Speaker Clustering of Speech Utterances Using A Voice Characteristic Reference Space

Wei-Ho Tsai, Shih-Sian Cheng and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China

{wesley,sscheng,whm}@iis.sinica.edu.tw

Abstract

This paper presents an effective technique for clustering speech utterances based on their associated speaker. In attempts to determine which utterances are from the same speakers, a prerequisite is to measure the similarity of voice characteristics between utterances. Since the vast majority of existing methods evaluate the inter-utterance similarity by taking only the information from the spectrum-based features of utterance pairs into account, the resulting clusters may not be well relevant to speaker, but instead likely to the environmental conditions or other acoustic classes. To compensate for this shortcoming, this study proposes to project utterances from their spectrum-based feature representation onto a reference space trained to cover the generic voice characteristics inherently in all of the utterances to be clustered. The resultant projection vectors naturally reflect the relationships between all the utterances and are more robust against the interference from non-speaker factors. We exemplarily present three distinct implementations for reference space creation.

1. Introduction

Speaker clustering refers to the task of unsupervised classification of speech utterances based on speaker voice characteristics. For more than one decade, interests and needs in speech-recognition community have provided a major motivation for the work on speaker clustering [1-3], in which speech data produced by the same speakers or speakers with similar voices are grouped together such that acoustic model adaptation can be carried out more effectively. However, there was still a dearth of study devoted to this problem. More recently, speaker-clustering research has enjoyed a renaissance [4-12], spurred by research activities in spoken document indexing for managing burgeoning collections of available speech data. It is desired that by clustering speech data from the same speakers, the human efforts required for documentation can be dramatically reduced or replaced.

Currently, the lion's share of speaker-clustering methods falls into a hierarchical clustering framework [4-10]. The method begins with a certain similarity computation for every pair of utterances, followed by a generation of cluster tree in an either bottom-up (agglomerative) or top-down (divisive) fashion according to some criteria on the similarity measure. The similarity computation is designed in such a way to produce larger values for similarities between utterances of the same speaker and smaller values for similarities between utterances of different speakers. Several similarity measures such as cross likelihood ratio [7], generalized likelihood ratio [6], and Bayesian information criterion [8,10] have been examined and compared in many literatures. However, these similarity measures are performed directly on the spectrum-based features which are known to carry various information besides the speaker voice characteristics, such as phonetic and environmental conditions. As a result, the generated clusters may not be well relevant to speaker attributes. In particular, speaker clustering might be vulnerable, when the utterances to be clustered are short and noisy.

To alleviate the above-mentioned problem, this study proposed a novel speaker-clustering framework which aims to exploit the underlying relationships of similarity between all the utterances to be clustered in a global sense rather than only measuring the inter-utterance similarity in a pairwise fashion. Such a framework enables that when attempting to judge if any pair of utterances belongs to the same speaker, some information from other utterances can be incorporated into the decision made for that pair of utterances, and hence provides a more reliable clustering than it can be done by taking only within-pair information into account. We implement this framework in three distinct ways, which are, respectively, based on utterance-individual Gaussian mixture modeling, utterance-universal vector clustering, and utterance-universal Gaussian mixture modeling followed by utterance-individual model adaptation.

2. Method overview

Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ denote N unlabeled speech utterances in a spectrum-based feature vector representation, each of which was produced by one of the P speakers, where $N \geq P$, and P is unknown. The aim of the speaker clustering is to partition the N utterances into M clusters such that $M = P$ and each cluster consists exclusively of utterances from only one speaker.

Fig. 1 shows the proposed speaker-clustering framework. Prior to the commonly-deployed similarity computation and cluster tree generation, a reference space, which aims to represent some generic characteristics of speaker voices, is constructed. The reference space is composed of K bases, where the basis is a general term referring to a representative of voice characteristics encoded in the spectrum-based features. The reference space can be created using either the utterances to be clustered or an extra speech data set other than the utterance set in question. The use of an external data resource might be able to cover more variety of voice characteristics and easy to perform on-line or incremental clustering, but on the other hand, might run the risk that a discrepancy of environmental and channel conditions exist between the external data set and the utterance set to be clustered. In this study, for performance comparison with other speaker clustering methods under a consistent evaluation condition, we only use the utterances to be clustered to train the reference space.

After a reference space is constructed, each of the N utterances, say $\mathbf{X}_i$, is converted from its spectrum-based feature representation into a K-dimensional *projection vector* $\mathbf{V}_i = [v(\mathbf{X}_i,\phi_1),\ v(\mathbf{X}_i,\phi_2),\ldots,\ v(\mathbf{X}_i,\phi_K)]'$ on the space, where prime denotes vector transpose, and $v(\mathbf{X}_i,\phi_k)$, $1 \leq k \leq K$, is a projection value that reflects the extent of how the utterance $\mathbf{X}_i$ can be characterized by the basis ϕ_k. It is hoped that, if two utterances, $\mathbf{X}_i$ and $\mathbf{X}_j$, are from the same speaker, say S_p, a majority of the projection values in $\mathbf{V}_i$ and $\mathbf{V}_j$ are relatively similar in some sense, resulting that $\mathbf{V}_i$ is closer to $\mathbf{V}_j$, compared to $\mathbf{V}_\ell$ of any utterance $\mathbf{X}_\ell \notin S_p$.

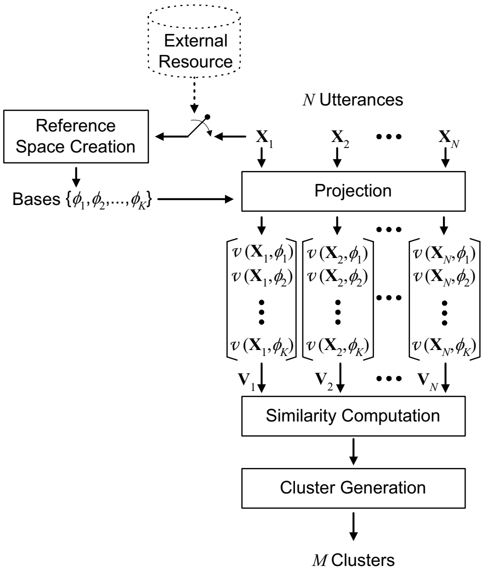


Figure 1: The proposed speaker-clustering framework.

By associating each utterance with a projection vector, the similarity between any two utterances, $\mathbf{X}_i$ and $\mathbf{X}_j$, can be computed straightforwardly using the cosine measure between $\mathbf{V}_i$ and $\mathbf{V}_j$:

$$S_u(\mathbf{X}_i,\mathbf{X}_j) = \frac{\mathbf{V}_i \cdot \mathbf{V}_j}{\|\mathbf{V}_i\| \|\mathbf{V}_j\|}. \tag{1}$$

Then, utterances deemed similar enough with each other are grouped into a cluster. In our implementation, cluster generation is performed in an agglomerative manner, which starts with each utterance in its own cluster and successively merging the most similar pair of clusters, say c_i and c_j, according to a *complete-linkage* cluster similarity defined by

$$S_c(c_i,c_j) = \min_{\substack{\mathbf{X}_m \in c_i \\ \mathbf{X}_n \in c_j}} S_u(\mathbf{X}_m,\mathbf{X}_n). \tag{2}$$

The output from the aggregation procedure above is a tree of clusters, and the final partition of the utterances is then determined by pruning the tree subsequently with only M leaves left. An appropriate value of M, which corresponds to the speaker population in the N utterances, can be estimated by applying the method described in [6].

3. Reference space creation

The effectiveness of the above speaker-clustering framework crucially depends on if a reference space is capable of summarizing the most relevant aspects of speaker voice characteristics inherent in the observed speech data. This section presents three distinct methods for the reference space construction along with the projection vector computation.

3.1 Utterance-individual Gaussian mixture modeling

Conventional speaker recognition, which aims to determine the identity of a talker from his/her voices, predominantly uses Gaussian mixture models (GMMs) to characterize the speaker-specific voice patterns. The main attraction of the GMM arises from its ability to provide smooth approximations to arbitrarily-shaped densities of long-term spectrum that are considered to be related to the characteristics of the speaker's voice rather than the specific linguistic message. Such a modeling technique can be applied in an unsupervised manner for the construction of a speaker-related reference space. To be specific, a GMM is created for each of the N utterances to be clustered, and the resulting N GMMs $\lambda_1, \lambda_2, \ldots, \lambda_N$ form a reference space with N bases $\phi_k = \lambda_k$, $1 \leq k \leq N$. For each utterance $\mathbf{X}_i$, the projection value on basis ϕ_k, $1 \leq k \leq N$, is then computed using

$$v(\mathbf{X}_i,\phi_k) = \log p(\mathbf{X}_i \mid \lambda_k)\cdot \tag{3}$$

Ideally, the value of $v(\mathbf{X}_i,\phi_k)$ is large if utterances $\mathbf{X}_i$ and $\mathbf{X}_k$ are from the same speaker, and is small otherwise. However, practically there is no guarantee about this behavior, since the GMMs may not be able to well characterize the speakers voices when the utterances involved are subject to very limited duration and diverse environmental conditions. It is hoped that through the use of a whole projection vector, the impact of those abnormal projection values could be diluted by other normal ones, and a more reliable similarity measure could, thus, be derived.

The concept of the above clustering method is basically the same as a prior study reported in [13]. A similar idea has also been presented recently from the viewpoint of the so-called *triangulation* [12], in which each utterance is modeled as a single Gaussian distribution. As learned from the speaker recognition task, a better performance of the speaker clustering may be obtained using a proper number of mixture Gaussian components rather than a single Gaussian density. However, determining the proper number of mixtures in GMMs is a sticky problem, especially in the case that the durations of the utterances involved might be rather diverse. To sidestep this problem, two alternative methods presented in the following subsections are further developed.

3.2 Utterance-universal vector clustering

Instead of using utterance-individual GMMs, a single, utterance-independent codebook having K codewords is created as a reference space using the entire feature vectors of the utterances to be clustered. The codebook can be considered as a universal model trained to cover the speaker-independent distribution of feature vectors. In our implementation, each codeword $\mathbf{w}_k$, $1 \leq k \leq K$, consists of a mean vector μ_k and a diagonal covariance matrix Σ_k. Training of the codebook is performed via k-means clustering algorithm, in which the distance between feature vectors is computed on the basis of Mahalanobis distance. The use of such a codebook-based reference space is motivated by the observation that although the codebook as a whole is a speaker-independent representation, a significant proportion of the individual codewords tend to be speaker-dependent, mainly because these codewords are the self-grouping results of the feature vectors largely from the same speakers. In other words, each of the speakers involving in the N utterances reflects his/her own set of favorable codewords. Thus, speaker clustering might be accomplished by examining and comparing the distribution of feature vectors of each utterance on the codebook.

After K codewords are generated, each feature vector of the utterances is explicitly assigned a codeword index. The projection value $v(\mathbf{X}_i,\phi_k)$ for utterance $\mathbf{X}_i$ with respect to basis ϕ_k, $1 \le k \le K$, is computed using

$$v(\mathbf{X}_i,\phi_k)=\frac{\#\text{ feature vectors in }\mathbf{X}_i\text{ assigned as }\mathbf{w}_k}{\#\text{ total feature vectors in }\mathbf{X}_i} \qquad (4)$$

3.3 Utterance-universal Gaussian mixture modeling followed by utterance-individual model adaptation

Alternatively, the problem of the diverse utterance duration as mentioned in Sec. 3.1 might be better handled by using some model-adaptation techniques developed in speech or speaker recognition research. Our basic strategy is to create an utterance-universal GMM using all the utterances to be clustered, followed by an adaptation of the utterance-universal GMM performed for each of the utterances using maximum *a posteriori* (MAP) estimation. This strategy resembles the GMM-UBM method [14] for speaker recognition, in which the required speaker-specific models are created by tuning the parameters of a universal speaker model pre-trained using speech data from plenty of speakers. The GMM-UBM method has been shown very effective, especially when only limited enrollment data is available. Such a merit could be taken advantage of in the speaker-clustering task for short utterances.

4. Experiments

4.1 Speech data

Speech data used in this study consisted of 197 utterances chosen from the test set of the *2001 NIST Speaker Recognition Evaluation Corpus* [15], which contains conversational cellular telephone speech collected by the Linguistic Data Consortium (LDC). The 197 utterances were spoken by 15 male speakers, and the number of utterances spoken by each speaker ranged from 7 to 12. Fig. 2 shows the duration histogram of the 197 utterances with non-speech regions removed. Speech features including 24 Mel-scale frequency cepstral coefficients (MFCCs) were extracted from these data for every 20-ms Hamming-windowed frame with 10-ms frame shifts.

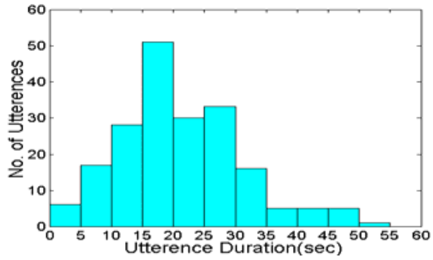


Figure 2: Duration histogram of the 197 utterances with non-speech regions removed.

4.2 Assessment methods

Performance of the speaker clustering was evaluated on the basis of two metrics: cluster purity [6] and Rand Index [16]. The cluster purity, which indicates the extent of agreement in a cluster, is defined by

$$\rho_m=\sum_{p=1}^{P}\left(\frac{n_{mp}}{n_{m*}}\right)^2, \qquad (5)$$

where ρ_m is the purity of the cluster m, n_{m*} is the total number of utterances in the cluster m, n_{mp} is the number of utterances in the cluster m that are from speaker S_p, and P is the total number of speakers involved. Eq. (5) follows that $n_{m*}^{-1} \le \rho_k \le 1$, in which the upper bound and lower bound reflect that all the within-cluster utterances are from the same speaker or completely different speakers, respectively. To evaluate the overall performance of an M-clustering for N utterances, an average cluster purity is computed using

$$\bar{\rho}=\frac{1}{N}\sum_{m=1}^{M} n_{m*}\rho_m . \qquad (6)$$

The Rand Index, which indicates the number of utterance pairs that are from the same speaker but are not grouped into the same cluster, and that are not from the same speaker but are grouped into the same cluster, is defined by

$$\gamma=\frac{1}{2}\sum_{m=1}^{M} n_{m*}^2+\frac{1}{2}\sum_{p=1}^{P} n_{*p}^2-\sum_{m=1}^{M}\sum_{p=1}^{P} n_{mp}^2, \qquad (7)$$

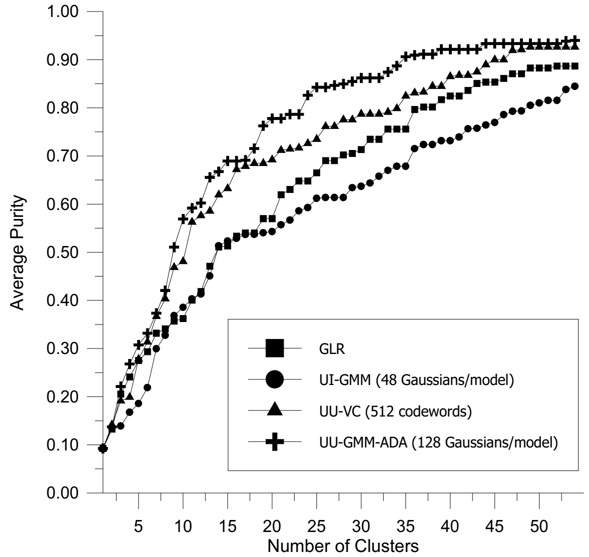
where n_{*p} is the number of utterances from speaker S_p. The lower the index, the better the clustering performs. A perfect clustering should produce an index of zero.

4.3 Experimental results

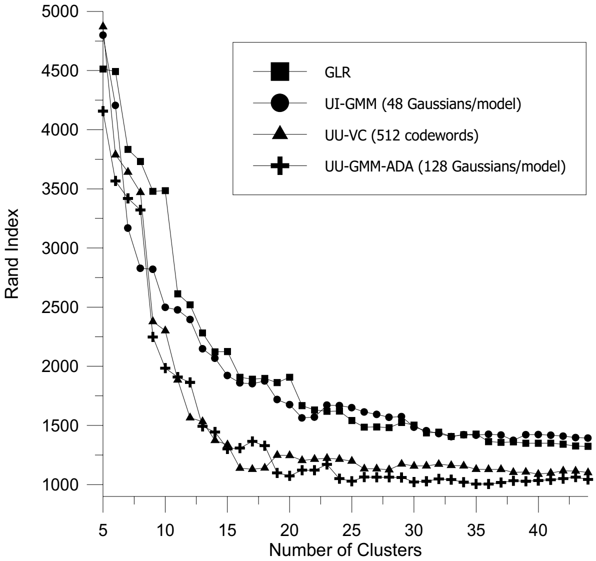
Fig. 3 shows the speaker-clustering results obtained with various methods. For better visualization, cluster purity and Rand Index were displayed as a function of number of clusters truncated. Here, “GLR” denotes the hierarchical clustering method using the generalized likelihood ratio as similarity measure [6], and its results served as a baseline for performance comparison. On the other hand, “UI-GMM”, “UU-VC”, and “UU-GMM-ADA” denote the proposed clustering methods, respectively, using utterance-individual Gaussian mixture modeling, utterance-universal vector clustering, and utterance-universal Gaussian mixture modeling followed by utterance-individual model adaptation. The codebook size used in the UU-VC method was 512, and the numbers of mixture components used in UI-GMM and UU-GMM-ADA were 48 and 128, respectively (empirically the most accurate configurations). All the Gaussian densities involved in this study used diagonal covariance matrices.

We can see from Fig. 3 that both UU-VC and UU-GMM-ADA clearly outperformed GLR in terms of both cluster purity and Rand Index. When the number of clusters is equal to the speaker population ($M = P = 15$), the best average cluster purity of 0.69 and Rand Index of 1302 were achieved with UU-GMM-ADA, which signifies a relative improvement of more than 35% compared to the cluster purity of 0.51 and Rand Index of 2124 obtained with GLR. In addition, the performance of the UI-GMM with 48 Gaussians per model was roughly equal to or slightly better than that of GLR, when appropriate numbers of clusters were generated. But, this is not the case when the cluster counts continue to grow.

We found that under this UI-GMM configuration, short utterances (less than 15 sec) tended to be falsely clustered. This is mainly attributed to the improper modeling caused by too many mixture components. By contrast, the GLR is computed on the basis of single Gaussian density, which could be well estimated even when utterances are short. Therefore, the GLR has less suffering from the clustering errors caused by short utterances, and hence, as shown in Fig. 3 (a), the average purity increased quickly with the number of clusters. The experimental results indicated that such a diverse utterance-duration problem can be circumvented by using either UU-VC or UU-GMM-ADA.



(a) Average cluster purity



(b) Rand Index

Figure 3: Speaker-clustering results obtained with various methods.

5. Conclusions

This study has investigated the methods of enhancing the inter-utterance similarity measurement for speaker clustering. Through the use of a voice characteristic reference space, the relationships of similarity between all of the utterances to be clustered can be exploited more effectively and reliably. We have shown that the performance of the existing hierarchical speaker clustering method can be boosted with the aid of various schemes for reference space creation.

Although fairly good performance improvement has been reported in this paper, one potential problem with our reference space creation is concerned with the correlation or overlap of voice characteristics between the reference bases. Future work will investigate if the speaker-clustering performance can be further improved by constructing a reference space having bases statistically independent with each other.

6. References

[1] Furui, S., “Unsupervised speaker adaptation method based on hierarchical spectral clustering”, *Proc. ICASSP’89*.

[2] Gish, Herbert, Siu, M.-H., and Rohlicek, R., “Segregation of speakers for speech recognition and speaker identification”, *Proc. ICASSP’91*.

[3] Kosaka, T. and Sagayama, S., “Tree-structured speaker clustering for fast speaker adaptation”, *Proc. ICASSP’94*.

[4] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M., “Automatic segmentation, classification and clustering of broadcast news audio”, *Proc. DARPA Speech Recognition Workshop*, 1997.

[5] Jin, H., Kubala, F., and Schwartz, R., “Automatic speaker clustering”, *Proc. DARPA Speech Recognition Workshop*, 1997.

[6] Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H., “Clustering speakers by their voices”, *Proc. ICASSP'98*, pp. 757-760.

[7] Reynolds, D. A., Singer, E., Carson, B. A., O’Leary, G. C., McLaughlin, J. J., and Zissman, M. A., “Blind clustering of speech utterances based on speaker and language characteristics”, *Proc. ICSLP’98*.

[8] Chen, S. S. and Gopalakrishnan, P. S., “Clustering via the Bayesian information criterion with applications in speech recognition”, *Proc. ICASSP’98*.

[9] Johnson, S. E., “Who spoke when? – Automatic segmentation and clustering for determining speaker turns”, *Proc. Eurospeech’99*.

[10] Zhou, B., and Hansen, J. H. L., “Unsupervised audio stream segmentation and clustering via the Bayesian information criterion”, *Proc. ICSLP’00*.

[11] Lapidot, I., Guterman, H., and Cohen, A., “Unsupervised speaker recognition based on competition between self-organizing maps”, *IEEE Trans. Neural Networks*, 13(4):877-887, 2002.

[12] Moh, Y., Nguyen, P., and Junqua, J.-C., “Towards domain independent speaker clustering”, *Proc. ICASSP'03*.

[13] Tsai, W. H., Chu, Y. C., Huang, C. S., and Chang, W. W., “Background learning of speaker voices for text-independent speaker identification”, *Proc. Eurospeech’01*.

[14] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, 10:19-41, 2000.

[15] *http://www.nist.gov/speech/tests/index.htm*

[16] Hubert, L., Arabie, P., “Comparing Partitions”, *Journal of Classification*, 2:193-218, 1985.