SPEECH TO LIP MOVEMENT SYNTHESIS BY HMM

Eli Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
Takayama 8916-5, Ikoma, Nara 630-0 1, Japan.
Tel.+81 743 72 5287, FAX:+81 74 3 72 5287,E-mail:{eli,nakamura,shikano}@is.aist-nara.ac.jp

ABSTRACT

Synthesized lip movement images can compensate lack of auditory information for hearing impaired people, and also contribute to realize a human-like face of computer agents. We propose a novel method to synthesize lip movement based on mapping from an input speech using HMM. This paper compares the HMM method and a conventional method using VQ or ANN to convert speech-to-lip movement images. In the experiment, error and time difference error between synthesized lip movement images and original ones are utilized for evaluation. The result shows that the error of the HMM method is 8.6% smaller than that of the VQ method. Moreover, the HMM method reduces time difference error by 34.8% than the VQ's. The result also shows that the errors are mostly caused by phoneme /h/ and /Q/. Since those phonemes are dependent on succeeding phoneme, the context-dependent synthesis on the HMM method is applied to reduce the error. The context-dependent HMM method realizes that the error(difference error) is reduced by 11.3%(8.9%) compared with the original HMM method.

1. INTRODUCTION

In human machine communication, it is quite important to realize natural and friendly interface. Speech recognition or computer lipreading have been developed at input means of communication. They are also important to provide natural and friendly interface at output means. While speech synthesis has been studied by many researchers, synthesis of lip movement has not been interested. However, the lip movement synthesis can take the significant role in human-machine communication. For example, it could be a useful tool for hearing impaired people to compensate lack of auditory information.

There are two approaches, 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 phoneme, coarticulation, duration and so on. The lip movement synthesis from speech seems to be more promising than synthesis from text.

As algorithms mapping from speech to lip movement, VQ based method and Neural-Network based method have been reported in [1][2][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) the frame-by-frame mapping doesn't take account of phoneme context. These problems might produce discontinuity in synthesized image sequence.

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

2. HMM METHOD FOR LIP MOVEMENT SYNTHESIS

2.1. HMM method using Viterbi alignment

We propose the HMM method for mapping from speech to lip movement images using Viterbi alignment. In speech recognition, when test speech is inputted, output phoneme corresponding an HMM of the maximum likelihood is selected. When the HMM of the maximum likelihood is specified, the Viterbi alignment of whole input speech can be calculated.
Figure 1. Viterbi alignment (assign HMM state on each frame deterministically)

Figure 2. Schematic Diagram of Lip Parameter Training by the HMM method

Here Viterbi alignment means a transition state sequence of HMMs to produce the maximum likelihood. Fig.1 shows each frame is assigned to each HMM state by Viterbi alignment. Consequently, if lip parameters are trained on each HMM state, they can be outputed along the HMM state sequence. The works applying Viterbi alignment to lip movement synthesis are also reported by [5][6]. The paper[5] doesn’t include estimation of lip parameters. The paper[6] only introduces a similar idea independently.

2.2. Algorithm of the HMM method

The basic idea of the proposed HMM method is mapping from HMM states to lip image parameters. Once HMMs are trained, the input speech can be transformed into HMM state sequence by HMM decoding. The lip movement sequence is obtained by concatenating the image parameters associated with the HMM state sequence. Fig.2 and Fig.3 shows a block diagram of the Viterbi HMM method. The training and synthesis algorithms are as follows.

Training

1. Prepare and parameterize audio and visual synchronous database.
2. Train HMMs using the training speech database.

Synthesis

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 CONDITION

Speech and image data are synchronously recorded in 125Hz. A lip image is parameterized to 12 three dimensional positions on the face including eight positions around lip outer contour. These parameters are finally transformed into 3D parameters such as height(X), width(Y) of lip outer contour and protrusion(Z). 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-word pause and post-word pause are used 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 216 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 difference error ΔE between the synthesized parameters $x_i = \{X^s, Y^s, Z^s\}$ and the original ones $x^o_i = \{X^o, Y^o, Z^o\}$.

$E = \left\{ \sum_i (x^s_i - x^o_i)^2 \right\}^{\frac{1}{2}}$

$ΔE = \left\{ \sum_i (Δx^s_i - Δx^o_i)^2 \right\}^{\frac{1}{2}}$. 

3. Align speech database into HMM state sequence according to Viterbi decoding.
4. Take an average of image parameters of the frames associated with the same HMM state.
Table 1. Errors and time difference errors by each method

<table>
<thead>
<tr>
<th>Method</th>
<th>$E$ cm</th>
<th>$\Delta E$ cm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>closed</td>
<td>open</td>
</tr>
<tr>
<td>VQ</td>
<td>1.11</td>
<td>1.16</td>
</tr>
<tr>
<td>HMM (correct)</td>
<td>1.03</td>
<td>1.05</td>
</tr>
<tr>
<td>HMM (error)</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>SV-HMM (correct)</td>
<td>0.85</td>
<td>0.91</td>
</tr>
<tr>
<td>SV-HMM (error)</td>
<td>0.87</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Figure 4. Synthetic lip parameters by the VQ method (X=height, Y=width, Z=protrusion)

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 difference error $\Delta E$ is adopted for evaluation of smoothness.

4. EXPERIMENT RESULT OF THE HMM METHOD

Table 1 results of our HMM-based speech-to-lip movement synthesis. The HMM method indicates two cases that one is composed by correct decoded transcriptions and another is composed by incorrect decoded transcriptions by phoneme recognition. In comparison to the VQ method, the reduction percentages of $E$ and $\Delta E$ of the HMM method (at error included case) are 8.6% and 34.8% respectively about open data. This result means the the HMM method can provide much smoother synthesized value than that of the VQ method. Fig.4, Fig.5 and Fig.6 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 method exposes many outbreak errors.

5. CONTEXT DEPENDENT HMM METHOD

The notable errors are found at /h/ and the silence of word beginning, because the lip con-

6. EXPERIMENT RESULT OF THE SV-HMM METHOD

The error distance of the SV-HMM method is also shown in Table 1. The reduction of the SV-HMM method shows 11.3% in $E$ and 8.9% in $\Delta E$ compared to the HMM method about open data including incorrect transcriptions. The example of image parameters for training data /saki+hodo/ is illustrated in Fig.7(HMM) and Fig.8(SV-HMM). The shading sections correspond the phoneme /h/. The SV-HMM
Figure 7. Synthetic lip parameters by the HMM method /saki+hodo/ (X=height, Y=width, Z=protrusion)

Figure 8. Synthetic lip parameters by the SV-HMM method /saki+hodo/ (X=height, Y=width, Z=protrusion)

method represents the remarkable reduction of errors for speech periods compared with the HMM method. The lip images in Fig. 9 show the configuration of /h/ by each method. The synthesized image by the SV-HMM method seems to be similar to the original one compared to that of the HMM method. Fig. 10 represents the errors of each phoneme with large error, where the white box means the HMM method and the black box means the context-dependent SV-HMM method. Those figures indicate that the context-dependent synthesis is very effective for speech to lip movement synthesis. Although the HMM method can consider contexts in the SV-HMM method easily, the conventional VQ method is difficult to deal with context information.

7. CONCLUSION

This paper proposes the speech-to-lip image synthesis method using HMMs. The succeeding viseme dependent HMM 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 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 method so as to utilize the succeeding context only with three viseme patterns.

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

REFERENCES