Acoustic space partition based on broad phonetic class for ensemble acoustic modeling

Xugang Lu[1]*, Yu Tsao*[2]*, Shigeki Matsuda*[1]*, Chiori Hori*[1]*, Hideki Kashioka*[1]

1.National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academic Sinica, Taiwan

Abstract

Ensemble acoustic modeling can be used to model different factors that cause variabilities of acoustic space, and provide different combination to improve the performance of automatic speech recognition (ASR). One of the main concerns is how to partition the training data set to several subsets based on which ensemble models are trained. Traditionally, the acoustic space is partitioned based on speaker variability which is modeled using Gaussian mixture model (GMM) of global acoustic distribution. We argue that modeling on global acoustic distribution may not be accurate enough to catch the speaker variabilities. For example, speaker variabilities, such as gender and accent information, may be encoded in local acoustic realizations of a few specific phonetic classes. Based on this consideration, we proposed an acoustic space partition method based on broad phonetic class (BPC) dependent modeling of speakers for ensemble acoustic modeling. With the principal component analysis (PCA) of the BPC based speaker representation, we designed two level hierarchical data partitions in the low dimensional speaker factor space. Ensemble acoustic models were trained on the partitioned data sets on both levels. Speech recognition experiments showed that using the proposed partition method, we obtained 9.64% and 32.23% relative improvements in character error reduction rate on the first and second level partitions, respectively.

Index Terms: Ensemble modeling, acoustic space partition, speaker supervector, speaker clustering.

1. Introduction

Ensemble acoustic modeling can be used to model different factors that cause variabilities of acoustic space, and provide different combination to improve the performance of automatic speech recognition (ASR) [1, 2, 3, 4]. The basic processing is shown in Fig. 1. There are two important parts in this ensemble modeling, one is how to partition the training data set to several subsets according to hidden factors that cause the acoustic variabilities. Based on the partitioned data subsets, ensemble acoustic models can be trained (as acoustic model 1 (AM 1), acoustic model 2 (AM 2), etc.). The second concern is how to integrate the ensemble models for a combined one for final speech recognition. In this study, we mainly focus on the former problem. In addition, differently from data partition methods used in machine learning, such as bagging, boosting, we try to manipulate data partition concerned with the factors for speaker variabilty which cause variability in acoustic for Chinese large vocabulary continuous speech recognition (LVCSR).

This work is support by MASTAR project of National Institute of Information and Communications Technology, Japan

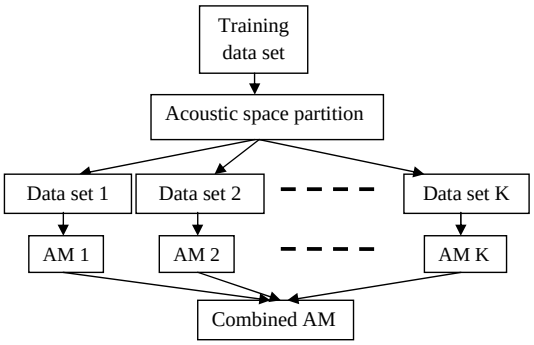

Figure 1: Ensemble acoustic modeling with data set partition and model combination.

A lot of factors cause speech acoustic variability including speakers' sex, accent, age, size, voice quality, speaking style or emotion. Among them, gender is one of the most important factors. Although accent is also identified as the second important factor [6], the accent actually has large uncertainty and is difficult to be labeled when collecting data corpus. Usually regional information is labeled which is correlated to accent information. The acoustic space partition method based on prior information of these important factors can be used to split the training data set, and train the ensemble acoustic models as those factor dependent models. However, the prior information, particularly the accent information, is difficult to be obtained or has large uncertainty if obtained. In this case, it is difficult to partition the data set for ensemble acoustic modeling.

Speaker or acoustic space clustering can be used to train multi-acoustic models in speech recognition without any concern of explicit physical meanings in modeling [5]. We argue that modeling the acoustic space with explicit physical meanings may help for a better clustering or partition of acoustic space, and improve the ASR performance. One interesting study concerned with accent speech modeling with explicit physical meaning was done in [6]. In their work, they analyzed the speaker variability factors concerned with gender and accent, and established gender and accent detection based speech recognition models. But in speaker modeling, they used a Gaussian mixture model (GMM) to model the global acoustic distribution without identifying the underlying phonetic context. As we know, in Chinese speech, the gender and accent information may be carried by several specific phonetic categories with local acoustic distribution. Gender and accent information is more easier to be identified with knowledge of phonetic context than that without phonetic context. Accurate phoneme based modeling can catch a lot of information, including linguistic information, as well as gender and accent information. However, in phoneme based modeling, a lot of training data is required.

And the model is not robust because there exists high confusion among different phoneme models.

Speaker variability may be resulted from different quality of voice or pronunciation of vowels and consonants, or realizing an utterance with different stress and prosody. This concerns with different manipulation of articulation organs that is closely relates to broad phonetic class (BPC), such as vowels, fricatives, etc. In addition, BPC based modeling is more robust than that of phoneme based modeling. Based on these considerations, we propose a speaker acoustic space partition method based on BPC dependent speaker modeling and clustering, and train ensemble acoustic models based on the partitions for Chinese LVCSR.

2. Speaker modeling based on broad phonetic class

Our speaker acoustic space partition method is inspired by the research of speaker recognition. Speaker variability is one of the most important aspects of acoustic variability. We want to partition the acoustic space based on speaker variability analysis. In order to analyze speaker variability, each speaker is represented using a speaker supervector. Differently from most studies in speaker modeling, we model each speaker with a set of BPC based GMM. In our study, the BPC is composed of vowels, stops, fricatives, nasals, and liquids which is classified based on their different articulation manners.

2.1. Broad phonetic class based speaker supervector

The extraction of speaker supervector is described in Fig. 2. The basic procedures are: (1) Training a GMM model using all

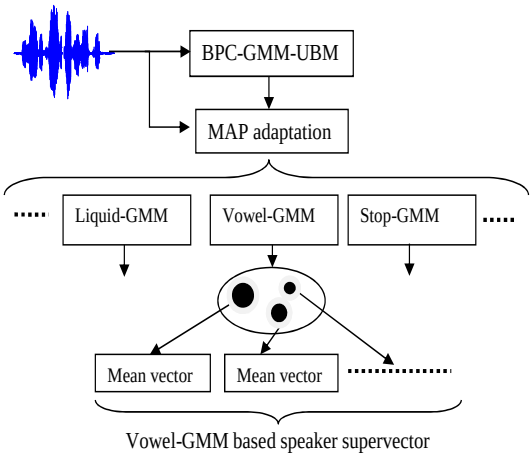

Figure 2: MAP adapted BPC-GMM speaker modeling.

training data sets for each BPC as a universal background model (UBM). After this processing, we have 5 GMM-UBMs; (2) Using maximum a posterior (MAP) adaptation method for each speaker data set, speaker BPC-GMMs are obtained; (3) Speaker supervectors are constructed as BPC dependent supervectors obtained by concatenating the mean vector of each GMM component [7].

2.2. Dimension reduction based on principal component analysis

Based on the processing in section 2.1, each speaker is characterized by a set of BPC supervectors. We may do speaker clustering for acoustic space partition based on the obtained supervectors. However, there are several disadvantages for clustering directly on the original supervectors. One is that there is no robust clustering method based on the high dimensional supervectors because of the curse of high dimensionality problem. The second is that it is difficult to identify the hidden factors with physical meanings from each speaker clusters. Based on this thinking, we do dimension reduction on the supervectors to find the variability of speaker acoustic space, and make data set partitions in a low dimensional space. Principal component analysis (PCA) is used in this study since PCA can find a low dimensional coordinate system corresponding to most of the maximum variations of the data distribution.

Suppose speaker supervector matrix is represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]$, where $\mathbf{x}_i \in R^d, i = 1, 2, ..., M$, d is the dimension of speaker supervector, and M is the total number of speakers. Then the eigenvalue decomposition of speaker covariance matrix is:

$$\begin{aligned} &\mathbf{C}_\mathbf{x}\mathbf{v}_i = \lambda_i \mathbf{v}_i; i = 1, 2, ..., M \\ &\mathbf{C}_\mathbf{x} \triangleq \frac{1}{M-1}\mathbf{X}\mathbf{X}^T \end{aligned} \quad (1)$$

In eq. (1), $\mathbf{C}_\mathbf{x}$ is speaker supevector matrix in which speaker supervectors are centered to be zero mean, T is a transpose operator, λ_i is the i-th eigenvalue corresponding to eigenvector $\mathbf{v}_i$. Then the speaker supervector can be projected on the first a few eigenvectors for dimension reduction as:

$$\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i, \quad (2)$$

where $\mathbf{x}_i$ is the supervector of speaker i, $\mathbf{W}$ is a matrix constructed from the first L eigenvectors of $\mathbf{C}_\mathbf{x}$ with top L eigenvalues. Then the new representation vector of a speaker is with L dimensions which is much smaller than the dimension of original supervector d. We can do factor analysis, data space partition and clustering in the low dimensional representation space.

3. Factor analysis and partition of speaker acoustic space

One Chinese accent speech data corpus is used in the factor analysis and partition. The data corpus is consists of 400 speakers with more than 160k utterances (200 speakers for male and female, respectively), totally about 180 hours acoustic speech. Based on speakers' regional information, they are classified into four groups, i.e., Beijing (BJ), Shanghai (SH), Canton (CT), and Taiwan (TW). Each group has 100 speakers (50 male and female speakers, respectively). For convenience of speech recognition experiments in later part, from the data set, we choose 8 speakers for testing (two speakers with one male and one female from each regional group were selected), and the left 392 speakers were used for analysis and training. Although all speakers are labeled with their regional information, their accents, however can not be classified based on this information. Based on listening test, there are large overlaps among different regional groups on their accents. It is difficult to partition the data set to train multiple accent dependent acoustic models according to their regional information.

3.1. Factor analysis of speaker acoustic space

We first analyze the factors that cause the variability of the speaker acoustic space on this data set based on the method introduced in section 2. Mel frequency cepstral coefficient (MFCC) feature is used in the analysis. 20 ms frame window with half overlap is used in feature extraction. 12 MFCCs and

log power energy, plus their first order time derivative, totally 26 dimension feature vector is used. In GMM representation of each speaker, 36 Gaussian mixture model (GMM) is used for each BPC model. Then each speaker is represented as a set of BPC dependent supervectors with $26 * 36 = 936$ dimensions for each. Based on each BPC supervector for speakers, we do PCA analysis of the speaker supervector matrix as explained in section 2.2. We first show the projection of the speaker supervectors on the first two PCA eigenvectors. The result for vowel BPC based representation is showed in Fig. 3. In this figure,

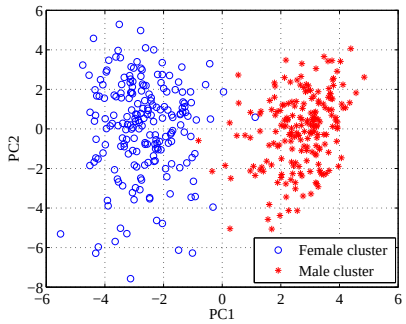

Figure 3: Projection of speaker supervectors on the first two principal eigenvectors based on BPC vowel representation.

each point represents the distribution location of one speaker in the speaker eigenspace. PC1 and PC2 represent the first and second principal components, respectively. From this figure, we can clearly see that there are two distinctive clusters of speakers which correspond well to their gender information. Because gender information is easy to be obtained with high certainty, we use gender recognition experiments to examine the accuracy of PCA based speaker clustering method. The results of gender recognition error rate based on each BPC-GMM clustering are shown as in table 1. We also did experiments based on GMM of

Table 1: Gender recognition error rate (RER) based on BPC-GMM speaker representation.

GMM-BPC	Vowel	Liquid	Stop	Fricative	Nasal
RER (%)	0.75	1.50	1.75	2.00	33.00

global acoustic distribution, and did PCA analysis, and clustering for gender recognition. We got recognition error rate about 1.25%. From these results, we can confirm that in BPC, vowels carry much more gender information than other categories in BPC. In addition, the BPC vowel based speaker clustering can get a better partition of speaker acoustic space according to gender information than that of based on GMM of global acoustic distribution. Then in ensemble acoustic modeling, we will partition the data set based on speaker supervectors of the BPC vowel representation.

In order to identify the second largest variability of the speaker acoustic space, we look into the distribution of each cluster in the PCA projection space. The female cluster is taken as an example, and the speakers from BJ and SH are labeled in the PCA space as shown in Fig. 4. In this figure, we can see that the second principal component seems to correspond to the regional or accent information but with large overlaps. Since we do not have a manually labeled accent information with high certainty on the data set, we can not do accent recognition experiments to examine the efficiency of the method. But for acoustic space partition for ensemble modeling, it is possible to use this information for speaker acoustic space partition to train ensemble acoustic models to improve the ASR performance.

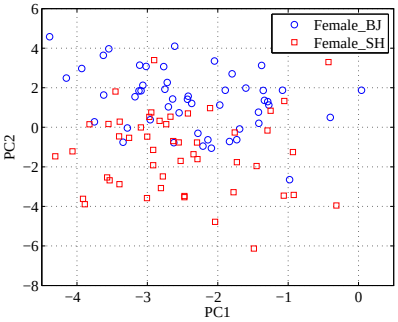

Figure 4: Projection of regional speaker supervectors on the first two principal components.

3.2. Speaker acoustic space partition

Based on the factor analysis, we do hierarchical data partition as: (1) Splitting the speaker data set to be two clusters which is concerned with gender information (the first level partition); (2) Further splitting each cluster to be two sub-clusters which is possibly concerned with accent information (the second level partition). All the partitions are done in a low dimensional speaker space using k-means clustering algorithm, and no label information is required. The low dimensional speaker space is constructed from BPC vowel speaker supervector based representation with PCA analysis (keeping the first two principal components).

4. Speech recognition experiments and evaluations

The HTK toolkit was used to build a Chinese LVCSR system. The feature vector was 26 dimension MFCC feature (the same as used in the analysis in section 2). A phonetic-decision tree based state-tying triphone hidden Markov model (HMM) was built for acoustic modeling (30 basic phoneme set was used in acoustic modeling). In recognition, a weighted finite state transducer (WFST)-based decoder was used [8]. A trigram language model trained using Chinese basic travel conversation text (BTEC) and a size of 50k lexicon were used in recognition. In ensemble modeling based recognition, multiple acoustic models are trained based on the partitioned data sets. Considering that speaker numbers in each partitioned cluster is not balanced, a MAP adaptation training strategy is used in training acoustic model for each cluster. The final acoustic model is a combination of all these multiple acoustic models (as shown in Fig. 1). Our focus in this study is data partition for ensemble acoustic modeling, therefore, a parallel decoding strategy is used in the last stage for model combination, i.e., the model with highest likelihood score was chosen as a decoding model. The final recognition result is measured as character error rate (CER).

In order to examine whether the testing data set consists of accent speech or not, we first did speech recognition experiments based on two types of acoustic models. The recognition results are showed in Fig. 5. In this figure, 'Standard model' and 'Mixed model' represent the recognition experiments based on acoustic models trained using standard Mandarin speech (speakers are from northern part of China) and regional mixed speech (speakers are from northern and southern parts of China), respectively. From this figure, we can see that in the testing data set, error rate for speakers from BJ is low compared with those from other regional groups. It is easy to understand because speakers from BJ belong to northern part

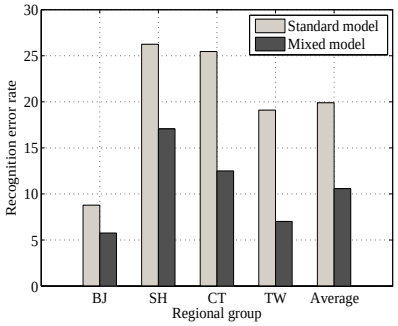

Figure 5: Recognition error rate based on standard acoustic model and regional mixed acoustic model.

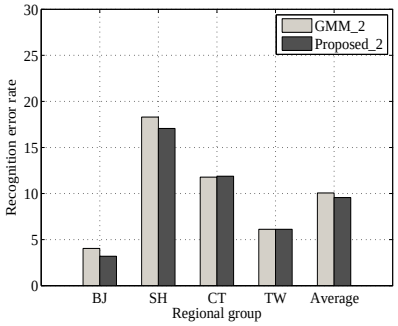

Figure 6: Recognition error rate for two acoustic models trained on data partition and clustering.

of China that have similar acoustic property as that of standard Mandarin. But speakers from SH, CT, and TW belong to southern parts of China that show different acoustic property compared with standard Mandarin, thus lead to high error rate in recognition. In addition, by comparing the two types of acoustic modeling, we can see that adding regional speakers' speech in training can make a better coverage of acoustic model thus greatly improve the recognition performance. But training a single model using acoustic speech from all regional groups is a statistical average of the training data set. We try to train ensemble acoustic models by splitting the data set based on their hidden factors which cause acoustic variability, and compare the performance on speech recognition. In comparison, the single model trained by using regional mixed speech is used as a baseline.

We have tested on random splitting of the training data set to two and four subsets, and trained ensemble models for speech recognition, but found no improvement. Based on the factor analysis in section 3, we designed hierarchical data partition method introduced in section 3.2. We trained the ensemble acoustic models using data subsets from the first and second level partitions, and did speech recognition experiments. For comparison, traditional speaker GMM clustering based on global acoustic distribution for two and four clusters are also designed for ensemble acoustic modeling. The results are shown in Fig. 6 and 7.

In these figures, 'GMM_2', 'Proposed_2', 'GMM_4' and 'Proposed_4' represent the recognition experiments based on acoustic models trained by traditional global acoustic GMM and proposed data partition modeling for two and four data clusters, respectively. Comparing Figs. 6 and 5, we can see that ensemble acoustic modeling, both the traditional GMM clustering and proposed partition modeling, significantly improved the performance. In addition, the proposed partition modeling performed better than traditional GMM clustering method.

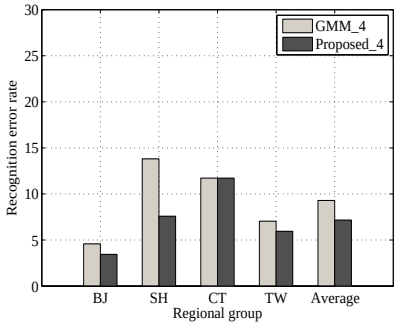

Figure 7: Recognition error rate for four acoustic models trained on data partition and clustering.

5. Discussion and conclusion

Considering that speaker variabilities , such as gender and accent information, may be carried by several specific phoneme categories rather than by all phonemes, we proposed a BPC-GMM supervector modeling for speaker representation. The gender information can be easily identified from the proposed speaker representation. Based on the BPC-GMM supervector, we designed two level hierarchical data partition algorithm in a low dimensional factor space for ensemble acoustic modeling. Compared with baseline performance, the proposed ensemble acoustic modeling using the first and second level data partitions provided 9.64% and 32.23% relative improvement in CER, respectively.

Several problems need to be studied in the future. In our study, because there is no accent label information on the data corpus, we can not do accent recognition experiments to examine the proposed data partition method. In the future, we will manually label the data corpus with accent information by listening test, and compare the labeled information with data partition results. In addition, in ensemble acoustic modeling, we only discussed the data partition for training ensemble models. How to combine the ensemble models to obtain an optimal model remains as one of our future work.

6. References

[1] Dietterich, T. G., "Ensemble methods in machine learning," Proc. of the First International Workshop on Multiple Classifier Systems, 1-15, 2000.

[2] Siohan, O., Ramabhadran, B. and Kingsbury, B., "Constructing ensembles of ASR systems using randomized decision trees," Proc. of ICASSP, I: 197-200, 2005.

[3] Xue, J., Zhao, Y., "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," IEEE Trans. SAP, 16 (3): 519-528, 2008.

[4] Tsao, Y., Lee, C. H., "An Ensemble Speaker and Speaking Environment Modeling Approach to Robust Speech Recognition," IEEE Transactions on Audio, Speech Language Processing, 17(5): 1025-1037, 2009.

[5] Padmanabhan, M. et al., "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems," Proc. of ICASSP96, II: 701-704, 1996.

[6] Huang, C., Chen, T., and Chang, E., "Accent issues in large vocabulary speech recognition," International Journal of Speech Technology, 7: 141-153, 2004.

[7] Campbell, W. M., Sturim, D. E., Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, 13(5): 308-311, 2006.

[8] Mohri, M., Pereira, F., and Riley, M., "Speech recognition with weighted finite-state transducers," Handbook of Speech Processing, Y. H. Jacob Benesty, Mohan Sondhi, Ed. Springer, 559-582, 2008.