Design of Ready-Made Acoustic Model Library by Two-Dimensional Visualization of Acoustic Space

Goshu Nagino and Makoto Shozakai

Information Technology Laboratory
Asahi Kasei Corporation
{g-nagino, makoto}@ljk.ag.asahi-kase.co.jp

Abstract

This paper proposes the technique enabling a design of ready-made library composed of high performance and small size acoustic models utilizing the method of visualizing multiple HMM acoustic models onto two-dimensional space (“COSMOS” method: aCOustic Space Map Of Sound), and providing one of these models without overburdening users. The acoustic space (as expressed in multi-dimensional future parameters) is partitioned into zones on two-dimensional space, allowing for the creation of highly precise acoustic models through the generation of acoustic models for respective zones of the acoustic space. A set of these acoustic models is called an acoustic model library. In an experiment of this paper, a plotted map (called the COSMOS map) featuring a total of 145 male speakers speaking in various styles was generated utilizing the COSMOS method. Through the COSMOS map, the distribution of each speaking styles was analyzed, thereby demonstrating the effectiveness of the COSMOS method in the analysis of acoustic space. The COSMOS map was then partitioned into concentric acoustic space zones to produce acoustic models representing each acoustic space zones. By selecting the acoustic model providing maximum likelihood score effectively using voice samples consisting of 5 words, the acoustic model, even if expressed in single Gaussian distribution, showed high performance comparable to speaker-independent acoustic model (called SI-model) expressed in 16 mixture Gaussian distributions. Furthermore, the acoustic model showed performance higher than SI-model adapted with voice samples of 30 words by the MLLR [2] method.

1. Introduction

Practical application of Automatic Speech Recognition (ASR) is increasing in embedded appliances such as vehicle navigation systems, personal digital assistants and cellular phones. Speaker-independent acoustic model (SI-model) are often implemented for these applications. However, such software is hardly universal in providing high speech-recognition performance for every user. In addition, as for such embedded appliances, processing power and available memory are generally restricted at a point of view of a cost. And not only ASR but also other application is operating in same appliance. Therefore, the number of parameters in SI-model should be reduced, and the acoustic model cannot demonstrate enough capability. Although the performance can be improved through the use of speaker-adaptive techniques such as the MAP [1] and the MLLR [2] methods, these require a significant quantity of voice samples in order to take effect. It is therefore important to substantiate compact acoustic models without compromising on performance and also to develop the technique enabling the provision of these models with a small quantity of voice samples to each user.

This paper proposes the technique enabling the design of ready-made acoustic model library utilizing the COSMOS method that visualizes multiple acoustic models onto a two dimensional map of acoustic space. The proposed method aims to establish and provide acoustic model library enabling highly accurate speech recognition for embedded applications without a burden to user.

2. COSMOS Method

In order to reduce the size of acoustic models while efficiently improving their accuracy, it is important to grasp the whole aspect of the target acoustic space represented by the acoustic model. The multi-dimensional scaling (MDS) method [3] utilizing the visual mapping of multi-dimensional information onto a lower order space consisting of two or three dimensions is extremely effective in enhancing the perceptibility of a multi-dimensional acoustic space. Without exception, the techniques shown in [3] utilize two-dimensional projections of the multi-dimensional vector information, and thus are useless in the mapping of information consisting of multi-dimensional Gaussian distributions. The technique based on the Principal Component Analysis (PCA) [4] does suggest a method of mapping acoustic models onto a two-dimensional plane by making use of primary and secondary components for concatenated vectors configured by the mean characteristic vectors of the acoustic models. However, the cumulative proportion (about 9%) of the primary and secondary components is significantly lower than 80% which is acknowledged as standard cumulative proportion for the PCA. The resulting scattered diagram can hardly be considered as an accurate reproduction of the spatial information of the original multi-dimensional Gaussian distribution. A procedure for mapping information containing multi-dimensional Gaussian distribution onto two-dimensional planes without causing losses in information needs to be devised.

The COSMOS method proposed handles the acoustic models as an approximated expression of the acoustic space representing a large amount of speech data. The method is proposed as an extension of the Sammon method [5] enabling a non-linear projection of aggregated acoustic models onto two-dimensional space.

2.1. Formulation

The Sammon method is a technique of non-linear projection featuring the optimization of mapped position coordinates
within lower order dimensions by the steepest descent method, thereby minimizing the difference \( E(t) \) between the summation of the mutual distances among the multi-dimensional information existing within the multi-dimensional space and the summation of the mutual Euclidean distances of the mapped position coordinates existing in the lower order space. The error function \( E(t) \) to be minimized is obtained from formula (1) below:

\[
E(t) = \frac{1}{\sum_{i \neq j} \left( \sum_{d \neq y} \left( d^*_{ij} - d_{ij} \right)^2 \right)}
\]

(1)

In general, an acoustic model is a generic designation for an aggregation consisting of multiple HMMs of acoustic units. The distance \( d^*_{ij} \) between acoustic models \( i \) and \( j \) is defined by formula (2) below:

\[
d^*_{ij} = \frac{1}{K} \sum_{k=0}^{K-1} d(i, j, k) \ast w(k)
\]

(2)

Whereas, \( w(k) \) signifies the weighted parameter for the acoustic unit \( k \), to be designated voluntarily depending on the purpose. Parameter \( d(i, j, k) \) signifies the mutual distance between the acoustic unit \( k \) within acoustic models \( i \) and \( j \). If HMM of the acoustic unit \( k \) of acoustic models \( i \) and \( j \) obeys mixed Gaussian distribution of the same structure, \( d(i, j, k) \) can be defined using formulas (3) - (6):

\[
d(i, j, k) = \frac{1}{S(k)} \sum_{s=0}^{S(k)-1} \frac{1}{L(j)} \sum_{l=0}^{L(j)-1} \frac{1}{M_i} \sum_{m_i=0}^{M_i-1} \frac{1}{M_j} \sum_{m_j=0}^{M_j-1} \text{pp}(i, j, k, s, l, m_i, m_j) \cdot \text{p}(i, k, s, l, m_i) \cdot \text{p}(j, k, s, l, m_j)
\]

(3)

\[
\text{dd}(i, j, k, s, l) = \sum_{w=0}^{M_i-1} \sum_{m_i=0}^{M_i-1} \text{p}(i, k, s, l, m_i) \cdot \text{p}(j, k, s, l, m_j)
\]

(4)

\[
c(i, j, k, s, l, m_i, m_j) = \frac{\left( \mu(i, k, s, l, m_i) - \mu(j, k, s, l, m_j) \right)^2}{\sigma(i, k, s, l, m_i) \cdot \sigma(j, k, s, l, m_j)}
\]

(5)

\[
\text{pp}(i, j, k, s, l) = \sum_{w=0}^{M_i-1} \sum_{m_i=0}^{M_i-1} \text{p}(i, k, s, l, m_i) \cdot \text{p}(j, k, s, l, m_j)
\]

(6)

\[
\mu(i, k, s, l, m), \sigma(i, k, s, l, m) \text{ and } p(i, k, s, l, m) \text{ respectively denote mean, variance and weight for the } m \text{ th Gaussian distribution of the acoustic model } i, \text{ the acoustic unit } k, \text{ the state } S \text{ and dimension } l. S(k) \text{ represents the number of states for the acoustic unit } k. L \text{ stands for the number of dimensions. } M_i \text{ and } M_j \text{ indicates the number of Gaussian mixtures of acoustic models } i \text{ and } j .
\]

In this paper, the acoustic parameters consist of 10 MFCCs, 10 delta MFCCs, and 1 delta energy. Therefore, \( L = 21 \). An acoustic model is composed of diphone-level HMM based on single Gaussian distribution in order to reduce the required processing power and memory consumption. Then, formula (3) through (6) are simplified as formula (7) below:

\[
d(i, j, k) = \frac{1}{S(k)} \sum_{s=0}^{S(k)-1} \frac{1}{L(j)} \sum_{l=0}^{L(j)-1} \frac{1}{M_i} \sum_{m_i=0}^{M_i-1} \frac{1}{M_j} \sum_{m_j=0}^{M_j-1} \left( \frac{\mu(i, k, s, l) - \mu(j, k, s, l)}{\sigma(i, k, s, l) \cdot \sigma(j, k, s, l)} \right)^2
\]

(7)

The resulting two-dimensional representation itself is called the COSMOS map, while the proposed mapping method is referred to as the COSMOS method. Respective acoustic model projected onto the COSMOS map is called STAR.

### 3. Analysis of speaking style

#### 3.1. Database

Variability based on the speaking styles was investigated using the COSMOS method. 145 Japanese males uttered a list of 176 words taken from the phoneme balanced word set (called ATR5240 in Japan) in the three speaking styles indicated in Table 1. The speech data was overlaid with background noise recorded at an exhibition hall at a Signal-to-Noise ratio of 20 dB. Sampling frequency is 11.025kHz.

#### 3.2. Visualization

All speech data (with the exception of the evaluation speaker’s data) was used in the training of a speaking-style-independent/speaker-independent acoustic model (SI-model). Then, this acoustic model was employed in turn as an initial model to retrain the speaking-style-dependent/speaker-dependent acoustic models (SD-model) using the Baum-Welch algorithm. Figure 1 shows the COSMOS map for the SD-model STARs. In this situation, weight \( w(k) \) for the acoustical unit \( k \) in formula (2) represents the frequency of occurrence of the diphones in the training-data. Respective STAR symbols correspond to Table 1. Within Figure 1, the distribution of the STARs indicates a distinct concentration of each group of speaking styles. This is considered to suggest the difference of acoustic features among respective speaking styles. The dotted arrow on the COSMOS map denotes the difference of acoustic features among each speaking style of the same speaker. It is suggested that the variance of speaking style is more significant than the variance of speaker.

In Figure 2, closed evaluation based on the SI-model was conducted for all speakers (open evaluation for evaluation speakers), symbolizing the STARs of speakers with recognition performance below 80% as “×” and those of speakers with recognition performance above 80% as “○”. According to Figure 2, speakers resulting in low recognition performance tend to be distributed in the peripherals of the COSMOS map.

<table>
<thead>
<tr>
<th>Name of speaking style</th>
<th>Instructions provided for sampling</th>
<th>Symbol of STAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Read utterance list at normal speed of conversation.</td>
<td>□</td>
</tr>
<tr>
<td>Fast</td>
<td>Read utterance list at faster than normal speed of speech.</td>
<td>□</td>
</tr>
<tr>
<td>High</td>
<td>Read utterance list at higher than normal tone of speech.</td>
<td>□</td>
</tr>
<tr>
<td>Whisper</td>
<td>Read utterance list at a level not to be overheard by near-by persons.</td>
<td>□</td>
</tr>
<tr>
<td>Loud</td>
<td>Read utterance list at a level to be heard by persons at some distance.</td>
<td>□</td>
</tr>
<tr>
<td>Lombard</td>
<td>Read utterance list among an ambient car noise.</td>
<td>□</td>
</tr>
<tr>
<td>Syllable enhanced</td>
<td>Read utterance list by enhancing the Japanese syllables.</td>
<td>□</td>
</tr>
</tbody>
</table>
4. Speaker clustering

Conventionally, speaker clustering method [6], EigenVoice (EV) method [7], cluster adaptive learning method [8] are proposed as procedures to segment acoustic space through clustering speaker-dependent models, these methods are very effective. However, none of the available methods of clustering is sufficiently effective for speakers whose recognition performance is low. From the previously described argument, it may be concluded that 1) acoustic models with similar characteristics are positioned at a relatively short distance to each other within the two-dimensional space, and 2) within a given distribution, the STARs with low recognition performance are located in the peripherals of the distribution on the COSMOS map. The second result also implies that additional efforts to enhance the performance are necessary, concerning the speakers located in the peripherals of the distribution on the COSMOS map.

Based on these considerations, speakers are clustered on the two-dimensional space obtained by the COSMOS method.

4.1. Clustering on the COSMOS map

As the STARs are congregated in a roughly circular configuration as is evident in Figures 1 and 2, with the requirement for precision at the peripherals of the distribution, the segmentation should be based on polar coordinates with finer partitioning to the peripherals of the Cosmos map. The segmentation reduces variations of acoustic features, resulting in sufficient representational capability even for a single Gaussian distribution.

Strips located at a given radius from the origin are called BANDs, and segments with a certain angular breadth as indicated in Table 2 are referred to as ZONEs. Each ZONE is partitioned into overlapping sectors. Setting of ZONEs in Table 2 depends on design guideline of an acoustic model library to be built. By using this clustering method, it becomes possible to correspond to diverse users with a special focus on speakers with relatively low recognition performance. Furthermore, this method of clustering facilitates the correspondence to speakers situated at the cluster boundary, which places the method at an advantage over the tree-structured speaker clustering method [6].

Additionally, the generation of new COSMOS map is possible by emphasizing the weight variable \(w(k)\) of acoustic unit difficult to be recognized. The emphasis is based on the frequency of occurrence for acoustic units contained in words with low recognition accuracy through the closed evaluation of training speakers. The newly generated COSMOS map shall also be subjected to the clustering procedure indicated above. This clustering pays attention to speakers who have lower recognition performance. A total of 134 ZONEs were thus generated. The acoustic models are constructed for each ZONEs. The set of these acoustic models are called the acoustic model library.

4.2. Selection of the acoustic model

As denoted in Figure 3, the acoustic models were searched in correspondence to the polar coordinate system. Starting with BAND_0 situated at the center of the distribution, model selection procedure is continued until the final BAND to find the acoustic model with highest likelihood score to the obtained voice samples as indicated in Figure 3.

5. Evaluation

Four evaluation speakers, two from central area and two from marginal area, of the distribution for each speaking style were selected. A total of 28 evaluation speakers were thus selected. The evaluation was executed with HTK [9], utilizing a parallel network consisting of the 176 words contained in the vocabulary for respective subjects of the evaluation.
The evaluation of the acoustic model library was conducted on the evaluation data obtained. The acoustic model selected from the acoustic model library according to the procedure described in section 4.2 is referred to as customized acoustic model. The performance of these customized acoustic models was compared to the baseline performance achieved by SI-model and the speaker adaptive acoustic models generated by a conventional speaker adaptive method, that is MLLR [2]. As it was considered appropriate to utilize the vocabulary words actually used in tasks during practical application for the selection process of the customized acoustic models, and also for the speaker adaptation process by the MLLR method, the voice samples for the selection and the adaptation were randomly selected from the evaluation data.

5.1. Results of experimentation
Recognition performance was compared for each subjects of evaluation by SI-model (baseline), speaker adaptive acoustic model (10, 30 samples for the adaptation – MLLR_N10, 30) and the customized acoustic models (5 samples for the selection – Custom_N5). The results of the comparison are as indicated in Figure 4. The average number of phonemes per utterance was 14.25. According to Figure 4, the performance of Custom_N5 was significantly higher than the results for SI-model, and showed a performance equivalent to or higher than MLLR_N30. SI-model is generally expressed in mixture of Gaussian distributions. Figure 5 provides a comparison of the recognition performance of Custom_N5 to SI-models expressed in 1 (baseline), 4 (Mix=4), 8 (Mix=8) and 16 (Mix=16) mixtures of Gaussian distribution. According to the results of the comparison, Custom_N5 showed a performance equivalent to those of the SI-model expressed in Mix=16 as regard to the ratio of speakers indicating recognition rate of less than 80%. Furthermore, the requirement for processing power for the model selection by likelihood evaluation with all acoustic models was reduced by approximately 80% by the model selection with employing the procedure described in section 4.2.

6. Conclusion
The COSMOS method of visualizing the acoustic space expressed by HMM acoustic models enables the intuitive perception of the vastness and structure of the acoustic space, demonstrating the effectiveness of the method in the analysis of the acoustic space. Speaker clustering based on the COSMOS method showed a recognition performance significantly superior to SI-model and speaker adaptive acoustic models with the same topology, and showed a performance equivalent to those of the SI-model expressed in 16 mixtures of Gaussian distributions. The proposed design method of ready-made acoustic model library was eminently effective in improving the recognition performance for speakers with recognition performance of less than 80% by conventional SI-model, and reduced the error rate by approximately 45% on average for the subjects of evaluation. Although the number of acoustic models constructed in this technique is excessively large, the appropriate procedure for selecting the acoustic model in this paper reduces drastically processing power of model selection. It may safely be concluded that the technique of speaker clustering based on the COSMOS method is capable of providing high performance and small size acoustic model without overburdening the user, even in the case of embedded applications with restrictions in the available memory and processing power.

7. References