

Mixed-Lingual Spoken Word Recognition by Using VQ Codebook Sequences of Variable Length Segments

Hiroaki KOJIMA

Kazuyo TANAKA

National Institute of Advanced Industrial
Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
hkojima@ni.aist.go.jp

University of Tsukuba / AIST
1-2, Kasuba, Tsukuba,
Ibaraki 305-8550, Japan
kaz.tanaka@aist.go.jp

Abstract

We are investigating unsupervised phone modeling. This paper describes a derivation method of VQ codebook sequences of variable length segments from spoken word samples, and also describes evaluation results by applying the method to mixed-lingual speech recognition tasks which include non-native speakers. The VQ codebook is generated based on a piecewise linear segmentation method which includes segmentation, alignment, reduction and clustering processes. Derived codebook sequences are evaluated by speaker independent recognition of a word set which is a mixture of English and Japanese word. Speech samples are uttered by both English and Japanese native speakers. The recognition rates of mixed-lingual 618 words by using a codebook consist of 128 codes are 89.7% for English native speakers and 79.4% for Japanese native speakers in average .

1. Introduction

The ultimate goal of this work is to develop a robust and efficient speech processing method. Recent progress of stochastic speech recognition systems based on HMMs and statistic language models achieve practical recognition performance for formally uttered typical speeches. But robustness to varieties of speakers, for example non-native speakers, still remains as a hot issue. Our studies deal with mixed-lingual speech recognition which includes non-native speakers. The approach is to utilize a language independent coding system based on intermediate phonetic codes[4][5]. The codes have intermediate preciseness between frame based VQ codes and orthographic or phonetic units. The advantages of incorporating intermediate codes are efficiency in matching process and effective adaptation to variation of utterances. Related researches focusing on unsupervised derivation of variable length sub-phonetic segments are for example:[1] [2][3], and focusing on mixed-lingual or multi-lingual speech processing are for example: [6][7][8][9]. Other researches trying to utilize universal common phonetic codes are for example: [10][11].

2. Generation of Codebook

Methods of generating intermediate codes can be categorized as follows: top-down, bottom-up and bi-directional.

Top-down method generates codes by breaking down each conventional phonetic unit into more precise intermediate codes. Bottom-up method generates codes inductively from samples. This paper mainly describes the bottom-up method. This method is motivated by the fact that a human infant seems able to achieve robust speech recognition by acquiring knowledge of his/her native phonological system properly, compared with HMM based speech recognizers. The task is to form sub phonetic units from spoken word samples without using any transcriptions except for the identification of each word in a lexicon[12]. Bi-directional method which integrates the aforementioned two methods is available to map the codes into conventional phonetic units.

3. Variable Length Segmentation

We use VQ codebook sequences of variable length segments as intermediate codebook sequences. In order to implement the bottom-up generation method, we adopt the “*piecewise linear segment lattice (PLSL)*” model as a framework for codes representation[13]. A spoken word sample is modeled by dividing it into several segments, each of which is represented as regression coefficients of feature vectors within the segment, that is, $\{a_k(k = 1, \dots, K), b_k(k = 1, \dots, K)\}$ in the following equation.

$$\hat{y}_k(t) = a_k(i, j)(x(t) - \overline{x(t)}) + b_k(i, j)$$

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

The feature vectors used in experiments consist of 12 cepstral coefficients and a log-power, at 5ms intervals.

An initial word model of PLSL is obtained by bundling the models of the samples which are belong to

the same word (**Fig.1(a)**). The lattice of a word model is then transformed to be a more phone-like structure by matching and aligning between the sequences of the segments.

The optimum segmentation of each sample which minimize the total distortion within the sample can be efficiently calculated using a dynamic programming (DP) algorithm, if the number of division is fixed. The optimum division into N segments is calculated with the following recurrent formulas as $g(N, J)$ where J is the length of a sample by frames.

$$\begin{aligned} g(1, j) &= d(1, j) \\ g(n, j) &= \min_i [g(n-1, i) + d(i, j)] \quad (\text{if } n > 1) \end{aligned}$$

The distortion within a segment from the i -th frame to the j -th frame in a sample is defined as follows:

$$d(i, j) = \frac{1}{K} \sum_{t=i}^j \sum_{k=1}^K (y_k(t) - \hat{y}_k(t))^2$$

The proper number of divisions is determined as the number N which minimize the following criteria. Assuming distributions of residual vectors $\mathbf{y}(t) - \hat{\mathbf{y}}(t)$ as a uniform normal distribution of variance Σ , the AIC (An information Criterion) and MDL (Minimum Description Length) are described as follows:

$$\begin{aligned} l_{AIC} &= \frac{1}{2|\Sigma|} g(N, T) + K \cdot N \\ l_{MDL} &= \frac{1}{2|\Sigma|} g(N, T) + K \cdot N \cdot \log T \end{aligned}$$

(Items independent from division are omitted.)

We modify these criteria as follows so that the number of division can be controlled arbitrary with the parameter α . (β is determined as the values L_{AIC} and L_{MDL} to be approximately equivalent.)

$$L_{AIC} = g(N, T) + \alpha \cdot N \quad (1)$$

$$L_{MDL} = g(N, T) + \alpha \cdot N \cdot \beta \cdot \log T \quad (2)$$

The matching score (distance) of a speech sample with a PLSL is defined as the total distortion of the sample with a sequence of segments within the PLSL by searching for the optimum division of the sample. This can also be efficiently calculated using DP.

The PLSL model has an ability to represent objects in arbitrary precision. 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 structure of different precision can be consistently integrated within a lattice.

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

4. Derivation Process of Codebook Sequences

4.1. Segmentation of Samples

First of all, each training sample is segmented into a sequence of piecewise linear segments by using the method described in the previous section. By using a set of sequences of segments from each word as a multi-template model, an unstructured initial model is generated as is shown in **Fig.1(a)**. Each sample in a test set is recognized by matching it with these sequences of segments using the DP beam search.

4.2. Alignment

Since the above models are unstructured and also the parameters of each model too much depends on each speaker, we attempt to organize phonetic structures by aligning the segments in each word model in the next step. 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 pattern. The number of segments, thus become the same for each word according to the initial segmentation of the reference pattern, as is shown in **Fig.1(b)**.

4.3. Reduction

In the next step, all the segments 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 aligned at the same position, and then a single sequence of the segments is derived from each aligned model. Consequently, as same number of sequences as initial models are generated by changing the reference pattern of aligned models for each word, as is shown in **Fig.1(b)** and **Fig.1(c)**.

4.4. Iteration

In order to maximize the fitness of the models, an iteration process as same as a maximum likelihood segmentation process is incorporated. Following to the reduction process, alignment process is performed by using each sequence of the reduced model as a reference pattern. Then, the aligned models are reduced again by calculating mean vectors of the segments in the same position.

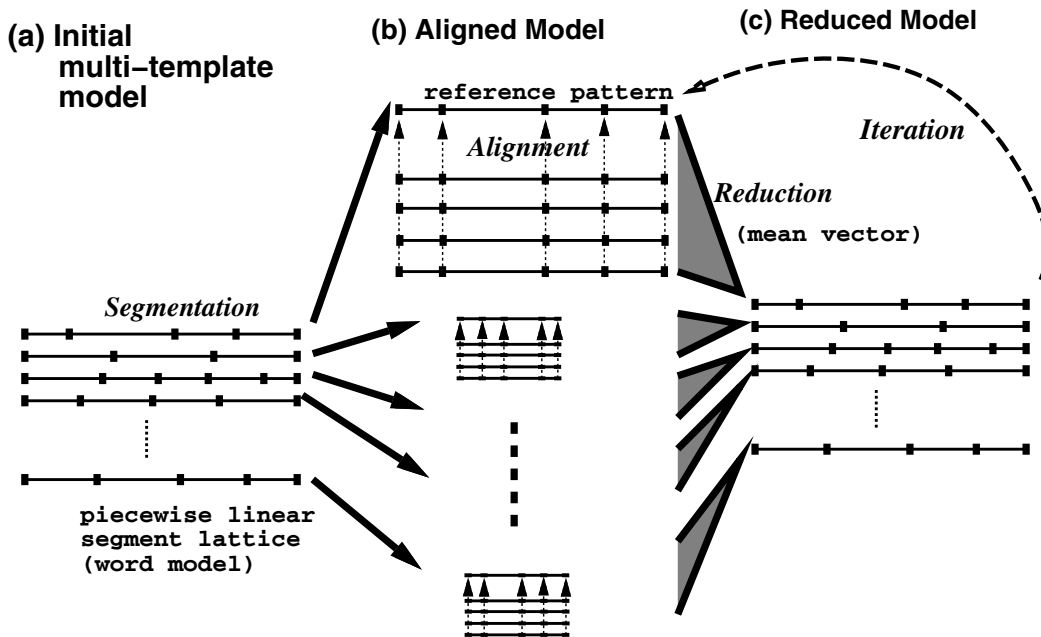


Figure 1: Derivation process of variable length segments.

4.5. Clustering into Codes

Intermediate phonetic codes are generated by clustering all the segments in all of the word models. The matching score of the input pattern with the clustered models is calculated by using the centroid values of the codes which correspond to the segments.

5. Experimental Results

5.1. Japanese Word Recognition

We have examined this model by applying it to speaker independent isolated word recognition tasks in order to evaluate this approach. At first we examined feasibility of the method by using Japanese word sets of 492 words and 1542 words. Each word samples are uttered once by 10 Japanese male speakers. The model is trained with the samples from 9 speakers, and tested with the samples from the other one speaker.

In this 1542 words corpus, the average number of initial segments per sample is 14.11, and the average length of a segment is 78.3ms using the AIC based criterion. By using the MDL based criterion, they are 14.13 and 78.2ms, accordingly.

By applying the iteration process, the recognition rates increase and saturate as is shown in **Table 1**.

The recognition results by using the models (a), (b), (c) and (c.2), and the results by applying clustering to each of them are shown in **Table 2**. ((c.2) stands for the models generated after 2 iteration process.) All of the models are AIC based in this experiment, .

These results show that the multi-template model of

Number of iterations		0	1	2
Recognition Rates of 492 words(%)	AIC based	94.1	98.2	98.4
	MDL based	93.9	94.1	94.7

Table 1: Recognition results by applying iteration process

Number of Codes		64	128	256	512	∞
Recognition Rates (%) of 492 words	(a)	71.8	81.1	82.9	83.3	84.2
	(b)	68.3	79.3	80.3	83.3	80.7
	(c)	89.6	94.3	93.7	96.1	94.1
	(c.2)	90.7	94.1	96.5	96.5	98.4
1542 words	(c)	—	85.2	89.6	90.8	91.5

Table 2: Recognition results by clustering all of the segments in all the models. (∞ : without clustering)

reduced models (c) is more robust to reduction of the codebook size by clustering, compared with the rudimentary models like (a) and (b), and that the (c) model keeps sufficient recognition rate by reducing the number of codes into 128.

5.2. English and Japanese Mixed Word Recognition

We also examined the method with a word set which is a mixture of English and Japanese words. The word set consist of 311 English word of station names in Britain and 307 Japanese words of company names in Japan.

Each word sample is uttered once by 7 English native

speakers and 7 Japanese native speakers. So, this corpus includes non-native speech. The model is trained with the samples from 12 speakers, and tested with the samples from the other two speakers.

Two kinds of methods in the alignment process are compared: “Common Alignment” and “Separate Alignment.” In the former, the alignment is performed without distinguishing English native samples and Japanese native samples. In the latter, the alignment is performed separately for English native samples and Japanese native samples.

In this experiment, samples from English native speakers and Japanese native speakers are recognized by the (c) model. The recognition rates, which are separately calculated for English word and Japanese word, are shown in **Table 3**.

Language	Native	Number of Codes				without Clustering
		64	128	256	512	
Common Alignment						
Ew	En	81.4	83.6	86.5	88.4	89.1
	Jn	62.1	70.7	76.2	73.3	71.1
Jw	En	91.9	95.8	95.1	94.1	91.2
	Jn	89.9	88.3	86.0	86.3	81.4
Separate Alignment						
Ew	En	81.0	84.2	88.4	88.1	90.7
	Jn	49.5	56.6	60.1	61.1	59.8
Jw	En	92.2	96.1	95.1	95.4	92.2
	Jn	64.5	63.8	61.9	64.2	76.9

Table 3: Mixed-lingual recognition results ([%], Ew:English word, Jw: Japanese word, En: English native speaker, Jn: Japanese native speaker)

The results show that this model also keeps sufficient recognition rate by reducing the number of codes into 128 even though including non-native speech. The recognition rates of 618 words in average using 128 codes from common alignment model are 89.7% for English native speakers and 79.4% for Japanese native speakers. The two kinds of alignment methods are approximately equivalent in this experiment. Compared with the recognition rates of Japanese words uttered by English native speakers, the rates of English words uttered by Japanese speakers are significantly lower. This suggests that English word samples uttered by Japanese native speakers are widely distributed.

6. CONCLUSION

This paper has described a bottom-up method to derive intermediate codes based on VQ codebook sequences of variable length sub-phonetic segments, and also reported experimental results of mixed-lingual speech recognition to evaluate the models. The results show that this

model keeps sufficient recognition rate by reducing the number of codes into 128 even though the target samples including non-native speech. As another project, we have started to apply the intermediate codebook models to speech data retrieval system[14]. For future work, we try to improve the method by constructing phonetically structured by extracting chunks from the sequence of codes[15], and by integrating top-down and bottom-up methods.

7. References

- [1] B. S. Atal, “Efficient coding of LPC parameters by temporal decomposition”. in *Proc. ICASSP83*, pp. 81–84, 1983.
- [2] S. Deligne, F. Yvon and F. Bimbot, “Variable-length sequence matching for phonetic transcription using joint multigrams”, in *Proc. EUROSPEECH 95*, pp. 2243–2246, 1995.
- [3] M. Bacchiani, M. Ostendorf, et al., “Design of a speech recognition system based on acoustically derived segmental units,” in *Proc. ICASSP-96*, 1996, pp. 443–446.
- [4] K. Tanaka and H. Kojima, “A speech recognition method with a language-independent intermediate phonetic codes,” in *Proc. ICSLP-2000*, 2000.
- [5] K. Tanaka and H. Kojima, “Between-word distance calculation in a symbolic domain and its applications to speech recognition,” *Information Sciences*, vol. 123, no. 1-2, pp. 25–41, Mar 2000.
- [6] L. Lamel and J. Gauvain, “Cross-lingual experiments with phone recognition,” in *Proc. ICASSP-93*, 1993, vol. II, pp. 507–510.
- [7] F. Weng and H. Bratt, “Study of multilingual speech recognition,” in *Proc. Eurospeech’97*, 1997, pp. 359–362.
- [8] P. Morin, T. Applebaum, R. Boman, Y. Zhao, and J.-C. Junqua, “Robust and compact multilingual word recognizers using features extracted from a phoneme similarity front-end,” in *Proc. ICSLP-98*, 1998, vol. 2, pp. 377–380.
- [9] T. Fegyó and P. Tatai, “Multi-lingual speech recognition based on demi-syllable subword units,” in *Proc. Eurospeech’99*, 1999, vol. 2, pp. 871–874.
- [10] T. Schultz and A. Waibel, “Fast bootstrapping of LVCSR systems with multilingual phoneme sets,” in *Proc. Eurospeech’97*, 1997, pp. 371–374.
- [11] F. Palou, P. Bravetti, O. Emam, V. Fischer, and E. Janke, “Towards a common alphabet for multilingual speech recognition,” in *Proc. ICSLP-2000*, 2000.
- [12] H. Kojima and K. Tanaka, “Formation of phonological concept structures from spoken word samples,” in *Proc. ICSLP-92*, 1992, pp. 269–272.
- [13] H. Kojima and K. Tanaka, “Organizing phone models based on piecewise linear segment lattices of speech samples,” in *Proc. EuroSpeech’97*, 1997, pp. 1219–1222.
- [14] K. Tanaka and H. Kojima, “Speech data retrieval system constructed on a universal phonetic code domain,” in *Proc. ASRU2001*, 2001.
- [15] H. Kojima and K. Tanaka, “Extracting phonological chunks based on piecewise linear segment lattices,” in *Proc. ICSLP-2000*, 2000.