SYLLABLE RECOGNITION USING SYLLABLE-SEGMENT STATISTICS AND SYLLABLE-BASED HMM

Nobutoshi TAKAHASHI and Seiichi NAKAGAWA

Toyohashi University of Technology
Department of Information and Computer Sciences
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi, 441-8580, JAPAN
{nobutosi, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT
In our previous research, we demonstrated the validity of segmental unit input hidden Markov model (HMM), which regards successive four frame MEL-cepstrum coefficients as a feature vector. The vector is reduced to lower dimensions using the KL transform. However, the model considers only the correlation between frames in a short section, but not the correlation between the frames over a long section.

In this paper, in order to represent the correlation over a long distance, we use the syllable-segment statistics that are calculated by the concatenation of feature vectors, corresponding to each state in a syllable based HMM. By combining this approach with a segmental-unit input HMM, the syllable recognition rate was improved to 87% from 83% for syllables taken from continuous speech, without using a language model. We also showed the effectiveness for continuous speech recognition.

1. INTRODUCTION
Hidden Markov models (HMMs) have been widely used for speech recognition. However, it is well known that they can not adequately express the correlation between successive feature vectors. To solve this problem, we previously studied a segmental unit syllable-based input HMM [1][2] and reported that it was comparable with a triphone-based HMM [3]. Although this model can represent the correlation between the successive feature vectors, it can not represent the correlation between the frames over a long section.

Howell proposed the method of combining HMM and a multi-layer perceptron [4]. This method normalizes the length of a speech pattern by HMM-based Viterbi alignment, and then the obtained fixed length pattern is inputted to a multi-layer perceptron. Howell did not attempt the integration of both likelihoods of HMM and perceptron (a posteriori probability).

In other research, a method of unifying HMM and a segment model was presented [5]. The method segments continuous speech into two successive phone pairs, and creates a segmental model by using a trajectory model. The extension of the trajectory model for a fixed length segment was reported [6], in which a traditional trajectory model was combined with statistics of model parameters.

In this paper, we use syllable-segment statistics combined with a traditional HMM. The use of syllable-segment statistics allows the model to express the correlation between the frames over a long distance (e.g., the correlation between a vector in the first state and a vector in the fourth state in a syllable-based HMM). By combining this approach with a segmental-unit input HMM, the syllable recognition rate was improved to 87% from 83% for syllables taken from continuous speech, without using a language model.

2. SYLLABLE SEGMENT MODEL
The syllable-segment model proposed in this paper uses a concatenative vector for each syllable, which is the concatenation of an averaged feature vector for each state in a syllable-based HMM.

First, using a syllable-based HMM, we obtain a mean vector for the feature vectors assigned for each state. Second, we concatenate these averaged vectors for every syllable, and third we estimate a Gaussian mixture model for every syllable using these concatenative vectors. Figure 1 illustrates this procedure.

Fig. 1. Syllable segment model
This syllable-segment model can capture the correlation between frames over a long section (for example, the feature parameters between the first state and the fourth state in a HMM).

In experiments, to create the syllable-segment models, we first conducted a forced Viterbi alignment against continuous speech using a conventional syllable-based HMM, and then segmented the continuous speech into syllable segments. Finally, we took syllable segments from the continuous speech.

Using such segmented syllables, we set up training or test samples consisting of concatenated feature vectors. As this concatenated feature vector consists of a high dimension, the dimension is reduced using the K-L transform.

3. DEVIATION OF SELECTED ELEMENT DISTRIBUTION IN GAUSSIAN MIXTURES

As the syllable-segment models use an averaged feature vector for each state, it is desirable that each state corresponds to a stationary part for a given utterance. In other words, the same element distribution in Gaussian mixtures for a state is preferable for a given utterance.

In our recognizer, we chose the most preferable element distribution instead of the summation of all probabilities (max operation instead of sum operation).

Specific patterns should be observed at the selection of a preferable element distribution under the condition of a specified speaker, a context, and so on. If there are such specific patterns, we can say that a syllable-segment model captures the correlation between frames over a long distance like trajectory models [6][7].

We therefore investigated whether a deviation of the distribution selection exists or not. We identified a part as stationary if the same element distribution was selected more than 50 percent of the time in all frames, which were aligned with the state. Such stationary parts existed in about 50 percent of all aligned states and syllables. In a 4-mixture HMM having 4 states, the possible number of patterns is $4^4 = 256$. For each syllable, there were about 16 specific patterns which consisted of selected element sequences in the stationary parts. For example, the third element distribution is selected for the first state in a syllable-based HMM, the fourth for the second state, the first for the third state and the third for the fourth state. In this result, it is expected that there is a deviation in the selection of mixture distributions; in other words, there is a correlation between feature vectors over a syllable-segment.

4. SEGMENT-UNIT INPUT HMM

The expression of the output probability computation of HMM for the input symbol sequence $y = y_1 y_2 \cdots y_T$ (T is the length of input sequence) and state sequence $x = x_1 x_2 \cdots x_T$ is given by the following [1] :

$$P(y_1 \cdots y_T) = \sum_x \prod_i P(y_i | y_{i-1} y_{i-2} x_{i-1} \cdots x_1 x_2 \cdots x_{i-1} x_i) \times P(x_i | x_{i-1} x_{i-2} \cdots x_{i-1})$$

$$\approx \sum_x \prod_i P(y_i | y_{i-3} y_{i-2} y_{i-1} x_{i-1} x_i) P(x_i | x_{i-1})$$

5. EXPERIMENT

5.1. Experimental Condition

In our experiments, continuous density HMMs (having full covariance matrices) with 5 states (4 output distributions) under duration control were used. They were trained by syllable-segmented data from A-J sets (50 sentences each) from an ATR speech database (utterances by 6 male speakers). For syllable categories with a small amount of data in the database, and 216 word data sets were additionally used. Afterward, the HMMs were retrained with MAP estimation using an Acoustic Society of Japan database with utterances by 30 male speakers (ASJ) (4,518 sentences) and a Japan Newspaper Article Sentences database with utterances by 125 male speakers (JNAS) (12,703 sentences).

The test data consisted of 939 newspaper article sentences spoken by 9 other male speakers.

The analysis conditions were as follows: sampling frequency of 12 kHz; Hamming window size of 21.33 ms; frame period of 8 ms; and LPC analysis of the 14th order, feature parameters of which were 10 mel.

Equation (1) or Eq.(2) is a conditional density HMM in 4-frame segments. Equation (3) or Eq.(4) is a conditional density HMM in 2-frame segments, and Eq.(5) is a segmental unit input HMM of 2-frame segments.

The segmental unit input HMM is an approximation of Eq.(2), that is, we use only the numerator of Eq.(2) as follows:

$$P(y_1 \cdots y_T) \approx \sum_x \prod_i P(y_i | y_{i-3} y_{i-2} y_{i-1} x_{i-1} x_i) P(x_i | x_{i-1})$$

Using segmental unit input HMM wherein several successive frames are inputted as one vector, since the dimension of the vector increases, results in lower precision in the estimation of the covariance matrix. Therefore, the K-L expansion was used to reduce the dimension in the experiments.
frequency cepstrum coefficients derived from Oppenheim and Johnson’s recursive way (denoted simply by LPCC) their first/second derivative, and ∆/∆∆ energy. In the case of MFCC, we used a sampling frequency of 16 kHz, Hamming window of 21.33 ms and frame period of 8 ms. The number of syllables used as a unit of acoustic modeling was 114.

The recognition rates using base-line HMM models are shown in Table 1.

Table 1. Syllable recognition performance by base-line HMMs [%]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>frm-HMM-CO</td>
<td>67.3</td>
<td>75.7</td>
<td>21.4</td>
<td>8.3</td>
<td>3.0</td>
</tr>
<tr>
<td>seg-HMM-CO</td>
<td>70.1</td>
<td>78.3</td>
<td>19.1</td>
<td>8.2</td>
<td>2.7</td>
</tr>
<tr>
<td>frm-HMM-AL</td>
<td>81.0</td>
<td>81.0</td>
<td>19.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>seg-HMM-AL</td>
<td>81.3</td>
<td>81.3</td>
<td>18.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

In the table, “frm” denotes a frame by frame input as with a conventional HMM, “seg” denotes a segmental unit input, “CO (continuous)” indicates the recognition rate of speaker-independent continuous syllable recognition, and “AL (alignment)” the recognition rate of syllables taken from continuous speech by the Viterbi alignment.

5.2. Result by Syllable Segment Model

The syllable segment models were built-in GMMs with diagonal covariance matrices and full covariance matrices, respectively. The numbers of mixtures are 1 and 4 for full covariance matrices, and 16 and 32 for diagonal covariance matrices. Furthermore, the concatenated feature vector corresponding to 4 states, i.e., 40 dimensions for syllable-segment statistics, was compressed into 20 dimensions by the K-L transform. Moreover, the concatenative vector using delta cepstrum coefficients was also constructed. Table 2 summarizes the syllable recognition rates for syllables taken from continuous speech by the Viterbi alignment when using syllable segment models for the case of LPCC.

Table 2. Syllable recognition rate using syllable segment models [%] and recognition/mis-recognition tendency in comparison with a base-line HMM model

<table>
<thead>
<tr>
<th>Model</th>
<th>COR [%]</th>
<th>CC</th>
<th>CS</th>
<th>SC</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>D16</td>
<td>47.7</td>
<td>16516</td>
<td>14125</td>
<td>1520</td>
<td>5675</td>
</tr>
<tr>
<td>KL-D16</td>
<td>55.9</td>
<td>19351</td>
<td>11290</td>
<td>1804</td>
<td>5391</td>
</tr>
<tr>
<td>KL-D32</td>
<td>55.5</td>
<td>19198</td>
<td>11443</td>
<td>1813</td>
<td>5382</td>
</tr>
<tr>
<td>KL-F1</td>
<td>66.8</td>
<td>23227</td>
<td>7414</td>
<td>2046</td>
<td>5149</td>
</tr>
<tr>
<td>KL-F4</td>
<td>68.9</td>
<td>23747</td>
<td>6894</td>
<td>2302</td>
<td>4893</td>
</tr>
<tr>
<td>KL-F1+∆</td>
<td>75.4</td>
<td>25919</td>
<td>4722</td>
<td>2595</td>
<td>4600</td>
</tr>
<tr>
<td>KL-F4+∆</td>
<td>78.4</td>
<td>26512</td>
<td>4129</td>
<td>3141</td>
<td>4054</td>
</tr>
</tbody>
</table>

In the table, “D” denotes a GMM with diagonal covariance matrices, “F” full covariance matrices, “KL” the KL transform, and “number” in D# or F# as the number of mixtures. ”COR” denotes the syllable recognition rate.

The recognition performance was improved by using full covariance matrices, delta-cepsrum and the K-L transform. The recognition performance was also improved by increasing the mixtures.

5.3. Recognition/Mis-recognition Tendency

We investigated the recognition and mis-recognition tendency of syllable segment models in comparison with base-line HMMs. Table 2 summarizes the results.

“CC” in the table denotes the number of syllables which were recognized correctly when using both a base-line HMM (frm-HMM-AL) and syllable-segment model. “CS” denotes the number of syllables which were recognized correctly using the HMM but mis-recognized using the syllable-segment model. “SC” denotes the number of syllables which were mis-recognized using the HMM and recognized correctly using the syllable-segment model, and ”SS” denotes the number of syllables which were mis-recognized by both the HMM and the syllable-segment models.

From these recognition/mis-recognition tendencies, we can expect that the syllable-segment model can compensate for the mis-recognition by the base-line HMM, and vice versa.

5.4. Combination of Syllable-based HMM and Syllable Segment Model

Recognition experiments combining the syllable-segment model with HMM were conducted. In order to normalize the two acoustic likelihoods, the likelihood score obtained from the syllable-segment model was multiplied by the number of frames. Furthermore, these likelihoods were averaged with a weight (weighted average) as follows:

\[ AS = \alpha AS_{HMM} + (1-\alpha) AS_{syllable\_segment} \times frames, \]

where AS is the acoustic likelihood score, \( \alpha \) a weight, and frames the numbers of frames in the syllable segment derived by a forced alignment against continuous speech using a conventional HMM.

Table 3 shows the syllable recognition rates. In the table, the weight \( \alpha \) is used as the best value. Figure 2 illustrates the recognition performance in terms of weights for combining both methods.

Recognition rates were improved by combining both models. With the combination of a segmental-unit HMM and syllable segmental statistics, we obtained a recognition accuracy of 87.1% for syllables, 95.1% for vowels, and 88.7% for consonants in continuous speech without any language model. The recognition accuracy was improved to 87.1% from 83.7%.

6. APPLICATION TO CONTINUOUS SPEECH RECOGNITION

In this section, we describe the result of continuous speech recognition by the combination of a standard HMM and the syllable-segment model. Our decoder produces N-best hypotheses, each of which consists of
Table 3. Syllable recognition rate by combination of syllable-based HMM and syllable-segment model
(a) LPC-melcepstrum coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>COR[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL-F4+Δ only</td>
<td>78.4</td>
</tr>
<tr>
<td>frm-HMM only</td>
<td>81.0</td>
</tr>
<tr>
<td>seg-HMM only</td>
<td>81.3</td>
</tr>
<tr>
<td>frm-HMM &amp; KL-F4+Δ</td>
<td>85.3</td>
</tr>
<tr>
<td>seg-HMM &amp; KL-F4+Δ</td>
<td>85.4</td>
</tr>
</tbody>
</table>

(b) MFCC

<table>
<thead>
<tr>
<th>Model</th>
<th>COR[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL-F4+Δ only</td>
<td>78.8</td>
</tr>
<tr>
<td>frm-HMM only</td>
<td>82.9</td>
</tr>
<tr>
<td>seg-HMM only</td>
<td>83.7</td>
</tr>
<tr>
<td>frm-HMM &amp; KL-F4+Δ</td>
<td>86.6</td>
</tr>
<tr>
<td>seg-HMM &amp; KL-F4+Δ</td>
<td>87.1</td>
</tr>
</tbody>
</table>

Fig.2 Syllable recognition rate by combination of syllable-based HMM and syllable-segment model (MFCC)

a syllable sequence. Therefore, by using the combination method, these hypotheses were rescored and the best one was chosen. For each hypothesis, the following processes were performed to calculate a new acoustic score.

First, the system creates syllable segmental features by using hypothesis’s syllable boundary information. Second, it calculates syllable segmental likelihood (score). And third, the scores by HMM and by syllable segment model are linearly interpolated and the obtained score is treated as an acoustic score.

The task is 100 person names spoken by 5 males. The average number of syllable in a name is 7.0. The system produced syllable sequences of 200-best hypotheses by one pass Viterbi decoding without language models, rescored 200-best hypotheses and chose the best one. Table 4 shows the continuous syllable recognition rates. The recognition accuracy was improved to 70.0% from 68.7% and the correct rate was improved to 74.5% from 74.0%.

Table 4. Continuous syllable recognition performance by rescoring based on a combination method [%]

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc</th>
<th>Cor</th>
<th>Sub</th>
<th>Ins</th>
<th>Del</th>
</tr>
</thead>
<tbody>
<tr>
<td>frm-HMM only</td>
<td>68.7</td>
<td>74.0</td>
<td>23.4</td>
<td>5.3</td>
<td>2.6</td>
</tr>
<tr>
<td>frm-HMM &amp; KL-F4+Δ</td>
<td>70.0</td>
<td>74.5</td>
<td>22.9</td>
<td>4.5</td>
<td>2.6</td>
</tr>
</tbody>
</table>

7. CONCLUSION

In this paper, we proposed a syllable segmental statistic model and a combination method with a traditional HMM. The statistics are modeled by a GMM and obtained from a set of the concatenation of feature vectors, corresponding to each state in a syllable-based HMM. For modeling and estimating, we conducted a forced Viterbi alignment against continuous speech using a conventional HMM, and then we segmented continuous speech into syllable segments. By combining HMMs and syllable-segment models, the syllable recognition rate was improved to 87.1% from 83.7% in segmented speech. Continuous syllable recognition rate was also improved to 70.0% from 68.7% in the accuracy rate and to 74.5% from 74.0 in the correct rate, respectively.

8. REFERENCES