ON MAXIMIZING THE WITHIN-CLUSTER HOMOGENEITY OF SPEAKER VOICE CHARACTERISTICS FOR SPEECH UTTERANCE CLUSTERING

Wei-Ho Tsai and Hsin-Min Wang
Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China
{wesley,whm}@iis.sinica.edu.tw

ABSTRACT

This paper investigates the problem of how to partition unknown speech utterances into clusters, such that the overall within-cluster homogeneity of speakers' voice characteristics can be maximized. The within-cluster homogeneity is characterized by the likelihood probability that a cluster model, trained using all the utterances within a cluster, matches each of the within-cluster utterances. Such probability is then maximized by using a genetic algorithm, which determines the best cluster where each utterance should be located. For greater computational efficiency, also proposed is an alternative solution that approximates the likelihood probability with a divergence-based model similarity. The method is further designed to estimate the optimal number of clusters automatically.

1. INTRODUCTION

Motivated by the increasing need for indexing and archiving the burgeoning amount of spoken data available universally, recent research on automatic classification of speech samples based on speakers' voice characteristics has been extended from the traditional task of speaker identification/verification [1] to an unsupervised paradigm. This paradigm generally involves two problems: segmenting an audio recording into speech utterances that contain only one speaker's voice, and grouping utterances from the same speaker into a cluster. These two problems are often addressed jointly and termed *speaker diarization* [2]. It is hoped that, by locating utterances from the same speaker, the human effort required for indexing speech data can be greatly reduced from having to listen to every long audio recording to only having to check a few utterances in each cluster. In this paper, we concentrate on the latter problem, referred to as *speaker clustering* hereafter. Given N speech utterances, each of which is assumed from one of the P unknown speakers, where $N \geq P$, our aim is to partition the N utterances into M clusters, such that $M = P$, and each cluster consists of utterances from only one speaker.

Currently, the most prevalent method for speaker clustering is a hierarchical clustering (HC) framework [3-8]. It computes the similarities of vocal characteristics between utterances, and then sequentially merges the utterances deemed similar to each other (agglomerative clustering), or alternatively, separates the utterances deemed dissimilar to each other (divisive clustering). During the agglomeration or division, it is aimed to maximize the similarities between all the utterances within a cluster. However, existing similarity measures, such as Kullback Leibler distance [4], cross likelihood ratio [5,7], and generalized likelihood ratio [3,6], are performed entirely on the spectrum-based features. Such features are known to carry various types of information besides a speaker's voice characteristics, e.g., phonetic and environmental information. As a result, there is no guarantee that the similarities between same-speaker utterances will always be larger than the similarities between different-speaker utterances, especially when the utterances are short and noisy. Since the similarity computation is independent of cluster generation, and the latter trusts the former completely, the inevitable errors of similarity computation can propagate down the whole process. In addition, cluster generation based on either agglomeration or division usually uses a nearest or farthest neighborhood selection rule to determine which utterances can be assigned to the same cluster. However, since the selection rule is commonly applied in a cluster-by-cluster manner, and no consideration is taken in respect of the interaction between clusters, HC can only make each individual cluster as homogeneous as possible, but cannot guarantee that the homogeneity for all the clusters can be summed to reach a maximum.

To overcome the HC's limitations, this study proposes finding the best partitioning of speech utterances by integrating the inter-utterance similarity computation and the cluster generation into a unified process. The process iteratively assigns utterances to a set of clusters and creates a stochastic model for each cluster, which attempts to maximize the similarities between each cluster model and the within-cluster utterances. To enable an efficient and effective search for the best partitioning, we apply a model adaptation technique, model similarity comparison method, and a genetic algorithm [9] to solve this problem.

2. MAXIMUM LIKELIHOOD CLUSTERING

The proposed method begins with specifying how many clusters are to be generated. Given a specified number of clusters, M, the task is to assign N utterances $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ to M clusters $c_1, c_2, \ldots, c_M$, where each utterance is represented by a frame-based feature vector stream, i.e., $\mathbf{X}_n = \{\mathbf{x}_{n,1}, \mathbf{x}_{n,2}, \ldots, \mathbf{x}_{n,Tn}\}$. Let h_n denote the index of the cluster that an utterance, $\mathbf{X}_n$, is assigned to, where h_n is an integer between 1 and M. The goal of optimal clustering, therefore, is to produce a set of cluster indices, $\mathbf{H}^* = \{h^*_1, h^*_2, \ldots, h^*_N\}$ satisfying $h^*_n = h^*_k$ for any utterances $\mathbf{X}_n$ and $\mathbf{X}_k$ from the same speaker. To this end, we first create a Gaussian mixture model (GMM) $\lambda^{(m)}$ for each cluster c_m, $1 \leq m \leq M$, by using all the feature vectors of the utterances assigned to c_m. Then, a certain level of agreement that the utterances assigned to the same cluster come from the same speaker is characterized by computing the likelihood probability, $\Pr(\mathbf{X}_n|\lambda^{(m)})$, $\forall\ h_n = m$. Conceivably, the larger the value of $\Pr(\mathbf{X}_n|\lambda^{(m)})$, the more suitable cluster c_m for utterance $\mathbf{X}_n$ will be. Thus, by taking the likelihood probabilities for all the utterances into account, $\mathbf{H}^*$ can be determined using

$$\mathbf{H}^* = \arg\max_{\mathbf{H}} \sum_{m=1}^{M} \sum_{n=1}^{N} \left[\log \Pr(\mathbf{X}_n \mid \lambda^{(m)}) - \log \Pr(\mathbf{X}_n \mid \lambda_n)\right] \delta(h_n, m), \quad (1)$$

where λ_n is a GMM trained using $\mathbf{X}_n$, and $\delta(\cdot)$ is a Kronecker Delta function. Eq. (1) can be considered as the maximization of the generalized likelihood ratio (GLR) for all the clusters.

However, although the solution to Eq. (1) exists, no close form can be derived from this equation directly. Moreover, since the cluster indices are not scalar objects, a gradient-based optimization cannot be used in this scenario. It is also infeasible to perform an exhaustive search, which would examine all possible solutions to determine the best one, because there are M^N possible combinations of cluster indices. Recognizing these difficulties, we propose applying the genetic algorithm (GA) [9] to find $\mathbf{H}^*$ by virtue of GA's global scope and parallel searching power.

The basic operation of the GA is to explore a given search space in parallel by means of iterative modifications of a population of chromosomes. Each chromosome, encoded as a string of alphabets or real numbers called genes, represents a potential solution to a given problem. In our task, a chromosome is exactly a legitimate $\mathbf{H}$, and a gene corresponds to a cluster index associated with an utterance. However, since the index of one cluster can be interchanged with another cluster's, multiple chromosomes may amount to an identical clustering result. For example, the chromosomes {1 1 1 2 2 3 3}, {1 1 1 3 3 2 2}, and {2 2 2 1 1 3 3} represent the same clustering result for grouping seven utterances into three clusters. Such a non-unique representation of solution would significantly increase the GA search space, and may lead to an inferior clustering result. To avoid this problem, we limit the inventory of chromosome to conform a baseform representation defined as follows. Let $I(\mathrm{c}_m)$ be the lowest index of the utterance in cluster, $\mathrm{c}_m = \{\mathbf{X}_n \,|\, h_n = m;\ 1 \le n \le N\}$. A chromosome is a baseform

$$\text{iff } \forall\ \mathrm{c}_m \text{ and } \mathrm{c}_l, \text{ if } m < l, \text{ then } I(\mathrm{c}_m) < I(\mathrm{c}_l). \tag{2}$$

As can be seen from the above example, chromosome {1 1 1 2 2 3 3} is a baseform, since the lowest index of the utterance in the 1st, 2nd, and 3rd clusters is 1, 4, and 6, respectively, which satisfies Eq. (2). On the contrary, both chromosomes {1 1 1 3 3 2 2} and {2 2 2 1 1 3 3} are not a baseform, since the lowest index of the utterance in the 1st, 2nd, and 3rd clusters does not satisfy Eq. (2). It is conceivable that all the non-baseform chromosomes can be converted into a unique baseform representation by interchanging between the cluster indices.

The GA optimization starts with a random generation of chromosomes according to a certain population size Z. Then, the fitness of all chromosomes is evaluated and ranked on the basis of the overall model likelihood, i.e.,

$$\mathcal{L}(\mathbf{H}) = \sum_{m=1}^{M}\sum_{n=1}^{N}\left[\log \Pr(\mathbf{X}_n \,|\, \lambda^{(m)}) - \log \Pr(\mathbf{X}_n \,|\, \lambda_n)\right]\delta(h_n, m) \tag{3}$$

As a result of this evaluation, a particular group of chromosomes is selected from the population to generate offspring by subsequent recombination. Next, crossover among the selected chromosomes proceeds by exchanging the substrings of two chromosomes between two randomly selected crossover points. A crossover probability is assigned to control the number of offspring produced in each generation. After crossover, a mutation operator is used to introduce random variations into the genetic structure of the chromosomes. This is done by generating a random number and then replacing one gene of an existing chromosome with a mutation probability. The procedure of fitness evaluation, selection, crossover, and mutation is repeated continuously, which hopes that the overall model likelihood will increase from generation to generation. When the maximum number of generations is reached, the best chromosome in the final population is taken as the solution, $\mathbf{H}^*$.

On the other hand, as the above optimization requires that $M \times Z$ GMMs be created during each GA iteration, the computational complexity can be too high to implement properly if the parameters of the GMMs are estimated via the expectation-maximization [10]. To overcome this problem, we use a model adaptation technique to generate cluster GMMs, instead of training them from scratch. Specifically, the method, stemming from the GMM-UBM method [11] for speaker verification, is to create a cluster-independent GMM λ using all the utterances to be clustered, followed by an adaptation of the cluster-independent GMM for each of the clusters using *maximum a posteriori* (MAP) estimation.

3. MINIMUM DIVERGENCE CLUSTERING

In addition to the training of cluster GMMs, another issue concerning the realization of Eq. (1) is the considerable complexity of likelihood computation. Specifically, the standard procedure for computing $\Pr(\mathbf{X}_n|\lambda^{(m)})$ is $\Pi_{t=1}^{T_n}\Sigma_{j=1}^{J} w_j^{(m)}\mathcal{N}(\mathbf{x}_{n,t};\boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)})$, which requires $T_n \times J$ computation of Gaussian density $\mathcal{N}(\cdot)$. Thus, each GA iteration involves $Z \times J \times (\Sigma_{n=1}^{N} T_n)$ computation of Gaussian density. When the number of utterances to be clustered is large, the whole clustering process can be extremely time consuming. To overcome this problem, we further propose a clustering method based on an approximation of the likelihood by a computationally more tractable metric, called *divergence* [12].

Recall that the likelihood $\Pr(\mathbf{X}_n|\lambda^{(m)})$ represents how well the cluster GMM $\lambda^{(m)}$ fits the distribution of the feature vectors of $\mathbf{X}_n$. If we characterize the distribution of the feature vectors of $\mathbf{X}_n$ by utterance GMM λ_n, the computation of $\Pr(\mathbf{X}_n|\lambda^{(m)})$ should be roughly equivalent to a certain similarity measurement between GMMs $\lambda^{(m)}$ and λ_n. Let $\{w_{n,i}, \boldsymbol{\mu}_{n,i}, \boldsymbol{\Sigma}_{n,i},\ 1 \le i \le J\}$ be the parameters of λ_n estimated via MAP adaptation from GMM λ. The similarity between GMMs $\lambda^{(m)}$ and λ_n can be measured by [13]

$$\mathcal{S}(\lambda^{(m)}, \lambda_n) = \sum_{j=1}^{J}\sum_{i=1}^{J} w_j^{(m)} w_{n,i} \exp[-\mathcal{D}(\boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)}; \boldsymbol{\mu}_{n,i}, \boldsymbol{\Sigma}_{n,i})], \tag{4}$$

where $\mathcal{D}(\boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)}, \boldsymbol{\mu}_{n,i}, \boldsymbol{\Sigma}_{n,i}) =$

$$\begin{aligned} &\frac{1}{2}\left(\boldsymbol{\mu}_j^{(m)} - \boldsymbol{\mu}_{n,i}\right)' \left(\boldsymbol{\Sigma}_j^{(m)^{-1}} + \boldsymbol{\Sigma}_{n,i}^{-1}\right)\left(\boldsymbol{\mu}_j^{(m)} - \boldsymbol{\mu}_{n,i}\right) \\ &+ \frac{1}{2}\mathrm{Tr}\left\{\left(\boldsymbol{\Sigma}_j^{(m)^{1/2}}\boldsymbol{\Sigma}_{n,i}^{-1/2}\right)\left(\boldsymbol{\Sigma}_j^{(m)^{1/2}}\boldsymbol{\Sigma}_{n,i}^{-1/2}\right)'\right\} \\ &+ \frac{1}{2}\mathrm{Tr}\left\{\left(\boldsymbol{\Sigma}_j^{(m)^{-1/2}}\boldsymbol{\Sigma}_{n,i}^{1/2}\right)\left(\boldsymbol{\Sigma}_j^{(m)^{-1/2}}\boldsymbol{\Sigma}_{n,i}^{1/2}\right)'\right\} - R, \end{aligned} \tag{5}$$

is the divergence between Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)})$ and $\mathcal{N}(\boldsymbol{\mu}_{n,i}, \boldsymbol{\Sigma}_{n,i})$, $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, and R is the dimension of the feature vectors. For greater computational efficiency, we keep the mixture weights unchanged during MAP adaptation, i.e., $w_j^{(m)} = w_{n,j} = w_i,\ 1 \le j \le J$. Since the mixture components of $\lambda^{(m)}$ and λ_n are aligned, Eq. (4) can be simplified as

$$\mathcal{S}(\lambda^{(m)}, \lambda_n) = \sum_{j=1}^{J} w_j \exp[-\mathcal{D}(\boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)}; \boldsymbol{\mu}_{n,j}, \boldsymbol{\Sigma}_{n,j})]. \tag{6}$$

A large value of $\mathcal{S}(\cdot)$ signifies a large level of homogeneity between the utterances within a cluster. Thus, speaker clustering can be converted into a problem of finding $\mathbf{H}^*$ satisfying

$$\mathbf{H}^* = \arg\max_{\mathbf{H}} \sum_{m=1}^{M} \sum_{n=1}^{N} \log S(\lambda^{(m)}, \lambda_n) \delta(h_n, m). \quad (7)$$

We refer to this clustering method as minimum divergence clustering. Since Eq. (7) is independent on the length of utterance, the computation complexity can be dramatically reduced, compared to the maximum likelihood clustering method.

4. ESTIMATION OF SPEAKER POPULATION SIZE

The proposed method described above is developed on the basis that a certain number of clusters is specified in advance. However, the optimal number of clusters, equal to the speaker population size, is unknown and must be estimated. To do this, we propose examining all the possible partitionings of *N* utterances with the numbers of clusters ranging from 1 to *N*, and then selecting one of the partitionings associated with the level of within-cluster homogeneity as high as possible and the number of clusters as small as possible. Such selection can be made with the Bayesian information criterion (BIC) [6,14].

The BIC scores a parametric model based on how well the model fits a data set, and how simple the model is:

$$\mathrm{BIC}(\Lambda) = \log \Pr(\mathbf{O} \mid \Lambda) - 0.5\, \gamma\, \#(\Lambda) \log |\mathbf{O}|, \quad (8)$$

where $\#(\Lambda)$ denotes the number of free parameters in model Λ, $|\mathbf{O}|$ is the size of the data set $\mathbf{O}$, and γ is a penalty factor. The larger the value of BIC(Λ), the better model Λ will perform. If we treat each of the possible partitionings as a parametric model for characterizing the speaker information of the utterances, the BIC for a model having *M* clusters can be conceptually computed by

$$\mathrm{BIC}(M \text{ Clusters}) \equiv \sum_{m=1}^{M} \sum_{n=1}^{N} \log S(\lambda^{(m)^*}, \lambda_n) \delta(h_n^*, m) - \frac{1}{2} \gamma\, M \log N, \quad (9)$$

where $\lambda^{(m)*}$ is the resulting GMM of cluster c_m after the optimization according to Eq. (7). In Eq. (9), we use the divergence-based similarity measurement to represent how well the model fits the data, which approximates the probability $\Pr(\mathbf{O}|\Lambda)$. The BIC value should increase with the increase of the *M* value initially, but will decline significantly after an excess of clusters is created. Thus, a reasonable number of clusters can be determined by choosing the partitionings that produces the largest value of BIC:

$$M^* = \arg\max_{1 \le M \le N} \mathrm{BIC}(M \text{ Clusters}). \quad (10)$$

5. EXPERIMENTS

Our speech data consisted of 197 utterances chosen from the test set of the *2001 NIST Speaker Recognition Evaluation Corpus* [15], which contains cellular telephone speech collected by the Linguistic Data Consortium. 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, were extracted from this data using a 20-ms Hamming-windowed frame with 10-ms frame shifts.

The performance of speaker clustering was evaluated on the basis of two metrics: cluster purity [5] and the Rand Index [16]. Cluster purity is the probability that if we pick any utterance from a cluster twice at random, with replacement, both of the picked utterances are from the same speaker. Specifically, the purity of cluster c_m is computed by

$$\rho_m = \sum_{p=1}^{P} \left(n_{mp} / n_{m*} \right)^2, \quad (11)$$

where n_{m*} is the total number of utterances in cluster c_m, and n_{mp} is the number of utterances in cluster c_m that were produced by the *p*-th speaker. The overall performance of *M*-clustering is evaluated by an average purity:

$$\bar{\rho} = \left(\sum_{m=1}^{M} n_{m*} \rho_m \right) \Big/ N. \quad (12)$$

The Rand Index, which indicates the probability that two randomly-selected utterances produced by the same speaker are grouped into different clusters, or that two randomly-selected utterances grouped into the same cluster are produced by different speakers, is defined by

$$\chi = \left[\sum_{m=1}^{M} \binom{n_{m*}}{2} + \sum_{p=1}^{P} \binom{n_{*p}}{2} - 2 \sum_{m=1}^{M} \sum_{p=1}^{P} \binom{n_{mp}}{2} \right] \Big/ \left[\sum_{m=1}^{M} \binom{n_{m*}}{2} + \sum_{p=1}^{P} \binom{n_{*p}}{2} \right] \quad (13)$$

where n_{*p} is the number of utterances from the *p*-th speaker. Note that the higher the value of purity, or the lower the value of Rand Index, the better the clustering performance is.

Our first experiment was conducted to assess the performance of our speaker-clustering methods under the condition that the number of clusters is specified as the speaker population size ($M = P = 15$). For performance comparison, a baseline system using GLR-based similarity computation followed by agglomerative clustering [3,5] (referred to as GLR-AC hereafter) was also evaluated. Table I shows the performance of the GLR-AC system. We examined GLR computed with different numbers of component densities in Gaussian mixture modeling. Except for the single-Gaussian models, which were full covariance structures and trained via maximum likelihood estimation, all the GMMs used in this study comprised diagonal covariance matrices and were trained via MAP-adaptation. It can be seen from Table I that speaker-clustering performance is less sensitive to the structure of Gaussian mixture modeling, but is rather sensitive to the choice of linkage for inter-cluster similarity computation. Overall, complete linkage performs the best, whereas single linkage performs the worst, and average linkage is between these two extremes. However, all methods were far from accurate for this task.

Table II shows the speaker-clustering results obtained by our proposed methods, namely, maximum likelihood clustering (MLC) and minimum divergence clustering (MDC). In GA optimization, the parameter values used for the maximum number of generations, the population size, the crossover probability, and the mutation probability were empirically determined to be 3000, 200, 0.3, and 0.1, respectively. Comparing Table I with II, it is clear that both MLC and MDC are superior to GLR-AC. On the whole, both MLC and MDC yield an average cluster purity above 0.7 and a Rand Index below 0.4, which signifies a relative improvement of more than 30%, compared to the cluster purity of 0.52 and the Rand Index of 0.53 obtained with GLR-AC. In addition, we can see from Table II that the performance of MLC is slightly better than that of MDC. However, as mentioned earlier, MLC is rather computationally extensive, due to the need to compute Gaussian densities frame by frame. Quantitatively, MLC required 2000 times the computational time of MDC for this clustering task, and took two weeks to complete a trial on a 3 GHz Pentium PC. This makes it difficult to use MLC to determine how many clusters should be generated if the speaker population size is not known in advance. Therefore, in the following experiments, we concentrated on examining the validity of MDC-based speaker clustering.

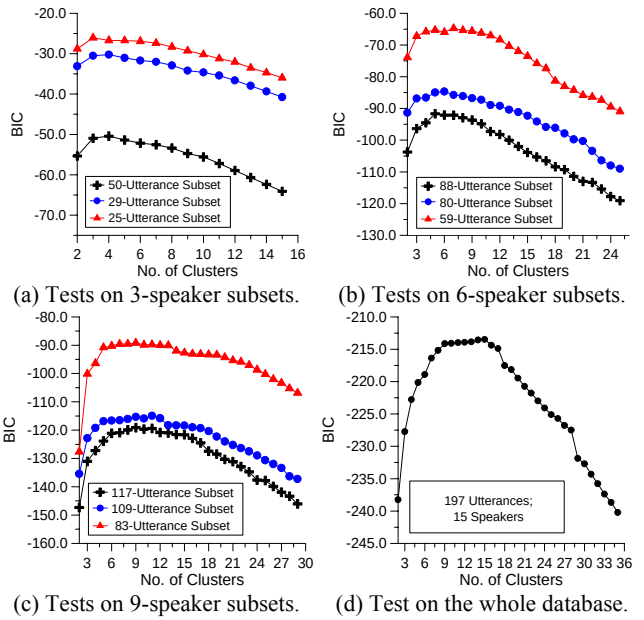
To investigate if the optimal number of clusters can be determined by Eq. (10), we divided the database into several subsets involving speaker population sizes of 3, 6, and 9. We then conducted clustering experiments using these subsets separately to examine if the automatically-determined numbers of clusters could be around 3, 6, and 9, respectively. For each of the speaker population sizes, we organized three subsets that were mutually distinct, as far as possible, in terms of the speakers involved. The penalty factor, γ in Eq. (9) was empirically set to 1.2 throughout this experiment. Fig. 1 shows the resulting BIC values as a function of the number of clusters. The peak of each curve in the figure indicates the optimal number of clusters according to the BIC. We can see that most of the peaks appeared near the actual number of speakers, and the BIC values declined significantly after an excess of clusters was created. This result validates the proposed method for estimating the speaker population size.

Table 1: Speaker-clustering results (Purity / Rand Index) obtained with the GLR-AC method.

No. of Gaussian Mixtures	Inter-cluster Similarity		
	Complete Linkage	Average Linkage	Single Linkage
1	0.51 / 0.54	0.38 / 0.69	0.17 / 0.83
2	0.48 / 0.63	0.36 / 0.75	0.17 / 0.84
4	0.50 / 0.57	0.39 / 0.68	0.16 / 0.84
8	0.52 / 0.53	0.35 / 0.78	0.16 / 0.84
16	0.52 / 0.54	0.34 / 0.79	0.16 / 0.84
32	0.51 / 0.55	0.33 / 0.79	0.16 / 0.84

Table 2: Speaker-clustering results (Purity / Rand Index) obtained with the proposed MLC and MDC methods.

No. of Gaussian Mixtures	Clustering Method	
	MLC	MDC
16	0.76 / 0.28	0.72 / 0.35
32	0.78 / 0.27	0.75 / 0.29
64	0.75 / 0.29	0.73 / 0.32



(a) Tests on 3-speaker subsets. (b) Tests on 6-speaker subsets.

(c) Tests on 9-speaker subsets. (d) Test on the whole database.

Fig. 2. BIC values as a function of number of clusters.

6. CONCLUSIONS

This paper has studied methods for clustering speech utterances so that the level of within-cluster homogeneity can be maximized in terms of speakers' voice characteristics. Such homogeneity has been characterized by either the likelihood probability that a cluster model tests for each of the within-cluster utterances, or the divergence-based similarity between a cluster model and each of the within-cluster utterance models. To maximize the overall within-cluster homogeneity, we have proposed applying a genetic algorithm to determine the best cluster where each utterance should be located. Experimental results show that the proposed method achieved a relative improvement of more than 30% in speaker-clustering performance, compared to the conventional method using GLR-based similarity measurement followed by agglomerative hierarchical clustering. In addition, the proposed clustering method incorporates the Bayesian information criterion to determine how many clusters should be generated. The experimental results show that the automatically-determined number of clusters approximates the actual speaker population size.

7. REFERENCES

[1] Campbell, J. P. "Speaker recognition: a tutorial," *Proc. IEEE*, 85(9):1437-1462, 1997.

[2] http://www.nist.gov/speech/tests/rt/rt2006/spring/.

[3] Gish, H., Siu, M. H., and Rohlicek, R. "Segregation of speakers for speech recognition and speaker identification," *Proc. ICASSP'91*.

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

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

[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] Tsai, W. H., Cheng, S. S., Chao, Y. H., and Wang, H. M., "Clustering speech utterances by speaker using eigenvoice-motivated vector space model", *Proc. ICASSP'05*.

[9] Goldberg, D. E. Genetic Algorithm in Search, Optimization and Machine Learning. New York: Addison-Wesley, 1989.

[10] Dempster, A., Laird, N., and Rubin, D. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 1977.

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

[12] Kullback, S. Information Theory and Statistics. New York: Dover, 1968.

[13] Huang, C. S., Wang, H. C., and Lee, C. H. "A study on model based error rate estimation for automatic speech recognition" *IEEE Trans. Speech and Audio Proc.*, 11(6):581-589, 2003.

[14] Schwarz, G. "Estimating the Dimension of a Model," *The Annals of Statistics* 6:461-464, 1978.

[15] http://www.nist.gov/speech/tests/spk/2001/index.htm

[16] Rand, W. M. "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, 66:846-850, 1971.