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

Takashi Masuko †, *Takao Kobayashi* †, *Masatsune Tamura* †, *Jun Masubuchi* †, *and Keiichi Tokuda* ‡

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 226-8503 Japan
‡Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan
E-mail:{masuko,tkobayas,mtamura,jmasu}@pi.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper presents a new technique for synthesizing visual speech from arbitrarily given text. The technique is based on an algorithm for parameter generation from HMM with dynamic features, which has been successfully applied to text-to-speech synthesis. In the training phase, syllable HMMs are trained with visual speech parameter sequences that represent lip movements. In the synthesis phase, a sentence HMM is constructed by concatenating syllable HMMs corresponding to the phonetic transcription for the input text. Then an optimum visual speech parameter sequence is generated from the sentence HMM in ML sense. The proposed technique can generate synchronized lip movements with speech in a unified framework. Furthermore, coarticulation is implicitly incorporated into generated mouth shapes. As a result, synthetic lip motion becomes smooth and realistic.

1. INTRODUCTION

Visual information as well as auditory information is important in perceptual recognition and understanding of spoken language. There have been proposed a variety of approaches to incorporating bimodality of speech into human-computer interaction interfaces. They include robust speech recognition by lip reading, lip synchronization in video telephony and video conferencing, dialog system with an animated talking agent, speech communication system for hearing impaired, etc. [1]-[7].

Although the use of multiple sources of information improves perception of speech generally, ambiguous or contradictory lip motion degrades intelligibility of auditory speech [8]. Therefore, lip synchronization and coarticulation are important issues in visual speech synthesis. One of the effective approaches to visual speech synthesis is the parametrically controlled polygon topology [4]-[6]. Recently statistical approaches based on Hidden Markov Models (HMMs) have also been proposed [9]-[12].

HMM-based techniques are divided into two general approaches: speech-driven and text-driven approaches. In the former approach, input auditory speech is classified into appropriate classes in a frame-by-frame basis using HMMs, then auditory information is converted into mouth shape using the statistics of the underlying HMM [9]-[11]. In the latter approach, visual speech is generated from a given phonetic specification [12].

In this paper, we present a new approach to text-to-audio-visual speech synthesis based on HMMs. We have proposed a text-to-speech synthesis system using the HMM-based speech parameter generation algorithm [13]-[15]. In this approach, the statistics of both static and dynamic features are taken into account when speech parameters are generated from HMMs. We have shown that the system can synthesize quite smooth and natural sounding speech with various voice characteristics [15],[16]. We apply this framework to visual speech synthesis. Excepting the feature parameters and synthesis speech units, the framework of the system is the same as the HMM-based audio speech synthesis system. Therefore, we can generate audio-visual speech in a unified framework simultaneously. This means that the synthesis system can generate synchronized lip movements with speech automatically. Furthermore, since generated parameter sequence reflects statistical information of both static and dynamic features of several phonemes before and after the current phonemes, coarticulation is implicitly incorporated into generated mouth shapes. As a result, synthetic lip motion becomes smooth and realistic.

The idea of using HMMs to generate speech parameters is similar to that of [12]. However, our approach is quite different from [12] which uses only static features. In addition, our approach requires neither quadratic curve fitting nor parameter smoothing in parameter generation.

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

Figure 1 illustrates a block diagram of the text-to-visual speech synthesis system based on HMMs. The synthesis system consists of two phases: training phase and synthesis phase. First, in the training phase, visual speech fea-

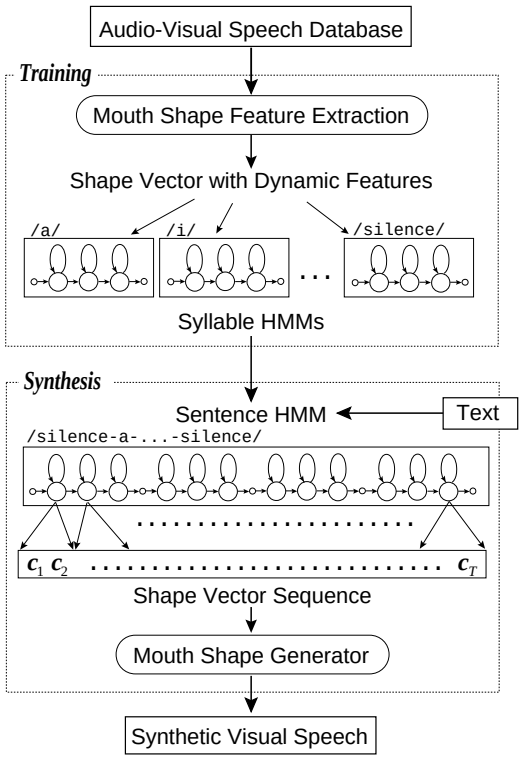

Figure 1: HMM-based text-to-visual speech synthesis system.

ture parameters, e.g. mouth positions or lip contours, are extracted from audio-visual speech database as the static features. Delta parameters, i.e., simple differences between consecutive two frames or linear regressions over a number of frames are also calculated from the extracted parameter vectors as the dynamic features. Then syllable HMMs are trained using obtained observation vectors consisting of static and dynamic features.

In the synthesis phase, arbitrary input text to be synthesized is transformed into a phonetic symbol sequence. According to the phonetic transcription, we construct a sentence HMM, which represents the whole text to be synthesized by concatenating syllable HMMs obtained in the training phase. From the sentence HMM, visual speech parameter vector sequence is generated using the ML-based parameter generation algorithm from HMM [13],[14] which will be described briefly in the next section. Finally, the generated parameter vector sequence is converted into visual speech such as lip animation and facial animation.

3. PARAMETER GENERATION FROM HMM

Let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ be a speech parameter vector sequence. We assume that the parameter vector $\boldsymbol{o}_t$ at frame t consists of the static feature vector $\boldsymbol{c}_t$, and its dynamic feature vector $\Delta\boldsymbol{c}_t$ that is, $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t']'$, where $\cdot'$ denotes matrix transpose. For example, static features are mouth positions for visual speech and cepstral coefficients for auditory speech. Dynamic features are delta coefficients given by

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L^-}^{L^+} w(\tau)\boldsymbol{c}_{t+\tau}, \tag{1}$$

where $w(\tau)$ is the weighting coefficient.

For a given continuous HMM λ with single Gaussian output distribution, we can obtain a speech parameter vector sequence $\boldsymbol{O}$ that maximizes $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ with respect to the state sequence $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$ and $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$ under the constraint of (1) [13],[14]. If the state sequence $\boldsymbol{Q}$ is explicitly known, the optimum parameter vector sequence is obtained by solving a set of linear equations.

Without dynamic features, (i.e., $\boldsymbol{o}_t = \boldsymbol{c}_t$) it is obvious that $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ is maximized when the parameter vector sequence is equals to the mean vector sequence which is determined independently of the covariances of the output distributions. On the other hand, by using delta parameters, generated parameter vector reflects both means and covariances of the output distributions of a number of states before and after the current state.

In addition, if we assume that the given HMMs is left-to-right models with no skips and explicit state duration densities $p_q(d_q)$, i.e., the probability of d_q consecutive observations in state q is known, then we can determine the state sequence explicitly [15].

4. VISUAL SPEECH SYNTHESIS EXPERIMENTS

We have developed a prototype text-to-visual speech synthesis system. The present system generates only 2-D inner lip contour animation. However, it can be applied to the synthesis of both inner and outer lip contours and 3-D mouth shapes with a slight modification on the choice of the shape parameters.

4.1. Audio-Visual Training Set

We used an audio-visual speech database consisting of 216 phonetically balanced Japanese words enunciated by a male speaker. Acoustic speech and the corresponding video images were recorded in parallel. The video images contain only mouth area and the tip of the nose. Speaker's lips and the tip of the nose were made-up in blue. NTSC video frames were digitized at 30 frames per second, 320 × 240 pixels, 24 bits per pixel. Acoustic speech was sampled at 10 kHz, 16 bits per sample. Captured images were phoneme labeled by hand according to the segmentation results of the acoustic speech.

4.2. Lip shape parameter extraction

Inner lip contours were extracted from captured images automatically and thereafter the errors were corrected by hand.

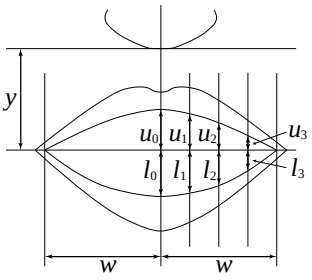

Figure 2: Mouth position parameters.

We use 10 position parameters shown in Fig. 2 to represent the lip shape: 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 contour $\{u_0, u_1, u_2, u_3\}$ and $\{l_0, l_1, l_2, l_3\}$ at 8 equally spaced points between the mouth corners. We assume here that mouth shape is symmetrical.

We composed 10 dimensional vector $\boldsymbol{c}_t = [y, w, \boldsymbol{c}_u', \boldsymbol{c}_l']'$ as the static feature vector, where $\boldsymbol{c}_u$ and $\boldsymbol{c}_l$ 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. Then delta parameters were calculated by (1) with $L^- = 1$, $L^+ = 0$, $w(0) = -w(-1) = 1$, namely, simple difference between current and preceding frames:

$$\Delta \boldsymbol{c}_t = \boldsymbol{c}_t - \boldsymbol{c}_{t-1}. \tag{2}$$

Consequently, each observation vector becomes 20 dimensional vector which consists of static and dynamic features.

4.3. Training of HMMs

Whereas we used triphone HMMs as the audio speech synthesis units [15], we use syllables as the visual speech synthesis units in the experiment. This is because it often occurs that one phoneme segment contains only one or two video frames. Thus we chose longer subword unit than phoneme. Fortunately, there is one-to-one correspondence between Japanese characters and syllables.

Table 1 shows all the syllables appeared in the database. It is known that Japanese syllables are classified into 42 visually distinct categories. However we treated syllables in Table 1 as distinct models because this enables us to synthesis audio-visual speech simultaneously using unified framework in which each audio-visual speech unit is represented by a single model.

We modeled each syllable by 3-state left-to-right model with single Gaussian diagonal output distribution and no skips. After the training of the syllable models, they were reestimated once with the embedded training version of the Baum-Welch algorithm. Finally, the training data was aligned to the models via the Viterbi algorithm to obtain the state duration densities. Each state duration density was modeled by a single Gaussian distribution.

Table 1: Syllables used in the system.

a, i, u, e, o, N, ka, ki, ku, ke, ko, sa, shi, su, se, so,
ta, chi, tsu, te, to, na, ni, nu, ne, no, ha, hi, fu, he, ho,
ma, mi, mu, me, mo, ya, yu, yo, ra, ri, ru, re, ro, wa,
ga, gi, gu, ge, go, za, ji, zu, ze, zo, da, de, do,
ba, bi, bu, be, bo, pa, pi, pu, pe, po,
kya, kyu, kyo, sha, shu, sho, cha, chu, cho, nya, nyu,
nyo, hya, hyu, hyo, mya, myu, myo, rya, ryu, ryo,
gya, gyu, gyo, ja, ju, jo, bya, byu, byo, pyu, pyo,
kka, kku, sse, tte, tto, ddo, ppa, kkyo, ssha, ccho, ppya
silence

5. RESULTS

Using the proposed visual speech synthesis system, we generated Japanese words and sentences which were not included in the training database. Fig. 4.3 shows an example of the generated lip shape images. Fig. 4.3(a) is a sequence of the inner lip contours which were extracted from real video images of a Japanese sentence /wa-N-na-u-to-ma-N-ru-i/, which means "the bases are loaded with one out" in English, uttered by the same speaker of the database. A portion around /to-ma/ of the entire sequence is shown in the figure. Fig. 4.3(b) shows a sequence of the inner lip contours generated from the synthesis system. It can be seen that synthetic lip motion is very smooth and resembles real lip motion.

Fig. 4.3(a) shows a comparison of the height of the interior opening of the lips $h = u_0 + l_0$ between real and synthetic visual speech for /wa-N-na-u-to-ma-N-ru-i/. It is again seen that the trajectory of the synthetic parameter is smooth and resembles that of the real parameter. It should be noted that no smoothing process was applied in the proposed system.

Although the synthetic lip motion looks realistic, it has been observed that little fluctuations in the portions of start and end of the mouth opening. To improve the performance, we added context dependent models to the /silence/ and /N/ models. Extra context dependent models were /silence-*/, /a-silence/, /i-silence/, /u-silence/, /e-silence/, /o-silence/, /N-silence/, and /N-{m,b,p}/, where /*/ denotes any phoneme. Fig. 4.3(b) is the result with the additional context dependent models. It can be seen that better performance is achieved in the portions of start and end of the mouth opening. In fact, the generated animation looks more realistic than that of Fig. 4.3(a).

6. CONCLUSION

We have proposed a new technique for generating visual speech from input text. The approach is based on the parameter generation algorithm from HMM with dynamic features. The effectiveness of the technique has been investigated by

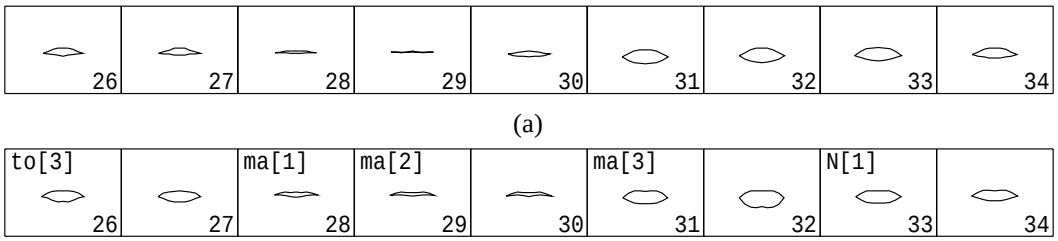

Figure 3: Comparison between real and synthetic lip shapes: (a) real, (b) synthetic. The number on the lower right corner represents the frame number. Name of syllable followed by state number of HMM is shown on the upper left corner.

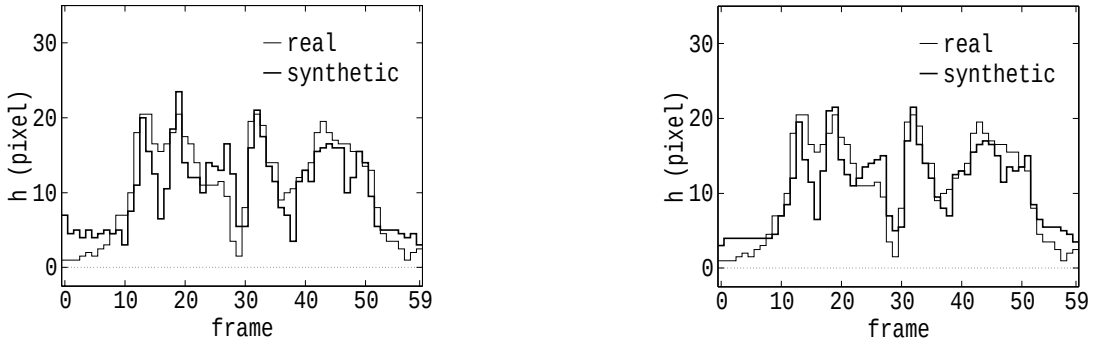

Figure 4: Trajectories of the height parameter $h = u_0 + l_0$: (a) without context dependent models, (b) with additional context dependent models.

experiments. It has been shown that generated visual speech is smooth and realistic. Although it is not presented here, we have also developed a speech-driven animated talking agent. Future work will be directed toward text-to-audio-visual speech synthesis based on HMMs.

REFERENCES

[1] D.G. Stork and M.E. Hennecke, *Speechreading by Humans and Machines*, Springer-Verlag, Berlin, 1996.

[2] F.I. Parke and K. Waters, *Computer Facial Animation*, ch.8 A K Peters, Wellesley, MA, 1996.

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

[4] M.M. Cohen and D.W. Massaro, "Modeling coarticulation in synthetic visual speech," in N.M. Thalmann and D. Thalmann, eds., *Models and Techniques in Computer Animation*, pp.139–156, Springer-Verlag, Tokyo, 1993.

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

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

[7] J. Beskow, K. Elenius, and S. McGlashan, "Olga — A Dialogue system with an animated talking agent," *Proc. Euro-Speech-97*, pp.1651–1654, Rhodes, Greece, Sep. 1997.

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

[9] A.D. Simons and S.J. Cox, "Generation of mouthshapes for a synthetic talking head," *Proc. Institute of Acoustics*, **12**, Pt.10, pp.475–482, 1990.

[10] T. Chen and R.R. Rao, "Audio-visual interaction in multimedia communication," *Proc. ICASSP-97*, Vol.I, pp.179–182, Munich, Apr. 1997.

[11] E. Yamamoto, S. Nakamura, and K. Shikano, "Speech to lip movement synthesis by HMM," *Proc. AVSP'97*, pp.137–140, Rhodes, Greece, Sep. 1997.

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

[13] K. Tokuda, T. Kobayashi and S. Imai, "Speech Parameter Generation From HMM Using Dynamic Features," *Proc. ICASSP-95*, pp.660–663, Detroit, 1995.

[14] 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. Euro-Speech-95*, pp.757–760, Madrid, Sep. 1995.

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

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