CLUSTERING SPEECH UTTERANCES BY SPEAKER USING EIGENVOICE-MOTIVATED VECTOR SPACE MODELS

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

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

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

ABSTRACT

This study investigates the problem of automatically grouping unknown speech utterances based on their associated speakers. The proposed method utilizes the vector space model, which was originally developed in document-retrieval research, to characterize each utterance as a *tf-idf*-based vector of acoustic terms, thereby deriving a reliable measurement of similarity between utterances. To define the required acoustic terms that are most representative in terms of voice characteristics, the eigenvoice approach is applied on the utterances to be clustered, which creates a set of eigenvector-based terms. To further improve speaker-clustering performance, the proposed method encompasses a mechanism of blind relevance feedback for refining the inter-utterance similarity measure.

1. INTRODUCTION

For more than two decades, automatic recognition of speaker based on vocal characteristics has received tremendous attention in facilitating human-machine communications and biometric applications. As more recently speech starts being exploited as an information source, the utility of recognizing speakers' voices are increasingly in demand in indexing and archiving the mushrooming amount of spoken data available. Traditional approaches to speaker recognition assume that some prior information or speech data are available from the speakers of concerned, while for the task of indexing or archiving, the basic strategy needs to be expanded to distinguish speakers in an unsupervised manner. As a result of this need, clustering speech utterances by speaker has emerged as a new challenging research problem [1-7], and the solutions to this problem are requiring to be further explored.

To date, most of the speaker-clustering methods in existence can amount to a hierarchical clustering framework [1-6]. This framework consists of three major components: a computation of inter-utterance similarity, a generation of cluster tree in either a bottom-up or top-down fashion according to some criteria on the similarity measure, and a determination of the number of clusters based on some termination conditions. Among the three components, the computation of inter-utterance similarity is of particular importance, because it crucially determines whether the generated clusters are related to speaker rather than other acoustic classes. Various methods based on cross likelihood ratio [3], generalized likelihood ratio [2], and Bayesian information criterion [4], etc., have been studied with the aim to produce larger values for similarities between utterances of the same speaker and smaller values for similarities between utterances of different speakers. However, since these similarity measures are performed entirely on the spectrum-based features, which are known to carry various information besides the speaker voice characteristics, such as phonetic and environmental conditions, the resultant clustering system might be vulnerable when the utterances addressed are short and noisy. In our prior work reported in [5], we show that a better similarity computation can be carried out on a reference space trained to cover the generic voice characteristics inherently in all of the utterances to be clustered. Because of incorporating out-of-pair information into the similarity computation for every pair of utterances, the clustering can be more robust against the interference from non-speaker factors.

As an extension of our prior work [5], this study further improves the speaker-clustering performance by primarily addressing one potential problem ignored in our prior work that the reference space is composed of intertwining voice characteristics rather than the most representative and statistically-independent ones. It is assumed that if the vocal characteristics of all the utterances to be clustered can be summarized as a set of the most representative and statistically-independent elements, utterances from the same and different speakers may be better distinguished by examining utterances with these elements. This idea is implemented by a means analogous to eigenvoice [8], which applies eigen decomposition on the parameters of models trained from a number of speakers. In addition, to further exploit various useful information for inter-utterance similarity computation, we re-formulate the speaker-clustering problem from a perspective of document retrieval. As will be shown below, some related concepts in document retrieval, such as relevance feedback, can be very useful as well for speaker clustering.

2. PROBLEM FORMULATION

Let $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_N$ denote N isolated speech utterances in a certain spectrum-based feature 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. If viewed as a problem of document retrieval, the partitioning could be done with an objective that when any of the N utterances, say $\mathbf{X}_k$, is chosen as an exemplar query to retrieve the relevant documents from the whole N-utterance set, the documents deemed most relevant in terms of same-speaker are the utterances within the cluster where $\mathbf{X}_k$ is located.

In solving the problem of document retrieval, vector space model (VSM) [9] is the most prevalent method and has consistently produced superior retrieval results in text material. The main attraction of the VSM is that it effectively structures the unstructured documents, making it easy to compare the similarity between documents and query. Under the VSM framework, each document or query is expressed by a vector of terms, with associated weights (vector elements) representing the importance of the terms in the document or query and within the whole document collection. A common approach to determine the weights is the so-called *tf-idf* method, in which the weight of a term is characterized by two factors: term frequency (*tf*) and inverse document frequency (*idf*). The *tf* accounts for how often or dense a term occurs in a given document or query, while the *idf* accounts for how particular a term occurs in the whole document collection. Specifically, the weight of a term k in document i is $w_{i,k} = tf_{i,k} \times idf_k$.

Applying the concept of VSM to the speaker-clustering problem, each utterance, $\mathbf{X}_i$, $1 \le i \le N$, is represented by a *tf-idf*-based vector, $\mathbf{W}_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,K}]'$, where K is the number of terms. The similarity between any two utterances, say $\mathbf{X}_i$ and $\mathbf{X}_j$, can be computed using the cosine measure between $\mathbf{W}_i$ and $\mathbf{W}_j$:

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

Then, utterances deemed similar enough with each other are grouped into a cluster. Following our previous implementation [5], cluster generation is performed in an agglomerative manner, which starts with each utterance in its own cluster, and then successively merges the most similar pair of clusters, say c_i and c_j, according to a *complete-linkage* cluster similarity defined by

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

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 [2].

3. EIGENVOICE-MOTIVATED TERM REPRESENTATION

In using the above VSM-based clustering method, a crucial issue is how to represent a term that is capable of characterizing some significant aspects of speaker voice. Unlike the text document retrieval, which instinctively uses character-composite keywords as terms, there is no visible term in speech audio available directly for representing speaker voice characteristics. One possibility, which motivated by our previous work [5], is to expediently treat each utterance as an individual term, and represent each term by a Gaussian mixture model (GMM) trained using the speech features of the associated utterance. Each utterance can then be converted into a vector with elements (weight of terms) assigned by the likelihoods that one utterance tests against all the utterance-dependent GMMs. However, since many utterance-dependent GMMs pertain to the same speakers, the voice characteristics carried by the above utterance-based terms can be largely overlapping with each other. This results in a twisting vector space, which may limit the discriminability between utterances from the same and different speakers. To circumvent this problem, this study proposes a term representation method based on the so-called eigenvoice, with the aim of minimizing the correlation between acoustic terms.

Eigenvoice is a basis vector, derived from a number of reference speakers' voices, for representing an *a priori* voice characteristic. In its original conception, a speaker-independent voice space, consisting of several eigenvoices, is constructed by applying a dimensionality reduction technique, such as principal component analysis (PCA), on a set of speaker-dependent models. When a new speaker is present, a speaker-specific model is generated for him/her from a linear combination of the eigenvoices according to the coordinate that the new speaker's voice is located. Since the voice data of new speakers is simply used for computing the coordinates, eigenvoice technique has been shown particularly effective for speaker adaptation in terms of computational efficiency and the requirement of adaptation data.

Relatedly, the eigenvoice technique has also been used to cluster speakers for improving speech-recognition performance [6]. However, in contrast to their work, which relies on a set of extra speech data for constructing the eigenvoice space, the proposed method fully uses the data from the utterances to be clustered. This avoids the risk of environmental and channel mismatch between the extra speech data and the set of utterances to be clustered, and on the other hand, enables us to make comparisons with other speaker-clustering methods under a consistent evaluation condition.

Fig. 1 shows the procedure for generating eigenvoice-based acoustic terms. To begin, a "universal GMM" is created using all the utterances to be clustered. The training method is based on the *k*-means clustering initialization followed by expectation-maximization (EM) [10]. An adaptation of universal GMM is then performed for each of the utterances using *maximum a posteriori* (MAP) estimation [11]. This gives N utterance-dependent GMMs $\lambda_1, \lambda_2, \ldots, \lambda_N$. The use of such a model adaptation instead of a direct EM-based training of GMM has two-fold advantages. One is to produce a more reliable estimate of the GMM parameters for short utterances than it can be done with direct EM-based training. The other is to force the mixtures of all the utterance-dependent GMMs to be in the same order. This uniformity of mixture index in all GMMs is necessary for the subsequent processing.

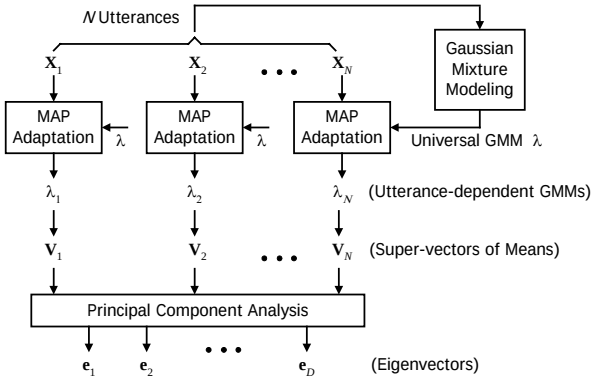


Fig. 1. Procedure for generating eigenvoice-based acoustic terms.

Next, all the mean vectors of each utterance-dependent GMM are concatenated in the order of mixture index to form a super-vector, with dimension of D. Then, PCA is applied to the set of N super-vectors, $\mathbf{V}_1$, $\mathbf{V}_2$, …, $\mathbf{V}_N$, obtained from N utterance-dependent GMMs. This yields D eigenvectors, $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_D$, ordered by the magnitude of their contribution to the between-utterance covariance matrix:

$$\mathbf{B} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{V}_i - \overline{\mathbf{V}})(\mathbf{V}_i - \overline{\mathbf{V}})', \tag{3}$$

where $\overline{\mathbf{V}}$ is the mean vector of all $\mathbf{V}_i$ for $1 \le i \le N$. The D eigenvectors constitute an eigenspace, and each of the super-vectors can be represented by a point on the eigenspace:

$$\mathbf{V}_i = \overline{\mathbf{V}} + \sum_{d=1}^{D}\phi_{i,d}\mathbf{e}_d, \tag{4}$$

where $\phi_{i,d}$, $1 \le d \le D$, is the coordinate of $\mathbf{V}_i$ on the eigenspace.

If each eigenvector is treated as an acoustic term, the importance of each term d, $1 \le d \le D$, with respect to an utterance $\mathbf{X}_i$ can be characterized by tf_{id} and idf_d, computed using

$$tf_{id} = \phi_{id}, \tag{5}$$

and

$$idf_d = \frac{1}{1 + e^{-\alpha(\sigma_d^2 + \beta)}}, \tag{6}$$

respectively, where σ_d is the standard deviation of $\phi_{i,d}$ for $1 \le i \le N$, and α and β are real constants for adjusting the sigmoid-based *idf*. Eq. (6) essentially discounts the acoustic terms which reflect less variation between utterances and hence are considered with little discriminating power.

4. BLIND RELEVANCE FEEDBACK

In analogy with document retrieval, a further improvement for speaker clustering may be made by applying *relevance feedback* (RF) [12], which refines queries using information from the documents considered relevant by users (explicit RF) or by system itself (blind RF). A typical RF method in text document retrieval is to append words from found documents known to be relevant to the keyword string of query, and then repeat the retrieval process based on the new query. Its intuitive counterpart in our task may be carried out by concatenating one utterance with others deemed similar to that utterance, and re-computing the similarity between the concatenated utterances. However, such an approach cannot control the amount of information appended from one utterance to another, and hence a severe propagation of error might happen whenever one utterance is concatenated with another different-speaker utterance. To apply RF more effectively, we propose to refine the *tf-idf*-based vectors of utterances, instead of using direct concatenation of utterances.

The basic idea is that the *tf-idf*-based vectors of utterances from an identical speaker are supposed to resemble each other, and therefore these vectors may be further rectified via a weighted average of multiple vectors deemed similar such that they can be more close to each other. To this end, let $R(i,k)$ denote the rank of inter-utterance similarity $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_k)$ among $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_1)$, $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_2)$, …, $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_N)$ in descending order, where $1 \le R(i,k) \le N$. A *tf-idf*-based vector of utterance $\mathbf{W}_i$ is rectified using

$$\hat{\mathbf{W}}_i = \sum_{k=1}^{N}\theta^{R(i,k)-1}\,\mathbf{W}_k, \tag{7}$$

where θ is a constant smaller than one. Implicit in Eq. (7) is that the new vector of utterance is a weighted sum of highly-ranked utterances' vectors. Using the rectified vectors, the inter-utterance similarity can be refined before clustering is performed.

5. EXPERIMENTAL RESULTS

Speech data used in this study consisted of 197 utterances chosen from the test set of the *2001 NIST Speaker Recognition Evaluation Corpus* [13]. The 197 utterances were spoken by 15 male speakers, and the number of utterances spoken by each speaker ranged from 5 to 39. 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.

Performance of the speaker clustering was evaluated on the basis of two metrics: cluster purity [2] and Rand Index [14]. The cluster purity, which indicates the extent of agreement in a cluster, is defined by

$$\rho_m = \sum_{p=1}^{P}\left(\frac{n_{mp}}{n_{m*}}\right)^2, \tag{8}$$

where ρ_m is the purity of the cluster c_m, n_{m*} is the total number of utterances in the cluster c_m, n_{mp} is the number of utterances in the cluster c_m that are from speaker s_p, and P is the total number of speakers involved. Eq. (8) 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

$$\overline{\rho} = \frac{1}{N}\sum_{m=1}^{M} n_{m*}\rho_m. \tag{9}$$

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

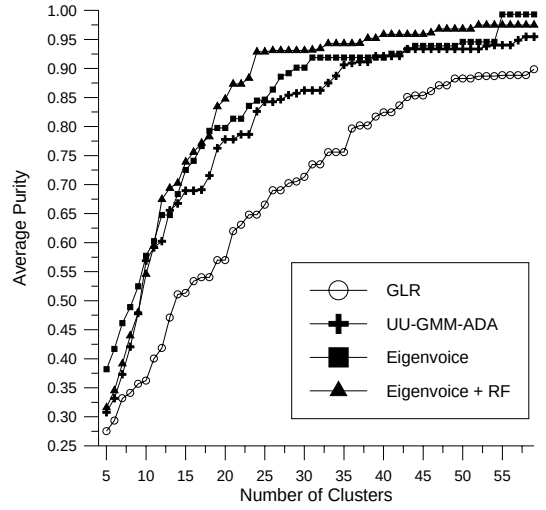
$$\gamma = \frac{1}{2}\sum_{m=1}^{M} n_{m*}^2 + \frac{1}{2}\sum_{p=1}^{P} n_{*p}^2 - \sum_{m=1}^{M}\sum_{p=1}^{P} n_{mp}^2, \tag{10}$$

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.

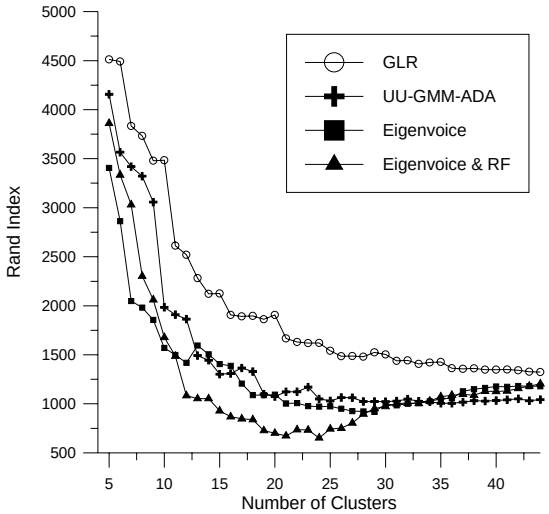
Fig. 2 shows the speaker-clustering results as a function of number of clusters. Here, "GLR" denotes the conventional hierarchical clustering method using the generalized likelihood ratio as an inter-utterance similarity measure [2]. "UU-GMM-ADA" is the best method presented in our previous work [6], which can be considered as representing an acoustic term by an utterance-dependent GMM. "Eigenvoice" denotes the proposed VSM-based clustering method using the eigenvoice-motivated term representation. The number of mixture components used in each of the utterance-dependent GMMs was empirically

determined to be 128. The values of α and β in Eq. (6) were 0.5 and 0, respectively. "Eigenvoice + RF" denotes the method "Eigenvoice" with blind relevance feedback described in Sec. 4 for refining the inter-utterance similarity measure. The value of θ in Eq. (7) was empirically set to be 0.3.

We can see from Fig. 2 that the proposed clustering methods consistently yielded higher cluster purity and lower Rand Index than the GLR-based method. Comparing the results obtained with "UU-GMM-ADA" and "Eigenvoice", it is clear that a better clustering performance can be achieved by representing utterances as vectors of acoustic terms derived from eigenvoices, instead of the utterance-dependent GMMs. It is also clear that the clustering performance can be further improved by applying the concept of relevance feedback to refine the similarity measure. When the number of clusters is equal to the speaker population ($M = P = 15$), we obtained the best cluster purity of 0.74 and Rand Index of 929, which signifies a relative improvement of more than 45%, compared to the cluster purity of 0.51 and Rand Index of 2124 obtained with the conventional GLR-based clustering method.



(a) Average cluster purity



(b) Rand Index

Fig. 2. Speaker-clustering results.

6. CONCLUSIONS

This study has presented an effective solution for speaker clustering by improving the similarity measure between speech utterances. The similarity measurement has been carried out by first converting utterances from their spectrum-based features into *tf-idf*-based vectors of acoustic terms, and then computing the cosine of vectors associated with each pair of utterances. In particular, to capture the most representative characteristics of speakers' voices, the acoustic terms have been represented as a set of eigenvectors obtained by applying the eigenvoice approach on the set of utterances to be clustered. Furthermore, through the use of blind relevance feedback, we have shown that the inter-utterance similarity measure can be further refined, and hence the performance of the hierarchical-based speaker clustering has been largely boosted in this study.

7. ACKNOWLEDGEMENT

This research was partially supported by the National Science Council, Taiwan, ROC, under Grant NSC93-2213-E-001-017.

8. REFERENCES

[1] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," *Proc. DARPA Speech Recognition Workshop'97*.

[2] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," *Proc. ICASSP'98*.

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

[4] S.S. Chen, and P.S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," *Proc. ICASSP'98*.

[5] W.H. Tsai, S.S. Cheng, and H.M. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," *Proc. ICSLP'04*.

[6] R. Faltlhauser, and G. Ruske, "Robust speaker clustering in eigenspace," *Proc. ASRU'01*.

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

[8] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio*, 8(6):695-707, 2000.

[9] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," *ACM Comm.*, 18(11):613-620, 1975.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39., pp. 1-38, 1977.

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

[12] G. Salton, Automatic text processing. Reading, MA: Addison-Wesley Publishing Company. 1989.

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

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