SPEECH-TO-LIP MOVEMENT SYNTHESIS BASED ON THE EM ALGORITHM USING AUDIO-VISUAL HMMS

Eli Yamamoto, Satoshi Nakamura, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science & Technology
8916-5 Takayama, Ikoma, Nara 630-01, JAPAN
Tel:+81-743-72-5287, Fax:+81-743-72-5289
e-mail:{eli-y,nakamura,shikano}@is.aist-nara.ac.jp

ABSTRACT

This paper proposes a method to re-estimate output visual parameters for speech-to-lip movement synthesis using audio-visual Hidden Markov Models(HMMs) under the Expectation-Maximization(EM) algorithm. In the conventional methods for speech-to-lip movement synthesis, there is a synthesis method estimating a visual parameter sequence through the Viterbi alignment of an input acoustic speech signal using audio HMMs. The HMM-Viterbi method produces the output visual parameters per HMM state specified by the decoded HMM states. However, the HMM-Viterbi method involves a substantial problem, which is caused by the deterministic decoding process to assign a single HMM state for an input audio frame. The deterministic process may output incorrect visual parameters due to incorrect HMM state alignment. The proposed method avoids the deterministic decoding process by the non-deterministic visual parameter estimation by the EM algorithm. The proposed method repeatedly estimates the visual parameter sequence while maximizing the likelihood of the audio-visual observation sequence using audio-visual HMMs. The objective evaluation shows that the proposed method is more effective than the HMM-Viterbi method especially for the bilabial consonants.

1. INTRODUCTION

Lip movement synthesis can play a significant role in human-machine communication. If lip movements are synthesized well enough to do lip-reading, hearing impaired people may be able to estimate auditory information from the visualized computer agent.

This paper investigates synthesis methods for realizing human-like lip movements by mapping from acoustic speech signals to visual parameter sequences. The lip movement synthesis from acoustic speech signals also permits lip-synchronization between input acoustic speech signals and a synthesized lip image sequences. Lip-synchronization is one of the techniques for human-like visualized computer agents in interactive communication systems.

Mapping algorithms from acoustic speech signals to lip movement sequences have been reported based on: Vector Quantization(VQ)[1], Artificial Neural Networks[2] and Gaussian Mixtures[3]. These methods are based on frame-by-frame(or frames-by-frames) mapping from acoustic speech parameters to visual parameters. These mapping algorithms have two major problems: 1) frame-by-frame mappings are fundamentally many-to-many, and 2) extensive training sets are required to account for context information.

A different approach utilizes speech recognition technique, such as phonetic segmentation[4] and Hidden Markov Model (HMM)[5][6][7][8]. These methods convert acoustic speech signals into visual parameter sequences based on information such as a phonetic segment, a word, a phoneme, an acoustic event and so on. These methods have the advantage that explicit phonetic information is available to handle coarticulation effects caused by surrounding phoneme contexts.

We have shown a synthesis method based on the Viterbi decoding algorithm using audio phoneme HMMs(We call the method the HMM-Viterbi method in the following) is more efficient than the VQ method[8]. However, the HMM-Viterbi method converts an audio parameter sequence to a visual parameter sequence through a deterministic single HMM state sequence. The deterministic process involves a substantial problem, which may give rise to an incorrect visual parameter sequence out of an incorrect HMM state sequence. For example, if bilabial consonant is decoded to other categories classified by place of articulation, the synthesized lip movement would generate a sense of incompatibility to a viewer. To solve the problem, we extend the HMM-Viterbi method with an un-deterministic process.

This paper presents a new method to estimate a visual parameter sequence from an input acoustic speech signal by applying the Expectation-Maximization algorithm(HMM-EM). The HMM-EM method repeatedly estimates the visual parameter sequence while maximizing the likelihood of the audio-visual observation sequence using audio-visual HMMs. The re-estimating operation is regarded as the auto-association of a complete pattern out of an incomplete pattern for time series. In experiments, the HMM-EM method is compared to the HMM-Viterbi method.

2. HMM-VITERBI SYNTHESIS METHOD

The first method is a baseline, HMM-Viterbi[8], which is composed of two processes, such as a decoding process which converts an acoustic speech signal to a most likely

HMM state sequence by the Viterbi algorithm and a look-up table process which converts an HMM state to corresponding visual parameters per frame. The synthesis algorithm of the HMM-Viterbi method is explained as the following with Figure 1.

step 1 Analyze and convert an input acoustic speech signal to an audio parameter sequence.
step 2 Align the audio parameter sequence into an HMM state sequence using the Viterbi alignment.
step 3 Retrieve the output visual parameter sequence associated with the HMM state sequence.
step 4 Synthesize a lip image sequence from the retrieved visual parameter sequence to visualize the lip movement.

The visual parameters per an audio HMM state are trained by taking average for all visual parameters assigned to the same audio HMM state.

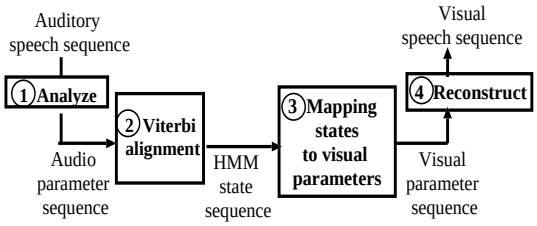

Figure 1: Algorithm of the HMM-Viterbi method

3. HMM-EM SYNTHESIS METHOD

3.1. Estimating Visual Parameters using the EM Algorithm

The quality of visual parameters synthesized by the HMM-Viterbi method depends on the accuracy of the Viterbi alignment. The incorrect HMM states assigned by the Viterbi alignment may produce wrong visual parameters. The proposed HMM-EM method does not depend on the deterministic Viterbi alignment.

The proposed method re-estimates the visual parameter sequence $\hat{O}^V = \hat{o}^V(1), \hat{o}^V(2) \cdots \hat{o}^V(T)$ for the given audio parameter sequence $O^A = o^A(1), o^A(2) \cdots o^A(T)$ by the EM algorithm using audio-visual HMMs. The observation sequence is the sequence of a vector consisting of n-dimension audio parameters and m-dimension visual parameters, $O^A \oplus O^V$. Although the visual parameter sequence does not exist initially, the required visual parameters are synthesized iteratively from initial values by re-estimation procedure maximizing the likelihood of the audio-visual observation sequence using audio-visual HMMs. The re-estimation of a visual parameter sequence is formulated as

$$\hat{O}^V = \arg\max_{O^V} P(O^A \oplus O^V | O^A, M^{AV}), \tag{1}$$

where $\hat{O}^V$ means an estimated visual parameter sequence.

The likelihood of the proposed method is derived by considering all HMM states at a time. To incorporate all states of all HMMs, the likelihood of the audio-visual observation sequence can be defined as following.

$$\begin{aligned} &\sum_{Q(all\ k)} P(M_k^{AV}) P(O^A \oplus O^V | Q, M_k^{AV}) \\ &= \sum_{Q(all\ k)} P(M_k^{AV}) \pi_{q_1}(M_k^{AV}) \\ &\times \prod_{t=1}^{T} a_{q_{t-1}q_t}(M_k^{AV}) b_{q_t}(o^A(t) \oplus o^V(t) | M_k^{AV}), \end{aligned} \tag{2}$$

where M_k^{AV} is the k-th audio-visual HMM, and $P(M_k^{AV})$ is the model probability. $\pi_j(M_k^{AV})$, $a_{ij}(M_k^{AV})$ and $b_j(o^A(t) \oplus o^V(t) | M_k^{AV})$ are the joint initial state probability, the joint transition probability and the joint output probability of audio-visual parameters, respectively. Q represents a state sequence. The summation of $Q(all\ k)$ considers all models M_k^{AV} at a time. In the next section, derivation of the re-estimation formula of visual parameter is described.

3.2. Algorithm of Visual Parameter Estimation

The re-estimation formula is defined to maximize the auxiliary function $A(\hat{O}^V | O^V)$ over an estimated visual parameter sequence $\hat{O}^V$.

$$\begin{aligned} &A(\hat{O}^V | O^V) \\ &= \sum_{Q(all\ k)} P(M_k^{AV}) P(O^A \oplus O^V | Q, M_k^{AV}) \\ &\times log P(M_k^{AV}) P(O^A \oplus \hat{O}^V | Q, M_k^{AV}) \end{aligned} \tag{3}$$

In the EM algorithm, the maximization of the auxiliary function is equivalent to increasing likelihood of an observation sequence. The re-estimation formula of visual parameter at time t is derived by differentiating the auxiliary function by the m-th visual parameter $\hat{o}_m^V(t)$. Let that the output probability density function is mixed Gaussian distributions with mean vector with $\mu_n^A(M_k^{AV}, j), \mu_m^V(M_k^{AV}, j)$ and covariance matrix $\Sigma(M_k^{AV}, j)$ with its components, $\sigma_{nn'}^{A,A}(M_k^{AV}, j)$, $\sigma_{mm'}^{V,V}(M_k^{AV}, j)$, $\sigma_{nm}^{A,V}(M_k^{AV}, j)$. n, m are the index of audio parameter dimension, and of visual parameter dimension. $|\Sigma(M_k^{AV}, j)|$ is the determinant of $\Sigma(M_k^{AV}, j)$. The re-estimation formula is derived as follows:

$$\begin{aligned} \hat{o}_m^V(t) = &\frac{\sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \frac{1}{|\Sigma(M_k^{AV}, j)|}}{\sum_k \sum_j P(M_k^{AV}) \gamma(t; M_k^{AV}, j) \frac{\Sigma'^{V,V}_{mm}(M_k^{AV}, j)}{|\Sigma(M_k^{AV}, j)|}} \\ &\times \left(\begin{array}{l} \mu_m^V(M_k^{AV}, j) \Sigma'^{V,V}_{mm}(M_k^{AV}, j) \\ -\sum_n (o_n^A(t) - \mu_n^A(M_k^{AV}, j)) \Sigma'^{A,V}_{nm}(M_k^{AV}, j) \end{array} \right) \end{aligned} \tag{4}$$

where $\gamma(t; M_k^{AV}, j)$ is the state occupation probability in state j of M_k^{AV} at time t. $\Sigma'(M_k^{AV}, j)$ means the adjoint

of $\Sigma(M_k^{AV}, j)$. Thus $\Sigma'^{A,V}_{nm}(M_k^{AV}, j)$ means the component of the n-th audio parameter and the m-th visual parameter. The formula (4) is derived under a constraint that the covariance $\sigma^{A,A}_{nn'}(M_k^{AV}, j) = 0$ at $n \neq n'$ and $\sigma^{V,V}_{mm'}(M_k^{AV}, j) = 0$ at $m \neq m'$. Furthermore, the re-estimation formula is simplified as follows if the covariance matrix is diagonal.

$$\hat{o}_m^V(t) = \frac{\sum_k \sum_j P(M_k^{AV})\gamma(t; M_k^{AV}, j) \frac{\mu_m^V(M_k^{AV}, j)}{\sigma^{V,V}_{mm}(M_k^{AV}, j)}}{\sum_k \sum_j P(M_k^{AV})\gamma(t; M_k^{AV}, j) \frac{1}{\sigma^{V,V}_{mm}(M_k^{AV}, j)}}. \quad (5)$$

The algorithm for the visual parameters re-estimating can be summarized as the following with Figure 2:

step 1 Set the initial value for visual parameter $o_m^V(t)$.

step 2 Calculate $\gamma(t; M_k^{AV}, j)$ for all frames under the Forward-Backward algorithm(EM algorithm for HMM). Estimate $\hat{o}_m^V(t)$ using formula (5) at each frame.

step 2' If a convergence condition is satisfied, go to the next step, otherwise return to step 2.

step 3 Synthesize a lip image sequence from the retrieved visual parameter sequence to visualize the lip movement.

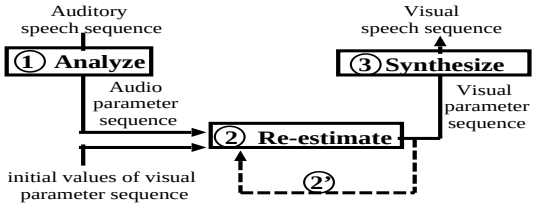

Figure 2: Algorithm of the HMM-EM method

4. LIP PARAMETER SYNTHESIS EXPERIMENTS

4.1. Experiment Condition

Speech and 3D lip position data for a female speaker of Japanese were recorded at 125Hz using the OPTOTRAK, 3D position sensoring system. These 3D positions were transformed into the visual parameters height(X), width(Y) of the outer lip contour and protrusion(Z) based on five parameters of the 3D lip model[9]. The audio parameter has 33 dimensions of 16-order mel-cepstral coefficients, their delta coefficients and the delta log power.

Fifty-four phonemes and two pauses were modeled by for audio HMMs of the HMM-Viterbi method and audio-visual HMMs of the HMM-EM method. The pause models are prepared separately for the word beginning and the word ending. Triphone HMMs are not adopted, because the triphone HMMs requires huge amounts of time synchronous training data. Each audio HMM and audio-visual HMM has left-to-right structure with 3 states, where an output probability on each state has 256 tied-mixture Gaussian distributions. HMMs are trained by the audio or audio-visual synchronous database composed by 326 Japanese words, which consists of phonetically balanced words. The other 100 words are prepared for testing.

Table 1: Compared Synthesis Methods

Method	Training params#		Synthesis params#		Initial visual parameters
	A	V	A	V	
HMM-V	33	—	33	—	—
HMM-EM-1	33	3	33	3	HMM-V
HMM-EM-2	33	6	33	3	HMM-V
HMM-EM-3	33	9	33	3	HMM-V
HMM-EM-4	33	9	33	3	Pause

The measure to evaluate synthesized lip movements is Euclidian error distance E between the synthesized visual parameters and the original parameters extracted from human movements.

In the HMM-EM method, the state occupation probabilities $\gamma(t; M_k^{AV}, j)$ are updated after re-estimation of all visual parameters for the utterance.

4.2. Compared Synthesis Methods

To verify the effect of the HMM-EM method, the five synthesis methods on Table 1 are compared in the experiment. The HMM-EM method can be implemented by various conditions. We tried to make the number of parameter vectors fluctuate taking account of a dependency on the quality of HMMs. In the HMM-EM-2 method, the visual parameter vector consists of 6 parameters of 3 visual parameters and their time differential parameters. Likewise the HMM-EM-3 method contains the acceleration part of visual parameters in addition to parameters of the HMM-EM-2 method. Note that in all HMM-EM methods the number of the tied-mixture distribution is fixed at 256 as well as that of HMM-Viterbi method. As for the initial values for visual parameters, the HMM-EM-1,2,3 methods use the visual parameters synthesized by the HMM-Viterbi method and the HMM-EM-4 method uses a visual parameters of the lip closure shape at pause.

4.3. Results

The results of objective evaluation of the five methods are shown in Table 2. Each column in Table 2 indicates the error distances averaged by all frames or correctly decoded, incorrectly decoded, and incorrectly decoded /p//b//m/ frames at the HMM-Viterbi method. In the errors averaged by all frames, the HMM-EM-3 method reduces the error distance by 1% against the HMM-Viterbi method. The HMM-EM-4 method gives a large error due to the flat start

Table 2: Error distances of synthesis methods

	E cm			
	All	All Correct	All Incorrect	/p//b//m/ Incorrect
HMM-V	1.066	1.062	1.075	1.701
HMM-EM-1	1.093	1.106	1.051	1.370
HMM-EM-2	1.063	1.077	1.021	1.392
HMM-EM-3	1.052	1.072	0.989	1.254
HMM-EM-4	1.207	1.231	1.134	1.061

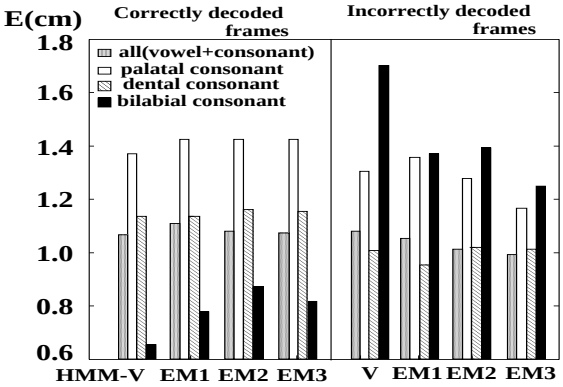

Figure 3: Errors by consonant category

of the lip closure.

We investigate errors of the HMM-EM-1,2,3 methods under incorrectly decoded frames at the HMM-Viterbi method. Their errors are compared with detailed three categories of palatal, dental and bilabial consonants in Figure 3. The HMM-Viterbi method shows the large error at the bilabial consonant category of incorrectly decoded frames. It is known that the bilabial consonants /p//b//m/ are quite sensitive for audience. For these phonemes, the errors of the HMM-EM-3 method is reduced by 26% compared to errors of the HMM-Viterbi method at incorrectly decoded frames.

An effect of the HMM-EM method is illustrated at Figure 4 and Figure 5. The figures show a test Japanese word /kuchibiru/. The horizontal axis means the number of frames corresponding to time. The vertical axis means visual parameters. The solid lines on the figures are the synthesized visual parameters, and the dotted lines are visual parameters by the original recorded human movement. The two vertical lines show the beginning and ending times of the utterance. The synthesized height visual parameter of the HMM-Viterbi method does not form the valley of the lip closure of /b/ because of the incorrectly Viterbi alignment at Figure 4. However the HMM-EM-3 method of Figure 5 shows the correct articulation.

5. CONCLUSION

This paper proposes a new method to re-estimate visual parameters from acoustic speech signals using audio-visual HMMs based on the EM algorithm. In the experiment, the HMM-EM method shows error reduction compared to the HMM-Viterbi method at incorrectly decoded bilabial consonants.

On the other hand in the correctly decoded frames by the HMM-Viterbi method, the HMM-EM method blurred the correct articulation of the HMM-Viterbi method. The influence of errors due to blur needs to be evaluated by the subjective test of the visualized lip images. In future works, the re-estimation formula with covariances between audio and visual parameters will be implemented. The correlation between audio and visual parameters will give more natural synthetic visual parameters for an input speech.

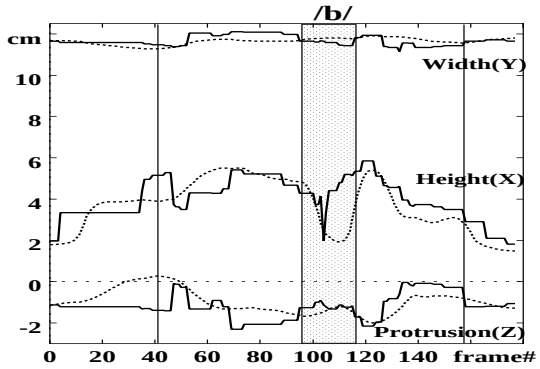

Figure 4: Visual parameters synthesized by HMM-Viterbi method

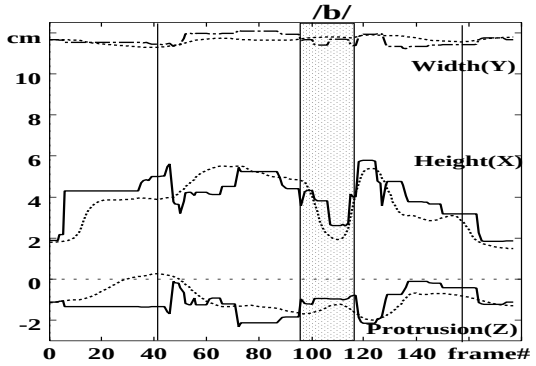

Figure 5: Visual parameters synthesized by HMM-EM-3 method

6. REFERENCES

1. Morishima, S. and Harashima, H.: A Media Conversion from Speech to Facial mage for Intelligent Man-Machine Interface, *IEEE Journal on sel. areas in Communications*, Vol. 9, No. 4, pp. 594–600 (1991).
2. Lavagetto, F.: Converting Speech into Lip Movements: A Multimedia Telephone or Hard of Hearing People, *IEEE Trans. on Rehabilitation Engineering*, Vol. 3, No. 1, pp. 90–102 (1995).
3. Rao, R.R. and Chen,T., "Cross-Modal Prediction in Audio-Visual Communication", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.4, pp.2056-2059(1996)
4. Goldenthal,W., Waters,K., Van Thong,J.M. and Glickman,O. "Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe!", Eurospeech'97 Proceedings,Vol.4, pp.1995-1998 1997)
5. Simons, A. and Cox, S.: Generation of Mouthshape for a Synthetic Talking ead, *Proc. of the Institute of Acoustics*, Vol. 12, No. 10 (1990).
6. Chou, W. and Chen, H.: Speech Recognition for Image Animation and Coding, *ICASSP 95*, pp. 2253–2256 (1995).
7. Chen, T. and Rao, R.: Audio-Visual Interaction in Multimedia Communication, *ICASSP 97*, pp. 179–182 (1997).
8. Yamamoto, E., Nakamura, S. and Shikano, K.: Speech-to-Lip Movement Synthesis by HMM, *ESCA Workshop of Audio Visual Speech Processing*, pp. 37–140 (1997).
9. Guiard-Marigny, T., Adjoudani, T. and Benoit, C.: 3D Models of the Lips and Jaw for Visual Speech Synthesis, *in "Progress in Speech Synthesis", J. van Santen et al.,Eds, Springer-Verlag* (1996).