How to Judge Reusability of Existing Speech Corpora for Target Task by Utilizing Statistical Multidimensional Scaling

Goshu Nagino [1,2], *Makoto Shozakai* [1], *Kiyohiro Shikano* [2]

[1] Speech Solutions, New Business Development, Asahi Kasei Corporation
[2] Graduate School of Information Science, Nara Institute of Science and Technology
{nagino.gb, shozakai.mb}@om.asahi-kasei.co.jp, shikano@is.naist.jp

10.21437/Interspeech.2007-400

Abstract

In order to develop a target speech recognition system with less cost of time and money, reusability of existing speech corpora is becoming one of the most important issues. This paper proposes a new technique to judge the reusability of existing speech corpora for a target task by utilizing a statistical multidimensional scaling method. In an experiment using twelve tasks in five speech corpora, our proposed method could show high correlation to the cross task recognition performance and judge the reusability of existing speech corpora correctly for the target task with lower cost.

Index Terms: statistical MDS, reusability, acoustic model, task dependency

1. Introduction

Recognition accuracy is still extremely sensitive to environmental conditions such as the speaker characteristic, the speaking style, the background noise and the task domain. These issues are called a task dependency. The task dependency has strong impact on a recognition performance of the Automatic Speech Recognition (ASR) in embedded appliances such as car-navigation systems, personal digital assistants and robots. In these appliances, processing power and available memory size are generally restricted at a cost-conscious point of view, as not only the ASR but also other applications are operating on a same platform. In such a case, a number of parameters contained in an acoustic model should be reduced. That's why the acoustic model cannot demonstrate enough performance even if it is trained from a huge speech corpus covering various tasks. So, an acoustic modeling optimized for a target task is expected. In recent research [4], a task dependency and reusability of four speech corpora have been investigated by a cross task recognition experiment. We can select speech data, which has closer acoustic characteristics, through a speech recognition experiment with a few target task speech data (development data). However, no one can judge whether the selected speech data is enough and has high reusability for the target task. If there is a technique of a judgment of the reusability for the target task, we can judge what we should invest the ASR system in. For instance, a collecting target speech data, an acoustic modeling, a language modeling, an evaluation and a system maintenance.

In this paper, how to judge reusability of existing speech corpora for a target task is described. In an experiment, 12 tasks contained in 5 Japanese speech corpora are evaluated by a statistical multidimensional scaling (MDS) method called as COSMOS (COmprehensive Space Map of Objective Signal) method [5] that visualizes aggregate of speech data within two or three dimensional space. The visualization is acknowledged as an effective technique to grasp the multidimensional space that humans cannot understand easily. It is expected that to comprehend the relationship between a target task speech data and existing speech corpora by using the visualization of their acoustic space is effective in order to analyze reusability of existing speech corpora.

In the next section, our proposed method is described. In Section 3, an overview of speech corpora is described. In Section 4, our proposed method is described and the effective ness is investigated trough a cross task recognition experiment. Finally, a summary and an outlook on a future work are given in Section 5.

2. Reusability Judgment

In this section, a new technique of applying a statistical multidimensional scaling method to judge reusability of existing speech corpus for a target task is described.

2.1. Framework

In Figure 1, a block diagram of proposed method is shown. At first, a small quantity of voice samples of a target task is collected as indicated in Block C. The amount of voice samples is T seconds. In this Block, the number of speakers is N. In Block D, the target task speaker dependent acoustic model called as target TSD-model is trained with the small quantity of voice samples. In Block A, N speakers are selected from existing speech corpora. The number of the existing speech corpora is M. As for selecting the speakers, there are several approaches such as a random selection and some speaker clustering methods. In Block B, a TSD-model is trained with T seconds voice samples which are selected at random from all voice samples for the each existing speech corpus and for each speaker. Block E has three steps. At first, distances among of all TSD-models of the target task and the existing speech corpora are computed. Then, a total number of TSD-model is $(M+1)\times N$ and the total number of a combination of distance between two TSD-models is $\{(M+1)\times N\}^2$. Next, the TSD-models are visualized by utilizing a conventional MDS method with the distance between TSD-models. Finally, the reusability of the existing speech corpora for the target task is judged from the visualized map. If a distribution of the target TSD-models are covered with the TSD-models trained from the existing speech corpora, the existing speech corpora have high reusability. Then, the target task (speaker independent) acoustic model will be trained with the existing speech corpora. As for a training method, the speaker adaptive training [6] and the selective training [7] are expected effective. On the other hand, if a distribution of the TSD-models trained from the existing speech corpora are separated from that of the target TSD-models, the existing speech corpora has low reusability. In latter case, it is necessary to collect a large amount of voice samples of the target task with a consuming cost. A technique of building an effective target speech corpus with lower cost consumption by utilizing statistical MDS method was already proposed [8].

2.2. Statistical multidimensional scaling

The multidimensional scaling (MDS) method [9] featuring a visual mapping of multidimensional information onto visible space of low order (one to three) dimension is extremely effective in enhancing perceptibility of multidimensional data space such as acoustic space encompassing speech data. As an extension of the conventional MDS, a statistical MDS handling statistical model such as GMM and HMM which are assumed as an approximated expression of the multidimensional data space represented by the speech corpus was proposed [5].

2.2.1. Formulation

In the Sammon method [10], which is one of the conventional MDS, the error function E_m in formula (1) is minimized iteratively by the steepest descent method. $D(i,j)$ denotes mutual distance between the vector i and j existing in higher order dimensional space. $D_m(i,j)$ denotes mutual Euclidean distances of the mapped lower order coordinates of the vector i and j at m th iteration. Generally, initial mutual distance $D_0(i,j)$ is computed from the initial position given by random value.

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

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

Here, vectors are replaced by statistical acoustic models based on HMM. In general, an acoustic model is a generic designation for an aggregation consisting of multiple models of acoustic units. Accordingly, the mutual distance $D(i,j)$ between acoustic model i and j is defined by the following:

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

Here, $d(i,j,k)$ denotes mutual distance between the acoustic unit k within the acoustic model i and the acoustic unit k within the acoustic model j. $w(k)$ represents weight value such as an occurrence frequency for the acoustic unit k. K is the total number of acoustic units. Assuming all acoustic models ($i = 1,\ldots,N$) share a common topology with one-on-one state alignment between respective acoustic models, $d(i,j,k)$ can be defined using formula (4).

$$d(i,j,k) \equiv \frac{1}{S(k)}\sum_{s=0}^{S(k)-1}\frac{1}{L}\sum_{l=0}^{L-1} dd(i,j,k,s,l) \qquad (4)$$

$S(k)$ represents the number of states in acoustic unit k. L stands for the dimension of acoustic feature parameters. In this paper, the Bhattacharyya distance [11] is adopted as the distance $dd(i,j,k,s,l)$ between Gaussian distributions.
This statistical MDS is called COSMOS (COmprehensive Space Map of Objective Signal) method. And, the resulting visualized map itself is called COSMOS map.

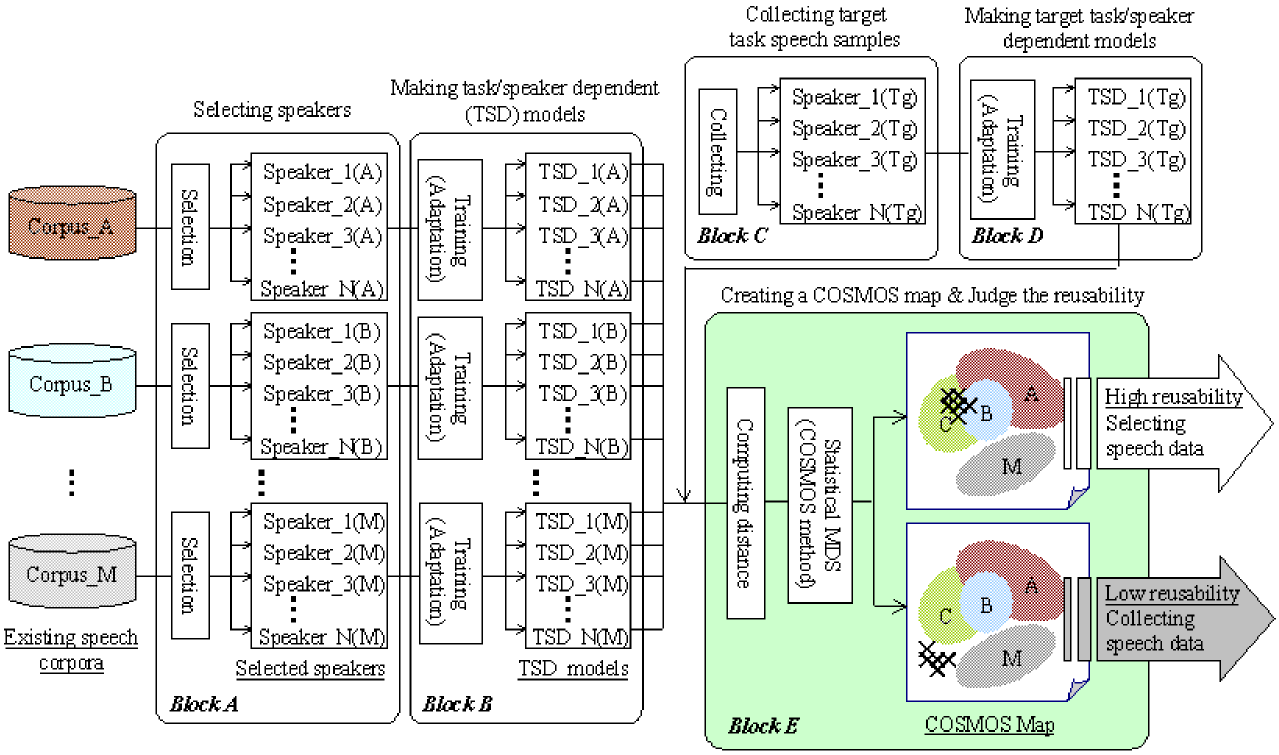

Figure 1: *Block Diagram of Reusability Judgment.*

2.3. Judgment

Reusability of existing speech corpora for a target task is judged based of a distance between a distribution of the target task and that of the each existing speech corpora on a COSMOS map. For each existing speech corpora, a two-dimensional single Gaussian distribution is defined by mapped position of the TSD-models. The Bhattacharrya distance measure is utilized for a distance between two Gaussian distributions. And, a threshold of the distance makes a decision of reusing the existing speech corpora or collecting the target task speech data.

3. Speech Corpora

In this section, an overview of the 5 Japanese speech corpora evaluated in this paper is described. Each speech corpus consists of several tasks. Total number of tasks is 12.

1) “JNAS” contains utterances of a phoneme-balanced continuous word sentence (*Jb*) and a newspaper (*Jn*).
2) “S-JNAS” contains utterances of a phoneme-balanced continuous word sentence (*SJb*), a newspaper (*SJn*) and an information retrieval (*SJi*) uttered by elder persons.
3) “CIAIR” contains utterances of a phoneme-balanced continuous word sentence (*Cdb*) and a dialog while driving in a car. In the dialog, a speaker talks with the ASR (*CdA*), a human (*CdH*), the wizard of OZ system (*CdW*). And, the task domain is an information retrieval.
4) “ATR-APP” contains of a phoneme-balanced continuous word sentence (*Ab*) recorded in several areas in Japan.
5) “CSJ” contains utterances of a continuous word sentence uttered by a natural speaking style in an academic meeting (*Ca*) and a lecture speech (*Cs*).

Each utterance is recorded in a clean room or with using a close-to-talk microphone. And, speakers are male only. Table 1 shows a data size of each task. 15 males are selected as an evaluation speaker at random in each task. In same speech corpus, one speaker utters often several tasks. A speaker selected as the evaluation speaker in one task is removed from training speakers in other task. Sampling frequency is 16kHz. This paper evaluates the reusability of the existing speech corpora having more various acoustic characteristics. Reusability of existing speech corpora containing 6 car navigation command tasks by utilizing this visualizing method was discussed in recent our work [12].

Table 1. *Data Size.*

Task	Speakers (male) train/total	Size [h] train/total
Jb	124/151	6.9/8.5
Jn	124/151	18.1/22.4
SJb	126/151	18.7/22.1
SJn	161/202	31.0/38.6
SJi	34/51	2.5/3.6
Ab	1364/1379	45.7/46.2
Cdb	261/314	5.1/6.1
CdA	245/297	1.9/2.3
CdW	247/298	3.2/3.8
CdH	258/310	4.8/5.7
Ca	785/804	70.8/72.9
Cs	774/805	57.6/62.7

4. Experiment

In this experiment, *Jn* as a dictation task, *CdH* as a natural dialog task in a car and *Cdb* as a general task in the car are evaluated as target tasks. First, reusability of existing speech corpora is analyzed by a cross task recognition experiment. Then, task matched recognition for a target task is performed to evaluate reusability of existing speech corpora. Next, in order to show an effectiveness of our proposed method, distributions of each existing speech corpora and the target task and the cross task recognition performance are compared. Here, an acoustic model structure is 43 mono-phones HMM containing 3 states. The acoustic feature parameters consist of 12 MFCCs, 12 delta-MFCCs and 1 delta-log power.

4.1. Cross task recognition

Every task dependent acoustic model is trained with task dependent training speech data. A number of Gaussian is 8. An evaluation network is a phoneme loop network. Table 2 shows a performance of a cross task recognition experiment. In task matched acoustic model (called TM-model) case, the performance is transcribed in a bold. In Table 2, a task dependency is clearly shown. The TM-model shows highest performance for task-matched case. As for reusability, *Jb* has higher reusability for *Jn*, *CdA* has higher reusability for *CdH*, *Jb* has higher reusability for *Cdb*. In case of the target task *Cdb*, however, a big performance gap between the TM-model (*Cdb*) and *Jb* which shows second highest performance. It suggests *Jb* doesn’t have higher reusability for *Cdb*. It seems that speech data in *Cdb* task should be newly collected to capture higher performance. However, it is not convincing to judge the reusability only by a cross task recognition experiment before building the target task speech corpus with lots of cost.

Table 2. *Cross Task Recognition Performance.*
(*Phoneme Accuracy [%]*)

Acoustic model	Target task		
	Target1(*Jn*)	Target2(*CdH*)	Target3(*Cdb*)
Jb	64.88	48.01	62.24
Jn	**65.44**	48.48	60.51
SJb	61.42	45.96	58.22
SJn	61.75	46.26	56.65
SJi	60.32	48.60	59.04
Ab	60.27	46.58	57.63
Cdb	58.54	48.48	**69.66**
CdH	45.87	**51.84**	52.93
CdW	44.91	51.49	52.05
CdA	46.22	51.52	55.02
Ca	51.13	44.94	50.95
Cs	54.80	45.86	49.68

4.2. Reusability Judgment

In Block C and D, 50 speakers utter 30 seconds voice samples in a target task. And, target TSD-models having single Gaussian per state are trained with these voice samples. The cost to collect voice samples is quite low. In Block A, 50 speakers are selected from each speech corpus among 11 tasks. As for a speaker selection, the speaker clustering method is applied. In Block B, TSD-models are trained with 30 seconds voice samples selected at random. In Block E, a distance between TSD-models of the target task and existing tasks is computed. Then, a total number of TSD-model is $(11+1)\times 50 = 600$. Next, these 600 TSD-models are visualized by utilizing our statistical MDS. Figure 2 shows visualized (COSMOS) map with these 600 TSD-models. A distribution of the target TSD-models are shown on the COSMOS map. In case of *Jn*, the distribution is almost covered with that of *Jb*. It means *Jb* has higher reusability for *Jn*. In case of *CdA*, the distribution is covered with that of *CdH*. It means *CdA* has higher reusability for *CdH*. In case of *Cdb*, the distribution is not covered with the distribution of TSD-models of existing tasks. It means no task has higher reusability for *Cdb*. Table 3 shows a distance between a distribution of each existing speech corpora and that of a target task. In higher reusability case, the distance is close. And, in lower reusability case, the distance is far.

A correlation between a performance and a distance is shown in Table 4. Then, the correlation is computed with only five closest tasks. For instance, in the target 1 (*Jn*) case, "*Ab*, *Jb*, *SJb*, *SJn*, *Cdb*" are used to compute a correlation. As for a judgment of reusability, a speech corpus having closer acoustic characteristics to a target task is important. In Table 4, a negative correlation is shown. And the correlation is high. In the target 2 (*CdH*) case, the correlation grows higher (from "-0.410" to "-0.827") by computing with four closer tasks.

Table 3. *Distance to Target Task*

Acoustic model	Target task		
	Target1(*Jn*)	Target2(*CdH*)	Target3(*Cdb*)
Jb	0.0375	3.1107	1.5989
Jn	-	2.5760	1.1198
SJb	0.4827	1.7491	3.5576
SJn	0.4471	1.7128	3.3964
SJi	1.3328	1.4825	5.6703
Ab	**0.0348**	2.0165	1.2296
Cdb	1.1198	1.3598	-
CdH	2.5760	-	1.3598
CdW	2.8814	**0.0023**	1.2826
CdA	2.4766	0.0581	**0.9084**
Ca	2.0636	0.2404	2.6823
Cs	2.4573	0.5596	3.8602

Table 4. *Correlation between Performance and Distance*

Target task		
Target1(*Jn*)	Target2(*CdH*)	Target3(*Cdb*)
-0.749	-0.410	-0.889

5. Summary and Future work

In this paper, we proposed a new technique of how to judge reusability of existing speech corpora for a new target task by utilizing a statistical MDS. In the experiment, our proposed method demonstrated more correctly the reusability of the existing speech corpora for the target task, which was not clearly shown by a cross task recognition experiment.

In our future work, we will investigate how to select proper data from existing speech corpora having high reusability and how to train a high performance acoustic model.

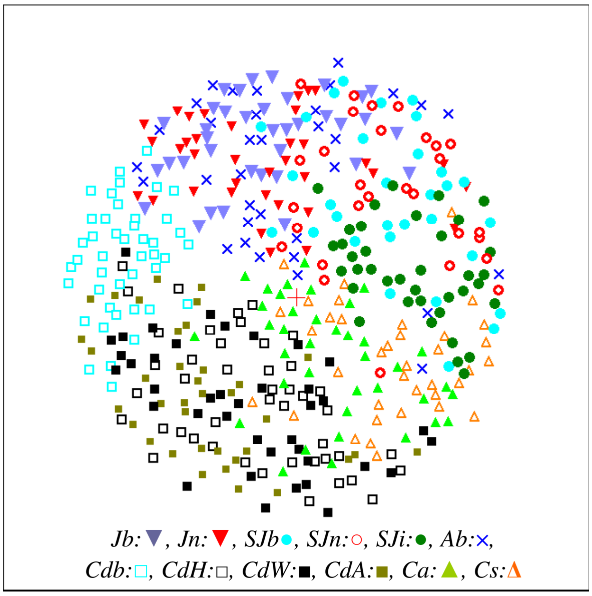

Figure 2: *Visualized (COSMOS) Map*

6. References

[1] 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.

[2] T.Kosaka et al., "Tree-structured speaker clustering for speaker-independent continuous speech recognition," Proc. ICSLP, pp.1375-1378, 1994.

[3] M. Shozakai et al, "A speech enhancement approach E-CMN/CSS for speech recognition in car environments," Proc. ASRU Workshop, pp.450-457, 1997.

[4] F. Lefeve et al., "Genericity and portability for tash-independent speech recognition," Computer speech and language 19, pp.345-363, 2005.

[5] M. Shozakai et al., "Acoustic space analysis method utilizing statistical multidimensional scaling technique," Proc. NSIP, May 2005.

[6] T.Anastasakos et al., "A compact model for speaker-adaptive training," Proc. ICSLP, pp. 1137–1140, 1996.

[7] T. Cincarek et al., "Selective EM Training of Acoustic Models based on Sufficient Statistics of Single Utterances," Proc. ASRU, pp. 168-173, 2005.

[8] G. Nagino et al., "Building an effective corpus by using acoustic space visualization (COSMOS) method," Proc. ICASSP, vol. I, pp.449-452, 2005.

[9] 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.

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

[11] K. Fukunaga, "Introduction to statistical pattern recognition (Second edition)," Academic Press, Inc., 1990.

[12] G. Nagino et al., "Analyzing Reusability of Speech Corpus Based on Statistical Multidimensional Scaling Method," Proc. ICSLP, pp.161-164, 2006.