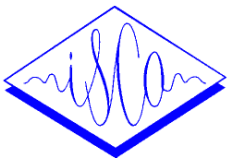

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

HMM-BASED TEXT-TO-AUDIO-VISUAL SPEECH SYNTHESIS

Shinji Sako†, *Keiichi Tokuda*†, *Takashi Masuko*††, *Takao Kobayashi*††*and Tadashi Kitamura*†

†Department of Computer Science
Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN

††Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Yokohama, 226-8502 JAPAN

10.21437/ICSLP.2000-469

ABSTRACT

This paper describes a technique for text-to-audio-visual speech synthesis based on hidden Markov models (HMMs), in which lip image sequences are modeled based on image- or pixel-based approach. To reduce the dimensionality of visual speech feature space, we obtain a set of orthogonal vectors (eigenlips) by principal components analysis (PCA), and use a subset of the PCA coefficients and their dynamic features as visual speech parameters. Auditory and visual speech parameters are modeled by HMMs separately, and lip movements are synchronized with auditory speech by using phoneme boundaries of auditory speech for synthesizing lip image sequences. We confirmed that the generated auditory speech and lip image sequences are realistic and synchronized naturally.

1. INTRODUCTION

It is well known that human speech is bimodal both in expression and perception. It is shown that human perception of auditory speech can be affected by the visual information of lip movements [1]. There have been proposed various approaches to incorporating bimodality of speech into human computer interaction interfaces. Audio-visual speech synthesis is one of the research topics in this area.

There exist two approaches to modeling lip movements, that is, model-based approach and image- or pixel-based approach. As one of the model-based approaches, a facial animation synthesis system using 3-D mouth shape have been developed [2]. We have also been proposed an audio-visual speech synthesis system [3] by applying a framework of HMM-based speech synthesis [4], [5], in which the lip shape is modeled by geometric parameters representing the lip contour. However, one of the difficulties in this approach is extracting positional parameters from a large amount of image sequences. On the other hand, in the image- or pixel-based approach, it is easy to handle data by processing intensities of pixels directly in the training and synthesis stages. Moreover, realistic lip image sequences, which include inner mouth parts such as teeth or a tongue, can be easily synthesized without creating computer graphics (CG).

In this paper, we adopt the latter approach, and propose a technique for text-to-audio-visual speech synthesis based on HMM. The synthesis system consists of auditory speech synthesis part and visual speech synthesis part. The auditory speech synthesis part is the same as [5]. The visual speech synthesis part also adopts the same framework except for the feature parameters. Lip movements are synchronized with auditory speech by using phoneme boundaries of auditory speech for synthesizing lip image sequences. Image-based visual speech synthesis systems based on HMM have also been proposed in [6], [7]. Our approach differs from them in that dynamic feature parameters are used in addition to static feature parameters, and the generated sequence reflects statistical information of both static and dynamic features of several phonemes before and after the current phoneme.

In the following, we summarize the HMM-based audio-visual speech synthesis system in Section 2. Experimental conditions and results are also given in Section 3 and 4, respectively, and concluding remarks are presented in the Section 5.

2. HMM-BASED AUDIO-VISUAL SPEECH SYNTHESIS SYSTEM

The block diagram of the HMM-based text-to-audio-visual speech synthesis system is illustrated in Figure 1. First, arbitrary input text to be synthesized is converted to a context-based label sequence. Then a sentence auditory HMM is constructed by concatenating context-dependent phoneme auditory HMMs according to the label sequence. State durations of the sentence auditory HMM are determined by the state duration densities [8], and a sequence of auditory speech parameter vectors is generated from the sentence auditory HMM by using a speech parameter generation algorithm [4]. Simultaneously, state durations of the sentence visual HMM are determined based on the obtained phoneme durations of synthetic speech. According to the

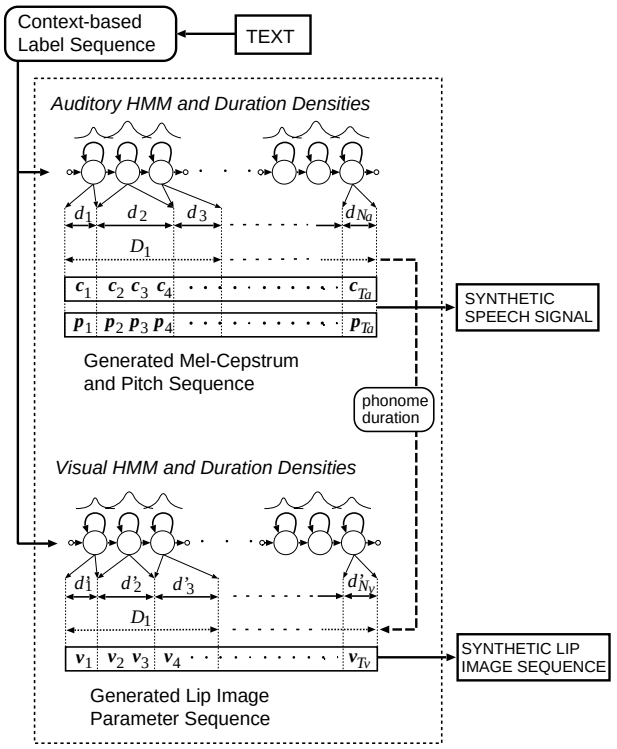

Figure 1: The synthesis part of HMM-based text-to-audio-visual-speech synthesis system.

state durations, a sequence of visual speech parameter vectors is generated in the same manner as auditory speech synthesis.

2.1. Audio Speech Synthesis

We use mel-cepstral coefficients as spectral parameters. Sequences of mel-cepstral coefficient vectors are modeled by continuous density HMMs. Pitch patterns and state durations are modeled by multi-space probability distribution HMMs [9] and multi-dimensional Gaussian distributions, respectively. The auditory feature vector consists of two streams, i.e., the one for spectral parameter vector and the other for pitch parameter vector, and each phoneme HMM has its state duration densities. To model variations of spectra, pitch and duration accurately, we take account of contextual factors such as phone identity factors, stress-related factors and locational factors. The distributions for spectral parameters, pitch parameters and state durations are clustered independently by using a decision-tree based context clustering technique [10].

A sequence of mel-cepstral coefficients and pitch values is generated from the sentence auditory HMM by using parameter generation algorithm summarized in the following.

For a given continuous HMM λ and a state sequence $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_T\}$, we obtain an auditory parameter sequence $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ that maximizes $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$. The output distribution of each state is assumed to be a single Gaussian distribution. If the output parameters are determined independently of preceding or succeeding frames, it is obvious that $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ is maximized when the parameter vector sequence is equal to the mean vector sequence. This may result in discontinuities at state boundaries in the generated parameter sequence.

To avoid this problem, we assume that feature parameter vector $\boldsymbol{o}_t$ consists of static and dynamic feature vectors, that is, $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$, and dynamic feature vectors were calculated as follows:

$$\Delta^{(n)}\boldsymbol{c}_t = \sum_{i=-L_-^{(n)}}^{L_+^{(n)}} w^{(n)}(i)\boldsymbol{c}_{t+i}, \quad n = 1, 2, \tag{1}$$

where $\Delta^{(1)}\boldsymbol{c}_t = \Delta\boldsymbol{c}_t$, $\Delta^{(2)}\boldsymbol{c}_t = \Delta^2\boldsymbol{c}_t$. Under these constrains, a sequence of static feature parameter vectors $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_T\}$ is determined by a set of linear equations $\partial \log P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)/\partial\boldsymbol{C} = \boldsymbol{0}_{TM}$, which can be solved by a fast algorithm derived in [4]. By using dynamic features, the generated parameter vector sequence reflects both means and covariances of the output distributions of a number of frames before and after the current frame.

Finally, speech is synthesized directly from the generated mel-cepstral coefficients and pitch values by using the MLSA (Mel Log Spectrum Approximation) filter [11], [12].

2.2. Visual Speech Synthesis

One of the problems of the image-based approach is high computational complexity of modeling and synthesizing lip images because of high dimensionality of the feature space. To cope with this problem, we apply principal component analysis (PCA) to lip images, and represent each lip image by a linear combinations of orthogonal vectors (eigenlips) in a manner similar to the eigenface [13]. By using a subset of coefficients associated with eigenlips as feature parameters, dimensionality of the feature space and computational complexity could be reduced significantly.

We used phoneme models as the visual speech synthesis units, so that lip movements can easily be synchronized with auditory speech. The basic idea for synchronizing lip movements with auditory speech is to use the identical phoneme durations in the auditory and visual speech synthesis parts. In this work, the state durations of visual HMMs are determined based on phoneme durations obtained in the auditory speech synthesis part. According to the obtained state durations, a sequence of PCA coefficients is generated from the sentence visual HMM in the same manner as auditory speech synthesis. A lip image sequence is reconstructed from generated PCA coefficients and eigenlips.

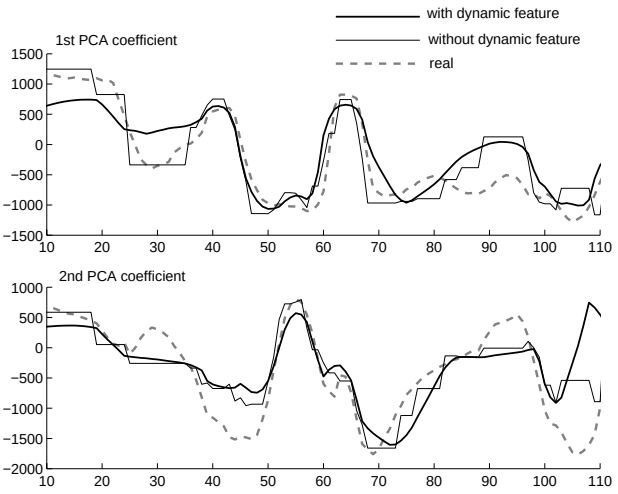

Figure 2: Trajectories of the synthesized PCA coefficient vectors. (top: 1st PCA coefficient, bottom: 2nd PCA coefficient)

3. EXPERIMENTAL CONDITIONS

We used phonetically balanced 450 sentences from the ATR Japanese speech database for training auditory HMMs. Auditory speech signals were sampled at 16kHz and windowed by a 25ms Blackman window with a 5ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis [11]. The auditory feature vector consists of spectral and pitch parameter vectors. The spectral parameter vector consists of 25 mel-cepstral coefficients including the 0th coefficient, and their first and second derivatives. The pitch parameter vector consists of log F_0 and its first and second derivatives.

We constructed an audio-visual speech database which consists of the same phonetically balanced Japanese 503 sentences as the ATR Japanese speech database. Auditory speech and the corresponding video images were recorded in parallel using a DAT recorder and a digital VCR. The video images contained only mouth area and the tip of the nose. NTSC video frames were digitized at 29.97 fps, 720×480 pixels, 24bits per pixel. Each frame was decomposed into two interlaced fields. As a result, we obtained about 60 lip images per second. Captured images were phoneme-labeled semi-automatically according to the segmentation results of the auditory speech. Location and intensity normalization were applied to training lip image sequences, then the lip regions were cut out to 176×144 pixels. For training visual HMMs, we used the same 450 sentences from this database as those used for training auditory HMMs. We obtained 1024 eigenlips by applying PCA to 1024 lip images selected randomly from 179281 frames in the training set, and represented all the training images by linear combinations of the eigenlips. The visual feature vector consisted of 32 coefficients associated with the top 32 eigenlips and their first and second derivatives. The top 32 eigenlips contained about 80% of the statistical variance of the ensemble.

Auditory and visual synthesis parts used 5-state left-to-right HMMs and 3-state left-to-right HMMs, respectively. The details of the contextual factors taken into account for modeling auditory speech are shown in [5], and the only phonetic contextual factors were taken into account for modeling visual speech.

4. RESULTS

We synthesized audio-visual speech of Japanese sentences which were not included in the training data. Figure 2 shows the trajectories of the synthesized lip image parameters for a sentence. Only a fragment corresponding to the phrase "/t-o-k-a-i-d-e-w-a/," which means "in a city" in English, is shown in the figure. The thick lines show the trajectories of PCA coefficients of the synthesized visual-speech with dynamic features and thin lines show those without dynamic features. Dashed lines show those of a real speech. The trajectories of the synthesized parameters with dynamic features are smooth and resemble those of the real parameters. It is noted that no additional smoothing process was applied in the proposed system.

Figure 3 shows generated spectra and lip image sequence for the same fragment as shown in Figure 2. It was observed that the synthesized speech and lip image sequence are smooth and realistic. The movie files[1] attached to this paper demonstrate our text-to-audio-visual speech synthesis system. Movie files [Movie 01692_01.MOV] and [Movie 01692_02.MOV] were generated with dynamic visual features and without dynamic visual features, respectively. In both of the movies, auditory speech were synthesized with dynamic features.

5. CONCLUSION

We proposed a technique for image-based audio-visual speech synthesis from an arbitrary input text by applying the framework of HMM-based speech synthesis. The experimental results showed that, by using the PCA coefficients and their dynamic feature parameters as visual features, we can synthesize smooth and realistic lip image sequences from visual HMMs, and that generated auditory speech signal and lip image sequences are synchronized naturally. Future work will be directed toward constructing an audio-visual speech synthesizer which can synthesize whole facial image sequence.

[1]The latest movie files can be found in our WWW site at: *http://kt-lab.ics.nitech.ac.jp/˜sako/lipsynthesis*

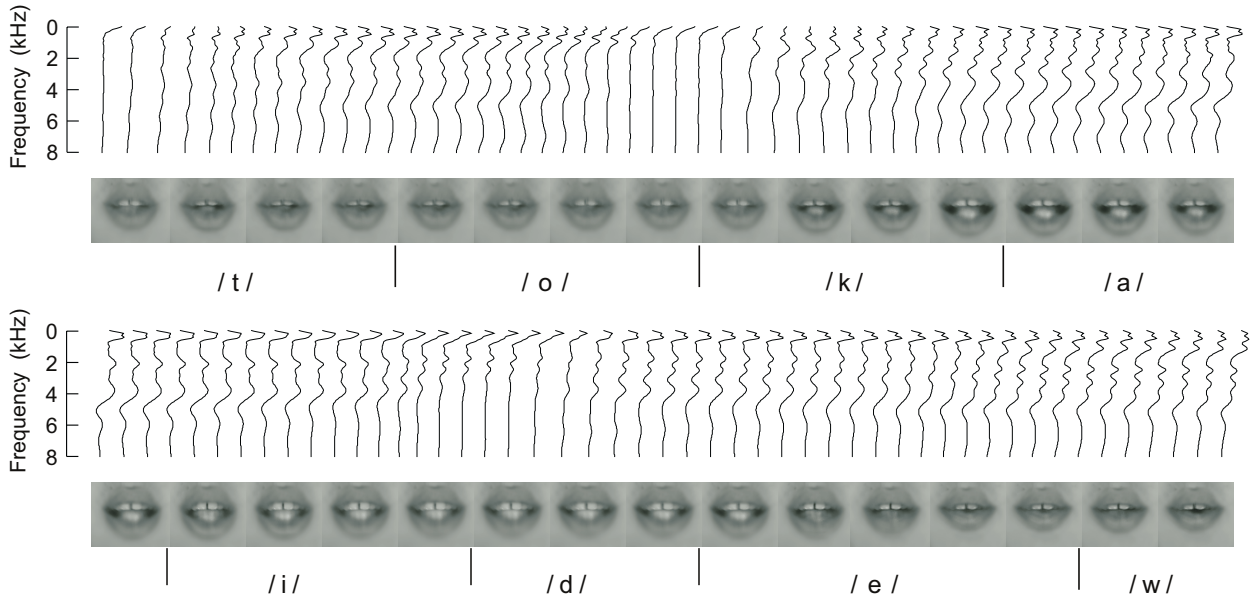

Figure 3: Generated spectra and image sequence for a sentence "/t-o-k-a-i-d-e-w-a/".

6. ACKNOWLEDGMENT

We would like to thank Mr. Takayoshi Yoshimura for providing us with some programs. This work was supported in part by the Ministry of Education, Science, Sports and Culture of Japan, Grant-in-Aid for Scientific Research 10555125, 1998 and 11878064, 1999, and in part by Regular Assistance Grant of the Hoso-Bunka Foundation, Inc.

7. REFERENCES

1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, 264, pp.746–748, Dec. 1976.

2. M. M. Cohen, J. Beskow and D. W. Massaro, "Recent developments in facial animation: An Inside View," Proc. AVSP, pp.201–206, 1998.

3. M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," Proc. of EUROSPEECH, Vol 2, pp.959–962, 1999.

4. K. Tokuda, T. Masuko, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," Proc. ICASSP, I, pp.389–392, 1996.

5. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH, vol.5, pp.2347–2350, 1999.

6. N. M. Brooke and S. D. Scott, "Two- and three-dimensional audio-visual speech synthesis," Proc. AVSP, pp.213–218, 1998.

7. J. J. Williams, A. K. Katsaggelos and M. A. Randolph, "A hidden Markov model based visual speech synthesizer," Proc. ICASSP, vol.4, pp.2393–2396, 2000.

8. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration modeling in HMM-based speech synthesis system," Proc. of ICSLP, vol.2, pp.29–32, 1998.

9. K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. of ICASSP, vol.1, pp.229–232, Mar. 1999.

10. J. J. Odell, "The use of context in large vocabulary speech recognition," PhD dissertation, Cambridge University, 1995.

11. T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP, vol.1, pp.137–140, 1992.

12. S. Imai, "Cepstral analysis synthesis on the mel frequency scale," Proc. of ICASSP, pp.93–96, 1983.

13. M. A. Turk and A. P. Pentland, "Face recognition using eigennfaces," Proc. of IEEE Computer Society Conf. on Computer Vision and Patter Recognition, pp.586–591, 1991.