Face Translation: A Multimodal Translation Agent

Max Ritter, Uwe Meier, Jie Yang, Alex Waibel,
{mritter, uwem, yang, ahw}@cs.cmu.edu

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

In this paper, we present Face Translation, a translation agent for people who speak different languages. The system can not only translate a spoken utterance into another language, but also produce an audio-visual output with the speaker's face and synchronized lip movement. The visual output is synthesized from real images based on image morphing technology. Both mouth and eye movements are generated according to linguistic and social cues. An automatic feature extracting system can automatically initialize the system. After initialization, the system can generate synchronized visual output based on a few pre-stored images. The system is useful for a video conference application with a limited bandwidth. We have demonstrated the system in a travel planning application where a foreign tourist plans a trip with a travel agent over the Internet in a multimedia collaborative working space using a multimodal interface.

1. Introduction

Natural and realistic face synthesis is essential for successful animation, film dubbing, computer talking head (avatar), video compression, multimedia entertainment, and speech-based interfaces. Recently there has been a significant interest in the area of face synthesis [1-7]. A large effort has been directed to developing autonomous software agents that can communicate with humans using speech, facial expression, and gestures.

The foci of different face synthesis systems can be divided into video manipulation, human-computer interaction and agents for human-human communication. Different tasks impose different requirements on naturalness (cartoon or realistic face), usability, and real-time implementation. Much attention has been paid to lip synchronization in face synthesis research. Most of those systems are based on a phonemic representation (phoneme or viseme). Typically, the phonemic tokens are mapped onto lip poses and the lips are synthesized from either real images (e.g., Video Rewriting [1]) or graphic approach (e.g., Baldi [7]). In this study, we are interested in developing a translation agent for Internet applications. The system can not only translate a spoken utterance into another language, but also produce an audio-visual output with the speaker's face and synchronized lip movement. The work is closely related to Video Rewriting [1] but different in several ways. Video Rewriting models vocal co-articulation via triphones. In language translation applications, triphone models are not available in another language. Face Translation uses image processing and morphing technologies to generate images between phonemes. Furthermore, Face Translation synthesizes not only lip movements based on translated text, but also eye gaze based on user's location. Face Translation also uses situation dependent eye movements and eye blinking to make the interaction more realistic. The system is designed for Internet applications. In the initialization phase, the user is asked to read a few sentences. The visemes are selected by phoneme segmentation from speech recognition and then mapped into the target language. Some facial expressions, such as eye gaze and eye blinking, are captured automatically by the system at the same time. These results are stored in a database. The database is transmitted to the receiving end. At the receiving end, the system can generate visual output based on a few pre-stored images. We have demonstrated the system in a travel planning application where a foreign tourist plans a trip with a travel agent over the Internet in a multimedia collaborative working space using a multimodal interface.

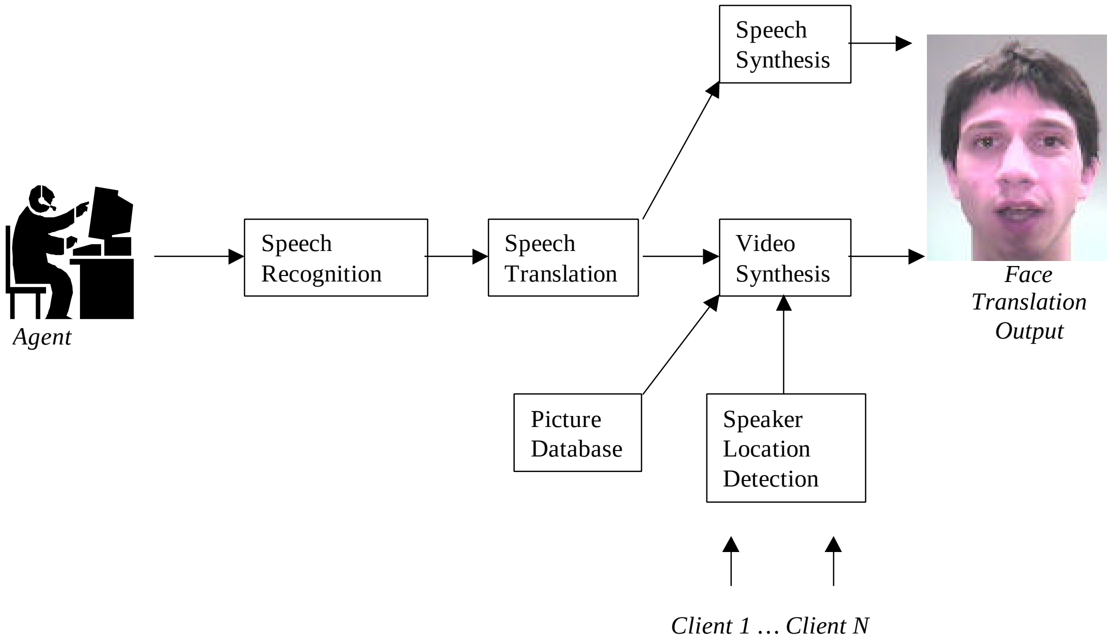

Figure 1 Audio Visual Synthesis for Speech to Speech Translation

2. System Overview

Figure 1 shows a scenario where a user communicates with other user(s) via Internet. We call this user “agent” and other user(s) at remote site “client(s).” The agent and client(s) speak different languages. But the agent can talk to the client(s) via the Face Translation system. The client will see the agent’s face speak the translated sentences with synchronized lip movements. The agent’s eyes will also look at the client during the conversation. The system works as follows. When the agent speaks to the system, the speech-to-speech translation module translates the spoken utterance into an intermediate language and then maps onto the target language. The string of the translated text is sent to the receiving end. At the receiving end, the system synthesizes synchronized acoustic and visual speech output based on text input. The eye gaze is determined by the location of the client detected by the location detector.

Speech Recognition and Translation

We use the JANUS system [8,9] for speech recognition and translation. The JANUS Speech Recognition Toolkit, developed in the Interactive Systems Labs, embodies various tools in an easily programmable platform. It has been successfully applied to many speech recognition tasks, such as dictation and telephone conversations.

The JANUS speech translation system translates spoken language, much like a human interpreter. It currently operates on a number of limited domains such as appointment scheduling, hotel reservation, or travel planning. The JANUS travel planning application can also access databases to automatically provide additional information such as train schedules or city maps to the user. All systems are designed to work without a conventional keyboard. Currently, 17 different languages can be handled.

Speech Synthesis

We use the FESTIVAL Text to Speech System [10] to generate the audio speech from a text string. Festival offers a full text to speech system through a number of APIs. It is easy to use and can be customized to a certain extent. In addition, we use the FESTIVAL system not only to generate the acoustic output, but also to convert text to phonemes. These phonemes (and their timestamps within the generated audio) are needed for the video synthesis.

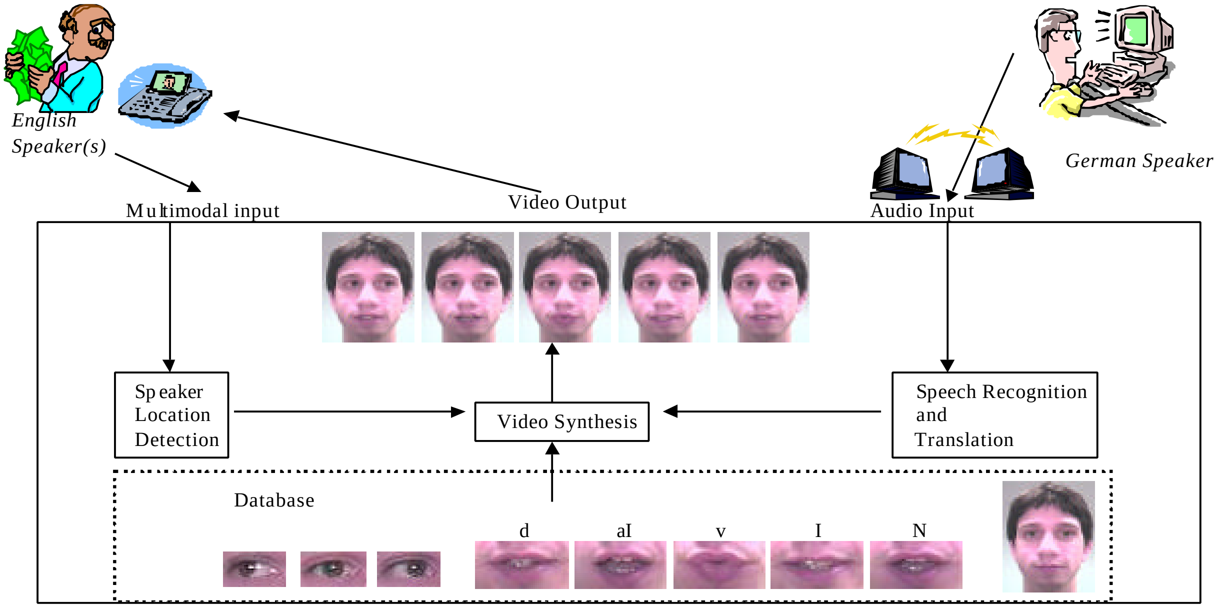

Figure 2 Face Translation

Video Synthesis

The video synthesis is based on an image of a frontal face, a database of images for different visemes, and a database of images for different eye gaze and eye blink. We will discuss how to obtain the database automatically in the next section. For the lip synthesis we retrieve a lip image for each viseme from the lip database. One viseme is chosen for each phoneme in the audio synthesis. The time alignment of these phonemes is used to synchronize the lip movement and the sound. The eye gaze is synthesized based on the result from the user location detector. The synthesized eye gaze is looking in the direction of the client's position in the room. This is achieved by selecting pre-stored eye images from the database and morphing them onto the base face image.

The face synthesis process is as follows. The system uses a frontal face as the base image and morphs mouth and eyes onto the base image at the appropriate positions. This is a reverse procedure of cutting mouth and eyes where the system needs a precise and stable feature locator to retrieve facial feature images. In the synthesis process, the system needs to put mouth and eyes back onto a face at the exactly right spot. They are resized to fit the destination region and copied onto it. The intensity of the synthesized images has to be adjusted because the components might be taken in different lighting conditions. This can be done by a brightness adaptation according to the brightness of the base image in that position. A smooth transition between the base and each of the new images is calculated, using an elliptical shape mask. Figure 2 shows an example of the video synthesis.

3. Database

In order to synthesize the face, it is necessary to create databases of the face, mouth, and eye images. Our objective is to make this procedure as automatic as possible. We have employed a real-time face tracker and a facial feature tracker developed in our lab [11]. First, the user is asked to speak a given sentence, which covers a set of visemes big enough to allow reasonable synthesis. While speaking, his/her face and voice are recorded, including synchronization information. This will provide the new lips images. The user is then asked to look at different directions. While the user is looking at different directions, the system records the face images. After recording these two video-sequences, the rest of the procedure is fully automatic. The system extracts images of mouth and eyes, labels them and stores them in the databases.

As a first approach, the user has to speak a sentence in the destination language, because in some cases visemes do not exist in the source language, e.g. there is no "th" viseme in German. For storage, visemes are labeled with the name of the phoneme uttered in the according image. However, the database does not need to contain lip images for every phoneme. If, for synthesis, the exact image does not exist, a preference table is used to find the best matching existing image. The advantage of this approach is the ability to make the best out of the existing data. As soon as more data becomes available, the synthesis can show more detailed lip movement. The idea is to have a system, which allows a quick initial registration and then learns and becomes better while being used.

To build up the lip database from the first recorded sequence, the acoustic signal and the known text label are fed to the JANUS recognizer. JANUS uses forced alignment to compute the timestamps of the phonemes in the acoustic signal. Using these timestamps and the synchronized video data from the recording, the system can determine which image in the recorded sequence provides which viseme. Our facial feature tracking system can locate the lip region. Labeling the images with the desired gaze positions from the second recording is easier, because the system assumes that the user follows the instruction during the recording. We are going to use a gaze tracker to make this process fully automatic. The lip and eye regions are extracted with reasonable regions around them, which define the image that is saved in the databases. The positions of these regions have to be very precise and consistent to produce natural looking output.

4. Application

In order to demonstrate the feasibility of the system, we have applied it to a travel-planning task. The demo setup is as follows. An agent who speaks only German helps, via Internet, two English-speaking clients to plan a trip to Germany. The agent sits in front of a workstation and the clients stand in a room with a SMART Board. The SMART Board is an interactive whiteboard that allows users to control their applications directly from the Board's large, touch-sensitive surface. The agent and the clients discuss the flight schedule, hotel reservation, car rental, and tour with the help of movies and panoramic images. With the SMART Board, the clients can easily manipulate multimedia information shared with the agent by speech, gesture and handwriting. The clients can see the synthesized face of the agent. The eye gaze of the synthesized face is directed at the location of the client who has made the most recent voice query, so that the client could feel he/she has been served.

The system has also been applied to a teleconference where a foreign speaker discusses problems with several people sitting in a meeting room. Again, the people in the meeting room can see the synthesized face of the foreign participant on the SMART Board. The lip movement of the foreign participant is synthesized based on the translated text and the eye gaze is directed to the target of the message.

5. Conclusion

We have presented a multimodal translation agent in this paper. The system translates both audio and video from one language to another language. In the current system audio and video translation work only for English and German. Speech to speech translation from and to other languages including Spanish, Japanese, and Korean is already available with our JANUS System. The use of the Face Translation system with these languages is conceptually the same as it is for English and German. What is needed is to generate a list of phonemes for the sentence to be spoken. If a language contains different phonemes, the phoneme to viseme mapping table has to be modified. Currently we are working on identifying the eye-gaze automatically. Using this technique, the system could simplify the process of database building. The eye database could then be built on the fly while the system is being used.

Acknowledgements

The Face Translation was developed within the C-Star Project. We would like to thank the Interactive Systems Speech People for their support with the JANUS Recognizer.

REFERENCES

1. Video Rewrite: Driving Visual Speech with Audio
Christoph Bregler, Michele Covell, Malcolm Slaney
Computer Graphics Proceedings, Annual Conference Series, 1997

2. Visual Speech Synthesis with Concatenative Speech. *Asa Hallgren, Bertil Lyberg.* AVSP98, Terrigal Australia

3. Kinematics-Based Synthesis of Realistic Talking Faces, *Takaaki Kuratate, Hani Yehia, Eric Vatikiotis-Bateson,* AVSP98, Terrigal, Australia

4. Generation of lip-synched Synthetic Faces From Phonetically Clustered Faces Movement Data *Francisco M. Galanes, Jack Unverferth, Levent Arslan, David Talkin* AVSP98, Terrigal, Australia

5. Real-time Talking Head Driven by Voice and its Application to Communication and Entertainment, *Shigeo Morishima* AVSP98, Terrigal, Australia

6. Visual Speech Synthesis Based on Parameter Generation From HMM: Speech-Driven and Text-and-Speech Driven Approaches. *M. Tamura, T. Masuko, T. Kobayashi, K. Tokuda* AVSP98, Terrigal, Australia

7. Recent Developments in Facial Animation: An Inside View *M.C. Cohen, J. Beskow and Dominic W. Massaro* AVSP98, Terrigal, Australia

8. JANUS-III: Speech-To-Speech Translation in Multiple Languages *Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld, Puming Zhan* IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, 1997

9. End-To-End Evaluation in JANUS: A Speech-To-Speech Translation System *Donna Gates, Alon Lavie, Lori Levin, Alex Waibel, Marsal Gavalda, Laura Mayfield, Monika Woszczyna, Puming Zhan* Proceedings of the ECAI 96, Budapest, 1996

10. Festival Speech Synthesizer *University of Edinburgh* http: //www.cstr.ed.ac.uk/projects/festival.htm

11. Visual Tracking for Multimodal Human Computer Interaction, *Jie Yang, Rainer Stiefelhagen, Uwe Meier, and Alex Waibel*, Proceedings of CHI 98, pp. 140-147