# Lip Movement Synthesis from Speech Based on Hidden Markov Models

Eli Yamamoto Satoshi Nakamura Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-01, JAPAN

## Abstract

*Speech intelligibility can be improved by adding lip image and facial image to speech signal. Thus the lip image synthesis plays a important role to realize a natural human-like face of computer agents. Moreover the synthesized lip movement images can compensate lack of auditory information for hearing impaired people. We propose a novel lip movement synthesis method based on mapping from input speech based on Hidden Markov Model (HMM). This paper compares the HMM-based method and a conventional method using vector quantization (VQ). In the experiment, error and time differential error between synthesized lip movement images and original ones are used for evaluation. The result shows that the error of the HMM based method is 8.7% smaller than that of the VQ-based method. Moreover, the HMM-based method reduces time differential error by 32% than the VQ's. The result also shows that the errors are mostly caused by phoneme /h/ and /Q/. Since lip shapes of those phonemes are strongly dependent on succeeding phoneme, the context dependent synthesis on the HMM-based method is applied to reduce the error. The improved HMM-based method realizes reduction of the error(differential error) by 10.5%(11%) compared with the original HMM-based method.*

## 1 Introduction

In human machine communication, it is quite important to realize natural and friendly interface. Speech recognition or computer lipreading has been developed as input means of communication. It is also important to provide natural and friendly interface as output means. While speech synthesis has been studied by many researchers, synthesis of lip movements has not been interested. However, the lip movement synthesis can take the significant role in human-machine communication. It has been reported that speech intelligibility in noisy environments is improved by adding information of face to acoustic signals degraded by noise[8]. Moreover, it could be a useful tool for hearing impaired people to compensate lack of auditory information.

There are two approaches for the lip movement synthesis, synthesis from text and synthesis from speech. This paper focuses on the synthesis from speech. The lip movement synthesis requires many information including phoneme, coarticulation, and duration. Just like text-to-speech synthesis, it is generally difficult to control all of these parameters only by text information. On the other hand, speech includes the information of phoneme, coarticulation, duration. Thus the lip movement synthesis from speech seems to be more promising than synthesis from text.

As a mapping algorithm from speech to lip movements, VQ based method and Neural-Network based method have been reported in [6][7][5][3]. These methods are based on frame-by-frame mapping from speech parameters to image parameters. The frame-by-frame mapping has problems such that 1) the frame-by-frame mapping is fundamentally many-to-many mapping, and 2) It is very difficult to take account of phoneme context in the frame-by-frame mapping from a viewpoint of training.

This paper proposes a new method based on HMM(Hidden Markov Model) that takes account of intra phoneme HMM contexts into mapping. The many-to-many mapping problem can be also reduced by the proposed method. The experiment of lip movement synthesis for Japanese words ensures that the proposed method is more accurate and natural than the conventional method. Moreover, the proposed method is extended to the case of inter phoneme HMM contexts. The effectiveness of the extended HMM-based method is clarified in the experiments.

0-8186-8344-9/98 $10.00 © 1998 IEEE

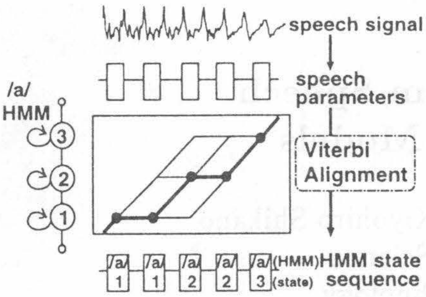

Figure 1: Viterbi alignments

# 2 HMM-based Lip Movement Synthesis

## 2.1 Synthesis Algorithm

The proposed HMM-based method is based on mapping from speech to lip movement images through HMM states along Viterbi alignment. The Viterbi alignment associates optimal HMM state to the input speech frame by maximizing likelihood. Fig.1 shows the Viterbi alignment between input frames and HMM states. The basic idea of the proposed HMM-based method is mapping from HMM states to lip image parameters. Once speech HMMs are trained, the input speech can be transcribed into HMM state sequence by Viterbi alignment. The lip movement images are obtained by concatenating the image parameters associated with HMM state sequence. The works applying Viterbi alignment to lip movement synthesis are also reported by [2][1]. The paper[2] doesn't include estimation of lip parameters. The paper[1] introduces a similar idea independently, however, it doesn't consider inter phoneme HMM context dependency. Fig.2 and Fig.3 show a block diagrams of the HMM-based method. The training and synthesis algorithms are as follows.

### Training Algorithm

1. Prepare and parameterize audio and visual synchronous database.
2. Train HMMs using the training speech database.
3. Align speech database into HMM state sequence according to the Viterbi alignment.
4. Take an average of image parameters of the frames associated with the same HMM state.

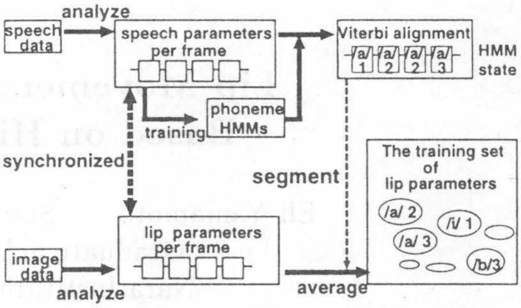

Figure 2: Schematic Diagram of Lip Parameter Training by the HMM-based method

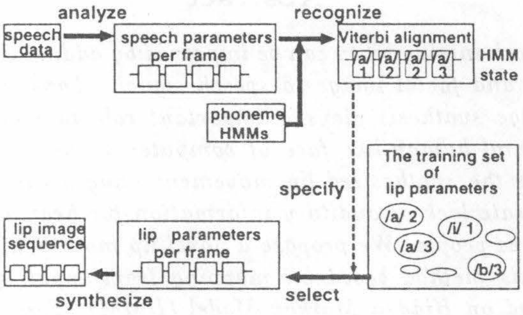

Figure 3: Schematic Diagram of Lip Movement Synthesis by the HMM-based method

### Synthesis Algorithm

1. Align input speech into the HMM state sequence by the Viterbi decoding.
2. Retrieve the image parameter associated with the HMM state along Viterbi alignment.
3. Concatenate the retrieved image parameters as a lip movement sequence.

# 3 Experiments

## 3.1 Experiment Condition

Speech and image data are synchronously recorded in 125Hz using OPTOTRAK 3020 shown in Fig.3.1. A lip image is parameterized to 12 three dimensional positions on the face including eight positions around the lip outer contour. These parameters are transformed into 3D parameters such as hight(X), width(Y) of the lip outer contour and protrusion(Z) shown in Fig.5 [4]. The speech data is parameterized to 16-order mel-cepstral coefficients, their delta coefficients and delta log power. Tied-mixture Gaussian HMMs for 54 phonemes, pre-utterance pause and post-

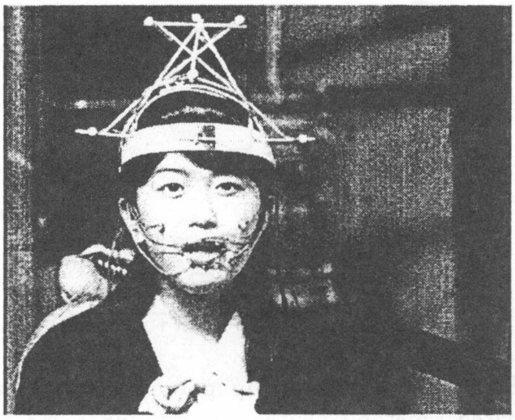

Figure 4: Marker locations (lip outer contour=8 points, cheek=3 points, jaw=1 point) ©ATR

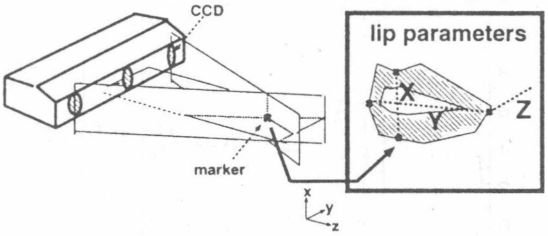

Figure 5: Sensing markers' 3D positions, and construction lip parameters

utterance pause are trained with 256, 256 and 128 distributions. The pause models are separately prepared for the word beginning and the ending. The audio and speech synchronous database are composed of 326 phonetically balanced Japanese word training data and another 100 word testing data. The synthesized lip parameters are evaluated by Euclidian square error distance $E$ and the time differential error $\Delta E$ between the synthesized parameters $x_i^s = \{X^s, Y^s, Z^s\}$ and the original ones, $x_i^o = \{X^o, Y^o, Z^o\}$.

$$E = \{\sum_i (x_i^s - x_i^o)^2\}^{\frac{1}{2}}$$

$$\Delta E = \{\sum_i (\Delta x_i^s - \Delta x_i^o)^2\}^{\frac{1}{2}},$$

where $\Delta x_i = x_i(t+1) - x_i(t)$. The same weights are assigned to each dimension of image parameters in this experiment. The time differential error $\Delta E$ is adopted for evaluation of smoothness. The phoneme recognition accuracy of speech HMMs for testing data in this experiment is 74.3% under the simple grammar of the consonant-vowel constraint of Japanese. The consonant-vowel constraint is imposed by the fact that the consonants of Japanese do not appear in the isolated phonemes separated to the vowels. The accuracy of HMMs is obtained by $100 \times (Total\# - Deletion\# - Substitution\# - Insertion\#)/Total\#$, where $Total\#$ means the number of the total phoneme labels, and $Deletion\#$, $Substitution\#$, and $Insertion\#$ indicates the deletion number, the substitution number, and the insertion number of the decoded phoneme labels, respectively. The accuracy of HMMs indicates implies the accuracy of the Viterbi alignments by the HMMs.

## 3.2 VQ based Method

Experiments using the VQ based mapping method are also carried out for comparison. The VQ method maps VQ code of an input speech to lip image parameters frame-by-frame. The following is the VQ based mapping method.

**Training Algorithm**

1. Generate VQ codebooks of 256 codewords both for speech and lip image parameters by the LBG algorithm.
2. Count correspondence histogram $w_{k,1}, \cdots, w_{k,256}$ between speech code $C_k^{sp}$ and lip image codes $C_1^{im}, \cdot, C_{256}^{im}$.
3. Assign the expected image parameters to the input speech VQ code. The expected image parameters are given by the following three methods.

$$(1)\quad C_k^{sp} \underset{\text{map}}{\longrightarrow} C_{k'}^{im} = C_l^{im} \qquad (l = argmax_l\, w_{k,l})$$

$$(5)\quad C_k^{sp} \underset{\text{map}}{\longrightarrow} C_{k'}^{im} = \frac{\sum_l^5 w_{k,l} C_l^{im}}{\sum_{l=1}^5 w_{k,l}} \qquad (l \in top5\ of\ w_{k,l})$$

$$(256)\quad C_k^{sp} \underset{\text{map}}{\longrightarrow} C_{k'}^{im} = \frac{\sum_{l=1}^{256} w_{k,l} C_l^{im}}{\sum_{l=1}^{256} w_{k,l}}$$

(1) is the case that lip image parameters with the highest count is assigned to the input speech VQ code. (5) is the case that the weighted combination of lip parameters of the top 5 count is assigned to the input speech VQ code. The last one is the expectation for over all lip image parameters. The synthesis algorithm is thus described as follows.

Table 1: Distance Evaluation by HMM-based method and VQ based method

| | $E$ cm | | $\Delta E$ cm | |
|---|---|---|---|---|
| | closed | open | closed | open |
| VQ(1) | 1.43 | 1.48 | 0.87 | 0.88 |
| VQ(5) | 1.15 | 1.22 | 0.37 | 0.38 |
| VQ(256) | 1.14 | 1.15 | 0.28 | 0.28 |
| HMM(correct) | 1.05 | 1.04 | 0.20 | 0.18 |
| HMM(recog) | 1.05 | 1.05 | 0.20 | 0.19 |

**Synthesis Algorithm**

1. Quantize an input speech frame by the speech VQ codebook into $C_k^{sp}$.
2. Retrieve the image parameter, $C_{k'}^{im}$ associated with the input speech VQ code $C_k^{sp}$.
3. Synthesize an output lip image by the lip image parameter $C_{k'}^{im}$

## 4 Results

Table 1 shows the results of our HMM-based speech-to-lip movement synthesis. The HMM-based method indicates two cases such that correct decoded transcriptions are given and decoded transcriptions are given by phoneme recognition. In the latter case, incorrect transcriptions are included. In comparison to the VQ method, HMM-based methods give smaller distance than that of VQ-based method. The reduction percentages of $E$ and $\Delta E$ of the HMM-based method(phoneme recognition case) to VQ(256) are 8.7% and 32% respectively for testing data. This result means the the HMM-based method can provide much smoother synthesized images than those of the VQ method. Fig.6, Fig.7 and Fig.8 show the actual difference of image parameters for a testing data /neQchuu/. In these figures, the solid lines indicate the synthesized image parameters and the dotted ones indicate the original ones. In figures, the vertical lines designate the start and the end time points of the utterance. The VQ-based method exposes many outbreak errors. The hatched regions in Fig.8 indicate incorrect transcription caused by phoneme recognition.

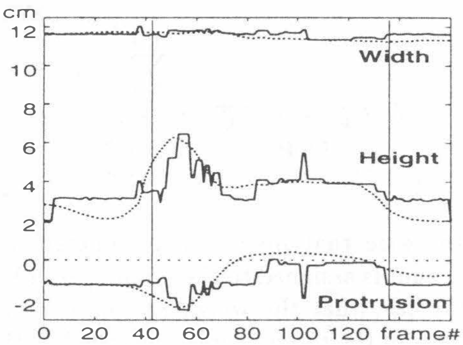

Figure 6: Synthetic lip parameters by the VQ-based method

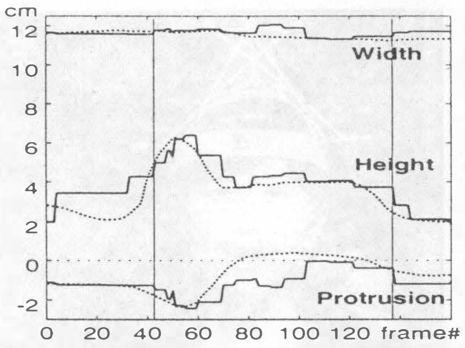

Figure 7: Synthetic lip parameters by the HMM-based method with correct phoneme sequence /neQchuu/

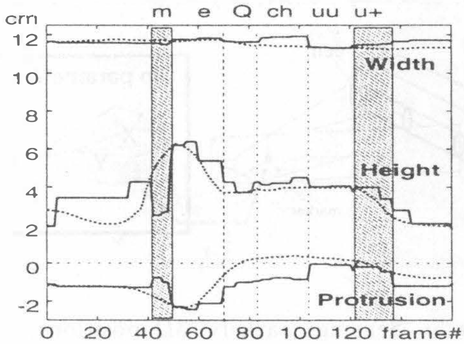

Figure 8: Synthetic lip parameters by the HMM-based method with error phoneme sequence /meQchuuu+/

## 5 SV-HMM-based Method with Context Dependency

The notable errors are found at /h/,/Q/ and the silence of word beginning, because the lip configurations of those phonemes depend on succeeding phoneme, especially on succeeding viseme. However, it is not easy to train the context dependent HMMs and the mapping from every state to lip image parameters. Then the HMM-based method consider-

Table 2: Viseme Clustering Results

| | |
|---|---|
| Viseme 1 | n,b,by,f,m,my,p,py,s,sh,t,ts,u,ue,ui,uu,w,y,z |
| Viseme 2 | Q,ch,d,g,gy,hy,j,k,ky,n,ny,o,oN,oo,ou,r,ry |
| Viseme 3 | a,a-,aN,aa,ai,ao,e,eN,ee,ei,h,i,iN,ii |

ing a succeeding phoneme, the Succeeding Viseme dependent HMM-based method(SV-HMM), is also proposed. The SV-HMM-based method is characterized at taking an average of image parameters per HMM, per state and per viseme of succeeding phoneme. Lip image synthesis can be carried out by looking ahead the succeeding phoneme information along the Viterbi alignment. Visemes in this paper are defined as 3 classes by merging image parameters of the first state of 54 phoneme HMMs by bottom-up clustering because of insufficient training data. The training algorithm of the SV-HMM-based method is the same as that of the HMM-based method except training data is taken average depending on not only HMM state but succeeding viseme. In the synthesis algorithm of the SV-HMM-based method the output image parameters are specified by the HMM state and the succeeding viseme looking ahead along Viterbi alignment as well as the training step. Table 2 shows the phoneme members belong to each viseme class. The viseme class 1 is considered as the lip closing shape. Likewise the viseme class 3 seems to be the lip opening shape. The viseme class 2 may take the middle position between the viseme class 1 and the viseme class 3.

# 6 Experiment Results of SV-HMM-based Method

The error distance of the SV-HMM-based method is shown in Table 3. The reduction of the SV-HMM-based method shows 10.5% in $E$ and 11% in $\Delta E$ compared to the HMM-based method for testing data including phoneme recognition errors. The example of image parameters for training data /saki+hodo/ is illustrated in Fig.9(HMM) and Fig.10(SV-HMM). The shading sections correspond the phoneme /h/. The SV-HMM-based method represents the remarkable reduction of errors for speech periods compared with the HMM-based method. The lip images in Fig.11 show the lip image of /h/ by the each method. The synthesized image by the SV-HMM-based method seems to be close to the original one compared to that of the HMM-based method. Fig.12 represents the errors of each phoneme with large error, where the white box means the HMM-based method and the black box means the context dependent SV-HMM-based method. Those figures indicates that the context dependent synthesis is very effective for speech to lip movement synthesis. Although the HMM-based method can consider inter phoneme HMM contexts in the SV-HMM-based method easily, the conventional VQ-based method is difficult to deal with context information from training point of view.

Table 3: Error distances by the HMM-based method and the SV-HMM-based method

| | $E$ cm | | $\Delta E$ cm | |
|---|---|---|---|---|
| | closed | open | closed | open |
| HMM(correct) | 1.05 | 1.04 | 0.20 | 0.18 |
| HMM(recog) | 1.05 | 1.05 | 0.20 | 0.19 |
| SV-HMM(correct) | 0.90 | 0.90 | 0.18 | 0.17 |
| SV-HMM(recog) | 0.91 | 0.94 | 0.19 | 0.17 |

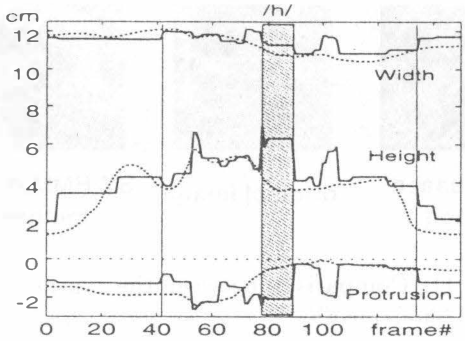

Figure 9: Synthetic lip parameters by the HMM-based method /saki+hodo/

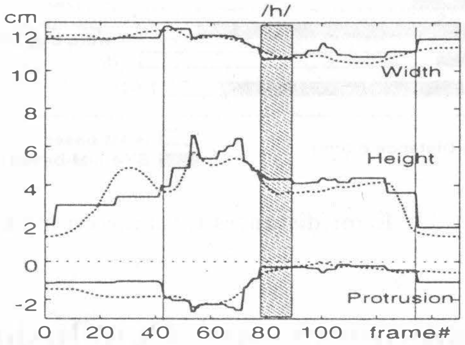

Figure 10: Synthetic lip parameters by the SV-HMM-based method /saki+hodo/

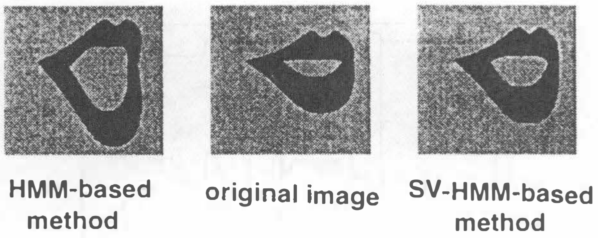

Figure 11: Comparison between the synthetic lip images of phoneme /h/

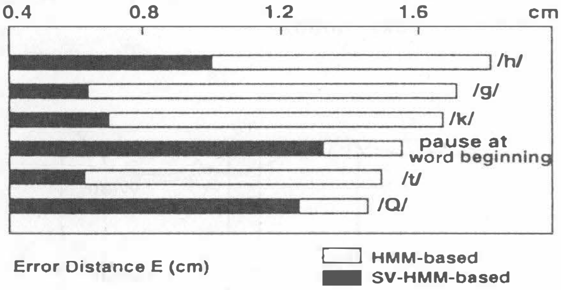

Figure 12: Error distances for large error phoneme

# 7 Summary and Conclusions

This paper proposes the speech-to-lip image synthesis method using HMMs. The succeeding viseme dependent HMM-based method is also proposed to improve large errors of /h/ or /Q/ phonemes. Evaluation experiments clarify the effectiveness of the proposed methods compared with the conventional VQ method.

As for the context dependent HMM-based method, it is natural to extend from monophone to biphone or triphone for HMMs. But it is impossible to construct biphones or triphones from very few training data. So we have limited the context dependent HMM-based method so as to utilize the succeeding context only with three viseme patterns.

In synthesis, the HMM-based method holds the intrinsic difficulty that the synthesis precision depends upon accuracy of the Viterbi alignment. Since the Viterbi alignment assigns the single HMM state for each input frame, it may synthesize a wrong lip image for the output because of incorrect Viterbi alignment. This problem would be solved by extending the Viterbi algorithm to the Forward-Backward algorithm that can consider HMM state alignment probabilistically in synthesis.

# References

[1] T. Chen and R. Rao. Audio-visual interaction in multimedia communication. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[2] W. Chou and H. Chen. Speech recognition for image animation and coding. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995.

[3] S. Curinga, F. Lavagetto, and F. Vignoli. Lip movement synthesis using time delay neural networks. *Proc. EUSIPCO96*, 1996.

[4] T. Guiard-Marigny, T. Adjoudani, and C. Benoit. A 3-d model of the lips for visual speech synthesis. In *Second ESCA/IEEE Workshop on Speech Synthesis*, New Palts, New York, 1994.

[5] F. Lavagetto. Converting speech into lip movements: a multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering*, 3(1), 1995.

[6] S. Morishima, K. Aizawa, and H. Harashima. An intelligent facial image coding driven by speech and phoneme. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1989.

[7] S. Morishima and H. Harashima. A media conversion from speech to facial image for intelligent man-machine interface. *IEEE Journal on Selected Area in Communications*, 9(4), 1991.

[8] W. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustic Society of America*, 26:212–215, 1954.