Adaptive Man-Machine Interfaces
MIT9904-15

Progress Report: July 1, 2002— December 31, 2002

Tomaso Poggio

Project Overview

In this project we aim to achieve three significant extensions of our recent work on developing a text-to-visual-speech (TTVS) system (Ezzat, Geiger, Poggio 2002). The existing *synthesis* module may be trained to generate image sequences of a real human face synchronized to a text-to-speech system, starting from just a few real images of the person to be simulated. We propose to 1) extend our morphing approach from video to audio to address issues of audio synthesis, 2) to improve the real-time performance of our system and 3) to extend the system to use morphing of 3D models of faces -- rather than face images -- to output a 3D model of a speaking face.

The main applications of this work are for virtual actors, video dubbing, and very-low-bandwidth video communication. In addition, the project may contribute to the development of a new generation of computer interfaces more user-friendly than today's interfaces.

An overview of our system is shown below:

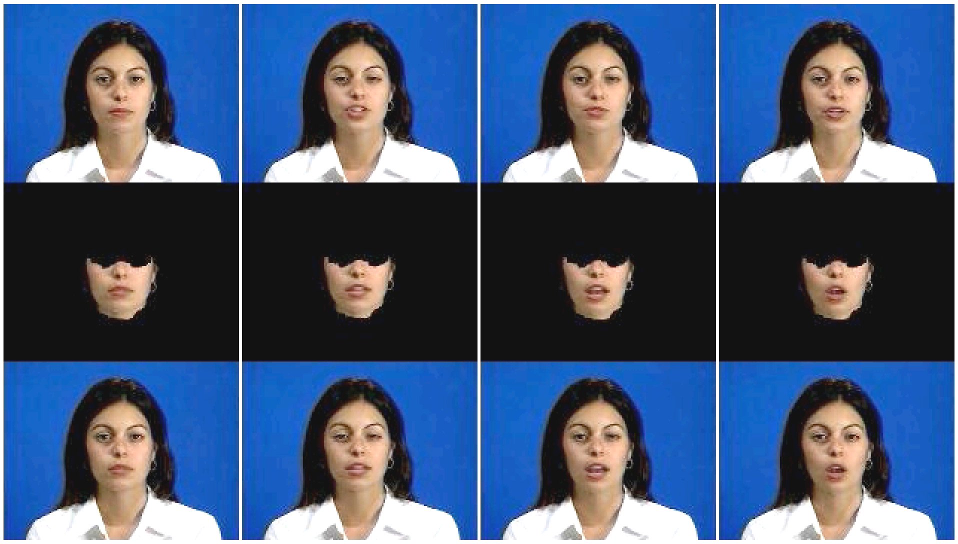


Figure 1: Top row: an original background sequence. Middle row: the synthetic mouth animations our system generates. Bottom row: the synthetic facial animations composited into the original background sequence.

Progress Through December 2002

In the audio morphing subproject, we have completed the data collection stage where we collected 15 minutes of a speaker uttering about 200 sentences. We identified 3 axes of variation which our audio morphing algorithm will control: 1) duration 2) pitch and 3) spectral content. In our initial experiments, we have implemented the TD-PSOLA (time domain pitch-synchronous overlap add) (Moulines and Charpentier 1990), and found it to be acceptable for modifying both duration and pitch. Shown in the data below are examples of a recorded phoneme whose duration and pitch are altered using TD-PSOLA. Current efforts are underway to identify a suitable spectral modification algorithm. Once this is identified, a morphing algorithm which can interpolate audio between different durations, pitches, and spectra will be developed.

🔊

Original phonetic audio

🔊

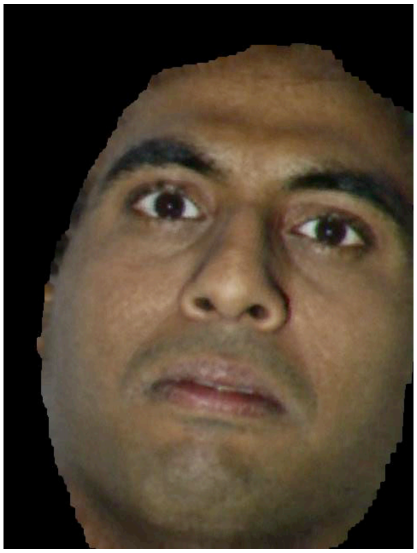
Audio Duration-normalized using TD-PSOLA

🔊

Audio Duration- and Pitch-normalized using TD-PSOLA

We have begun preliminary steps towards improving the speed and efficiency of our Mary101 facial animation system. This stage will be performed in collaboration with the **Spoken Language Systems (SLS) group** at MIT, led by **Dr. James Glass**. Preliminary steps performed in this regard are the purchase of a fast 2 Gighahertz computer on which to deploy the facial animation system, and the transfer of the Mary101 system from our group to SLS.

On the third topic, we have acquired an **Eyetronics** structured light scanner with which to record human subjects in 3D as they speak. We have also identified the various subtasks which need to be performed for the 3D facial animation system as 1) acquisition of data 2) cleaning up of data 3) establishing of 3D correspondence between the scans 4) building the 3D Multidimensional morphable model 5) Reanimation of the face to new audio. Attached is a sample 3D texture scan from our Eyetronics scanner.



Research Plan for the Next Six Months

We plan in the next six months to:

1) Continue developing our audio morphing approach
2) Continue improving our system so that it works in near real-time.
3) Extend the system described above by acquiring 3D models of faces -- rather than face images – in order to build a 3D morphable model capable of outputting 3D model of a speaking face.

References:

[1] Beymer, D. and Poggio, T. Image Representation for Visual Learning. Science, 272, 1905-1909, 1996

[2] Blanz, V., Vetter, T. A Morphable Model for the Synthesis of 3D Faces. In: Computer Graphics Proceedings SIGGRAPH'99, pp. 187--194, Los Angeles, 1999

[3] Eyetronics, www.eyetronics.com

[4] Ezzat, T, and T. Poggio. Visual Speech Synthesis by Morphing Visemes, International Journal of Computer Vision, 38, 1, 45-57, 2000.

[5] Ezzat, T., Geiger G, and T. Poggio. Trainable Videorealistic Facial Animation. In Proceedings of SIGGRAPH 2002

[6] Moulines and Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication 9, 453-467

[7] Vetter, T. and Blanz, V. Estimating coloured 3d face models from single images: An example based approach. In Burkhardt and Neumann, editors, Computer Vision -- ECCV'98 Vol. II, Freiburg, Germany, 1998. Springer, Lecture Notes in Computer Science 1407.