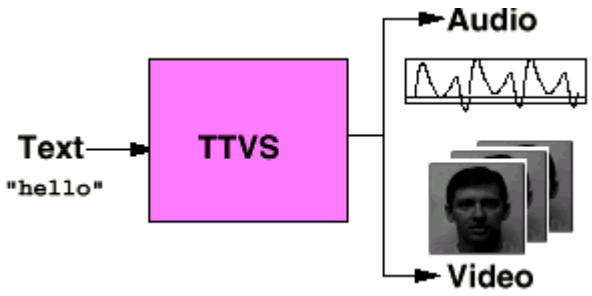
Adaptive Man-machine Interfaces
MIT9904-15

Progress Report: January 1, 2000—June 30, 2000

Tomaso Poggio

Project Overview

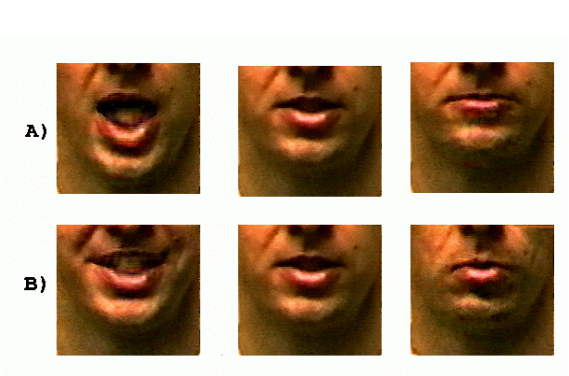
We proposed two significant extensions of our recent work on developing a text-to-visual-speech (TTVS) system (Ezzat, 1998). The existing *synthesis* module may be trained to generate image sequences of a real human face synchronized to a text-to-speech system, starting from just a few real images of 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.

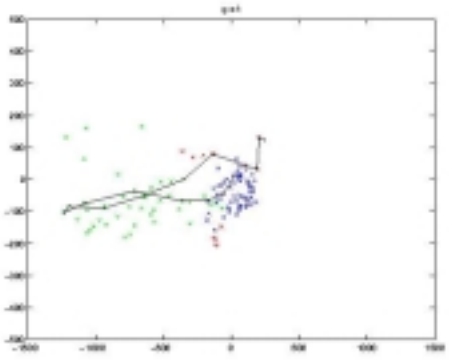


Progress Through June 2000

Volker Blanz has made preliminary tests to examine the feasibility of morphing between different 3D visemes for the 3D talking face project. Shown below in the first 2 images are two 3D visemes recorded using a Cyberware scanner. The third image is an intermediate image generated by morphing between the first two. The correspondence between the two visemes are generated using a semi-automatic algorithm that uses optical flow applied to 3D. These results are very encouraging and we plan to record more 3D visemes to capture the whole range to mouth movements.

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 principal components (PCA). Shown in the image below on the left are the first two lip principal components: one axis represents degree of mouth opening and closing, while the other represents smiling-rounding. The technique also allows us to obtain trajectories of mouth movement, as shown in the image on the right. In general, we plan on using this technique to learn coarticulation models from video data, which may then be applied to 2D or 3D talking faces.





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 generate the parameters of the morphable model from the phonetic time series, and which will implicitly incorporate coarticulation knowledge.

2) develop further the 3D talking face approach by recording more 3D visemes and morphing between them.