BUILDING AN EFFECTIVE CORPUS BY USING ACOUSTIC SPACE VISUALIZATION (COSMOS) METHOD

Goshu Nagino and Makoto Shozakai

Information Technology Laboratory, Asahi Kasei Corporation
Atsugi AXT Main Tower 22F, Okada 3050, Atsugi, Kanagawa, 243-0021, Japan
{g-nagino, makoto}@ljk.ag.asahi-kasei.co.jp

ABSTRACT

This paper proposes the technique of building an effective corpus with lower cost by using the method of visualizing multiple HMM acoustic models into a two-dimensional space ("COSMOS" method: COmprehensive Space Map of Objective Signal, previously aCOustic Space Map Of Sound) method. In an experiment of this paper, adapted acoustic models of 533 male speakers are made with a small quantity of voice samples (10 words) per each speaker. Then a plotted map (called COSMOS map) featuring a total of 533 male speakers is generated utilizing the COSMOS method. A corpus was built by selecting 200 male speakers located only in the periphery of the distribution in the COSMOS map and by collecting voice samples (165 words) per each speaker. The acoustic model trained from the corpus showed higher performance than the one trained from other corpus built with 200 male speakers selected randomly in the COSMOS map or all of 533 male speakers in the COSMOS map.

1. INTRODUCTION

Practical application of Automatic Speech recognition (ASR) is accelerated in embedded appliances such as vehicle navigation systems, personal digital assistants and humanoid robots. Speaker-independent acoustic model (SI-model) is often implemented for these applications. However, building a large-scale corpus is indispensable for training of SI-model to consume enormous cost. For example, the cost of collecting voice samples per each speaker is about 400US$ and the cost increases in proportion to the number of the speakers.

Generally, speakers are often selected without any inspection for building corpus. However, the corpus might include a lot of speakers having similar acoustic features each other. Once voice samples of statistically enough number of speakers in an acoustic space have been already collected, even if voice samples of new speaker located in the same acoustic space are collected additionally, they will not contribute to improving the speech recognition performance. It wastes a cost idly. It is possible to build a corpus with lower cost if there is a technique of selecting the speakers that contribute to improving the speech recognition performance with lower cost before collecting voice samples with high cost. In addition, it is expected that the corpus built by selecting the speakers that contribute to improving the speech recognition performance is more effective than the database of the speakers selected randomly in terms of speech recognition performance.

In this paper, we describe the COSMOS (COmprehensive Space Map of Objective Signal, previously aCOustic Space Map Of Sound) method [1][2] that visualizes multiple acoustic models into a two-dimensional map of the acoustic space in Section 2. In Section 3, we describe the technique of analyzing the acoustic space that contributes to improving the speech recognition performance by using the COSMOS method. We propose the technique of building a corpus with lower cost by using the COSMOS method in Section 4. In Section 5, we discuss our summary.

2. COSMOS METHOD

The multidimensional scaling (MDS) method [3] featuring a visual mapping of multidimensional information onto a lower order space consisting of two or three dimensions is extremely effective in enhancing the perceptibility of the 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] suggests a method of mapping acoustic models onto a two-dimensional space 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 required 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 needs to be devised for mapping information containing multi-dimensional Gaussian distribution onto a two-dimensional space with as minimal information loss as possible.

The proposed COSMOS method [1][2] handles the acoustic models as an approximated expression of the acoustic space representing a large amount of speech samples. The method enabling a nonlinear projection of aggregated acoustic models onto two-dimensional space is proposed as an extension of the Sammon method [5].

2.1. Formulation

The Sammon method is a technique of nonlinear projection of multidimensional vectors featuring the optimization of the mapped coordinates within two-dimensional space by the

steepest descent method, thereby minimizing the error function E_m between the summation of the mutual distances $D(i,j)$ among the multidimensional vector i and j existing in the higher order space and the summation of the mutual Euclidean distances $D_m(i,j)$ of the mapped coordinates of the multidimensional vector i and j at m th iteration of the steepest descent method. The error function E_m to be minimized is obtained from formula (1) and (2) below;

$$E_m \equiv \frac{1}{c}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\left\{D(i,j)-D_m(i,j)\right\}^2 / D(i,j)\right] \quad (1)$$

$$c \equiv \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D(i,j) \quad (2)$$

In general, an acoustic model is a generic designation for an aggregation consisting of multiple models of acoustic units (diphone). Accordingly, the mutual distance $D(i,j)$ between acoustic model i and j is defined by the following;

$$D(i,j) \equiv \sum_{k=1}^{K} d(i,j,k) * w(k) \Big/ \sum_{k=1}^{K} w(k) \quad (3)$$

Here, $d(i,j,k)$ denotes the mutual distance between the acoustic unit k within the acoustic model i and the acoustic unit k in the acoustic model j. $w(k)$ represents occurrence frequency for the acoustic unit k. K indicates total number of acoustic units. Respective acoustic model projected onto the COSMOS map is called STAR.

The following is applied to diphone acoustic models based on HMM having single Gaussian distribution per state in order to reduce the required processing power and memory consumption. Although publicly acknowledged distance measures might be applicable, the Euclidian distance of mean vectors normalized by standard deviation vectors shall be adopted as $d(i,j,k)$ within this paper. Assuming all acoustic models share a common topology with one-on-one state alignment between respective acoustic models, $d(i,j,k)$ may be expressed using the following equation;

$$d(i,j,k) \equiv \frac{1}{S(k)}\sum_{s=0}^{S(k)-1}\frac{1}{L}\sum_{l=0}^{L-1}\frac{\left(\mu(i,k,s,l)-\mu(j,k,s,l)\right)^2}{\sigma(i,k,s,l)*\sigma(j,k,s,l)} \quad (4)$$

$\mu(i,k,s,l)$ and $\sigma(i,k,s,l)$ denote the mean value and the standard deviation value of the single Gaussian distribution at dimension l for the state s of the acoustic unit k within the acoustic model i. $S(k)$ represents the number of states of the acoustic unit k. L signifies the dimension size of the acoustic feature. In this study, the acoustic features consist of 10 MFCCs, 10 delta MFCCs and 1 delta energy. Therefore, L equals 21. It is possible to expand formula (4) to mixed Gaussian distribution [2].

3. ACOUSTIC SPACE ANALYSIS

3.1. Corpus

Two Japanese males uttered a list of 5240 words taken from the phoneme balanced word set (called ATR5240 in Japan). This database is used for a seed acoustic model (Seed-model) for re-training or adaptation. 561 Japanese males uttered a list of 175 words taken from ATR5240 in one of speaking styles indicated in Table 1. The data of 533 Japanese males are used for training and those of 28 Japanese males are used for evaluation. The speech data is overlaid with background noise recorded at an exhibition hall at a Signal-to-Noise ratio of 20 dB. Sampling frequency is 11.025kHz.

Table 1 : *Speaking Style*

Speaking style	Instructions provided for recording	STAR
normal	Read utterance list at normal speed of conversation.	×
fast	Read utterance list at faster than normal speed of speech.	Δ
high	Read utterance list at higher than normal tone of speech.	○
whisper	Read utterance list at a level not to be overheard by near-by persons.	●
loud	Read utterance list at a level to be heard by persons at some distance.	▲
Lombard	Read utterance list among an ambient car noise.	□
syllable enhanced	Read utterance list by enhancing the Japanese syllables.	■

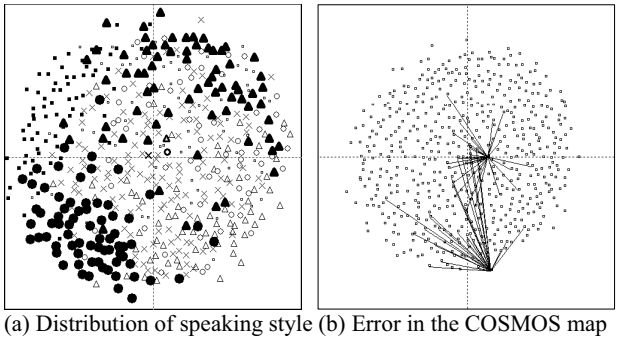
3.2. Visualization

Speaker-dependent acoustic models (SD-models) are retrained using the EM algorithm based on the Seed-model. Figure 1(a) shows the COSMOS map where the SD-models are mapped by using the COSMOS method. In this situation, $w(k)$ in formula (3) represents occurrence frequency of the diphone k in the training data. Respective STAR symbols correspond to Table 1. Figure 1(b) shows an example of a mapping error of the COSMOS method. In Figure 1(b), one STAR of speaker located in the center of the distribution and one STAR of speaker located in the periphery of the distribution in the COSMOS map are connected with 30 neighborhood speakers in the original multidimensional space respectively. The STAR of speaker located in the periphery of the distribution in the COSMOS map is connected not only with neighborhood STARs of speakers located in the periphery of the distribution in the COSMOS map, but also with the STARs of speakers located in the center of the distribution in the COSMOS map. It means that the COSMOS map has mapping error and that the STARs of speakers located in the periphery of the distribution in the COSMOS map have acoustic features of the STARs of speakers located both in the periphery and in the center of the distribution in the COSMOS map. We have already found three characteristics of the COSMOS map. First of all, the models that have similar acoustic features are located close to each other. Secondly, as for the STARs of speakers located in the center of the distribution in the COSMOS map, the acoustic features tend to be the average of those of the STARs of all speakers and the acoustic features change continuously as the positions change from the center to the periphery. Finally, the STARs of speakers that show lower performance seem to be located in the periphery of the distribution more often than in the center of the distribution in the COSMOS map. Therefore, we suppose that the acoustic space of the periphery of the distribution in the COSMOS map contributes to improving the speech recognition performance.

3.3. Evaluation

In this section, we prove the supposition in Section 3.2. We evaluate the performance of three corpora. Each corpus consists of N speakers, 1)who are selected randomly, 2)who are located in the center of the COSMOS map and 3)who are located in the periphery of the COSMOS map. In an experiment of this paper, N is set to be 100, 150, 200, 250 and 300. The corpus including speakers located in the center or periphery of the distribution in the COSMOS map is called *Center* or *Periphery* respectively. The corpus including speakers selected randomly is called *Random*. The corpus including all 533 training speakers (N=533) is called *Baseline*. Next, acoustic models are retrained with each corpus. The evaluation is executed with HTK, utilizing a parallel network consisting of the 175 words contained in the vocabulary of voice samples of each speaker for evaluation.

Figure 2 shows that the corpus *Periphery* has higher performance than the corpus *Random* and the corpus *Center* and the corpus *Baseline*. The similar result is obtained in the corpus optimization technique with PCA [4]. Thus, grasping an acoustic space contributing to improve the speech recognition performance is effective to optimize the corpus and to improve the speech recognition performance. When using the COSMOS method optimizes the corpus, an acoustic space that contributes to improving the speech recognition performance means a periphery of the distribution in the COSMOS map.



(a) Distribution of speaking style (b) Error in the COSMOS map

Figure 1 : *Speaking style COSMOS map*

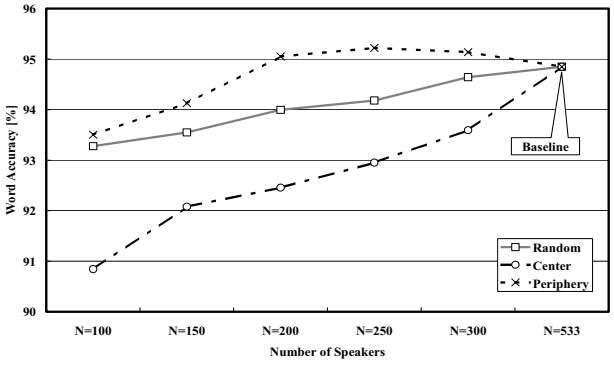


Figure 2 : *The relation between a size and a performance of the corpus (1)*

4. BUILDING A CORPUS

4.1. Proposed method

According to the result in Section 3.3, we propose a technique of building a corpus effectively by using the COSMOS method. In this section, the technique of selecting speakers that contributes to improving the speech recognition performance with a small quantity of voice samples is described. The Block diagram of the proposed method is shown in Figure 3.

At first, a small quantity of voice samples of the speaker is collected as indicated in Block A. The vocabulary depends on a task. In an experiment of this paper, it is assumed that the cost of collecting voice samples in Block A is smaller enough than that of collecting voice samples in Block E. In Block B, Seed-model is adapted using the MLLR method [6] with a small quantity of voice samples. The adapted acoustic model (Adapt-model) is made as an approximation model of SD-model. Block A and B are performed for enough quantity of speakers. In Block C, all Adapt-models are mapped onto a two-dimensional space by using the COSMOS method. Then, $w(k)$ in formula (3) represents the occurrence frequency of the diphones k in the task vocabulary. In Block D, speakers located in the periphery of the distribution in the COSMOS map are selected. Finally, in Block E, the corpus is built by collecting large enough quantity of voice samples of the selected speakers.

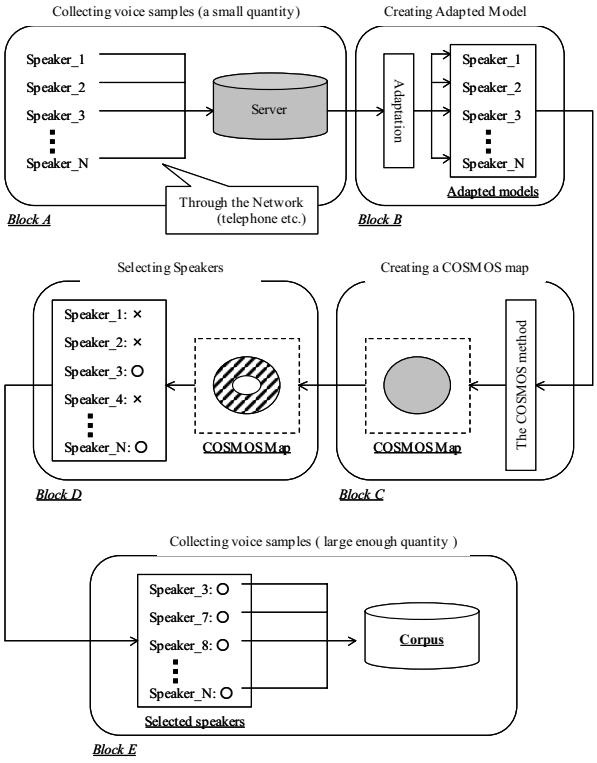


Figure 3 : *The Block diagram of building a corpus*

4.2. Evaluation

An experiment is performed according to the Figure 3. In Block A, voice samples of 10 words for each 533 male speakers are collected. In Block E, the number of the voice samples to be collected is 165 words excluding 10 words already collected in Block A. The COSMOS map of 533 Adapted-models generated in Block C is called Adapted-model COSMOS.

Here, we investigate whether it is possible to select speakers that contribute to improving the speech recognition performance by the Adapted-model COSMOS. Figure 4 shows two kinds of the COSMOS map. Figure 4(a) shows the COSMOS map where the SD-models retrained in Section 3.2 are mapped (called SD-model COSMOS), and Figure 4(b) shows the Adapted-model COSMOS. In Figure 4(a), the STARs of speakers located in the periphery of the distribution in the COSMOS map are symbolized as "×", and in Figure 4(b) the corresponding STARs of speakers are also denoted as "×". As can be seen, although there is a difference between the two COSMOS maps, we can see that many of the STARs of speakers located in the periphery of the distribution in Figure 4(a) are located in the periphery in Figure 4(b) as well. Therefore, it is expected that the Adapted-model that is adapted with only a small quantity of voice samples (10 words) is an effective model to judge whether the speaker is located in the acoustic space that contributes to improving the speech recognition performance or not.

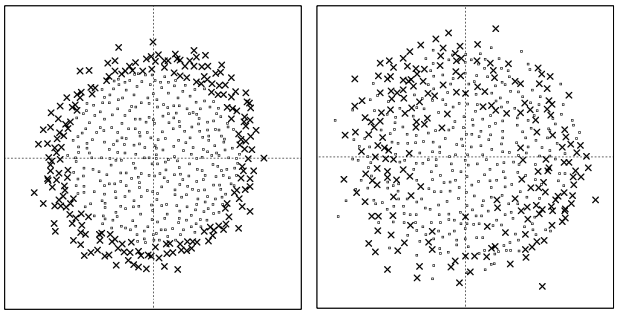
We evaluate the performance of the corpus built by using the proposed method. The database is called *Proposed*. It is compared to two corpora, *Random* and *Periphery*, built in Section 3.3. *Random* is evaluated as baseline and *Periphery* as an upper limit in this experiment. *Periphery* means the optimized corpus built with the SD-model COSMOS in which the SD-models of the speakers having more significant acoustic information are mapped.

Figure 5 shows a relation between a size of the corpus and a performance for each corpus. The performance means the word accuracy of the acoustic model that is expressed in 8 mixture Gaussian distributions and trained with each corpus. In Figure 5, it is shown that the performance of *Proposed* is equivalent to that of *Periphery* and is better than that of *Random*. In addition, *Proposed* built with voice samples of 200 speakers shows higher performance than the same as *Baseline*. It means that more than 60% of cost reduction is realized. As the number of speakers N in the corpus is increasing in the range of more than 200, the difference of the performance between *Proposed* and *Periphery* is decreasing. The reason is that the number of speakers collected in Block A is limited to 533 in this experiment. In the practical use, the total number of speakers collected in Block A is unlimited. Therefore it is expected that the difference of performance between *Proposed* and *Periphery* will increase more and more.

5. CONCLUSION

In this paper, we proposed the technique of building a corpus with lower cost. At first, we described that the COSMOS method was effective to analyze an acoustic space that contributes to improving the speech recognition performance. In addition, we presented that the technique of collecting speakers by using the COSMOS method was effective to build a corpus that showed higher performance. As a result, we demonstrated that the proposed method was cost-effective to establish technology of building corpus

In this paper, we did not discuss about a procedure of collecting voice samples in Block A. Collection through a telephone is reasonable as collection procedure with lower cost. We will work on the case using telephone to collect voice samples to select speakers in order to establish practical efficiency of our proposed method.



(a) SD-model COSMOS (b) Adapted-model COSMOS

Figure 4 : *Comparison of the distribution*

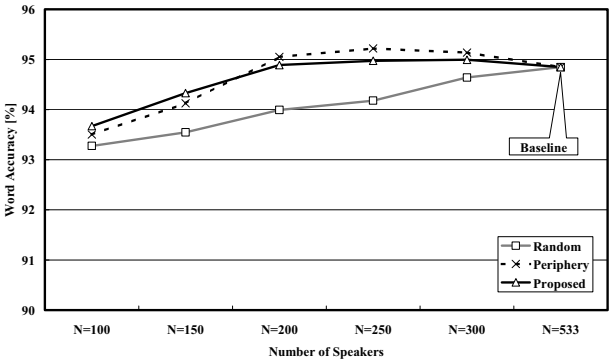


Figure 5 : *The relation between a size and a performance of the corpus (2)*

6. REFERENCES

[1] M. Shozakai et al., "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," Proc. ICSLP, vol.1, pp.717-720, 2004.

[2] G. Nagino et al., "Design of ready-made acoustic model library by two-dimensional visualization of acoustic space," Proc. ICSLP, vol.4, pp.2965-2968, 2004.

[3] A. K. Jain et al., "Statistical pattern recognition: a review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp.4-37, 2000.

[4] A. Nagorski et al., "Optimal selection of speech data for automatic speech recognition system, " Proc. ICSLP, vol.4, pp.2473-2476, 2002.

[5] J. W. Sammon, "A nonlinear mapping for data structure analysis, " IEEE Trans. Computers, vol. C-18, no.5, pp.401-409, May 1969.

[6] C. J. Leggretter et al., "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models, " Computer Speech and Language, vol.9, pp.171-185, 1995.