Automatic Selection of Visemes for Image-based Visual Speech Synthesis

Jie Yang, Jing Xiao, Max Ritter
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{yang+, jxiao, mritter}@cs.cmu.edu

Abstract

An image-based approach provides an efficient way for visual speech synthesis. In an image-based visual speech synthesis system, a few lip images, namely visemes, are used for generating an arbitrary new sentence. Many approaches select visemes manually. In this paper, we propose a method for a system to automatically select visemes by minimizing the synthesis error. The feasibility of the proposed method has been demonstrated by experiments. We describe an application of image-based visual speech synthesis to a multimodal communication agent for a translation task where two people, who speak different languages, can talk to each other over the Internet.

1. Introduction

Face synthesis has been an active research area recently [2, 3, 4, 5, 6, 7, 9]. A large effort has been directed to developing autonomous software agents that can communicate with humans using speech, facial expression, and gestures. 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 [2]) or graphic approaches (e.g., Baldi [3]). However, different tasks impose different requirements on naturalness (cartoon or realistic face), usability, and real-time implementation. In this research, we are interested in developing a multimodal communication agent for Internet applications [11]. 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 movements. The work is closely related to Video Rewriting [2] but different in several ways. The Video Rewriting models vocal co-articulation via triphones. In language translation applications, triphone models might be not available in another language. Our system employs image processing and morphing technologies to generate images between phonemes. Furthermore, our system synthesizes not only lip movements based on translated text, but also eye gaze based on user's location. The system also uses eye blinking and other facial expressions 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.

Many image-based visual synthesis approaches [2, 4, 11] select visemes manually or randomly. The quality of image synthesis is then dependent on the experience of the user or without control. In this paper, we propose a method for an image-based visual speech synthesis system to automatically select visemes by minimizing the synthesis error. In fact, a synthesis system could find the boundaries of visemes with the help of a speech recognizer. The system could then select the key frame so that it can minimize the synthesis error for the corresponding visemes. In this way, the system can automatically select visemes with minimum synthesis error. Experimental results demonstrate the feasibility of the proposed method.

2. Automatic Selection of Visemes

An image-based visual speech synthesis system uses visemes to generate an arbitrary new sentence. But selection of visemes is not trivial. A major challenge is the lack of a clear definition of visemes. A general practice is to segment video by audio. The process can be described as follows. First, the user is asked to speak one or a few given sentences, which covers a set of visemes enough for reasonable lip synthesis. While speaking, his/her face and voice are recorded with synchronization information. The information will be used to create a database necessary for lip synthesis. To build up the database, the acoustic signal and the known text label are fed to a speech recognizer. The speech recognizer uses forced alignment to compute

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

the time-stamps of the phonemes in the acoustic signal. Although there is no one-to-one mapping between phoneme and viseme boundaries, phoneme boundaries are commonly used as an approximation for viseme boundaries. Therefore, with these time-stamps and the synchronized video data from the recording, the system can determine which images in the recorded sequence provide which phoneme. In general, there are more than one images for each viseme. However, it is impossible to store all these images in the database and it is also unnecessary to store all of them. Then the task is to optimally select one of the images, which can best represent the corresponding viseme, i.e., from which the other images for this viseme can be synthesized. In the previous approaches, there is no systematic way to select visemes. Visemes are selected either randomly or manually.

We propose to automatically select a set of visemes by minimizing the synthesis error. The idea can be implemented as follows

- Define a synthesis error
- Design a proper cost function based on the synthesis error
- Minimize the cost function

The synthesis error can be measured from different viewpoints such as image intensity and geometry distortion, or even human feeling. In this paper, we assume the synthesis error can be measured automatically. Furthermore, there are many different method available for synthesizing a new image sequence, such as linear and nonlinear mappings. Without losing generality, we use intensity linear interpolation to synthesize images between visemes. Therefore, we could employ the sum of squared differences between the pixel color values of the original and synthesized image sequences as the synthesis error. The cost function is then

$$\min_j E = \min_j \sum_{i=0}^{n} (I_i - \hat{I}_i(f(j)))^2,$$

where I is the pixel color values of the original image sequence for the current viseme, $\hat{I}$ is the pixel color values of the synthesized image sequence for this viseme, i is image index, and $f()$ is the synthesis mapping, $f(j)$ means that the jth image in the original sequence is used as the viseme for synthesizing the new images.

In order to compute the synthesis error, we need to consider two cases: boundary viseme and mid-visemes. First, we consider the first phoneme of the acoustic signal. Suppose there are m frames of image in the corresponding sequence of this phoneme. Since the mouth should be closed before the phoneme was spoken, an image with closed mouth can be used as the reference image. We use each frame within the segmented boundaries to interpolate its lip region with the lip region of the reference image linearly and generate m images. We then compute errors for all these synthesis sequences. There is, at least, one frame with the smallest synthesis error, which is the optimal viseme. We then consider the other visemes. With the first optimal viseme as the reference image, the second viseme can be selected in the same way. So does the third viseme. Repeating this process for all the corresponding sequences of visemes, we can obtain all the optimal visemes from the original image sequence. These optimal visemes can then be labeled and stored into the database. The whole process can be fully automatic.

3. Experimental Results

We have tested the proposed method on a lipreading database. The database contains both male and female speakers. Figure 1 shows an example of speech signal with the time stamp by a male speaker. The sentence is "PUT IT WITH THE BOXES PLEASE COME HOME BEFORE DARK." In this project, we use the JANUS system [8, 10] for speech recognition and translation. The JANUS Speech Recognition Toolkit, developed in the Interactive Systems Labs, embodies various tools in an easily programmable platform. With the JANUS, The phonemes are segmented as follows:

{SIL 0 3} {P 4 14} {UH 15 20} {T 21 30} {IH 31 32} {T 33 37} {SIL 38 39} {W 40 41} {IH 42 43} {DH 44 60} {SIL 61 64} {DH 65 72} {AH 73 79} {SIL 80 81} {B 82 86} {AA 87 102} {K 103 109} {S 110 118} {IX 119 138} {Z 139 149} {SIL 150 204} {P 205 224} {L 225 226} {IY 227 228} {Z 229 238} {K 239 247} {AH 248 250} {M 251 252} {HH 253 285} {OW 286 287} {M 288 289} {SIL 290 291} {B 292 296} {AX 297 300} {F 301 311} {OW 312 315} {R 316 317} {SIL 318 329} {D 330 344} {AA 345 347} {R 348 349} {K 350 374} {SIL 375 382}

Within braces, the first element is the name, the second is the starting time in millisecond, and the third is the ending time in millisecond. Based on the time alignments, the system can select visemes. Here we illustrate results for the first two visemes, "P" and "UH". There are 11 frames of images corresponding to "P," from 4th frame to 14th frame. Its reference image is the silence "SIL." Figure 2 shows the bar chart of normalized synthesis errors. It is obvious that the 9th image generates the minimum error.

There are 6 frames corresponding to "UH," from 15th frame to 20th frame. Figure 3 shows chart of normalized synthesis errors for "P-UH." Compared to the worst case,

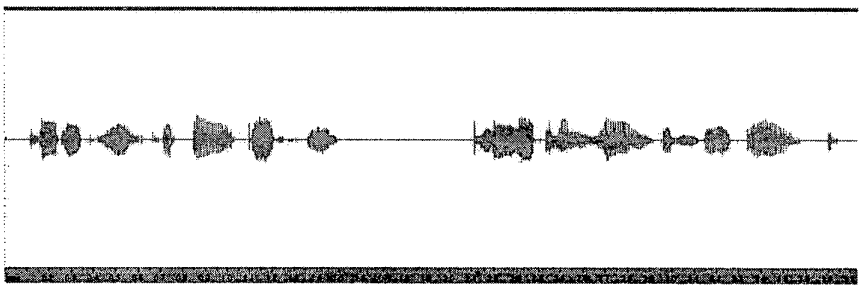

Figure 1. The speech signal of example sentence with the time stamp

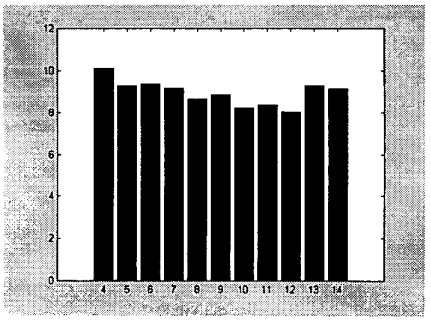

Figure 2. Error chart for image sequence "SIL-P"

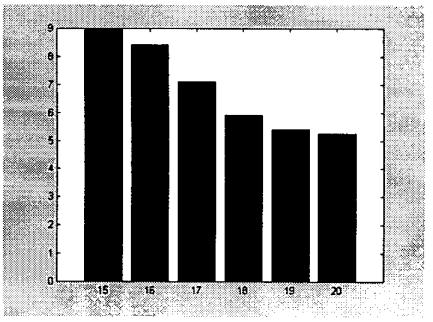

Figure 3. Error chart for image sequence "P-UH"

which possibly occurs in the previous methods, the new method produces much smaller synthesis errors, because the proposed method guarantees extracting the optimal visemes from sequences. Figure 4 compares normalized synthesis errors produced by a user and the system for the first 10 visemes of a test sentence. The solid line represents the system's performance and the dotted line is the user's errors. The curve of user selection is always above the one of the automatic selection.

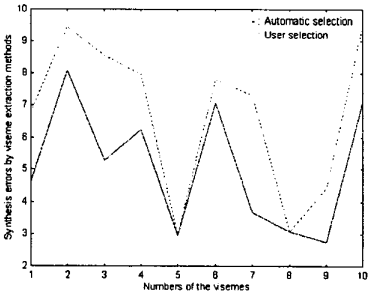

Figure 4. Comparison of synthesis errors

4. An Application Example

The motivation of this research is to improve naturalness of the multimodal communication agent developed in our lab [11]. The multimodal communication agent consists of an image-based synthesis system, speech recognition/translation speech synthesis softwares. The system is able to

- Speak a given text and direct the gaze to a certain direction
- Receive the text from another application over the network
- Allow a new user to register his/her face with acceptable effort (as short as 10 minutes).
- Provide a possibility to use the system in low-bandwidth teleconferencing

The system was written in C++ and divided into several modules which were glued by the Microsoft Foundation Classes (MFC) interface. Communication with other programs was implemented with socket connections, and Tcl or Perl scripts in some cases to keep the communication more flexible.

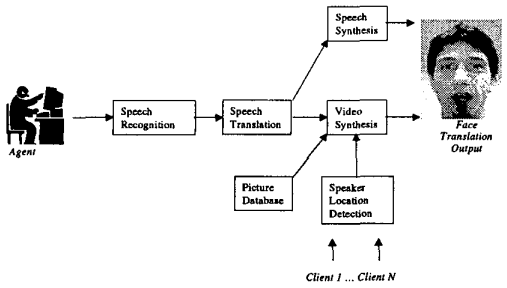

Figure 5. System overview

An advantage of the multimodal communication agent is to allow a user to easily add his/her own face and facial features to the database. The system offers three different modes for user registration. A user can add his/her face and facial features by three different modes: automatic mode, on-line interactive mode and off-line interactive mode.

With supporting from speech recognition/translation and synthesis softwares. The multimodal communication agent can help people who speak different languages to talk to each other with both video and audio via Internet. We use the JANUS system for speech recognition and translation. 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. It can handle as many as 17 different languages. We use the FESTIVAL Text to Speech System [1] to generate the audio speech from text string. Festival offers a full text to speech system through a number of APIs.

Figure 5 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(s) "client(s)." The agent and client(s) speak different languages. But the agent can talk to the client(s) via the system. The client will see the agent's face speaking the translated sentences with synchronized lip movements. The agent's eye 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 the text input. The eye gaze is determined by the location of the client detected by the location detector.

5. Conclusion

We have presented an approach to automatically select optimal visemes for visual speech synthesis. With the proposed method, it is possible to minimize the synthesis errors. We described the development of an image-based multimodal communication agent. The system is capable of generating audio-visual output from a pre-stored image database and text input. The system translates both audio and video from one language to another language. In the current system audio and video translation works for English and German. We have applied the system to a translation task where two people, who speak different languages, can talk with each other over the Internet.

Acknowledgements

We would like to thank Hua Yu for his help on phoneme segmentation in the experiments. We would also like to thank our colleagues in the Interactive Systems Laboratories for their support and discussion on this research.

References

[1] A. W. Black, P. Taylor, and R. Caley. Festival. `www.cstr.ed.ac.uk/projects/festival.html`, 1998. The Centre for Speech Technology Research (CSTR) at the University of Edinburgh.

[2] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Computer Graphics Proceedings, Annual Conference Series*, pages 353–360, aug 1997.

[3] M. Cohen, J. Beskow, and D. W. Massaro. Recent developments in facial animation: An inside view. In *Proceedings of Auditory-Visual Speech Processing (AVSP 98)*, dec 1998.

[4] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. In *MIT A.I Memo No. 1658*, May 1999.

[5] F. M. Galanes, J. Unverferth, L. Arslan, and D. Talkin. Generation of lip-synched synthetic faces from phonetically clustered face movement data. In *Proceedings of Auditory-Visual Speech Processing (AVSP 98)*, pages 191–194, dec 1998.

[6] A. Hällgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *Proceedings of Auditory-Visual Speech Processing (AVSP 98)*, pages 181–183, dec 1998.

[7] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In *Proceedings of Auditory-Visual Speech Processing (AVSP 98)*, pages 185–190, dec 1998.

[8] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldà, T. Zeppenfeld, and P. Zhan. Janus-iii: Speech-to-speech translation in multiple languages. In *Proceedings of ICASSP*. IEEE, 1997.

[9] S. Morishima. Real-time talking head driven by voice and its application. In *Proceedings of Auditory-Visual Speech Processing (AVSP 98)*, pages 195–199, dec 1998.

[10] L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna. Testing generality in janus: A multi-lingual speech translation system. In *Proceedings of ICASSP*. IEEE, 1992.

[11] M. Ritter, U. Meier, J. Yang, and A. Waibel. Face translation: A multimodal translation agent. In *Proceedings of Auditory-Visual Speech Processing (AVSP 99)*, aug 1999.