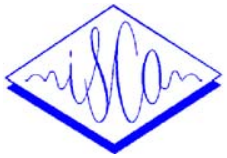


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TEXT-TO-AUDIO-VISUAL SPEECH SYNTHESIS BASED ON PARAMETER GENERATION FROM HMM

Masatsune Tamura, Shigekazu Kondo, Takashi Masuko, and Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
mtamura@ip.titech.ac.jp, masuko@ip.titech.ac.jp, Takao.Kobayashi@ip.titech.ac.jp

10.21437/Eurospeech.1999-234

ABSTRACT

This paper describes a technique for synthesizing auditory speech and lip motion from an arbitrary given text. The technique is an extension of the visual speech synthesis technique based on an algorithm for parameter generation from HMM with dynamic features. Audio and visual features of each speech unit are modeled by a single HMM. Since both audio and visual parameters are generated simultaneously in a unified framework, auditory speech with synchronized lip movements can be generated automatically. We train both syllable and triphone models as the speech synthesis units, and compared their performance in text-to-audio-visual speech synthesis. Experimental results show that the generated audio-visual speech using triphone models achieved higher performance than that using syllable models.

1. INTRODUCTION

Incorporating bimodality of speech into human-computer interaction interfaces generally enhances speech perception and understanding by both humans and computers. From this point of view, recently, there have been proposed various approaches to synthesizing audio-visual speech [1]-[5]. Although some techniques can generate smooth facial animation and auditory speech with acceptable quality, there still exist many problems to be solved. One of these problem is to synthesize speech with arbitrarily speaker individuality and various speaking styles.

In this paper, we present a new approach to text-to-audio-visual speech synthesis based on hidden Markov model (HMM). We have proposed a speech synthesis system using an HMM-based speech parameter generation algorithm [7]. We have also proposed a text-to-visual speech synthesis system by applying this framework to visual speech synthesis [10]. Furthermore, this approach has been extended to speech-driven and text-and-speech-driven visual speech synthesis [11]. Here we apply this approach to text-to-audio-visual speech synthesis, that is, simultaneous generation of both auditory speech and lip motion from a given text in a unified framework.

In the proposing approach, audio and visual features for each speech unit are modeled by a single HMM and both audio and visual parameters are generated in the same framework simultaneously. Therefore, the synthesis system can generate synchronized lip movements with auditory speech automatically. Since the obtained parameter sequence reflects statistical information of both static and dynamic features of several phonemes before and after the current phonemes, synthetic audio-visual speech becomes smooth and natural without requiring additional parameter smoothing. Furthermore, since it has been shown that we can change voice characteristics of synthetic speech by applying speaker adaptation techniques to the speech unit HMMs [8][9], the similar technique can be applied to vary speaker individuality of the synthetic audio-visual speech.

2. HMM-BASED TEXT-TO-AUDIO-VISUAL SPEECH SYNTHESIS SYSTEM

Figure 1 illustrates a block diagram of the text-to-audio-visual speech synthesis system. Excepting the feature parameters and speech units, the framework of the system is the same as the auditory and visual text-to-speech synthesis systems based on HMM [7][10].

2.1. Training of Speech Unit HMMs

Visual speech feature parameters and auditory speech feature parameters are extracted from audio-visual speech database. We use mel-cepstral coefficients and mouth position parameters as the static auditory and visual features respectively. Delta parameters for both auditory and visual parameters, $\Delta \boldsymbol{c}_{\boldsymbol{a}t}$ and $\Delta \boldsymbol{c}_{\boldsymbol{v}t}$, are calculated using the extracted static features $\boldsymbol{c}_{\boldsymbol{a}t}$ and $\boldsymbol{c}_{\boldsymbol{v}t}$. The auditory feature vector $\boldsymbol{o}_{\boldsymbol{a}t} = [\boldsymbol{c}'_{\boldsymbol{a}t}, \Delta \boldsymbol{c}'_{\boldsymbol{a}t}]'$ and the visual feature vector $\boldsymbol{o}_{\boldsymbol{v}t} = [\boldsymbol{c}'_{\boldsymbol{v}t}, \Delta \boldsymbol{c}'_{\boldsymbol{v}t}]'$ are combined into a single audio-visual observation vector $\boldsymbol{o}_t = [\boldsymbol{o}'_{\boldsymbol{a}t}, \boldsymbol{o}'_{\boldsymbol{v}t}]'$.

Using this audio-visual observation vector $\boldsymbol{o}_t$, we train speech unit HMMs, i.e., syllable or phoneme HMMs. In the training of HMMs, we regard an observation sequence to be divided into two streams, namely, auditory and visual parameter streams. In the case of phoneme HMMs, we use triphone HMMs and apply a decision tree based clustering procedure [12] to share the states. Since influential contextual factors are different

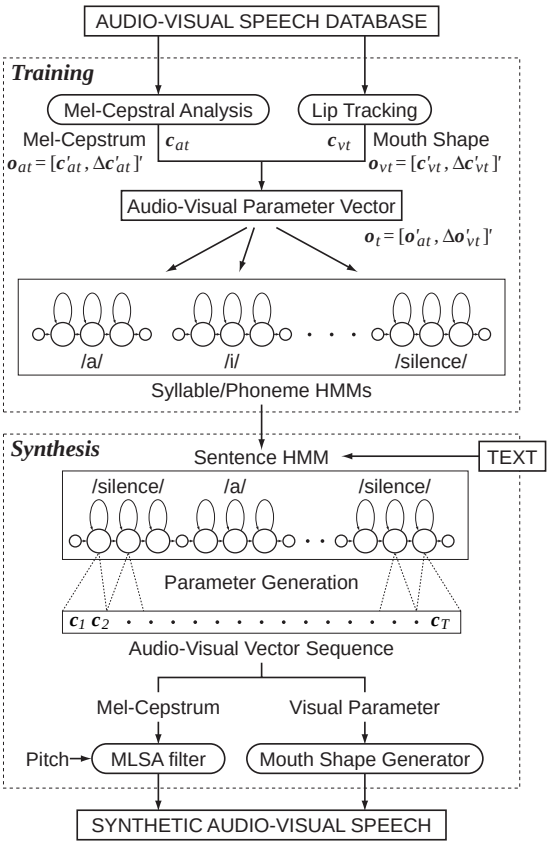


Figure 1: HMM-based text-to-audio-visual speech synthesis system.

between auditory and visual parameters, we construct distinct decision trees for auditory and visual parameter streams. By descending the obtained decision tree, we can synthesize triphones that are not observed in the training data.

2.2. Parameter Generation and Synthesis

In the synthesis phase, arbitrary input text to be synthesized is transformed into a phonetic symbol sequence. According to the phonetic transcription, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating syllable or phoneme HMMs. From the sentence HMM, an audio-visual speech parameter vector sequence is generated using the parameter generation algorithm [6].

In the parameter generation algorithm, a speech parameter vector sequence $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ is obtained maximizing $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ with respect to the state sequence $\boldsymbol{Q}$ and the static feature vector sequence $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$ where $\boldsymbol{c}_t = [\boldsymbol{c}'_{\boldsymbol{a}t}, \boldsymbol{c}'_{\boldsymbol{v}t}]'$ for a given continuous HMM λ and a sequence length T. If the state sequence $\boldsymbol{Q}$ is explicitly known, the optimum parameter vector sequence is obtained by solving a set of linear equations. By using dynamic features, the generated parameter vector reflects both means and covariances of the output distributions of a number of frames before and after the current frame.

Finally, auditory speech is synthesized from the mel-cepstrum sequence using MLSA filter [14]. At the same time, the visual parameter vector sequence is converted into visual speech such as lip animation. Then auditory speech signal and lip animation are combined as the output audio-visual speech.

3. IMPLEMENTATION OF AUDIO-VISUAL SPEECH SYNTHESIS SYSTEM

3.1. Audio-Visual Training Set

In the previous work [10][11], we used an audio-visual speech database consisting of 216 phonetically balanced Japanese words. Since this database is not sufficient for modeling phoneme models, in this work, we made larger audio-visual speech database consisting of 100 phonetically balanced Japanese sentences enunciated by a male speaker. 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 30 fps, 640 $\times$ 480 pixels, 24 bits per pixel. Further each frame was decomposed into two interlaced fields. As a result, we obtained 60 lip shape images per second. Since one phoneme segment often contains only one or two video frames in a rate of 60 fps, visual feature vectors are linearly interpolated to 120 fps. Captured images were phoneme labeled automatically according to the segmentation results of the auditory speech. Auditory speech was sampled at 12 kHz, 16 bits per sample.

3.2. Audio-Visual Feature Parameters

We used mel-cepstral coefficients as auditory speech features. The mel-cepstral coefficients were obtained by a mel-cepstral analysis technique [13] [14] on each 33.3 ms frame of speech with a Blackman window every 8.3 ms. The static feature vector $\boldsymbol{c_a}$ consists of 26 mel-cepstral coefficients including the zeroth coefficient. Then delta parameters were calculated by simple difference between current and preceding frames. Thus, each auditory feature vector becomes a 52 dimensional vector.

To represent the lip shape, we used 10 position parameters [10][11]. They are vertical distance from the nose to the corner of the mouth y, horizontal opening of inner contour $2w$, vertical distances from horizontal axis, which is the line joining mouth corners, to the inner upper contour $\{u_0, u_1, u_2, u_3\}$ and inner lower contour $\{l_0, l_1, l_2, l_3\}$ at 4 equally spaced points between the center and the right corner of the mouth. We extracted these position parameters from captured images by hand. As the static visual feature vector, we used a 10-dimensional vector $\boldsymbol{c_v} = [y, w, \boldsymbol{c}'_{\boldsymbol{v}u}, \boldsymbol{c}'_{\boldsymbol{v}l}]'$ where $\boldsymbol{c_{v}}_u$ and $\boldsymbol{c_{v}}_l$ are DCTs of $\boldsymbol{u} = [u_0, u_1, u_2, u_3]'$ and $\boldsymbol{l} = [l_0, l_1, l_2, l_3]'$, respectively. As well as the auditory feature parameters, delta parameters were calculated by simple difference between current and preceding frames. Consequently, each visual

feature vector becomes a 20-dimensional vector which consists of static and dynamic features.

3.3. Models, Training, and Synthesis

In the HMM-based auditory speech synthesis, phoneme models are used as the speech synthesis units [7]-[9]. On the other hand, we used CV syllable HMMs in the visual speech synthesis [10][11]. Here, we trained both syllable and phoneme models as the speech synthesis units, and compared their performance in text-to-audio-visual speech synthesis.

We modeled each Japanese syllable by a 7-state left-to-right model with single Gaussian diagonal output distribution and no skips. A total number of 144 syllables were appeared in the database. After the training of the syllable models, they were reestimated with the embedded training version of the Baum-Welch algorithm.

For each phoneme, we trained a 4-state left-to-right model with single Gaussian diagonal output distributions and no skips. First, we estimated 45 monophone HMMs and copied to triphone HMMs. The number of triphones which is observed in the training database is 2029. Then we reestimated HMMs and constructed decision trees for auditory stream and visual stream. Four sets of phoneme models were obtained by changing the threshold for decision tree construction. Resultant phoneme models consisted of 3031 / 993 PDFs, 3031 / 440 PDFs, 1055 / 993 PDFs, and 1055 / 440 PDFs for auditory / visual streams, respectively. After constructing trees, we again reestimated HMMs.

In addition, since we analyzed and modeled audio-visual speech with a frame rate of 120 fps, the generated visual parameter sequence has the same frame rate. To synthesize lip animation with 30 fps, we downsampled the generated visual parameter sequences by a factor of 4.

4. EXPERIMENTAL RESULTS

Since the audio-visual database that we used was still small and thus available training data was limited, we chose one sentence from the database arbitrarily and used it as a testing sentence. Then the remainder of the sentence, namely 99 sentences out of the 100-sentence set, were used as the training data. We generated both training and testing sentences using the proposed audio-visual speech synthesis technique.

Table 1 shows a comparison of the spectral distortion and lip shape distortion for training data when different speech unit models were used. Spectral distortion is the rms mel-cepstral distance between spectra of real utterance and synthetic speech. Lip shape distortion is the rms difference between heights of mouth opening of the real video images and synthetic lip shapes. In this experiment, we used pitch contour and duration information obtained from the database. It is shown that phoneme models give higher performance than the syllable model.

Table 1: Comparison of synthesis performance for training sentences.

Model	No. of A / V PDFs	Distortion in dB / Pixel
Phoneme	1055 / 440	3.8 / 14.3
	1055 / 993	3.8 / 13.1
	3031 / 440	3.4 / 14.6
	3031 / 993	3.5 / 13.4
Syllable	1008 / 1008	4.1 / 14.5

Figure 2 shows comparison of generated spectral envelopes and lip movements for a portion of the testing sentence. In the figure, (a) extracted from real utterance, (b) synthetic speech generated from phoneme models with 3031 and 993 PDFs for auditory and visual parameters, and (c) synthetic speech generated from syllable HMMs are shown, respectively. From this figure, it can be seen that synthetic audio-visual speech is smooth and resembles real one.

Through the informal listening tests, it was shown that the synthetic auditory speech using phoneme models provide higher quality than that using the syllable HMMs. However, it was observed only slight differences between the phoneme models and the syllable model in generated lip movements.

5. CONCLUSION

We have proposed a technique for generating audio-visual speech from arbitrary input text. The approach is based on the parameter generation algorithm from HMM with dynamic features. It has been shown that triphone models give higher performance than syllable models. Furthermore generated audio-visual speech is quite smooth and natural. Although we generated only spectral parameters for auditory speech synthesis, we have already developed a speech synthesis system in which spectral information and prosodic information are modeled in the same frame work simultaneously [15]. Audio-visual synthesis with various speaker individuality is our future work.

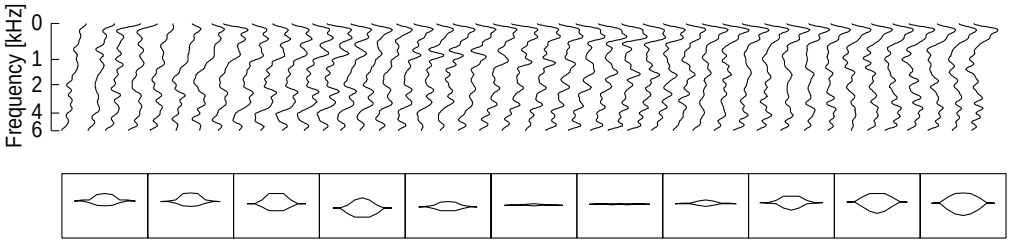
ACKNOWLEDGMENT

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.

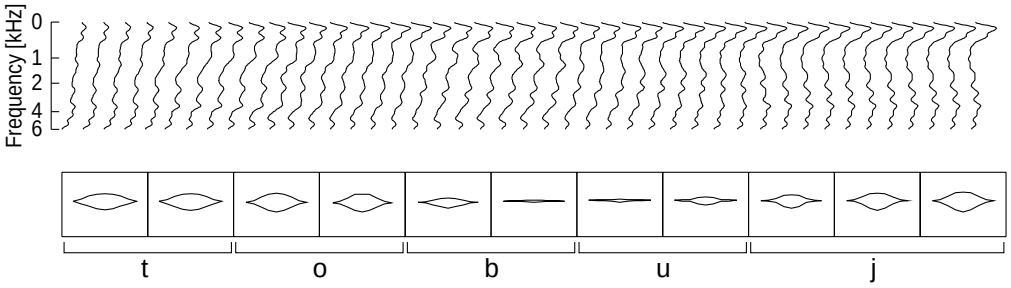
REFERENCES

[1] D.R. Hill, A. Pearce, B. Wyvill, “Animating speech: an automated approach using speech synthesized by rule,” *The Visual Computer*, **3**, pp.277–289, 1988.

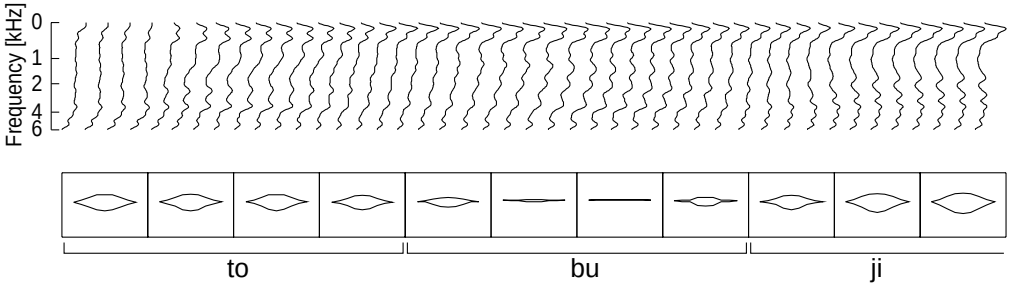
[2] K. Waters and T.M. Levergood, “DECface: an automatic lip-synchronization algorithm for synthetic faces,” *Technical Report CRL* 93/4, DEC Cambridge Research Laboratory, Cambridge, MA, Sep. 1993.



(a) Real utterance



(b) Synthetic speech generated from 4-state phoneme HMMs



(c) Synthetic speech generated from 7-state syllable HMMs

Figure 2: Spectral envelopes and lip shapes for a Japanese phrase /t-o-b-u-j-i-y-u-u/ *in testing sentence.*

[3] N.M. Brooke and S.D. Scott, "Computer graphics animations of talking faces based on stochastic models," *Proc. IEEE ISSIPNN*, pp.73–76, Apr. 1994.

[4] B. Goff and C. Benoît, "A text-to-audiovisual speech synthesizer for french," *Proc. ICSLP'96*, pp.2163–2166, Oct. 1996.

[5] J. Beskow, K. Elenius, and S. McGlashan, "Olga — A Dialogue system with an animated talking agent," *Proc. EUROSPEECH'97*, pp.1651–1654, Sep. 1997.

[6] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," *Proc. EUROSPEECH'95*, pp.757–760, Sep. 1995.

[7] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," *Proc. ICASSP-96*, I, pp.389–392, May 1996.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," *Proc. ICASSP-97*, pp.1611–1614, Apr. 1997.

[9] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *Proc. 3rd ESCA/COSCOSDA Workshop on Speech Synthesis*, pp.273–276, Nov. 1998.

[10] T. Masuko, T. Kobayashi, M.Tamura, J. Masubuchi, K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," *Proc. ICASSP'98*, pp.3745–3748, May. 1998.

[11] M. Tamura, T. Masuko, T. Kobayashi, K.Tokuda, "Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches," *Proc. AVSP'98*, pp.219–224, Dec. 1998.

[12] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *Proc. ARPA Human Language Technology Workshop*, pp.307–312, Mar. 1994.

[13] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel-generalized cepstral analysis — A unified approach to speech spectral estimation," *Proc. ICSLP'94*, pp.1043–1046, Sep. 1994.

[14] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP'92*, pp.I-137–I-140, Mar. 1992.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. EUROSPEECH'99*, Sep. 1999.