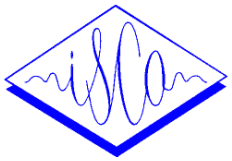

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

VISUAL SPEECH SYNTHESIS BASED ON PARAMETER GENERATION FROM HMM: SPEECH-DRIVEN AND TEXT-AND-SPEECH-DRIVEN APPROACHES

Masatsune Tamura, Takashi Masuko, Takao Kobayashi, and Keiichi Tokuda †

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
†Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan
E-mail: mtamura@ip.titech.ac.jp, masuko@ip.titech.ac.jp, tkobayas@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper describes a technique for synthesizing synchronized lip movements from auditory input speech signal. 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. Audio-visual speech unit HMMs, namely, syllable HMMs are trained with parameter vector sequences that represent both auditory and visual speech features. Input speech is recognized using the syllable HMMs and converted into a transcription and a state sequence. A sentence HMM is constructed by concatenating the syllable HMMs corresponding to the transcription for the input speech. Then an optimum visual speech parameter sequence is generated from the sentence HMM in ML sense. Since the generated parameter sequence reflects statistical information of both static and dynamic features of several phonemes before and after the current phonemes, synthetic lip motion becomes smooth and realistic. We show experimental results which demonstrate the effectiveness of the proposed technique.

1. INTRODUCTION

The use of multiple sources of information, such as auditory information and visual information, generally enhances speech perception and understanding by both humans and computers. There have been proposed various approaches to incorporating bimodality of speech into human-computer interaction interfaces. Visual speech synthesis with synchronized lip movements is one of the research topics in this area [1]-[7].

We have also proposed an alternative approach to text-to-visual speech synthesis based on hidden Markov model (HMM) [8]. In this approach, first, syllable HMMs are trained with visual speech parameter sequences that represent lip movements. Next, in the synthesis stage, 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 using the parameter generation algorithm from HMM with dynamic features [9],[10]. It was shown that the proposed HMM-based approach can generate smooth and realistic lip motion.

In this paper, we apply this framework to visual speech synthesis from auditory speech signal. We present both speech-driven and text-and-speech-driven approaches. In the training phase, audio-visual speech unit HMM, namely, syllable HMMs are trained with parameter vector sequences that represent both auditory speech features and visual speech features. In the synthesis phase of the speech-driven approach, input speech is recognized and converted into a transcription and a state sequence using the trained HMMs. In contrast, for the text-and-speech-driven approach, since the transcription of the input speech is known, only a state sequence is determined by the Viterbi algorithm. Using the recognition results, a sentence HMM is constructed by concatenating the syllable HMMs corresponding to the transcription for the input speech. Then an optimum visual speech parameter sequence is generated from the sentence HMM in the same manner of [8].

Modeling the lip movements with HMM is not a novel idea. In fact, there have been proposed many approaches using HMMs in lipreading or speechreading area [15]. Moreover, the idea of using HMMs to generate visual speech is similar to those of [11]-[14]. That is, input auditory speech is classified into appropriate classes in a frame-by-frame basis using

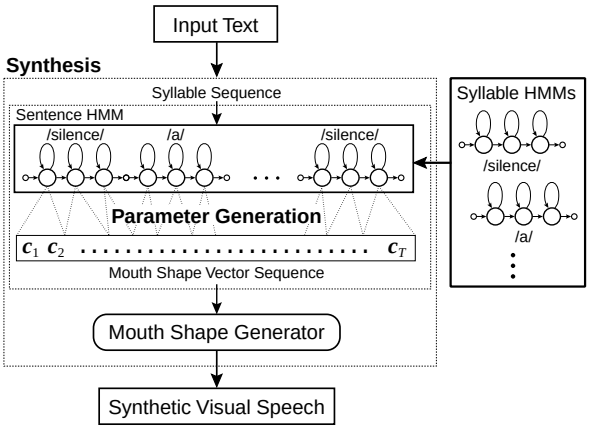

Figure 1: HMM-based text-to-visual speech synthesis system.

HMMs, then auditory information is converted into mouth shape using the statistics of the underlying HMM.

However, a distinct feature of our approach is to generate visual parameter sequence from HMM using dynamic features. 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 without requiring parameter smoothing.

2. OVERVIEW OF HMM-BASED VISUAL SPEECH SYNTHESIS

2.1. Text-to-Visual Speech Synthesis System

A block diagram of the text-to-visual speech synthesis system [8] is illustrated in Figure 1. Excepting the feature parameters, the framework of the system is the same as the auditory text-to-speech synthesis system based on HMM [16],[17]. Whereas mel-cepstral coefficients are used as the feature parameters in the auditory speech synthesis system, mouth position parameters are used in the visual speech synthesis system. Visual speech feature parameters are extracted from audio-visual speech database and an HMM is trained for each syllable using obtained visual features consisting of both 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, a sentence HMM is constructed, which represents the whole text to be synthesized, by concatenating syllable HMMs. From the sentence HMM, visual speech parameter vector sequence is generated using the ML-based parameter generation algorithm from HMM [9],[10]. Finally, the generated parameter vector sequence is converted into visual speech such as lip animation.

2.2. 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 transpose. For example, static features are mouth positions for visual speech and mel-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 distributions, 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) [9],[10]. 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 equal to the mean vector sequence which is determined independently of the covariances of the output distributions. On the other hand, by using delta parameters, the generated parameter vector reflects both means and covariances of the output distributions of a number of frames before and after the current frame.

2.3. Visual Speech Synthesis from Auditory Speech Input

We apply here the above framework to synchronized visual speech synthesis from auditory speech signal. Proposing system consists of two phases: training phase and synthesis phase.

Figure 2 shows a block diagram of the training phase. First, visual speech feature parameters, i.e., mouth positions or lip contours, are extracted from audio-visual speech database as the static features. At the same time, auditory speech feature parameters, i.e., mel-cepstral coefficients, are extracted using mel-cepstral analysis. Delta parameters for both auditory and visual parameters are also calculated from (1) using the extracted static features.

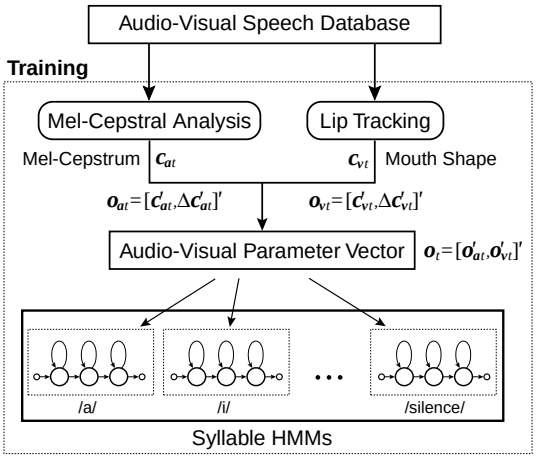

Figure 2: Training phase in HMM-based visual speech synthesis system.

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 one audio-visual observation vector $\boldsymbol{o}_t = [\boldsymbol{o}'_{\boldsymbol{a}t}, \boldsymbol{o}'_{\boldsymbol{v}t}]'$. Using this observation vectors extracted from audio-visual speech database, we train syllable HMMs. In the training of HMMs, we regard an input observation sequence to be divided into two streams, namely, auditory and visual parameter streams.

The synthesis phase is divided into two stages: the first stage is syllable-based recognition of auditory input speech and the second stage is synthesis of visual speech. A block diagram of the synthesis phase is shown in Figure 3.

In the recognition stage, we obtain mel-cepstral coefficients from auditory input speech using a mel-cepstral analysis. Then speech recognition is performed based on syllable HMMs which are obtained in the training phase. In this stage, we use the auditory parameter stream only on likelihood calculation for HMMs. As a result, we obtain the syllable sequence and state durations for the input speech.

The synthesis stage is almost the same as the text-to visual speech synthesis system described in **2.1**. According to the obtained syllable sequence, we construct a sentence HMM, which represents the whole text of the input speech, by concatenating syllable HMMs obtained in the training phase. From the sentence HMM with the information on state durations, a visual speech parameter vector sequence is generated using the ML-based parameter generation algorithm from HMM described in the previous section. It is noted that we use the visual parameter stream only in this stage.

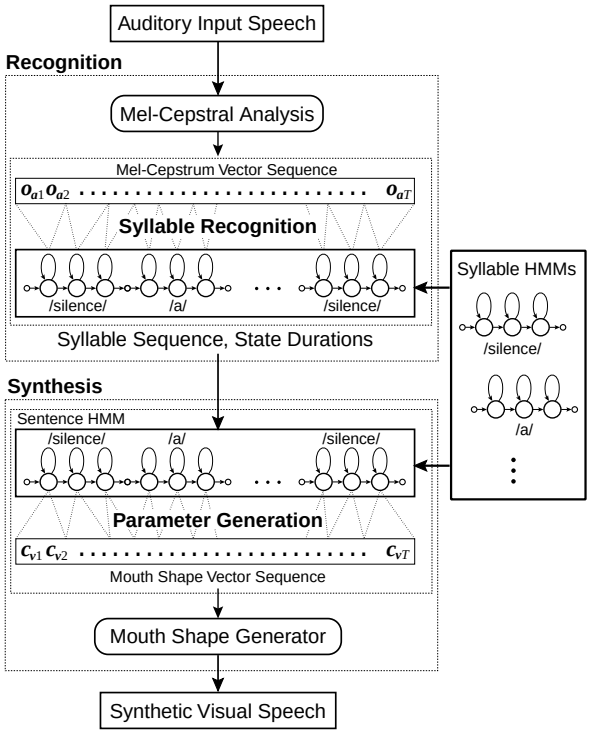

Figure 3: Synthesis phase in HMM-based visual speech synthesis system.

2.4. Text-and-Speech-Driven Visual Speech Synthesis

When both input speech and corresponding text information are available, the recognition stage becomes much simpler. In this case, given text is transformed into a syllabic symbol sequence. According to the syllabic transcription, a sentence HMM is constructed by concatenating syllable HMMs. Then only thing to do here is obtaining the state durations by the Viterbi algorithm. The synthesis stage is the same as the speech-driven case described in the previous section.

3. IMPLEMENTAION OF VISUAL SPEECH SYNTHESIS SYSTEM

3.1. Audio-Visual Training Set

We used an audio-visual speech database consisting of 216 phonetically balanced Japanese words 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 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, 640 $\times$ 480 pixels, 24 bits per pixel. Further each frame was decomposed into two interlaced

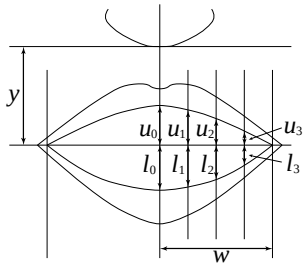

Figure 4: Mouth position parameters.

fields. As a result, we obtained 60 lip shape images per second. 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 Parameter

Since the present visual speech synthesis system that we have developed can generates only 2-D inner lip contour animation, we use 10 position parameters shown in Figure 4 to represent the lip shape. 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 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. Here we assume that mouth shape is symmetrical. These parameters were extracted from captured images automatically and thereafter the errors were corrected by hand.

We used a 10-dimensional vector $\boldsymbol{c_v} = [y, w, \boldsymbol{c}'_{\boldsymbol{v}u}, \boldsymbol{c}'_{\boldsymbol{v}l}]'$ as the static feature vector, where $\boldsymbol{c_{vu}}$ and $\boldsymbol{c_{vl}}$ 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 $L^- = 1, L^+ = 1, \{w(-1), w(0), w(1)\} = \{-1/2, 0, 1/2\}$. Consequently, each visual feature vector becomes a 20-dimensional vector which consists of static and dynamic features.

We used mel-cepstral coefficients as auditory speech feature. The mel-cepstral coefficients were obtained by a mel-cepstral analysis technique [18],[19] on each 25ms frame of speech sampled at 12kHz with a Blackman window every 16.7ms. The auditory speech frames were synchronized with lip shape images. The static feature vector $\boldsymbol{c_a}$ consists of 18 mel-cepstral coefficients in which the 0th coefficient is not included. Then dynamic features were calculated by (1) with $L^- = 1$, $L^+ = 1$, $\{w(-1), w(0), w(1)\} = \{-1/2, 0, 1/2\}$. Delta log energy was also used as a dynamic feature. Thus, each auditory feature vector becomes a 37 dimensional vector.

3.3. Models, Training and Synthesis

Whereas we used triphone HMMs as the auditory speech synthesis units [16], we used syllables as the visual speech synthesis units. Since one phoneme segment often contains only one or two video frames in a rate of 60 fps, we chose longer subword unit than phoneme. Fortunately, there is one-to-one correspondence between Japanese syllabic symbols and Japanese CV syllables.

There exists a total of 112 syllables including /silence/ in the database [8]. Although it is possible to classify Japanese syllables into fewer visually distinct categories, we treated all syllables appeared in the database as distinct models. This is because we modeled auditory and visual features of each syllable in a single HMM simultaneously. Furthermore, to improve the quality of synthetic visual speech, we added context dependent models to the /silence/ and /N/ models [8]. Extra context dependent models are /silence-$*$/, /a-silence/, /i-silence/, /u-silence/, /e-silence/, /o-silence/, /N-silence/, /N-{m,b,p}/, and /N-{$*$−{m,b,p}}, where /$*$/ denotes any phoneme. Consequently, we used 119 syllable HMMs in total.

For each syllable, we trained a 4-state left-to-right model with single Gaussian diagonal output distributions and no skips. After the training of the syllable models, they were reestimated once with the embedded training version of the Baum-Welch algorithm.

In the recognition stage, we realized a simple syllable-based continuous speech recognition system with no language models. In addition, since we analyzed and modeled audio-visual speech with a frame rate of 60 fps, the generated visual parameter sequence has the same frame rate. To synthesize lip animation with 30 fps, we downsampled the generated parameter sequences by a factor of 2.

4. EXPERIMENTS AND RESULTS

Using the proposed visual speech synthesis system, we generated Japanese words and sentences which were not included in the training data. Since the audio-visual database that we used was small and thus available training data was limited, we chose one word from the database arbitrarily and used it as a testing word. Then the remainder of the words, namely 215 words out of the 216-word set, were used as the training data.

Figure 5 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 a Japanese word /ni-se-mo-no/, which means “an imitation” in English.

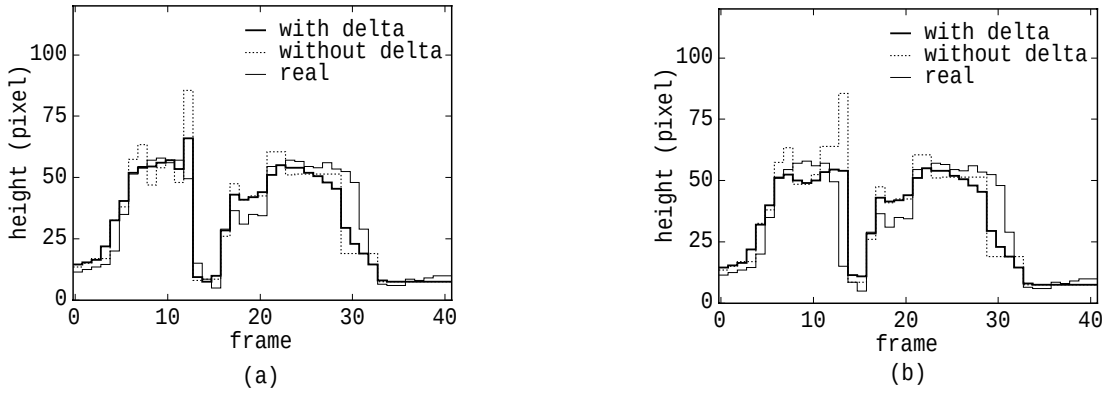

Figure 5: Trajectories of the height parameter $h = u_0 + l_0$ for an utterance /ni-se-mo-no/: (a) speech-driven approach, (b) text-and-speech-driven approach.

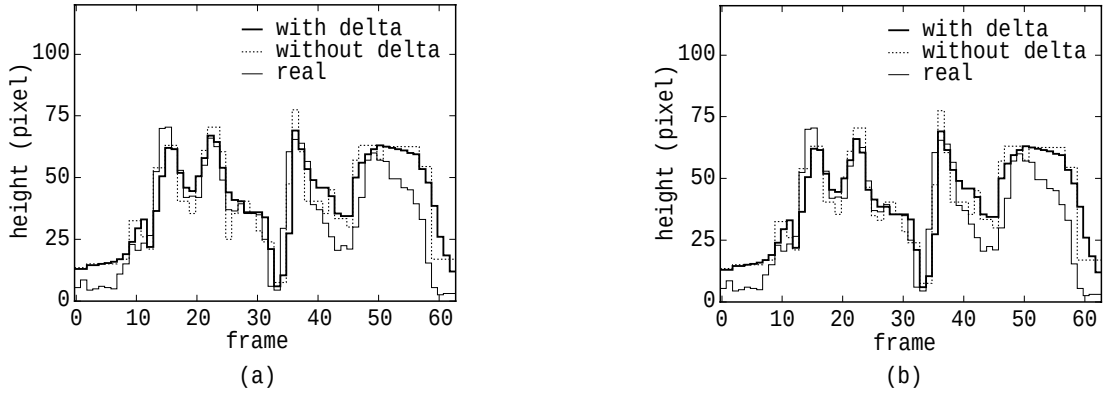

Figure 6: Trajectories of the height parameter $h = u_0 + l_0$ for for an utterance /wa-N-na-u-to-ma-N-ru-i/: (a) speech-driven approach, (b) text-and-speech-driven approach.

Figure 5(a) shows the result of speech-driven approach. The thick line is the trajectory of h for the synthetic speech with dynamic features and thin line is that of the real speech. In this case, since we performed syllable-based continuous speech recognition rather than word recognition, there can exist substitution, deletion, and insertion errors in the resultant transcriptions. In fact, the obtained transcription became /ni-chi-ya-mo-no/, in which the syllable /se/ was recognized as /chi-ya/ However, the trajectory of the synthetic parameter is very smooth and resembles that of the real parameter. It should be noted that no additional smoothing process was applied in the proposed system. In the figure, the result without using the dynamic features is also shown by the dotted line. As we described in **2.2**, the generated sequence becomes the mean vector sequence when the dynamic features are not used. As a result, the trajectory has rapid changes and this causes jerky motion in lip animation.

Figure 5(b) shows the result of text-and-speech-driven approach. In this case, the correct transcription for the input speech is known and thus the recognition error does not arise. It is again seen that the trajectory of the synthetic parameter is smooth and resembles that of the real parameter.

Figure 6 shows the result for a 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. The trajectory of a portion around /ru-i/ is slightly different from that of the real speech. However, it was observed that synthetic lip motion looks still smooth and natural.

To evaluate the quality of the generated visual speech, we conducted DMOS tests. Testing utterances consisted of five words and one sentence which were not included in the training data. Subjects were six males and one female.

Figure 7 shows the results of subjective tests. The speech-driven approach achieved almost the same performance as the text-and-speech-driven approach. It is also seen that the scores for the proposed framework

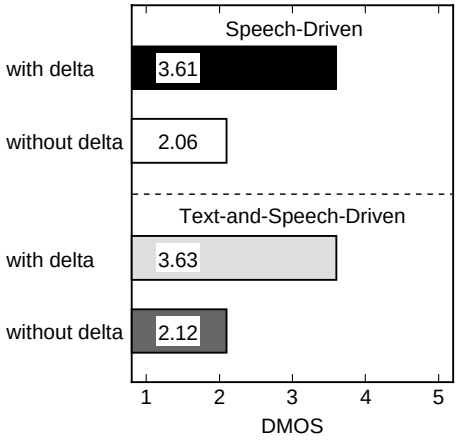

Figure 7: DMOS score for synthetic visual speech.

are much higher than the cases that the dynamic features were not used. In fact, the generated animation looks smooth and realistic in the propose approaches. In contrast, without using dynamic features, lip motion looks jerky.

5. CONCLUSION

We have proposed a technique for generating visual speech from auditory input speech. The approach is based on the parameter generation algorithm from HMM with dynamic features. The effectiveness of the technique has been investigated by experiments. It has been shown that generated visual speech is smooth and realistic. Although the current visual speech synthesis system generate simple inner lip contour animation only, 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. Future work will be directed toward text-to-audio-visual speech synthesis based on HMMs.

6. ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Encouragement of Young Scientists, 0780226, 1998.

7. REFERENCES

1. D.R. Hill, A. Pearce, and B. Wyvill, "Animating speech: an automated approach using speech synthesised by rule," *The Visual Computer*, **3**, pp.277–289, 1988.
2. F.I. Parke and K. Waters, *Computer Facial Animation*, ch.8 A K Peters, Wellesley, MA, 1996.
3. F.I. Parke, "Parameterized models for facial animation," *IEEE Computer Graphics and Applications*, **2**, 9, pp.61–68, Nov. 1982.
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. J. Beskow, "Rule-based visual speech synthesis," *Proc. EUROSPEECH'95*, pp.299–302, Madrid, Sep. 1995.
7. B. Goff and C. Benoît, "A text-to-audiovisual speech synthesizer for french," *Proc. ICSLP'96*, pp.2163–2166, Philadelphia, Oct. 1996.
8. 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, Seattle, May. 1998.
9. K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP'95*, pp.660–663, Detroit, 1995.
10. 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, Madrid, Sep. 1995.
11. 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.
12. N.M. Brooke and S.D. Scott, "Computer graphics animations of talking faces based on stochastic models," *Proc. 1994 Int. Symposium Speech, Image Processing and Neural Networks*, pp.73–76, Hong Kong, Apr. 1994.
13. T. Chen and R.R. Rao, "Audio-visual interaction in multimedia communication", *Proc. ICASSP'97*, Vol.I, pp.179–182, Apr. 1997.
14. E. Yamamoto, S. Nakamura, and K. Shikano, "Speech to lip movement synthesis by HMM," *Proc. AVSP'97*, pp.137–140, Rhodes, Greece, Sep. 1997.
15. D.G. Stork and M.E. Hennecke, *Speechreading by Humans and Machines*, Springer-Verlag, Berlin, 1996.
16. 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.
17. 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.
18. 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, Yokohama, Japan, Sep. 1994.
19. 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, San Francisco, Mar. 1992.