EXTRACTING PHONOLOGICAL CHUNKS BASED ON PIECEWISE LINEAR SEGMENT LATTICES

Hiroaki KOJIMA and Kazuyo TANAKA
Electrotechnical Laboratory, AIST, MITI
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{hkojima, ktanaka}@etl.go.jp

ABSTRACT

The task of our research is to form phone-like models and a phoneme-like set from spoken word samples without using any transcriptions except for the lexical identification of each word in a vocabulary. This framework is derived from two motivations: 1) automatic design of optimal speech recognition units and structures of phone models, and 2) multi-lingual speech recognition based on language-independent intermediate phonetic codes. The procedure consists of two steps: 1) constructing a VQ codebook of sub-phonetic segments from speech samples, and 2) extracting phonological chunks from sequences of the codes. Segment model is represented with "piecewise linear segment lattice" model, which is a lattice structure of segments, each of which is represented as regression coefficients of feature vectors within the segment. Phonological chunks are extracted with a criterion based on Kullback-Leibler divergence between the distribution of individual VQ codes. The recognition rate yields approximately 90% on the 1542 words task with 128 VQ codes.

1. BACKGROUND

The task of our research is to form phone-like models and a phoneme-like set from spoken word samples without using any transcriptions except for the lexical identification of each word in a vocabulary[1]. The procedure consists of two steps: 1) constructing a VQ codebook of sub-phonetic segments from speech samples, and 2) extracting phonological chunks from sequences of the codes.

This framework is derived from two motivations.

The first is to design optimal speech recognition units and structures of phone models. In the paradigm of traditional stochastic method of speech recognition, the processes to improve performance tend to follow a wrong spiral of making more precise models and increasing the size of training samples. On the other hand, a human infant seems able to acquire knowledge of his/her native phonological system properly. By analogy of this situation, we adopt the task to acquire phonological units and their structures inductively from speech samples. In our approach, we assume that the essential factor in acquiring a phonological system is to discover the relationships between utterances and their meanings. In order to make this practical for experimental purposes, we simplified them into the relationships between isolated spoken word samples and their lexical identification. The basis of this idea is that an appropriate model should be better formed through interactive communication in a flexible task in which the least necessary knowledge is provided, than be fully defined a priori. Related researches seem to be getting more active for example: [2][3][4][5][6].

The second motivation is toward multi-lingual speech recognition. We are developing a speech recognition method based on language-independent intermediate phonetic codes[7]. In this case, the segmental units described in this paper can be used as intermediate phonetic codes, and the phonological chunks can be used as phonemes of the language. The advantage of this approach is that it is available for languages whose phonetic and phonological systems are not investigated sufficiently.

2. DERIVING SEGMENTAL UNITS

2.1. PLSL model

In order to represent structure of segmental units, we have proposed the "piecewise linear segment lattice (PLSL)" model[8].

A spoken word sample is modeled by dividing it into several segments, each of which is represented by regression coefficients of feature vectors within the segment, that is, \(\{a_k(k = 1, .., K), b_k(k = 1, .., K)\}\) in the following equation:

\[
\hat{y}_k(t) = a_k(i, j)(x(t) - \bar{x}(t)) + b_k(i, j) \tag{1}
\]

where \(\hat{y}_k(t)\) is the least square estimation of the \(k\)-th component of feature vector \(y(t)\) at the \(t\)-th frame, \(x(t) = S \cdot t\) (\(S\) is a constant), and \(\bar{x}(t)\) is a mean value of \(x(t)\).

An initial word model of PLSL is obtained by bundling the models of the samples which belong to the same word.
(Fig.1(a)). The lattice of a word model is then transformed to a more phone-like structure by matching and aligning between the sequences of the segments.

The optimum segmentation of each sample which minimizes the total distortion within the sample can be efficiently calculated using a dynamic programming (DP) procedure, if the number of divisions is fixed. The proper number of divisions is determined as the number \( N \) which minimizes a modified AIC criterion. Matching distance is defined as the total distortion of a sample with a PLSL, which can also be efficiently calculated using DP.

The PLSL model has an ability to represent objects with arbitrary precision. And compared with typical stochastic models, PLSL has the following advantages:

1) model parameters can be stably estimated with less samples,
2) its structure can be dynamically changed with less calculation,
3) hierarchical structures of different precision can be consistently integrated within a lattice.

All these computational characteristics are crucial points to derive phone-like structures.

2.2. Segmentation of Samples

We have examined this model by applying it to speaker independent isolated word recognition tasks. The spoken word samples used here consist of 1542 Japanese words uttered once by 10 male speakers. In preliminary experiments, we used a reduced set of 492 words selected from the vocabulary.

The model is trained with the samples from 9 speakers, and tested with the samples from the other one speaker (i.e. speaker independent). The feature vectors consist of 12 cepstral coefficients and a log-power, at 5ms intervals.

First of all, each sample in the training set is segmented into a sequence of piecewise linear segments as described in the previous section. In this sample set, the average number of segments per sample is 12.1, and the average length of a segment is 72.4 ms.

Then the unstructured multi-template models are constructed as shown in Fig.1(a). Each sample in the testing set is recognized by matching it with these models using the DP beam search. The recognition rate of 84.2\% for the 492 word is saturated at a beam width of 128. Accordingly, we use this beam width in the following experiments. The recognition rate for the 1542 word with this model is 64.3\%.

2.3. Alignment and Reduction

The above model is unstructured and not suitable for organizing phone-like structures. By aligning the segments in each word model, this model has rudimentary structures. One sequence selected from the training set is used for a reference pattern of each word model. All the other samples of that word in the training set are segmented
by aligning them to the reference. The number of segments, thus become the same for each word (Fig.1(b)). The recognition rate for the 492 words with this model is 80.7%.

When the matching path for the input sample is allowed to cross over the segments in the different sequences of the original samples in this model (b), the recognition rate for the 492 words is 91.5%.

In the next step, all the segments which are aligned at the same position in the model (b) are bundled and reduced into a single segment. Reduction is performed by calculating mean vector of the segments at the same position, and then a sequence of the segments is derived for each alignment group shown in Fig.1(c). The recognition rate is 94.1% for the 492 words, and 91.5% for the 1542 words. In order to obtain more stable value of feature vectors of the segments, it is possible to go back to the alignment step and iterate the processes until they saturate.

2.4. VQ Codebook of Segments

We try to generate VQ codebook of segment patterns by clustering all segments in all of the word models. The clustering is applied not only to the model (c) but also to (a) and (b) for comparison. Table 1 shows the recognition rates according to the number of clusters.

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>( \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition (a)</td>
<td>71.8</td>
<td>81.1</td>
<td>82.9</td>
<td>83.3</td>
<td>84.2</td>
</tr>
<tr>
<td>(b)</td>
<td>68.3</td>
<td>79.3</td>
<td>80.3</td>
<td>83.3</td>
<td>80.7</td>
</tr>
<tr>
<td>492 words(%)</td>
<td>89.6</td>
<td>94.3</td>
<td>93.7</td>
<td>96.1</td>
<td>94.1</td>
</tr>
<tr>
<td>1542 words(%)</td>
<td>-</td>
<td>83.2</td>
<td>80.6</td>
<td>90.8</td>
<td>91.5</td>
</tr>
</tbody>
</table>

Table 1: Recognition results by clustering all of the segments in all the models. (\( \infty \): without clustering)

These results show that the multi-template model of blocking models (c) is more robust to reduction of codebook size, compared with the rudimentary models like (a) and (b), and that these models yield sufficient performance by using as many number of units as ordinary phonetic units.

3. EXTRACTING PHONOLOGICAL CHUNKS

In the next step, we attempt to extract phonological chunks from the sequences of VQ codes which are generated in the experiments above. Because one of the most essential features of phonemes is their description of words, we determine a criterion of extracting chunks based on the feature that it appears frequently within the same word compared with the distribution in the whole vocabulary set. According to this argument, we determined the criterion as the following equation based on the Kullback-Leibler divergence between the distribution of each code within a word and that in the whole words.

\[
\rho = \sum_{w \in W} \sum_{c \in C} P_w(c) \log \frac{P_w(c)}{Q(c)}
\]

where \( W \) is a vocabulary set, \( C \) is a set of sequences of codes, \( P_w(c) \) is the existence probability of segment pairs in a word model \( w \) which have the same code \( c \), and \( Q(c) \) is the existence probability of segments of code \( c \). When estimating \( P_w(c) \), pairs of segments within a sample are not counted, whereas pairs of segments in different samples are counted.

The set \( C \) which maximizes \( \rho \) can be extracted by selecting sequences \( c \) according to the value of \( \kappa(c) \):

\[
\kappa(c) = \sum_{w \in W} P_w(c) \log \frac{P_w(c)}{Q(c)}
\]

We conducted experiments to extract chunks by applying the criterion to the VQ code sequences in the models which are generated in the previous subsection. We use two codebooks \( B_1 \) and \( B_2 \) which are generated from the models (a) and (c) respectively. Their codebook sizes are both 128. Table 2 shows the top 5 chunks for each sequence length from 2 to 5. In this table, the chunks are transcribed with the demiphonemic labels[9] instead of code number, considering human intuitive perception.

This result shows that \( B_2 \) generally exceeds \( B_1 \) on both criterion \( \kappa \) and on occurrences. The reason for this is considered to be that the feature of segments within a word is normalized by alignment and blocking processes.

In order to improve the proposed criterion, the following points must be considered:

1) Integrating the lengths of chunks with the criterion.
2) Integrating the distances between codes, except for Hamming distance used here.

For item 2), the following equation is an example of a distance definition:

\[
P_w(c) = \frac{1}{N_S} \sum_{s \in S} e^{-\frac{d(s, \pi(s))^2}{2\sigma^2}}
\]

where \( S \) is a set of segments within a model of word \( w \). \( \pi(s) \) is a code of a segment \( s \), \( d(x, y) \) is a distance between the centroids of code \( x \) and \( y \), and \( \sigma \) is a constant based on the variance of the code.

We expect that this criterion has the potential to improve the accuracy of extraction.

The feasibility of these chunks must be investigated by applying them to speech recognition tasks. In this case, the
word models are constructed by integrating the selected chunks. Because the chunks do not necessarily cover the segments in an entire word sample, we may need to develop a partial matching method, which is, for example, incorporating a ghe model inserted between successive chunks and matches any patterns.

4. CONCLUDING REMARKS

We have reported on the method and experiments extracting phonological chunks. The experimental results are summarized as follows:

1) The recognition rate is significantly improved by alignment and reduction of the segments which are at the same position in the time axis.

2) Reduced segment models yields sufficient performance with 128 phone-like units without using any transcriptional knowledge.

3) Phonological chunks are extracted based on the Kullback divergence.

Application of the extracted chunks to speech recognition tasks remain for future work. We also continue to investigate the optimal segmentation method which integrates temporal division and clustering, and the optimal method of extracting chunks.

Table 2: Extracted chunks (Top 5). (κ: criterion, “Within a word” and “whole words”: number of occurrences)

<table>
<thead>
<tr>
<th>Length</th>
<th>Codebook ( B_1 ) : model (a)</th>
<th></th>
<th>Codebook ( B_2 ) : model (c)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extracted chunks</td>
<td>( \kappa )</td>
<td>Within a word</td>
<td>Whole words</td>
</tr>
<tr>
<td>2</td>
<td>silE3-NN</td>
<td>2.133</td>
<td>278</td>
<td>501</td>
</tr>
<tr>
<td>2</td>
<td>silE3-eh</td>
<td>1.241</td>
<td>214</td>
<td>304</td>
</tr>
<tr>
<td>3</td>
<td>qg2-er</td>
<td>1.108</td>
<td>165</td>
<td>273</td>
</tr>
<tr>
<td>4</td>
<td>N#-silE3</td>
<td>1.070</td>
<td>141</td>
<td>370</td>
</tr>
<tr>
<td>5</td>
<td>silE3-re</td>
<td>0.811</td>
<td>132</td>
<td>242</td>
</tr>
<tr>
<td>3</td>
<td>qe#-po-silE3</td>
<td>0.341</td>
<td>51</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>oo-qg2-er</td>
<td>0.201</td>
<td>29</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>qg2-er-silE3</td>
<td>0.191</td>
<td>28</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>qg2-shik-silE3</td>
<td>0.159</td>
<td>23</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>tt-qg2-er</td>
<td>0.145</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>silE3-eh-ch-ia-#di</td>
<td>0.030</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>t#-qg2-er-silE3</td>
<td>0.027</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>ia-qch-po-silE3</td>
<td>0.025</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>oo-qg2-er-silE3</td>
<td>0.023</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>silE3-hit-qch-jo</td>
<td>0.023</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>qg2-a#-do-hnk#-di</td>
<td>0.012</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>silE3-eh-ch-lu-qg2-shik</td>
<td>0.012</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>#a-e#2-tt-qg2-er</td>
<td>0.012</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>silE3-re-#a-e#2-en2</td>
<td>0.012</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>silE3-NN-qg4-qh3-qh</td>
<td>0.011</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

REFERENCES