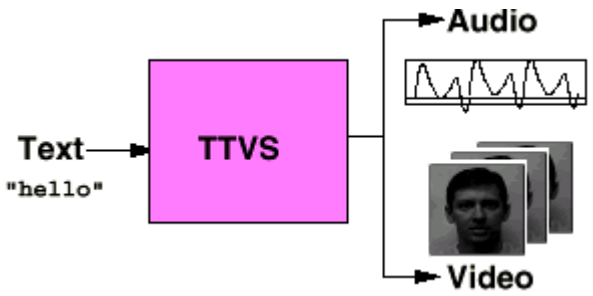
Adaptive Man-machine Interfaces
MIT9904-15

Progress Report: July 1, 1999—December 31, 1999

Principal Investigators

Project Overview

We proposed two significant extensions of our recent work on developing a text-to-visual-speech (TTVS) system (Ezzat, 1998). 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 proposed 1) to extend the system to use morphing of **3D models** of faces -- rather than face images -- and to output a 3D model of a speaking face and 2) to **enrich the context** of each viseme to deal with coarticulation issues.
The main applications of this work are for virtual actors and for 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. Our text-to-audiovisual speech synthesizer is called MikeTalk. MikeTalk is similar to a standard text-to-speech synthesizer in that it converts text into an audio speech stream. MikeTalk also produces an accompanying visual stream composed of a talking face enunciating that text. An overview of our system is shown in the figure.



Progress Through December 1999

We have been successful to attract, as we planned to try to do, Volker Blanz as a part-time postdoc. However, our progress has been slower than expected on the 3D subproject because Volker, who will be mainly responsible for it working with Tony Ezzat, has been delayed in joining CBCL (he has still to defend his thesis at the Max Planck Institute in Tuebingen). Volker is the coauthor – with Thomas Vetter – of a SigGraph 99 paper, which attracted a lot of attention, about a technique, based on previous work in our group, to synthesize 3D face models from single images. The plan is that he will be a part-time postdoc (shared with the University of Freiburg where Thomas Vetter is now Professor) at CBCL; the 3D NTT project will also involve Thomas Vetter. We see below a picture of Volker and the 3D synthetic face estimated by his and Thomas' algorithm from the single image on the left…

On the second subproject Tony Ezzat has made significant progress. He has recorded a training corpus of a human speaker uttering various sentences naturally, and obtained a low-dimensional parameterization of the lip shape using a new extension of the morphable model of Jones (1998) based on statistical shape-appearance techniques. The results are encouraging and yield a synthesis module with a surprisingly small number of parameters.

Research Plan for the Next Six Months

We plan in the next six months to:

1) develop further our approach to deal with the coarticulation problem. As we described we have recorded a training corpus of a human speaker uttering various sentences naturally, and obtained a low-dimensional parameterization of the lip shape. We will now use learning algorithms to estimate the parameters of the morphable model from the phonetic time series.

2) start the work to extend our system to 3D models of faces and produce as output a 3D face, complete of texture. The work will be done in collaboration with Thomas Vetter and Volker Blanz. The plan is to eventually record a 3-dimensional face as it dynamically utters the same visual corpus we have designed and extract 3D visemes. We will do preliminary experiments to evaluate the quality of the data and our capability to synthesize new 3D visemes, starting with an initial visit of Volker Blanz in March 2000.

3) begin an extension of our system to Japanese. At present we plan to begin this project with visits of Ms. Minako Sawaki from NTT to our lab.