 2D face mesh to a user's face image. 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, which takes about one to five minutes to adjust the mask for a person already using the system for more than two times.



(a)original image (b)the generic mesh mask (c)adjusted mesh mask

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

2.3 Speech Driven Face Synthesis

After the 2D face mesh is adjusted, it can be used to animate facial expressions driven by speech. Figure 3 is the flow diagram.

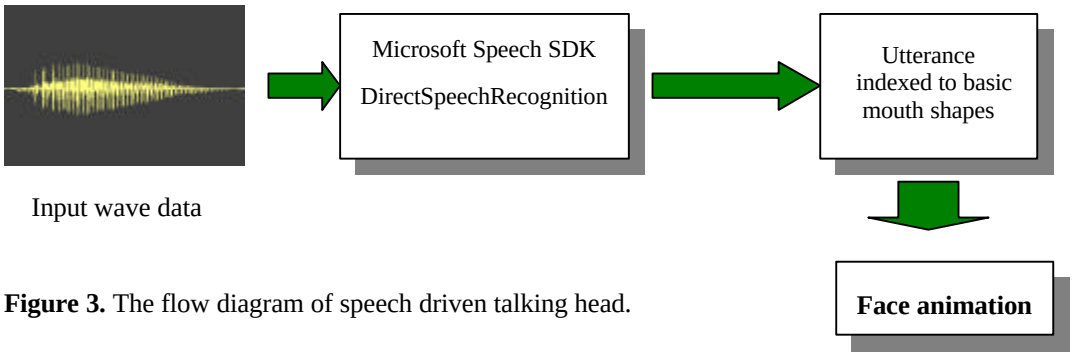


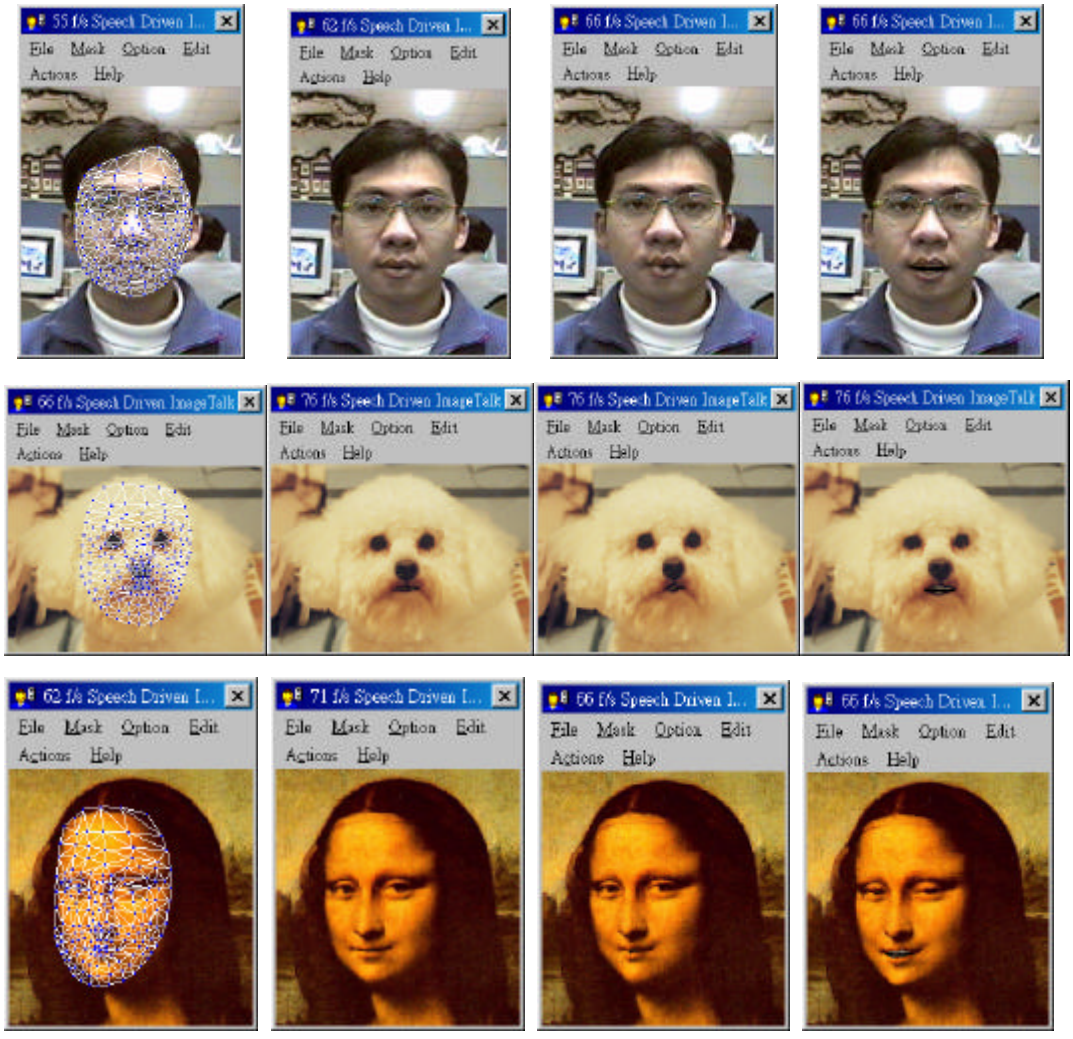
Figure 3. The flow diagram of speech driven talking head.

First, we feed the speech wave data to Microsoft DirectSpeechRecognition API [5] that helps us to get pronunciations and conjectures of the speech. For first-time users, they are encouraged to have a training session to increase the recognition rate by reading some texts for 10 minutes. We have defined a context free grammar describing the basic 408 Mandarin Chinese utterances [6] for the API, and the most matched utterances can therefore be generated. A table mapping from utterances to element phonetic notations is used to get basic facial expression. Thus, we can get a sequence of basic facial expressions corresponding to the input data. For example, a Mandarin Chinese word "good" pronounced as /hau/, is converted to be /h/+/au/, and the corresponding mouth shape is from "h" and moving to "au".

3. Results and Applications

The current performance of the proposed talking head system is about 30 frame/sec in CIF format on a Pentium II 233MHz PC. The speech recognition is processed off-line using Microsoft's DirectSpeechRecognition API in Speech SDK 4.0. It costs about two times of input speech length to recognize input speech.

One immediate application of the talking head system is video e-mail. Compared to frame-based video e-mails, e.g. Video Live Mail of Cyberlink corp. [7], which requires about 400Kbytes to 4Mbytes per minute, depending on the video quality, the proposed system can achieve 100Kbytes/Min (voice is compressed using GSM610) with an additional face image (about 40Kbytes compressed using JPEG). Another application is the idea of "VR-talk" used in Internet chat rooms. Users can use new face images such as movie stars or even animals to represent themselves in chatrooms. This feature would satisfy people's intension of concealing and disguising themselves on the Internet for fun or for other purposes.



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

Figure 4. Three examples of pronouncing a Chinese syllable "wen" (u+en).

Reference:

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

[2] 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.

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

[4] 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"* PacificGraphics 98.

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

[6] 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.

[7] Cyberlink Corporation. http://www.cyberlink.com.tw