interspeech 2008 incorporating SST 08

A Real-Time Text to Audio-Visual Speech Synthesis System

Lijuan Wang[1], Xiaojun Qian[2], Lei Ma[1], Yao Qian[1], Yining Chen[1], and Frank Soong[1]

[1]Microsoft Research Asia, Beijing, China
[2]Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[1]{lijuanw, lema, yaoqian, ynchen, frankkps}@microsoft.com; [2]qian_xj@hotmail.com

10.21437/Interspeech.2008-596

Abstract

In addition to speech, visual information (e.g., facial expressions, head motions, and gestures) is an important part of human communication. It conveys, explicitly or implicitly, the intentions, the emotion states, and other paralinguistic information encoded in the speech chain. In this paper we present a multi-language, real-time text-to-audiovisual speech synthesis system, which automatically generates both audio and visual streams for a given text. While the audio stream is generated by our new HMM-based TTS engine, the visual stream is rendered by incorporating multiple animation channels, which control a cartoon figure parameterized in a 3D model simultaneously. The challenges in synthesizing, synchronizing, and integrating multiple-channel information sources are investigated and methods of generating natural, realistic animations are developed. The result of rendering all available or learned information is an expressive audio-visual synthesis module for user-friendly, human-machine communication applications.

Index Terms: audio-visual, speech synthesis, HMM-based, animation generation

1. Introduction

Speech in essence is a stream of synchronized audio and visual signals. Previous researches on speech synthesis are mainly focused on handling the audio part. As computers become more and more powerful, integrating visual signal into speech is possible. Adding visual information makes speech easier to understand, which is especially helpful to those of impaired hearing or in a noisy environment [1]. Moreover, visual information, like facial expressions and body gestures, is more effective to convey emotions, which has important applications in virtual reality, games, video conferencing, etc.

The visual signal of speech is composed of facial expression, head movement, and body gesture. Previous research on phoneme production have mainly focused the movements of lip and tongue. McGurk showed in [5] the strong dependency of speech perception on lips. In his experiment, the utterance of /b/ is combined with the lip movement of /g/. The non-matched lip movement makes the subject perceive /d/ instead of /b/. However, other facial expressions, like cheek/eye/eyebrow movements, are also vital to speech perception [1,2,4]. Even non-facial movements (e.g. head rotations [3,11] and body gestures) need to be closely synchronized with the utterance. One example is the nod of head synchronized with the utterance of 'yes'. Therefore, a text to audio-visual speech synthesis should automatically generates the facial (lip, eye, cheek, etc.), head, and body motions that naturally accompany the audio utterances.

This paper introduces a text-to-audiovisual speech synthesis system, which automatically provide synthesized audio and visual streams, synchronization, and integration. The audio stream is generated by our new HMM-based speech engine. The visual stream is obtained by synthesizing multiple animation channels, which respectively control a 3D parametric model named Walter. These channels, like lip-sync, head movement, eye blink, facial emotions and limb gesture, correspond to different sets of parametric controllers (muscles, virtual muscles and bones) that drive the 3D model and make Walter alive and expressive. The final audiovisual sequence is obtained by integrating these channels and synchronizing the visual stream with the audio stream. The submitted video clip is a sample sequence automatically synthesized by our system.

The rest of the paper is organized as follows. Section 2 introduces the infrastructure of our system. Section 3 presents the text-to-speech synthesis component. Section 4 shows the facial expression and body animation synthesis component. Section 5 draws the conclusions.

2. System Overview

2.1. Flow of Audio-Visual Synthesis

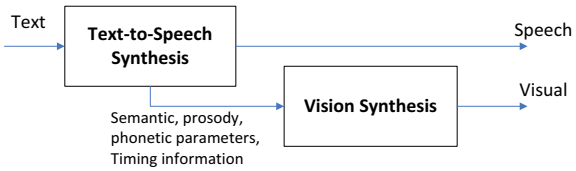

Figure 1: *Two-step text to audio-visual speech synthesis.*

The proposed text to audio-visual speech synthesis has two stages, as shown in Fig. 1. In the first step, the audio signal, as well as the semantic, prosody, phonetic and synchronized timing information, are obtained. In the second stage, the semantic, prosody, phonetic and timing information is used to generate natural, expressive, and synchronized facial, head, and body motions. The animation rendering is also done. Thus, the visual stream is obtained and the synchronization between audio and visual streams is guaranteed.

Fig. 2 shows a more detailed synthesis flow. First, the text analysis module analyzes the text input, produces syntactic and semantic features, and predicts prosody and phoneme sequences. Then, the HMM-based speech synthesizer generates the spectrum and prosody coefficients by maximizing the likelihood of the pre-trained HMM models for the spectrum, pitch and duration. The speech waveform is synthesized from the generated coefficients. The semantic and prosody features, including phonetic parameters, timing, part-of-speech (POS), word accent and stress, pitch contour and energy contour, will be used in the animation synthesis for multiple channels. These channels include lip-sync, head movement, eye blink, facial emotions and hand gesture. Two

kinds of animation generation methods are adopted. One is a key-frame based method which is used for lip-sync, facial emotions, and eye blink channels. The other is HMM-based method which is used in head motion synthesis. The final audiovisual sequence is obtained by integrating these channels and synchronizing the visual stream with the audio stream.

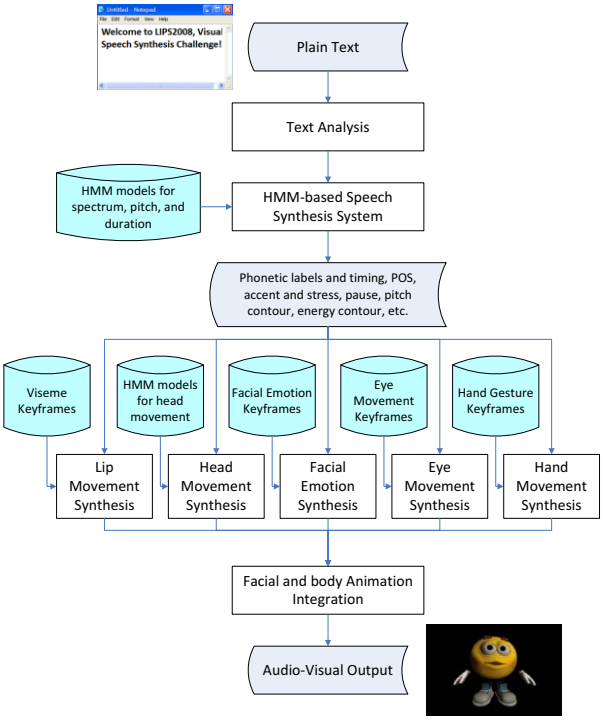

Figure 2: *Flowchart of text to audio-visual speech synthesis.*

2.2. System features

The proposed system has the following features. The more detailed explanations of the underlying modules are given in the next sections.

- Real-time, synchronized audio-visual stream synthesis
- Support multiple languages: American English and Mandarin Chinese
- Support multiple voice fonts of both male and female
- Changeable characters (3D models other than Walter can be used)
- Automatic lip movement synchronization
- Automatic head motion synthesis
- Support multiple facial emotions
- Automatic eye blink, eye gaze
- Support limb gestures

3. HMM-based Text-to-Speech Synthesis

The text-to-audio synthesis is done by a HMM-based speech engine, which consists of the training and synthesis processes [10]. The training process is similar to what is done during speech recognition. The main difference is that both spectrum (e.g., LSP coefficients and their dynamic features) and excitation (e.g., log $F0$ and its dynamic features) parameters are extracted from a speech database, which are modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). To properly model log $F0$ sequence that includes unvoiced regions, multi-space probability distributions are used to model the state output stream for log $F0$. Each HMM has its state duration density to model the temporal structure of speech. As a result, the audio system models spectrum, excitation, and durations in a uniform framework.

The synthesis process is the inverse of speech recognition. First, the text to be synthesized is converted to a context-dependent label sequence. Then the HMM of the audio stream is constructed by concatenating the context dependent HMMs according to the label sequence. Secondly, state durations of the HMM are determined based on the state duration probability density functions. Thirdly, the speech parameter generation algorithm generates the sequence of LSP coefficients and log $F0$ values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated LSP coefficients and $F0$ values with mixed excitation.

4. Animation trajectory generation

Three-dimensional models provide the most powerful means to generate animations. This section firstly explains how to generate various motions (like lip movement and head rotation); then introduces two methods of synthesizing animations: key-frame-based method and HMM-based method.

4.1. Basic animation controller: muscle and bone

Animation is a sequential changing of 3-D model parameters over time, which can be achieved in different ways, such as shape/morph target, bone/cage, skeleton-muscle based systems, motion capturing of points on face, and knowledge based solver deformations. Our system uses the skeleton-muscle based physical model to deform the parameterized 3D model.

In the skeleton-muscle model, there are different types of controllers: facial muscles for facial expression, virtual muscles for eye blink, and skeleton for head and hand motion. Each controller controls the manner and displacement of the vertices on the associated meshes. As Fig. 3 shows, the lip muscles control how wide the mouth opens. Similarly, as shown in Fig 4., body gestures can be achieved by controlling the spatial positions of bones on the skeleton.

Figure 3: *Lip motion achieved by controlling virtual muscle.*

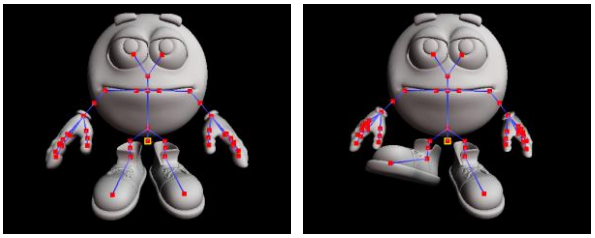

Figure 4: *Body gesture achieved by bone (skeleton) control.*

4.2. Key-frame-based animation generation

In the traditional drawing-based animation production, the key artist only draws some key-frames. Then, after testing and approving the rough animation, he hands over the rest to his

assistant. The assistant does the clean-up and makes up the necessary in between frames. The key-frame based animation generation method proposed in this study is basically the same. Firstly, a set of key-frames need to be prepared for each channel. These key-frames can be obtained by capturing the motions of a real human, or manual drawing of an artist. Then, the temporal position and weight of the key-frames are decided. Finally, interpolation fills in the gap between key-frames and the animation is generated. Linear interpolation would cause abrupt, unnatural motions. So, spline interpolation is used to smooth the movements.

Following channels in our system use the key-frame based method: lip-sync, eye blink, and facial emotions. For these channels, there are two key problems. The first is preparation of key-frames. For cartoon models, the facial expressions are more exaggerated. The key-frames are therefore designed by artists. The second problem is how to decide the position and weight of key-frames in the animation.

The animation of each phoneme is a consequence of specific articulatory movements. Motions of lips, tongue, teeth and jaw correspond to the specific phoneme. However, several phonemes look similar. The visually indistinguishable phonemes are typically grouped to the same *viseme category* (or viseme) [4]. In lipreading studies, viseme categories can be defined by clustering perceptually similar phonemes. Currently, in our model, there are separate visemes for each English phoneme, part of them are shown in Fig. 5. After the audio stream is synthesized, the phonetic sequence and their timing information are obtained. With this information, the viseme key-frames can be optimally inserted at the center of each phoneme. This is under the assumption that the distinctive motion (key-frame) of a phoneme occurs at the mid-point of the phoneme.

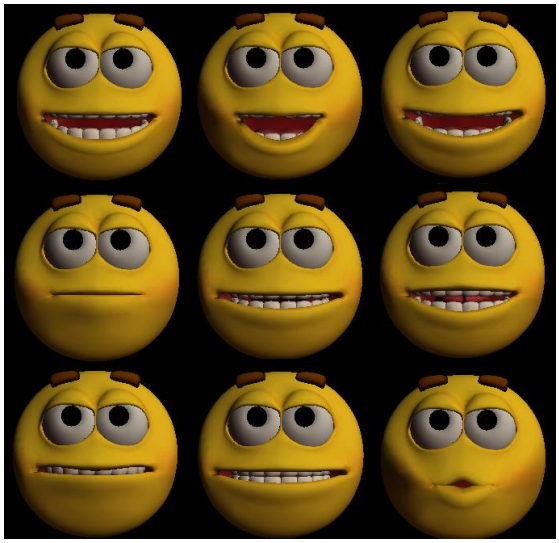

Figure 5: *Key-frames of viseme ("aa", "aw", "ay", "b", "ch", "dh", "f", "s", and "uw")*

Figure 6: *Key-frames of eye blink (BothOpen, OneClosed, BothClosed)*

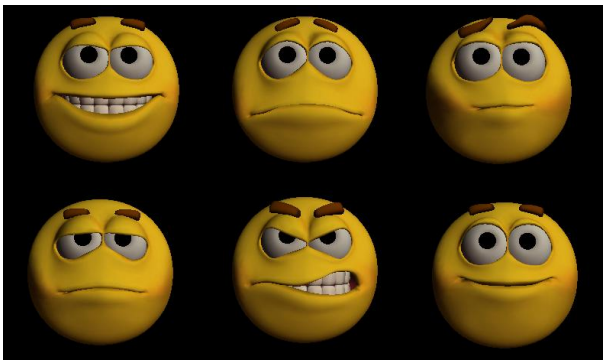

Figure 7: *Facial emotion key-frames (grin, sad, sarcastic, sleepy, angry, smirk)*

Similarly, two sets of key-frames for eye blink (Fig. 6) and facial emotion (Fig. 7) are designed respectively. In current system, the animations of the two channels are randomly generated under the constraint of some rules.

4.3. HMM Model-based animation generation

A stationary head without any rotation during speaking is unrealistic. This subsection shows how to generate natural head motions by using HMM-based motion generation method [3,11].

We extract the head motions from the Voice-of-America (VOA) video broadcasting because of the expressive speaking and the natural head motion of the female announcer. By using the head pose tracking technology [9], as shown in Fig.8 and Fig. 9, the Euler angels of the head on three dimensions is obtained. The head motions, which are synchronized with the speech prosody features (F0 and energy), are used as our training data.

We use HMM to model head motion because it provides a suitable and natural way to represent the temporal relation between acoustic prosodic features and head motions. HMMs are used to generate the most frequent head motion sequences based on the given observation (prosodic features).

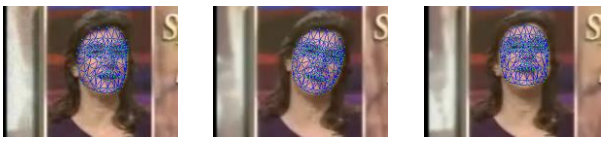

Figure 8: *Capturing 3D head motions (Euler angels on three dimensions) from a video clip.*

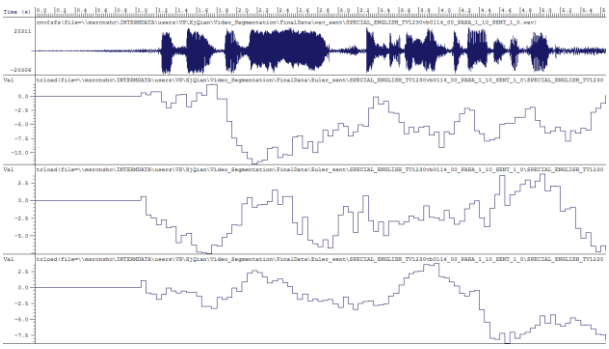

Figure 9: *Head motions (in Euler angles) on three dimensions during speech ("This is the VOA special English development report.").*

4.4. Integration of different channels

The visual stream is synthesized by generating multiple animation channels. These channels may not be independent of each other. Since each channel controllers multiple parameterized controllers (muscles, virtual muscles, bones), one controller can be shared by multiple channels. For example, the muscles around the mouth are controlled by both the lip-sync channel and the facial emotion channel. To handle the overlapped control of parameters by multiple channels, the maximum parameter value given by a controller is used as the final parameter value.

Similar problem may happen at vertex level. Since each controller controls the manner and displacement of the vertices on its associated meshes, a vertex can be influenced by multiple controllers. In that case, the influence is accumulated.

Fig. 10 shows the snapshots of the automatically synthesized audio-visual stream. The complete video clip has been submitted to the conference.

5. Conclusions

This paper introduces a real-time text-to-audiovisual speech synthesis system, which automatically generates audio and visual streams for given text. The details of synthesizing, synchronizing, and integrating the audio and visual streams are explained. Two methods of generating natural, emotional animations are proposed.

6. References

[1] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133-137, February 2004.

[2] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, Washington, D.C., USA, May 2002, p. 396-401.

[3] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283-290, July 2005.

[4] M. Brand, "Voice puppetry," in *Proc. of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, New York, NY, USA, 1999, pp. 21-28.

[5] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, December 1976.

[6] S. King, et al., "Creating Speech-Synchronised Animation," *IEEE Transactions on visualization and computer graphics* 11(3), 341-352 (2005)

[7] S. Fagel, "Merging methods of speech visualization," ZAS Papers in Linguistics 40, 19-32 (2005)

[8] Z.Y. Wu, S. Zhang, L.H. Cai, H.M. Meng, "Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a Three-dimensional Avatar," In *Proc. of Interspeech 2006*, pp. 1802–1805.

[9] Q. Wang, W. Zhang, X. Tang, and H.Y. Shum, "Real-Time Bayesian 3-D Pose Tracking," *IEEE Transaction on Circuits and Systems for Video Technology, Vol 16, 1533-1541, Dec. 2006.*

[10] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp/.

[11] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," in *Proc. of SIGGRAPH*, 2007.

[12] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2492–2499, 2005.

[13] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," in *Proc. of Eurospeech*, 1999, pp. 959–962.

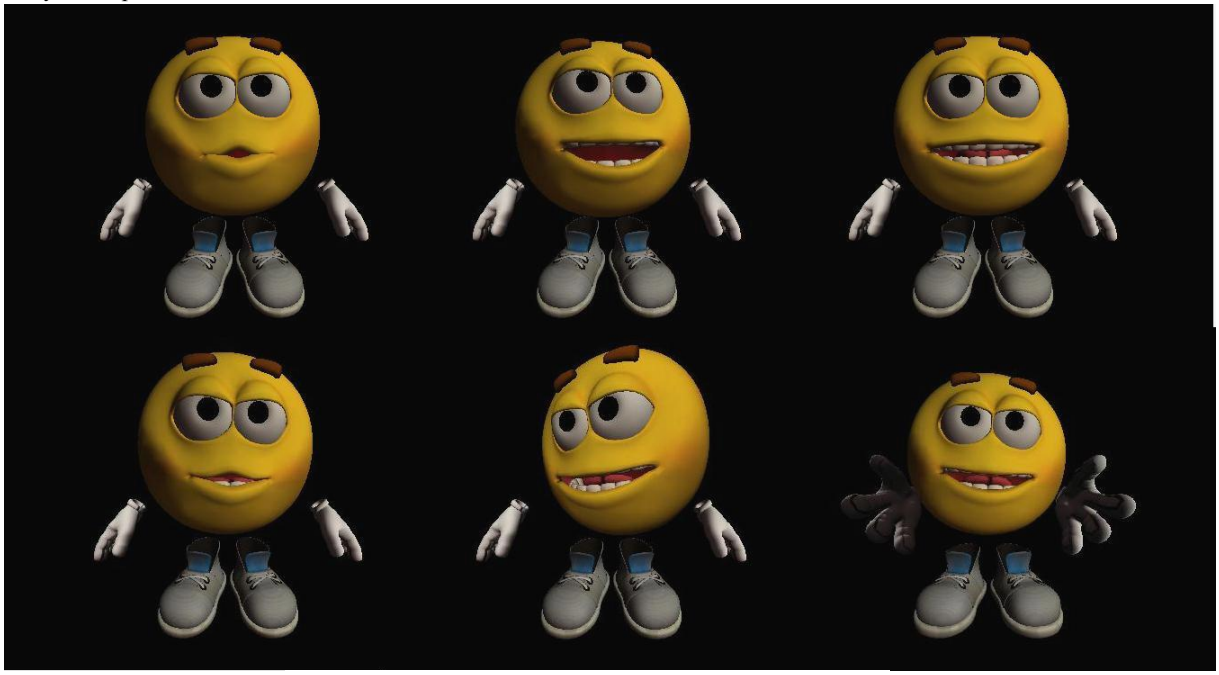

Figure 10: Snapshots of synthesized audio-visual stream.