SPEAKER CLUSTERING USING VECTOR REPRESENTATION WITH LONG-TERM FEATURE FOR LECTURE SPEECH RECOGNITION

*Chien-Lin Huang, Chiori Hori, Hideki Kashioka, Bin Ma**

National Institute of Information and Communications Technology, Kyoto, Japan
*Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
{chien-lin.huang, chiori.hori, hideki.kashioka}@nict.go.jp, mabin@i2r.a-star.edu.sg

ABSTRACT

Speaker clustering has been widely adopted for clustering the speech data based on acoustic characteristics so that an unsupervised speaker normalization and speaker adaptive training can be applied for a better speech recognition performance. In this study, we present a vector space speaker clustering approach with long-term feature analysis. The supervector based on the GMM mean vectors is adopted to represent the characteristics of speakers. To achieve a robust representation, total variability subspace modeling, which has been successfully applied in speaker recognition for compensating channel and session variability over the GMM mean supervector, is used for speaker clustering. We apply a long-term feature analysis strategy to average short-time spectral features over a period of time to capture the speaker traits that are manifested over a speech segment longer than a spectral frame. Experiments conducted on lecture style speech show that this speaker clustering approach offers a better speech recognition performance.

Index Terms— Speaker clustering, speech recognition, long-term feature, total variability

1. INTRODUCTION

Automatic speech recognition (ASR) has been widely applied in the automatic transcription of voice recordings while the acoustic mismatches including speaker variation and noises still constitute one of the major challenges to a reliable speech recognition system. To achieve the robust speech recognition, many techniques have been proposed to address the acoustic mismatch caused by the speaker variation [1]–[5] and speaker clustering is one of the effective approaches. Through the speaker clustering process, the voice recordings are segmented and clustered into homogeneous segments based on the speaker or other acoustic characteristics. Then the speaker clustering results can be used for the speaker-based cepstral mean normalization (CMN) [3, 4] and speaker adaptive training (SAT) [5] in the acoustic modeling. As a result, a better ASR performance is expected to be achieved.

Recently, Tsai et al. [6] presented an automatic speaker clustering algorithm using a voice characteristic reference space and maximum purity estimation, with the aim of maximizing the similarities between utterances within clusters. Tang et al. [7] applied a complete treatment for a partially supervised speaker clustering to assist the unsupervised speaker clustering process. They proposed to perform the speaker clustering based on the cosine distance metric and linear spherical discriminant analysis. In addition, speaker clustering is an essential part of speaker diarization. Nwe et al. [8] designed a strategy for the speaker diarization system considering speaker clustering. They used a consensus based cluster purification method that removed impure speaker segments in speaker clusters before the speaker modeling was conducted in the cluster purification process. Ishiguro et al. [9] formulated the speaker clustering problem as the clustering of sequential audio features generated by an unknown number of latent mixture components for efficient speaker identification. They employed a probabilistic model assuming time sensitive speaker mixtures at every time frame.

In this study, we study the speaker clustering techniques and apply the unsupervised speaker clustering for improving speech recognition. Figure 1 illustrates the process of the proposed speaker clustering. A novel vector space representation is proposed to capture the speaker discriminative characteristics based on the long-term feature analysis and total variability subspace modeling. Experiments were conducted on the lecture speech, TED[1] which provides streaming speech to spread ideas on the topics of Technology, Entertainment, and Design. The lecture speech is commonly with applause, laughter, music, etc. In a TED talk, it is not always monologue. There might be interviews or conversations in a talk. We apply the vector space strategy to represent spoken utterances and the speaker clustering is conducted to cluster the spoken utterances into a number of speaker clusters in each talk. For a better acoustic modeling, even in a talk involving only one speaker, the speaker clustering technique can also be applied to cluster the speech data based on different acoustic

[1] www.ted.com

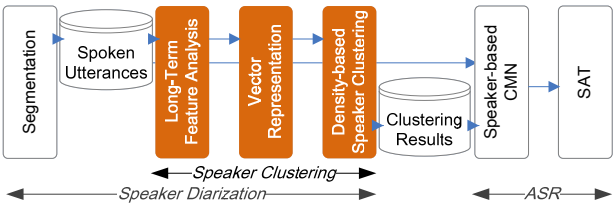

Fig. 1. The process of speaker clustering using vector representation with long-term feature for speech recognition.

environments such as different speaking types, noise levels, etc. In the reminder of the paper, we present the proposed speaker clustering techniques for lecture speech recognition in Section 2. Section 3 shows experiments in detail. We conclude with a summary of findings in Section 4.

2. VECTOR SPACE REPRESENTATION FOR SPEAKER CLUSTERING

2.1. Long-Term Feature Analysis

Feature extraction is an important step to estimate a numerical representation from speech samples and to characterize speakers. Mel-frequency cepstral coefficient (MFCC) is an effective speech feature analysis [10]. Speech signal is conventionally represented as a sequence of frames for short-term analysis and these frames are small enough to ensure that frequency characteristics of the magnitude spectrum are relatively stable. However, speech timbre and prosody are manifested over a speech segment of multiple short-term spectrums through phonetic units, such as vowels and consonants. To capture the spectral statistics over a long period of time, we proposed the speaker discriminative feature extraction using long-term feature (LTF) analysis [11]. An overlapping long-term window was applied on short-term features to average G short-term MFCC frames into N LTF frames with $N = (G - L)/Z + 1$. L denotes the size of the long-term window and Z is the step of the long-term window shift. The advantage of the LTF features is that they can simultaneously take account of short-term frequency characteristics and long-term resolution at the same time. LTF features with the mean of every four frames of MFCC features $(L = 4, Z = 2)$ is used for vector representation of speaker clustering and denoted as LTF-4 in this study.

2.2. Total Variability for Vector Representation

The vector space representation has been widely applied in the speaker recognition for the speaker modeling with Gaussian mixture model (GMM) which is trained by using spectral features of the spoken utterances of a speaker. A GMM supervector shows an effective way to represent the speaker characteristics in the spoken utterance. The GMM mean supervector $\mathbf{x}$ is obtained by stacking the mean vectors of all the Gaussian components $\mathbf{x} = \left[\mathbf{u}_1^T, \mathbf{u}_2^T, ..., \mathbf{u}_M^T\right]^T$, where $\mathbf{u}_m$ is the mean vector of the m-th Gaussian component. To solve the data sparseness problem, maximum a posteriori (MAP) estimation is normally used to adapt the speaker model from a universal background model (UBM) [12].

In real-world application, speech is recorded with different types of channels, sessions, and speakers. To compensate channel variability over GMM mean supervectors, a total variability space modeling has been proposed to represent both the speaker and channel variability simultaneously [13], in which the speaker and channel dependent GMM mean supervector $\mathbf{s}$ is given by

$$\mathbf{s} = \mathbf{x}_{ubm} + \mathbf{T}\mathbf{v}, \tag{1}$$

while $\mathbf{x}_{ubm}$ denotes the supervector of the concatenation of the UBM mean vectors, $\mathbf{T}$ is a rectangular low-rank matrix representing R bases spanning subspace with the important variability in the GMM mean supervector space, and $\mathbf{v}$ is a normally distributed random vector of size R that is learned from samples. The weighing vector $\mathbf{v}$ has been named i-vector which provides an elegant way to project the high dimensional vector space to a low dimensional vector space while retaining most of the speaker relevant information. With i-vectors, we can represent speaker dependent information in a low-dimensional space to suppress channel and session variability using popular statistical techniques. The i-vector representation is also able to capture the speaker characteristics with very short utterances and isolate the information of the target speaker from other unwanted variability with session compensation [14]. These characteristics are very suitable for vector representation in the speaker clustering.

2.3. Density-based Speaker Clustering

The speaker clustering is the spoken utterance assignments into similar speaker groups. In this study, we adopt density-based measurement for the speaker clustering in which each spoken utterance is regarded as an object represented by a vector. The distribution of these vectors can be viewed as a kind of density measure. For each object in a cluster, the neighborhood of a given radius ε has to contain at least a minimum number, *MinPts*, of objects. The density-based clustering has two main properties, density-reachable and density-connected [15]. We select a vector and measure its density according to the number of vetors (*MinPts*) in the radius distance ε and define the radius distance based on experiments as

$$\varepsilon = \frac{(d_{\text{avg}} + d_{\text{min}})}{2}, \tag{2}$$

where d_{avg} and d_{min} denote the average and minimum distances between vectors, respectively. *MinPts* is set as 6 in the experiments. The benefit of density-based clustering is that it is robust to outlier or noise detection. We can discover clusters of arbitrary shape. The density-based clustering does not need to pre-define the number of clusters. Since vector based representations show strong directional scattering patterns [7], we measure the similarity $sim(\mathbf{x},\mathbf{y})$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ using the cosine distance,

$$sim(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\mathbf{y}^T}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n}\mathbf{x}_i \times \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n}\mathbf{x}_i^2}\times\sqrt{\sum_{i=1}^{n}\mathbf{y}_i^2}}, \qquad (3)$$

to construct the speaker clusters. Among $sim(\mathbf{x},\mathbf{y})$, a value close to 1 means that two vectors are similar, whereas a value near 0 denotes two vectors are dissimilar.

2.4. Subspace Transformation and Normalization

The techniques of principal component analysis (PCA) and linear discriminant analysis (LDA) are commonly applied to reduce the dimension and collect discriminative information by projecting the data onto the pairwise linear discriminants [16]. In addition, we apply the within-class covariance normalization (WCCN) [17] to normalize speaker and channel effect in i-vector space to find the transformed vector $\hat{\mathbf{v}} = \mathbf{B}^T\mathbf{v}$. The transform matrix $\mathbf{B}$ is derived from the Cholesky decomposition of $\widehat{\mathbf{W}} = \mathbf{B}\mathbf{B}^T$, where $\widehat{\mathbf{W}}$ is the within-speaker covariance matrix estimated by

$$\widehat{\mathbf{W}} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}(\mathbf{v}_i^s - \mathbf{u}_s)(\mathbf{v}_i^s - \mathbf{u}_s)^T, \ \ \mathbf{u}_s == \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{v}_i^s, \ (4)$$

where S means the number of speakers that each has the number of N_s i-vectors in the training dataset. After the subspace transformation of LDA and PCA on GMM mean supervectors and WCCN on i-vectors, we further apply the length normalization to deal with the non-Gaussian behavior of vectors so that normalized vectors can better fit to the Gaussian assumptions in modeling. The vector representations are normalized to the unit length by capturing their directions as $\overline{\mathbf{v}} = \mathbf{B}^T\mathbf{v} / \left\|\mathbf{B}^T\mathbf{v}\right\| = \hat{\mathbf{v}} / \left\|\hat{\mathbf{v}}\right\|$.

3. EXPERIMENTS

3.1. Experimental Framework

We collected 760 talks from online TED website as the training data. Speech segmentation and word alignment of original talks were conducted by using SailAlign in which HTK was used as the recognizer [18, 19]. We removed non-speech data, rectified for the offset of subtitles to their acoustic equivalent, and spoken utterances were specified by the segments between pauses. The collected talks consisted of about 204 hour audio before the alignment procedure. We tested 8 TED talks with 818 utterances for the speech recognition evaluation. The Kaldi toolkit [20] was adopted for the ASR experiments. Based on the speaker clustering, the speaker-based CMN and SAT were applied to compensate the channel and speaker variations. The acoustic model is attributed to both phonetic variation and variation among speakers of the training population and these two variation sources are decoupled. SAT was used to jointly annihilate the inter-speaker variation and estimate the HMM parameters of speaker independent acoustic models [5]. HMM models were with 4,000 tied states and 120,000 Gaussian mixture components. We extended 39 phones of CMU pronunciation dictionary to 336 monophones based on accent and position information. The trigram language model was trained on the TED text transcript. The speech decoding process was based on the weighted finite state transducers (WFST) while we used the OpenFST tools [21].

3.2. Evaluation Metrics

We report the results of speaker clustering based on the clustering accuracy defined by

$$\text{clustering accuracy} = \frac{1}{N}\sum_{n=1}^{N}[c_n = l_n]\times 100\% , \qquad (5)$$

where N is the number of spoken utterances, l_n denotes the true label of the testing utterance n, c_n means the recognized cluster label which is output of the speaker clustering. The indication function [.] gives 1 if the argument is true and 0 otherwise [7]. The speech recognition performance is evaluated in terms of the word-error-rate (WER),

$$\text{WER} = \frac{ins + del + sub}{num}\times 100\% , \qquad (6)$$

considering errors of insertion (*ins*), deletion (*del*), and substitution (*sub*). *num* denotes the number of words in the testing utterances [22].

3.3. Evaluation of Speaker Clustering

First, we compare the conventional MFCC and the proposed long-term feature analysis (LTF) using the GMM-SVM-NAP speaker recognition system on the NIST Speaker Recognition Evaluation (SRE) 2008 corpus with evaluation metrics of equal error rate (EER) and minimum detection cost function (DCF) [11] in Table 1. Each frame of the

Table 1. Comparison of LTF and MFCC on NIST SRE-2008

Feature	Male		Female		All	
	EER	$100\times$DCF	EER	$100\times$DCF	EER	$100\times$DCF
MFCC	3.11%	1.18	3.45%	1.43	3.34%	1.38
LTF-4	3.14%	1.13	2.84%	1.43	**2.96%**	**1.33**

Table 2. Comparison of vector representations with long-term feature analysis and density-based speaker clustering

Vector Representation	Clustering Accuracy
GMM mean supervector	79.05%
GMM mean supervector (PCA)	72.68%
GMM mean supervector (LDA)	81.55%
Our vector representation	**88.75%**
Our vector representation (*K*-means clustering)	87.20%
Our vector representation (hierarchical clustering)	71.99%

speech data is represented by a 36-dimensional feature vector, consisting of 12 MFCC, together with their deltas and double-deltas. The proposed LTF-4 shows a good speaker discrimination and outperforms the conventional MFCC in speaker recognition. The speaker discriminative long-term features were used for speaker clustering. In the rest of experiments we focused on the proposed vector representation consisted of LTF-4, i-vector with WCCN and the unsupervised density-based speaker clustering. Note that MFCC was still used for speech recognition.

Table 2 shows evaluations of the density-based speaker clustering using various vector representations. A speaker independent UBM with 64 Gaussians was used for generating GMM mean supervectors. We applied the NIST SRE 2004, 2005, 2006 [11] telephone data, NIST SRE 2005, 2006 microphone data, and Switchboard II data to estimate the total variability matrix **T** in Eq. (1). The matrices of PCA, LDA and WCCN were estimated using the NIST SRE 2004, 2005, 2006 telephone data and NIST SRE 2005, 2006 microphone data. To evaluate the performance of speaker clustering, 30 independent trials were conducted, each of which involved a random selection of two TED talks. There were 84 spoken utterances in a talk on average. The average duration was 9.52 seconds per utterance. In Table 2, the speaker clustering based on the proposed vector representation provides a better performance than GMM mean supervectors and the projected vectors based on PCA and LDA. In the experiments, the dimensions of our vector representation, GMM supervectors with PCA and LDA were all set to 200. We also compared the *K*-means and hierarchical clustering techniques [7] with the density-based speaker clustering based on our vector representation. The density-based clustering gave the best performance.

3.3. ASR Evaluation on TED Talks

Table 3 shows the ASR results with two baselines: One is the speaker independent baseline, and the other is the speaker dependent baseline with an assumption of only one speaker in a talk. Because speaker information is effective for the speaker-based CMN and SAT, the speaker dependent baseline gives a better performance than the speaker independent baseline with both maximum likelihood (ML) and SAT. The WERs of the speaker dependent baseline with ML and SAT training were 23.23% and 21.59%, respectively. In addition, by assuming one or two speakers in a talk and applying the unsupervised speaker clustering, we evaluated different vector representations based on GMM mean supervector and the proposed vector representation. Compared with the speaker dependent baseline, the proposed unsupervised speaker clustering methods provided further improvements.

Table 3. Speech recognition evaluation (in WER) using the unsupervised speaker clustering for ML and SAT training

Speaker Clustering \ Acoustic Training	ML	SAT
Speaker independent	23.53%	22.76%
Speaker dependent	23.23%	21.59%
GMM mean supervector	23.31%	21.45%
GMM mean supervector (PCA)	23.14%	21.56%
GMM mean supervector (LDA)	23.26%	21.48%
Our vector representation	**23.07%**	**21.17%**

Table 4. Improvements by adding different techniques based on the proposed unsupervised speaker clustering for ASR

Systems	WER	Reduction
Speaker independent ML	23.53%	-
+ Speaker-based CMN	**23.07%**	**1.95%**
+ SAT	**21.17%**	**8.24%**
+ MMI	19.91%	6.33%

Table 4 summarizes the step-by-step WER reductions with the proposed speaker clustering for speaker-based CMN and SAT, and discriminative MMI training [23]. WER reductions of 1.95% and 8.24% were achieved for speaker-based CMN and SAT using the proposed unsupervised speaker clustering for TED lecture speech recognition. We found that the proposed method offered more than 10% WER reduction on average.

4. CONCLUSION

In summary, we have presented the advantages of using the vector space representation in speaker clustering for the lecture speech recognition. On the one hand, using the total variability subspace modeling with speaker discriminative long-term feature achieves a low-dimensional representation to suppress channel and session variability and contributes to a better speaker clustering performance than the original GMM mean supervectors. On the other hand, the density-based unsupervised speaker clustering is used for speaker-based CMN and SAT training bringing more accurate speech recognition than the conventional speaker dependent baseline. WER reductions of 1.95% and 8.24% have been achieved for speaker-based CMN and SAT using the proposed unsupervised speaker clustering for TED lecture speech recognition.

5. REFERENCES

[1] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[2] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] S. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in Speech Transcription at IBM under the DARPA EARS Program," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[4] G. Saon and J.-T. Chien, "Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

[5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 2, pp. 1137–1140, 1996.

[6] W.-H. Tsai, S.-S. Cheng, and H.-M. Wang, "Automatic Speaker Clustering Using a Voice Characteristic Reference Space and Maximum Purity Estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1461–1474, 2007.

[7] H. Tang, S. M. Chu, M. Hasegawa-Johnson, and T. S. Huang, "Partially Supervised Speaker Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 959–971, 2012.

[8] T. L. Nwe, H. Sun, B. Ma, and H. Li, "Speaker Clustering and Cluster Purification Methods for RT07 and RT09 Evaluation Meeting Data," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 461–473, 2012.

[9] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 447–460, 2012.

[10] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall, 1993.

[11] C.-L. Huang, H. Su, B. Ma, and H. Li, "Speaker Characterization Using Long-Term and Temporal Information," in *Proc. Interspeech*, pp. 370–373, 2010.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation", in *Proc. ICASSP*, 2006.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, "I-vectors in the Context of Phonetically-Constrained Short Utterances for Speaker Verification," in *Proc. ICASSP*, 2012.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. KDD*, pp. 291–316, 1996.

[16] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Englewood Cliffs, NJ: Prentice Hall, 2001.

[17] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *Proc. ICSLP*, pp. 1471–1474, 2006.

[18] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein and S. Narayanan, "SailAlign: Robust Long Speech-Text Alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[19] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge, England, Cambridge University, 2000.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.

[21] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFST: A General and Efficient Weighted Finite-State Transducer Library," in *Proc. International Conference on Implementation and Application of Automata (CIAA)*, pp. 11–23, 2007.

[22] C.-L. Huang and C.-H. Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *IEEE Trans. Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.

[23] L. R. Bahl, P. F. Brown, P. V. Souza, and L. R. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *Proc. ICASSP*, pp. 49–52, 1986.