Automatic Speaker Clustering Using A Voice Characteristic Reference Space And Maximum Purity Estimation

Wei-Ho Tsai[1], *Shih-Sian Cheng*[2], and *Hsin-Min Wang*[2]

[1]Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan
E-mail: whtsai@en.ntut.edu.tw, Phone: +886-2-27712171 ext. 2257, Fax: +886-2-27317120

[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: whm@iis.sinica.edu.tw, Phone: +886-2-27883799 ext. 1714, Fax: +886-2-27824814

Abstract

This paper investigates the problem of automatically grouping unknown speech utterances based on their associated speakers. In attempts to determine which utterances should be grouped together, it is necessary to measure the voice similarities between utterances. Since most existing methods measure the inter-utterance similarities based directly on the spectrum-based features, the resulting clusters may not be well-related to speakers, but to various acoustic classes instead. This study remedies this shortcoming by projecting utterances onto a reference space trained to cover the generic voice characteristics underlying the whole utterance collection. The resultant projection vectors naturally reflect the relationships of voice similarities among all the utterances, and hence are more robust against interference from non-speaker factors. Then, a clustering method based on maximum purity estimation is proposed, with the aim of maximizing the similarities between utterances within all the clusters. This method employs a genetic algorithm to determine the cluster to which each utterance should be assigned, which overcomes the limitation of conventional hierarchical clustering that the final result can only reach the local optimum. The proposed clustering method also incorporates a Bayesian information criterion to determine how many clusters should be created.

Index Terms: speaker clustering, maximum purity estimation, genetic algorithm

EDICS: SPE-SPKR Speaker Characterization and Recognition

I. INTRODUCTION

Speaker clustering refers to the task of grouping together unknown speech utterances based on a speaker's voice characteristics. For more than a decade, the interests and needs of the speech-recognition community have provided a major motivation for the work on speaker clustering [1]-[3], in which the major purpose is to group together speech data produced by the same speaker or speakers with similar voices such that the adaptation of acoustic models can be carried out more effectively. However, since speaker clustering simply serves as a supplementary process in speech recognition, there is still a dearth of studies devoted to this problem. More recently, speaker-clustering research has enjoyed a renaissance [4]-[19], spurred by research into spoken document indexing for managing burgeoning collections of available speech data. The main purpose of such an emerging research topic is that, by grouping speech data from the same speakers, the human effort required for documentation can be dramatically reduced, or even replaced.

Speaker clustering can be viewed as an unsupervised speaker-recognition problem, in which the process of speaker recognition [20] is concerned with determining the identity of a speaker (speaker identification) or determining if a speaker is who he/she claims to be (speaker verification). However, in contrast to the conventional speaker-recognition approach, which assumes that some prior information or speech data is available and can be modeled from the speakers concerned, speaker clustering must work without any knowledge of who the possible speakers are and how many are involved in the utterances to be clustered. Consequently, solutions to the speaker-clustering problem should be capable of extracting and comparing the voice characteristics underlying the utterance collections in an unsupervised manner.

A related task is speaker segmentation [5] [10] [11], which aims to locate the boundaries when there is a change of speaker in an audio stream containing multiple persons' speech utterances. In tandem with speaker clustering, speaker segmentation breaks up the continuous input into discrete utterances that are easy to process in speech/speaker recognition, and is, therefore, an essential step in spoken document indexing. Viewed from another angle, speaker segmentation may be accomplished with the aid of speaker clustering [14]. This is done by first segmenting an audio stream uniformly into a sequence of short regions that can be considered homogeneous in terms of their associated speaker, and then clustering the short regions and assigning each of them

a cluster index representing a speaker attribute. A change of speakers may occur between two adjacent short regions that have different cluster indices.

Currently, most speaker-clustering methods follow a hierarchical clustering framework [4]-[11], comprised of three major components: computation of inter-utterance similarities, generation of a cluster tree, and determination of the number of clusters. The similarity computation is designed 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 the Kullback Leibler (KL) distance [5], the cross likelihood ratio (CLR) [8], and the generalized likelihood ratio (GLR) [2][6][17], have been examined and compared in much of the literature. The generation of a cluster tree is done in either a bottom-up (agglomerative) or a top-down (divisive) fashion, according to some criteria derived from the similarity measure. The bottom-up approach starts with each utterance as a single cluster, and then successively merges the clusters in a pairwise manner until one cluster contains all the utterances. In the top-down approach, however, all the utterances start in a single cluster, and the clusters are successively split until each cluster contains exactly one utterance. The resulting cluster tree is then cut via an estimation of the number of clusters to retain the best partition. Representative methods for estimating the optimal number of clusters are based on the *BBN Metric* [6] and the Bayesian Information Criterion [7].

Among the three components in the above clustering framework, the computation of inter-utterance similarities is of particular importance, which determines whether the generated clusters are related to various speakers, rather than other acoustic classes. However, existing similarity measures, based on KL distance, CLR, or GLR, are performed entirely on the spectrum-based features, which are known to carry various types of information besides a speaker's voice characteristics, for example, phonetic and environmental information. As a consequence, speaker-clustering systems based on these similarity measures are usually vulnerable when the utterances to be clustered are short and noisy. To alleviate the problem, this study proposes a novel inter-utterance similarity measurement, which is carried out by projecting the utterances to be clustered onto a voice characteristic reference space, and then examining the degree of coincidence between the projection results of the utterances. As will be illustrated in the following sections, the reference space is trained to cover the generic voice characteristics inherent in all the utterances to be clus-

tered; hence, the resulting similarity measurement will be more robust against interference from non-speaker factors.

In addition to developing a more reliable inter-utterance similarity measurement, we also investigate how to optimally generate the clusters such that all the within-cluster utterances are from the same speaker. Conventional approaches based on either top-down or bottom-up hierarchical clustering use a nearest neighborhood selection rule to determine which utterances should be assigned to the same cluster. However, during the procedure of splitting one cluster or merging two clusters, the nearest neighborhood selection rule is applied in a cluster-by-cluster manner, rather than in a global manner that considers all the clusters. As a result, hierarchical clustering can only make each individual cluster as homogeneous as possible, but cannot attain the ultimate goal of maximizing the overall homogeneity. To solve this problem, we propose a new clustering method that explicitly aims to maximize the total number of within-cluster utterances from the same speakers. This is done by estimating the so-called *cluster purity* [6], in conjunction with an optimization process based on a *genetic algorithm* [21], to find the best partitioning of utterances that achieves maximal cluster purity.

The rest of this paper is organized as follows. Section II describes the specific problem we address, an overview of the clustering framework we propose, and the performance assessment method we use in this study. Section III introduces several methods for creating a reference voice space that represents an utterance as a projection vector, thereby measuring the similarities between utterances. In Section IV, we describe how to generate clusters in accordance with the criteria derived from the inter-utterance similarities. Section V discusses the problem of how to automatically determine the appropriate number of clusters. Section VI presents our experimental results. Finally, in Section VII, we present our conclusions, and discuss the direction of future works.

II. TASK DEFINITION AND METHOD OVERVIEW

Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ denote N unlabeled speech utterances in a certain spectrum-based feature representation, each of which is produced by one of the P speakers $s_1, s_2, \ldots, s_P$, where $N \geq P$,

and P is unknown. The aim of speaker clustering is to partition the N utterances into M clusters $c_1, c_2, \ldots, c_M$, such that $M = P$ and each cluster consists exclusively of utterances from only one speaker. For those utterances that contain multiple speakers, the partitioning is preferably performed after the utterances are pre-segmented into speaker-homogeneous regions [1]. However, in order to focus on the fundamental techniques for speaker clustering, this study does not investigate the speaker-segmentation problem and only deals with utterances containing a single speaker.

The performance [2] of speaker clustering is evaluated on the basis of cluster purity [6], which indicates the level of agreement in a cluster. For a cluster c_m, the purity ρ_m is computed by

$$\rho_m = \sum_{p=1}^{P} \frac{n_{mp}^2}{n_m^2}, \tag{1}$$

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 are produced by speaker s_p. From Eq. (1), it follows that $n_m^{-1} \leq \rho_m \leq 1$, in which the upper bound and lower bound reflect that all the within-cluster utterances were produced by the same speaker or completely different speakers, respectively. To evaluate the overall performance of M-clustering, we compute an average purity:

$$\bar{\rho} = \frac{1}{N} \sum_{m=1}^{M} n_m \rho_m. \tag{2}$$

Fig. 1 shows the proposed speaker-clustering framework. Prior to the inter-utterance similarity computation, a reference space, which represents some generic characteristics of speakers' voices, is constructed. The reference space is composed of K bases, where the basis is a general term referring to a representative of the voice characteristics encoded in the spectrum-based features. The reference space can be created using either the set of utterances to be clustered or an extra speech database. The use of the latter might allow a greater variety of voice characteristics to be covered and make it easier to perform on-line or incremental clustering; however, there might be a risk of environmental or channel mismatch between the reference space and the set of utterances to be clustered. In this study, we do not use an extra speech database to train the reference space,

[1] Interested readers are referred to [9] for the study of clustering multi-speaker utterances.

[2] Depending on the application, there are a number of other ways to evaluate the speaker-clustering performance, such as the misclassification rate [17], clustering efficiency [8], and the Rand Index [22]. This study chooses cluster purity, because its computation is more application-independent and its scale is more easily perceivable.

for the sake of comparing the performance of our method with that of other speaker clustering methods under a consistent evaluation condition.

After a reference space is constructed, each of the N utterances, say $\mathbf{X}_i$, is converted 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 $\mathbf{V}_j$ would be relatively similar in some sense, resulting in $\mathbf{V}_i$ being closer to $\mathbf{V}_j$, instead of $\mathbf{V}_l$ for any utterance $\mathbf{X}_l$ not from s_p.

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

$$\mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j) = \frac{\mathbf{V}_i \cdot \mathbf{V}_j}{\|\mathbf{V}_i\| \|\mathbf{V}_j\|}. \tag{3}$$

Accordingly, utterances deemed similar enough to each other can be grouped into a cluster. Since no information about the speaker population size P is available beforehand, we first produce a set of possible partitionings with the numbers of clusters ranging from 2 to N, and then choose the best partitioning using a method for estimating the optimal number of clusters.

Comparing with most existing speaker-clustering methods, the proposed system tries to improve clustering performance from two perspectives. One is to exploit "out-of-pair" spectral information to help measure the similarity between each pair of utterances. In contrast to most existing methods, which compute the inter-utterance similarities by considering only the spectral information from a pair of utterances at a time, the proposed system measures the similarity between each pair of utterances by referring to all the utterances to be clustered. Since the reference space in our system is built to capture the generic characteristics of speakers' voices, the resulting inter-utterance similarity measurement can be less text- and environment- dependent. From an alternate perspective, we attempt to enhance clustering performance by generating clusters in a global fashion, rather than in a piecemeal manner used in most existing systems. As will be illustrated in Sec. IV, the cluster generation is done with the aim of maximizing the total number of with-cluster utterances from the same speakers.

III. REFERENCE SPACE CREATION

The effectiveness of the above speaker-clustering framework crucially depends on whether a reference space is capable of summarizing the most relevant aspects of speakers' voice characteristics inherent in the observed speech data. This section presents four possibilities for reference space construction, namely, utterance-dependent Gaussian mixture modeling, utterance-independent vector clustering, utterance-independent Gaussian mixture modeling followed by MAP adaptation [23], and eigenvoice-motivated reference space [24].

A. Utterance-dependent Gaussian mixture modeling

For the conventional speaker-recognition problem, Gaussian mixture modeling is the predominant method for characterizing speaker-specific voice patterns [26]. The main attraction of the Gaussian mixture model (GMM) is its ability to provide smooth approximations of arbitrarily-shaped densities of a 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 $\phi_k, 1 \leq k \leq N$, is then computed using

$$v(\mathbf{X}_i, \phi_k) = \log \Pr(\mathbf{X}_i|\lambda_k) - \log \Pr(\mathbf{X}_i|\lambda_i). \tag{4}$$

Eq. (4) is the normalized likelihood probability that utterance $\mathbf{X}_i$ comes from the speaker characterized by GMM ϕ_k. Ideally, the value of $v(\mathbf{X}_i, \phi_k)$ would be large if utterances $\mathbf{X}_i$ and $\mathbf{X}_k$ are from the same speaker, and would be small otherwise. In practice, however, this cannot be guaranteed, since the GMMs may not always be capable of characterizing the speakers' voices well, especially when the utterances are of very limited duration and subject to diverse environmental conditions. In view of such imperfections, we hope that by using a whole projection vector, the impact of some abnormal projection values can be diluted by other normal ones, and a more reliable similarity measure can be derived. Fig. 2 shows an example of projection carried out on a collection of nine utterances from three speakers. Dark regions in the resulting likelihood

pattern represent large likelihood values, while light regions represent small values. We can see from the likelihood pattern that the whole projection vectors pertaining to the same speakers are more similar than those pertaining to different speakers.

The concept of the above projection method is adapted from a prior study reported in [27]. A similar idea has also been presented recently from the viewpoint of so-called *triangulation* [16], in which each utterance is modeled as a single Gaussian distribution. It is clear from speaker-recognition research that a better speaker clustering performance may be obtained by using a proper number of Gaussian components in a mixture, rather than a single Gaussian density. However, determining the proper number of Gaussian components in GMMs is a difficult problem, especially when the duration of the utterances might be rather diverse. Specifically, choosing a larger number of Gaussian components is advantageous for modeling the voice characteristics of long utterances more accurately, but it is disadvantageous for the short utterances that lack data for GMM parameter estimation. Meanwhile, choosing too few Gaussian components may make it difficult to distinguish between different-speaker utterances. We therefore develop several alternative methods in the following subsections to sidestep this problem.

B. Utterance-independent vector clustering

Instead of using utterance-dependent GMMs, we can create a single, utterance-independent codebook with R codewords as a reference space using all the 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 $\mathbf{w}_k, 1 \leq k \leq R$, consists of a mean vector $\boldsymbol{\mu}_k$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_k$. 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. Specifically, we found that after vector clustering, the feature vectors pertaining to a particular speaker do not spread uniformly over all the clusters, but distribute primarily in certain clusters. In other words, each of the speakers reflects his/her own set of favorable codewords. This might be because in unsupervised training

of the codebook, there is usually more than one codeword to represent a certain type of phonetic realization. Many codewords that correspond to identical phonetic fragments are generated to cover the variations of different speakers or environmental conditions. This results in codewords that are phonetically-related, as well as speaker-related.

Viewing each codeword as a basis of the reference space, the similarity between utterances can be measured by comparing the distribution for the feature vectors of each utterance in the codebook. To do this, each feature vector is assigned an index of the closest codeword in terms of the Mahalanobis distance. The projection value $v(\mathbf{X}_i, \phi_k)$ for each utterance $\mathbf{X}_i$ with respect to basis $\phi_k, 1 \leq k \leq R$, can then be computed by using

$$v(\mathbf{X}_i, \phi_k) = \frac{\text{Number of the feature vectors in } \mathbf{X}_i \text{ assigned as } \mathbf{w}_k}{\text{Number of total feature vectors in } \mathbf{X}_i}. \tag{5}$$

C. Utterance-independent Gaussian mixture modeling followed by MAP adaptation

Alternatively, the problem concerning diverse utterance duration mentioned in Sec. III-A might be handled better by using some model-adaptation techniques developed in speech or speaker recognition research. Our basic strategy is to create an utterance-independent GMM using all the utterances to be clustered, followed by an adaptation of the utterance-independent GMM for each of the utterances using maximum *a posteriori* (MAP) estimation.

Let $\lambda = \{\omega_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, 1 \leq j \leq J\}$ denote the parameter set of an utterance-independent GMM having J Gaussian components, where ω_j is the mixture weight, $\boldsymbol{\mu}_j$ the mean vector, and $\boldsymbol{\Sigma}_j$ the covariance matrix. For each utterance $\mathbf{X}_i$, with T_i feature vectors $\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,T_i}\}$, we compute the *a posteriori* probability of each feature vector $\mathbf{x}_{i,t}$, in the Gaussian component j of GMM λ:

$$\Pr(j|\mathbf{x}_{i,t}) = \frac{\omega_j \mathcal{N}(\mathbf{x}_{i,t}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{J} \omega_l \mathcal{N}(\mathbf{x}_{i,t}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}, \tag{6}$$

where $\mathcal{N}(\cdot)$ is a Gaussian density function. Then, an adapted GMM, $\lambda_i = \{\omega_{i,j}, \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}, 1 \leq j \leq J\}$, for each utterance $\mathbf{X}_i$ is obtained from

$$\boldsymbol{\mu}_{i,j} = \tau_{i,j} E_j(\mathbf{X}_i) + (1 - \tau_{i,j}) \boldsymbol{\mu}_j, \tag{7}$$

$$\boldsymbol{\Sigma}_{i,j} = \tau_{i,j} E_j(\mathbf{X}_i \mathbf{X}_i') + (1 - \tau_{i,j})(\boldsymbol{\mu}_j \boldsymbol{\mu}_j' + \boldsymbol{\Sigma}_j) - \boldsymbol{\mu}_{i,j} \boldsymbol{\mu}_{i,j}', \tag{8}$$

$$\omega_{i,j} = [\tau_{i,j} \zeta_{i,j} / T_i + (1 - \tau_{i,j}) \omega_j] \vartheta, \tag{9}$$

where ϑ is a scale factor that ensures all the mixture weights sum to unity, and $\tau_{i,j}$, $\zeta_{i,j}$, $E_j(\mathbf{X}_i)$, and $E_j(\mathbf{X}_i\mathbf{X}_i')$ are computed using

$$\tau_{i,j} = \frac{\zeta_{i,j}}{\zeta_{i,j} + \varepsilon}, \tag{10}$$

$$\zeta_{i,j} = \sum_{t=1}^{T_i} \Pr(j|\mathbf{x}_{i,t}), \tag{11}$$

$$E_j(\mathbf{X}_i) = \frac{1}{\zeta_{i,j}} \sum_{t=1}^{T_i} \Pr(j|\mathbf{x}_{i,t})\mathbf{x}_{i,t}, \tag{12}$$

$$E_j(\mathbf{X}_i\mathbf{X}_i') = \frac{1}{\zeta_{i,j}} \sum_{t=1}^{T_i} \Pr(j|\mathbf{x}_{i,t})\mathbf{x}_{i,t}\mathbf{x}_{i,t}', \tag{13}$$

where ε is a relevance factor.

This strategy stems from the GMM-UBM method [28] for speaker recognition, in which the required speaker-specific models are created by tuning the parameters of a universal speaker model pre-trained by using speech data from many speakers. The GMM-UBM method has proven very effective, especially when only limited enrollment data is available, and could be advantageous in speaker clustering.

D. Eigenvoice-motivated Reference Space

We have explored three possibilities for constructing a reference space, with the aim of better covering the voice characteristics of the utterances to be clustered. However, one potential problem with the three methods is that the bases of the reference spaces are not statistically-independent of each other. The characteristic overlapping between bases may lead to a twisted reference space, which limits the ability to discriminate between utterances from the same and different speakers. To deal with this problem, we now describe a reference space creation method that applies the technique of *eigenvoice* [29] to minimize the characteristic overlapping between bases.

Eigenvoices are derived from a number of reference speakers' voices, to represent an *a priori* voice characteristic. In its original form, 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 on which the new speaker's voice is located. Since the voice data of new speakers is only used for computing coordinates, the eigenvoice technique has proven particularly effective for speaker adaptation in terms of computational efficiency and the requirements of adaptation data. The technique has also been applied to cluster speakers to improve speech-recognition performance [13]. In contrast to the work in [13], which relies on a set of extra speech data to construct the eigenvoice space, the proposed method fully utilizes the data from the utterances to be clustered.

The procedure for reference space creation begins with MAP adaptation, described in Sec. III-C, to generate N utterance-dependent GMMs. All the mean vectors of each GMM are concatenated in the order of the Gaussian component index to form a super-vector with the dimension of D. Then, PCA is applied on the set of N super-vectors, $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N$, obtained from the N GMMs. This yields D eigenvectors, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D$, ordered by the degree of their contribution to the between-utterance covariance matrix:

$$\mathbf{B} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{U}_i - \bar{\mathbf{U}})(\mathbf{U}_i - \bar{\mathbf{U}})', \tag{14}$$

where $\bar{\mathbf{U}}$ is the mean vector of all $\mathbf{U}_i$ for $1 \leq i \leq N$. To capture the most representative voice characteristics, we only retain low-order K ($K < D$) eigenvectors with larger eigenvalues that reflect more variation between utterances. The K eigenvectors, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K$, serve as the bases of an eigenvoice-motivated reference space; hence, the projection value $v(\mathbf{X}_i, \phi_k)$ for utterance $\mathbf{X}_i$ with respect to basis $\phi_k, 1 \leq k \leq K$, can be computed using the inner product of the zero-meaned super-vector $(\mathbf{U}_i - \bar{\mathbf{U}})$ and the eigenvector $\mathbf{e}_k$, i.e.,

$$v(\mathbf{X}_i, \phi_k) = < \mathbf{U}_i - \bar{\mathbf{U}}, \mathbf{e}_k > . \tag{15}$$

IV. CLUSTER GENERATION

After converting each utterance into a projection vector, the similarity between utterances can be computed using Eq. (3). The next step is to group similar utterances into clusters. This section discusses two methods for generating clusters, namely, hierarchical clustering and the proposed maximum purity clustering [25].

A. Hierarchical Clustering

Hierarchical clustering [30] is a common approach to determining which utterances should be grouped together. In our implementation, clusters are generated in an agglomerative manner, starting with each utterance in its own cluster and successively merging the most similar pairs of clusters. The similarity between two clusters, say c_m and c_l, can be measured in several ways:

1) *Complete-linkage*

$$\mathcal{S}_c(c_m, c_l) = \min_{\mathbf{X}_i \in c_m, \mathbf{X}_j \in c_l} \mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j), \tag{16}$$

2) *Single-linkage*

$$\mathcal{S}_c(c_m, c_l) = \max_{\mathbf{X}_i \in c_m, \mathbf{X}_j \in c_l} \mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j), \text{or} \tag{17}$$

3) *Average-linkage*

$$\mathcal{S}_c(c_m, c_l) = \frac{1}{n_{(m,l)}} \sum_{\mathbf{X}_i \in c_m, \mathbf{X}_j \in c_l} \mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j), \tag{18}$$

where $n_{(m,l)}$ is the number of utterance pairs involved in the summation. The outcome of the aggregation procedure above is a tree of clusters. The final partition of the utterances is then determined by pruning the tree that only has M leaves left.

B. Maximum Purity Clustering

The principle behind hierarchical clustering is to make the agreement within a newly generated cluster as large as possible. However, when each agglomeration is performed, the hierarchical clustering can not guarantee that the overall within-cluster agreement will be maximized, since its decision does not consider the interaction between the new cluster to be generated and the existing clusters. Therefore, this clustering method is not optimal. In particular, some mis-clustering errors, arising from grouping different-speaker utterances together, can propagate down the whole process, and limit the clustering performance. To overcome this limitation, we propose a new clustering method, which considers how to assign utterances to clusters in a global fashion such that the overall within-cluster agreement can be maximized.

Let h_i denote the index of the cluster where an utterance $\mathbf{X}_i$ is located, and o_i denote the true speaker index of utterance $\mathbf{X}_i$. Note, h_i is an integer between 1 and M when the number

of clusters M is specified *a priori*, and o_i is an integer between 1 and P if there are P speakers involved. Our aim is to find a set of cluster indices $\mathbf{H} = \{h_1, h_2, \ldots, h_N\}$ assigned for N utterances to be clustered that maximizes the average cluster purity defined by Eq. (2), i.e.,

$$\begin{aligned}\mathbf{H}^* &= \arg\max_{\mathbf{H}} \frac{1}{N}\sum_{m=1}^{M} n_m \left(\sum_{p=1}^{P}\frac{n_{mp}^2}{n_m^2}\right) \\ &= \arg\max_{\mathbf{H}} \frac{1}{N}\sum_{m=1}^{M} \frac{\sum_{p=1}^{P}\left[\sum_{i=1}^{N}\delta(h_i,m)\delta(o_i,p)\right]^2}{\sum_{i=1}^{N}\delta(h_i,m)},\end{aligned} \tag{19}$$

where $\delta()$ is a Kronecker Delta function.

However, as the computation of cluster purity requires that the true speaker of each utterance is known in advance, it is impossible to find $\mathbf{H}^*$ from Eq. (19) directly. To make this equation solvable, we need to estimate the term $\sum_{p=1}^{P}\left[\sum_{i=1}^{N}\delta(h_i,m)\delta(o_i,p)\right]^2$ in the absence of the ground truth. Since

$$\begin{aligned}\sum_{p=1}^{P}\left[\sum_{i=1}^{N}\delta(h_i,m)\delta(o_i,p)\right]^2 &= \sum_{p=1}^{P}\left[\sum_{i=1}^{N}\delta(h_i,m)\delta(o_i,p)\right]\left[\sum_{j=1}^{N}\delta(h_j,m)\delta(o_j,p)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{P}\delta(h_i,m)\delta(o_i,p)\delta(h_j,m)\delta(o_j,p) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\delta(h_i,m)\delta(h_j,m)\left[\sum_{p=1}^{P}\delta(o_i,p)\delta(o_j,p)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\delta(h_i,m)\delta(h_j,m)\delta(o_i,o_j),\end{aligned} \tag{20}$$

the estimation of $\sum_{p=1}^{P}\left[\sum_{i=1}^{N}\delta(h_i,m)\delta(o_i,p)\right]^2$ hinges on how to determine the term $\delta(o_i,o_j)$ when utterances $\mathbf{X}_i$ and $\mathbf{X}_j$ are located in the cluster c_m. Motivated by Solomonoff *et al.*'s work [6], we determine $\delta(o_i,o_j)$ by using the following approximation:

$$\hat{\delta}(o_i,o_j) = \begin{cases} 1, & \text{if } i = j \\ \frac{\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)}{\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_{\xi_i})}, & \text{if } i \neq j, \text{ and } \mathcal{R}\left[\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)\right] \leq n_m \\ 0, & \text{otherwise} \end{cases}, \tag{21}$$

where $\mathcal{R}\left[\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)\right]$ denotes the rank of inter-utterance similarity $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)$ among $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_1)$, $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_2)$, $\cdots$, $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_N)$ in descending order, and $\mathbf{X}_{\xi_i}$ is the utterance most similar to $\mathbf{X}_i$, i.e., $\mathcal{R}\left[\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_{\xi_i})\right] = 2$. Implicit in Eq. (21) is the idea that two utterances, $\mathbf{X}_i$ and $\mathbf{X}_j$, can only be considered as being from the same speaker if the similarity $\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)$ is high enough to satisfy $\mathcal{R}\left[\mathcal{S}_u(\mathbf{X}_i,\mathbf{X}_j)\right] \leq n_m$. In addition, to avoid a possible misjudgement arising from an over-sized

n_m, we approximate $\delta(o_i, o_j)$ as the probability that utterances $\mathbf{X}_i$ and $\mathbf{X}_j$ belong to the same speaker. This possibility is measured by comparing the similarity $\mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j)$ with that of the two utterances that most likely belong to the same speaker, i.e., $\mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_{\xi_i})$. With this approximation, an optimal $\mathbf{H}^*$ may be found according to

$$\mathbf{H}^* = \arg\max_{\mathbf{H}} \frac{1}{N} \sum_{m=1}^{M} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \delta(h_i, m)\delta(h_j, m)\hat{\delta}(o_i, o_j)}{\sum_{i=1}^{N} \delta(h_i, m)}. \tag{22}$$

We note that the solution to Eq. (22) remains non-trivial, since a gradient-based optimization cannot be used in this scenario. Moreover, it is infeasible to perform an exhaustive search that would examine all possible solutions to determine the best one, because there are M^N possible combinations of cluster indices. To overcome these difficulties, we apply the genetic algorithm (GA) [21] to find $\mathbf{H}^*$ by using its 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 that of another cluster, multiple chromosomes may lead to an identical clustering result. For example, the chromosomes $\{1, 1, 1, 2, 2, 3, 3\}$, $\{1, 1, 1, 3, 3, 2, 2\}$, $\{2, 2, 2, 1, 1, 3, 3\}$, $\{2, 2, 2, 3, 3, 1, 1\}$, $\{3, 3, 3, 2, 2, 1, 1\}$, and $\{3, 3, 3, 1, 1, 2, 2\}$ represent the same clustering result of grouping seven utterances into three clusters. Such a non-unique representation of the solution would significantly increase the GA search space, and could lead to an inferior clustering result. To avoid this problem, we limit the inventory of chromosomes to conform to a *baseform* representation defined as follows. Let $\mathcal{I}(c_m)$ be the lowest index of the utterance in the m-th cluster, $c_m = \{\mathbf{X}_i | h_i = m, 1 \leq i \leq N\}$. A chromosome is a baseform iff

$$\forall\ c_m \text{ and } c_l,\ \text{if } m < l,\ \text{then } \mathcal{I}(c_m) < \mathcal{I}(c_l). \tag{23}$$

As the above example shows, chromosome $\{1, 1, 1, 2, 2, 3, 3\}$ is a baseform, since the lowest indices of the utterances in the first, second, and third clusters are 1, 4, and 6, respectively, which satisfies Eq. (23). In contrast, chromosome $\{1, 1, 1, 3, 3, 2, 2\}$ is not a baseform, since the lowest indices of the utterances in the first, second, and third clusters are 1, 6, and 4, respectively, which

does not satisfy Eq. (23). Likewise, the other chromosomes, $\{2,2,2,1,1,3,3\}$, $\{2,2,2,3,3,1,1\}$, $\{3,3,3,2,2,1,1\}$, and $\{3,3,3,1,1,2,2\}$ are not baseforms. Even so, it is conceivable that all the non-baseform chromosomes could be converted into a unique baseform representation by interchanging the clusters' indices.

Fig. 3 shows a block diagram of GA-based optimization. It starts with a random generation of baseform chromosomes according to a certain population size. The fitness of all chromosomes is then evaluated and ranked on the basis of the estimated average purity, i.e.,

$$\bar{\rho}(\mathbf{H}) = \frac{1}{N}\sum_{m=1}^{M}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\delta(h_i,m)\delta(h_j,m)\hat{\delta}(o_i,o_j)}{\sum_{i=1}^{N}\delta(h_i,m)}. \tag{24}$$

As a result of this evaluation, a particular group of chromosomes is selected from the population to generate offspring by subsequent recombination. The selection reflects the fact that chromosomes with superior fitness have a higher chance of being included in the next generation than those with inferior fitness. To prevent premature convergence of the population, this study employs the linear ranking selection scheme [31], which sorts chromosomes in increasing order of fitness, and then assigns the expected number of offspring according to their relative ranking.

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 ratio of the number of offspring produced in each generation to the population size. 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 resulting chromosomes that do not conform the baseform representations are converted into their baseform counterparts. The procedure of fitness evaluation, selection, crossover, and mutation is repeated continuously, which follows the principle of survival of the fittest, to produce better approximations of the optimal solution. Accordingly, it is hoped that the average purity of the clustering 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 of $\mathbf{H}^*$.

V. ESTIMATION OF THE SPEAKER POPULATION SIZE

In general, the more clusters we generate, the larger the value of purity we can obtain. However, if we generate too many clusters, a single speaker's utterances would be split across multiple clusters, so the speaker clustering would be incomplete. Clearly, the optimal number of clusters is equal to the speaker population size, which is unknown and needs to be estimated.

Our basic strategy for estimating the speaker population size is to define a score for assessing a partitioning of the utterances based on how a large average purity can be achieved at the expense of increasing the number of clusters. This problem may be tackled from the standpoint of model selection. Specifically, if each of the possible partitionings with different numbers of clusters is considered as a model for characterizing the speaker information of the utterances, we choose the model that can produce the largest average purity and has the smallest number of clusters. Viewed in this manner, the Bayesian information criterion (BIC) [32], which is popular for solving model-selection problems, could be used to assess the clustering.

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

$$\mathrm{BIC}(\Lambda) = \log \Pr(\mathbf{O}|\Lambda) - \frac{1}{2}\gamma \#(\Lambda) \log |\mathbf{O}|, \tag{25}$$

where γ is a penalty factor generally set to one, $\#(\Lambda)$ denotes the number of free parameters in model Λ, and $|\mathbf{O}|$ is the size of the data set $\mathbf{O}$. The larger the value of $\mathrm{BIC}(\Lambda)$, the better model Λ will perform. In another work on speaker clustering [7], BIC is applied to score a partitioning of an utterance collection, in which a cluster is represented by a uni-Gaussian density estimated from the feature vectors of the utterances, and the model Λ is a set of Gaussian densities. Since we convert each utterance from the feature vectors into a projection vector, our work differs from [7] by the way clusters are modeled, which is directly related to the clustering performance.

Consider a model Λ, consisting of M parameters for classifying a set of N utterances from P unknown speakers. Each of the parameters represents an integer index to tag each of the utterances. The model is designed with such an aim that, by having all the utterances tagged, the utterances belonging to the same speakers are tagged with the same index. Thus, the likelihood $\Pr(\mathbf{O}|\Lambda)$, which measures how well the model fits the data, is concerned with the probability that,

given N indices $h_1, h_2, \ldots, h_N$ for the N utterances, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, the utterances tagged with the same indices come from the same speakers. Suppose that the true speakers of N utterances, $o_1, o_2, \ldots, o_N$, are statistically independent of each other. We compute the likelihood $\Pr(\mathbf{O}|\Lambda)$ by

$$\begin{aligned}\Pr(\mathbf{O}|\Lambda) &= \prod_{i=1}^{N} \Pr\left(o_i = O(\mathbf{X})|h_i = C(\mathbf{X})\right) \\ &\cong \left[\Pr\left(O(\grave{\mathbf{X}}) = O(\mathbf{X})|C(\grave{\mathbf{X}}) = C(\mathbf{X})\right)\right]^N, \end{aligned} \tag{26}$$

where $O(\mathbf{X})$ denotes the true speaker of an arbitrary utterance $\mathbf{X}$ tagged with index $C(\mathbf{X})$, and $\Pr\left(o_i = O(\mathbf{X})|h_i = C(\mathbf{X})\right)$ represents the probability that, given an arbitrary utterance $\mathbf{X}$ tagged with the same index as utterance $\mathbf{X}_i$, the true speakers of utterances $\mathbf{X}_i$ and $\mathbf{X}$ are the same. For computational efficiency, all the probabilities $\Pr\left(o_i = O(\mathbf{X})|h_i = C(\mathbf{X})\right)$, $1 \leq i \leq N$, are approximated by $\Pr\left(O(\grave{\mathbf{X}}) = O(\mathbf{X})|C(\grave{\mathbf{X}}) = C(\mathbf{X})\right)$, which is the probability that any two utterances $\mathbf{X}$ and $\grave{\mathbf{X}}$ tagged with the same index come from the same speaker.

Assume that there are n_m utterances tagged with m, and among these n_m utterances there are n_{mp} utterances produced by speaker s_p. If we pick one of the n_m utterances twice at random, with replacement, the probability that both of the picked utterances come from s_p is $(n_{mp}/n_m) \times (n_{mp}/n_m)$. Thus, the probability that two utterances tagged with m come from the same speaker is $\sum_{p=1}^{P}(n_{mp}/n_m)^2$, which is also the cluster purity ρ_m defined in Eq. (1). Since the probability that one utterance tagged with m is n_m/N, we can estimate the probability that any two utterances $\mathbf{X}$ and $\grave{\mathbf{X}}$ tagged with the same index come from the same speaker by

$$\Pr\left(O(\grave{\mathbf{X}}) = O(\mathbf{X})|C(\grave{\mathbf{X}}) = C(\mathbf{X})\right) = \sum_{m=1}^{M} \frac{n_m}{N} \left(\sum_{p=1}^{P}(n_{mp}/n_m)^2\right) = \bar{\rho}. \tag{27}$$

Approximating $\bar{\rho}$ as $\bar{\rho}(\mathbf{H})$ in Eq. (24), the likelihood $\Pr(\mathbf{O}|\Lambda)$ can be obtained with $\bar{\rho}(\mathbf{H})^N$. Accordingly, we can score a partitioning of N utterances having M clusters via

$$\text{BIC}(M\text{-clustering}) = N \log \bar{\rho}(\mathbf{H}) - \frac{1}{2}\gamma M \log N. \tag{28}$$

The BIC-based score should increase with the increase of the M value in the beginning, but will decline significantly after an excess of clusters is created. A reasonable number of clusters can thus be determined by

$$M^* = \arg \max_{2 \leq M \leq N} \text{BIC}(M\text{-clustering}). \tag{29}$$

VI. EXPERIMENTS

A. Speech Data

Our speech database comprised two subsets, respectively, extracted from two corpora released by the Linguistic Data Consortium [33]: *the 1998 HUB-4 Broadcast News Evaluation English Test Material* (Hub4-98), which consists of broadcast news speech recorded at a 16 kHz sampling rate, and *the 2001 NIST Speaker Recognition Evaluation Corpus* (SRE-01), which consists of cellular telephone speech recorded at an 8 kHz sampling rate. The first subset contained 428 speaker-homogeneous utterances obtained by segmenting the episode "h4e-98-2" of Hub4-98, according to the annotation file. This subset involved 89 speakers, in which the number of utterances spoken by each speaker ranged from 1 to 27. The second subset, which stems from the test set of SRE-01, contained 197 speaker-homogeneous utterances spoken by 15 randomly-selected male speakers [3]. The number of utterances spoken by each speaker ranged from 5 to 39. Speech features, each consisting of 24 Mel-scale frequency cepstral coefficients (MFCCs), were extracted from these utterances for every 20-ms Hamming-windowed frame with 10-ms frame shifts. Prior to MFCC computation, voice active detection [34] was applied to remove salient non-speech regions that may be included in an utterance.

B. Experimental Results

1) Comparison of different inter-utterance similarity measures

Our first experiment was conducted to assess the performance of various inter-utterance similarity measures proposed in this study. Since a good similarity measure should consistently produce larger values for similarities between utterances of the same speaker and smaller values for similarities between utterances of different speakers, the assessment can be done by examining if the produced values are well distinguished for these two conditions. To do this, we compute two types of error probability for each inter-utterance similarity measure. One is the probability that two same-speaker utterances are falsely judged as being from different speakers (Type-I error), and the

[3]Empirical evidence shows that male speakers usually can be well distinguished from female speakers. We therefore focused on investigating the speech from single gender speakers.

other is the probability that two different-speaker utterances are falsely judged as being from the same speaker (Type-II error). These two types of error probability are subject to tradeoff, which can be represented by the Detection Error Tradeoff (DET) curve [35]. This representation takes as input the values of inter-utterance similarities, and produces as output the corresponding Gaussian deviates of the error probability distribution. The resulting DET curves are approximately straight lines, making it easier to visually compare the performance of different methods.

Figs. 4 and 5 show the DET curves obtained by testing subsets "SRE-01" and "h4e-98-2", respectively. In the figures, "UD-GMM", "UI-VC", "UI-GMM-ADA", and "EV" denote the methods used to create the reference spaces, respectively, by utterance-dependent Gaussian mixture modeling (Sec. III-A), utterance-independent vector clustering (Sec. III-B), utterance-independent Gaussian mixture modeling followed by MAP adaptation (Sec. III-C), and the eigenvoice-motivated approach (Sec. III-D). In addition, "GLR" denotes the generalized likelihood ratio-based similarity measure, which served as a baseline for performance comparison. Briefly, GLR between two utterances $\mathbf{X}_i$ and $\mathbf{X}_j$ is computed using

$$\mathcal{S}_u(\mathbf{X}_i, \mathbf{X}_j) = \frac{\Pr(\mathbf{X}_{ij}|\lambda_{ij})}{\Pr(\mathbf{X}_i|\lambda_i)\Pr(\mathbf{X}_j|\lambda_j)}, \tag{30}$$

where $\mathbf{X}_{ij}$ is the concatenation of $\mathbf{X}_i$ and $\mathbf{X}_j$, and λ_i, λ_j, and λ_{ij} are parametric models trained using $\mathbf{X}_i$, $\mathbf{X}_j$, and $\mathbf{X}_{ij}$, respectively. Each similarity measure was examined with varied configurations, i.e., varied numbers of Gaussian components, codewords, and eigenvectors. Although different numbers of Gaussian components should be used for utterances of different durations, we only examined each similarity measure in terms of the average performance by using the same number of Gaussian components. In addition, except for the single-Gaussian models, which were full-covariance structures, all the GMMs used in this study comprised diagonal covariance matrices. The thick line in each figure indicates the best discriminating performance that a similarity measure can achieve. The best performances for all the similarity measures were summarized in Figs. 4(f) and 5(f).

We can see from Figs. 4(f) and 5(f) that the eigenvoice-motivated approach performed best, while the GLR-based approach performed worst in distinguishing between same-speaker utterances and different-speaker utterances. The other three approaches, "UD-GMM", "UI-VC", and "UI-GMM-ADA" were between these two extremes. The results indicate that the proposed inter-

utterance similarity measures are superior to the GLR-based approach. This is attributed to the benefit of using a voice reference space to incorporate out-of-pair information into the similarity computation for every pair of utterances. Figs. 4(f) and 5(f) also show that both "UI-VC", and "UI-GMM-ADA" performed better than "UD-GMM". As pointed out in Sec. III-A, the major weakness of "UD-GMM" is that each basis of the reference space is generated from one utterance only, leading to the sensitivity of performance to the utterance duration or speech quality. Although the performance of "UD-GMM" may be improved by using different numbers of Gaussian components for utterances of different durations, automatically determining the optimal number of Gaussian components according to the utterance duration remains an unsolved issue and may be costly too expensive to fit the speaker-clustering task. By contrast, since each of the bases in "UI-VC", and "UI-GMM-ADA" is generated with the contribution of multiple utterances, the resulting similarity measures are more robust than that using "UD-GMM". In addition, we can see that "UI-GMM-ADA" yielded a slightly better result than "UI-VC". The performance of "UI-GMM-ADA" can be further improved by converting the bases into a statistically-independent set of eigenvectors through the use of the eigenvoice technique.

2) Comparison of different speaker-clustering methods

Experiments were then conducted to examine the performance of speaker clustering based on the best configuration for each of the inter-utterance similarity measures. We employed the agglomerative hierarchical clustering described in Sec. IV-A as a benchmark test. Figs. 6 and 7 show the average purity of speaker clustering as a function of the number of clusters, in which (a), (b), and (c) are the results obtained with complete linkage, average linkage, and single linkage, respectively. We can see that, as expected, the average purity increases as the number of clusters increases. Overall, the eigenvoice-motivated approach performed best, and "UI-GMM-ADA" and "UI-VC" came second and third, respectively. This result is roughly consistent with the result shown in Figs. 4 and 5. In addition to the choice of inter-utterance similarity measures, the performance of agglomerative speaker clustering is heavily dependent on the way that the similarity between clusters is represented. Comparing (a), (b), and (c) of Figs. 6 and 7, we can see that complete linkage is the best choice in this task, whereas single linkage is unusable. Focusing on Figs. 6(a) and 7(a), we can see that when the number of clusters is equal to the speaker population size, the

best average cluster purity obtained with the eigenvoice-motivated approach was 0.72 and 0.76 for "SRE-01" and "h4e-98-2", respectively. This demonstrates a notable improvement, compared to 0.51 and 0.72 obtained with the GLR-based approach.

Next, we examined if the speaker-clustering performance can be further improved by using the proposed maximum purity clustering method. In the GA optimization, the parameter values used for the maximum number of generations, the chromosome population size, the crossover probability, and the mutation probability were empirically determined to be 4000, 5000, 0.5, and 0.1, respectively. Fig. 8 shows the results obtained with agglomerative hierarchical clustering and maximum purity clustering, in which the inter-utterance similarity was computed on the basis of the eigenvoice-motivated approach. It is clear that maximum purity clustering outperforms agglomerative hierarchical clustering. When the number of clusters was specified as equal to the speaker population size, we can see that the average purity was improved from 0.72 ("SRE-01") and 0.76 ("h4e-98-2"), yielded by the agglomerative hierarchical clustering, to 0.81 ("SRE-01") and 0.82 ("h4e-98-2"), yielded by the maximum purity clustering.

3) Automatic determination of the speaker population size

Finally, the problem of automatically determining the speaker population size was investigated. We computed the BIC-based scores with respect to different numbers of clusters using Eq. (28). Fig. 9 shows the resulting scores obtained with the penalty factor γ set to be equal to, slightly greater than, and slightly smaller than one, respectively. The arrowed peak of each curve in the figures indicates the optimal number of clusters determined by the criterion of Eq. (29). We can see from Fig. 9(a) that most of the peaks appeared near the actual speaker population size, and the scores declined significantly after an excess of clusters was created. In general, the smaller the value of penalty factor, the larger the estimated optimal number of clusters, and vice versa. The results show that the speaker population size in subset "SRE-01" was estimated very well. However, Fig. 9(b) shows that the speaker population size in subset "h4e-98-2" tends to be underestimated. We speculate that this underestimation is mainly because, among the total 89 speakers in subset "h4e-98-2", there were 30 speakers who spoke only one utterance, and many of these speakers' utterances were shorter than five seconds, which leads to the tendency that these speakers are ignored. Despite this, we observe from the figure that we can mitigate such

underestimation by setting the penalty factor to be slightly less than one. This result indicates the feasibility of automatically determining the speaker population size.

VII. 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 among the utterances to be clustered can be exploited more effectively and reliably. We have presented several implementations for reference space creation. In particular, to capture the most representative characteristics of speakers' voices, the reference space has been represented as a set of eigenvectors obtained by applying the eigenvoice technique to the set of utterances to be clustered. This achieved a notable improvement in the speaker-clustering performance, compared to the common inter-utterance similarity measure based on the generalized likelihood ratio.

In addition, we have studied the method beyond the conventional hierarchical clustering for generating clusters such that all the within-cluster utterances can be, as far as possible, from one single speaker. This requirement has been formulated as a problem of estimating and maximizing the overall cluster purity. By representing cluster purity as a function of inter-utterance similarity, and applying the genetic algorithm to find the solution of this function, we have demonstrated a further improvement in the speaker-clustering performance, compared to the conventional agglomerative hierarchical clustering. Furthermore, the clustering method has been incorporated with the Bayesian information criterion to determine how many clusters should be generated. Experimental results show that the automatically-determined number of clusters can approximate the actual speaker population size.

With regard to usability, our future work will extend the current speaker-clustering methods to deal with speech data containing multiple non-simultaneous or simultaneous speakers. The clustering of non-simultaneous-speaker utterances may be done by either assigning each utterance to multiple related clusters [9] or pre-segmenting utterances into small speaker-homogeneous regions and then clustering these small regions. For handling simultaneous-speaker utterances, specific techniques for detecting and analyzing "overlapping speech" may be required.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants NSC93-2213-E-001-017 and NSC94-2422-H-001-007. The authors are very grateful to the anonymous reviewers and the associate editors, Dr. Ananth Sankar and Dr. Timothy J. Hazen, for their careful reading of this article and their constructive suggestions.

References

[1] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 1989.

[2] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 1991.

[3] T. Kosaka and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 1994.

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

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

[6] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 1998.

[7] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICSLP*), 1998.

[8] 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. International Conference on Spoken Language Processing* (*ICSLP*), 1998.

[9] J. McLaughlin, D. Reynolds, E. Singer, and G. C. O'Leary, "Automatic speaker clustering from multi-speaker utterances," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 1999.

[10] S. E. Johnson, "Who spoke when? - Automatic segmentation and clustering for determining speaker turns," *Proc. European Conference on Speech Communication and Technology* (*Eurospeech*), 1999.

[11] B. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," *Proc. International Conference on Spoken Language Processing* (*ICSLP*), 2000.

[12] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *PROCEEDINGS OF IEEE*, 88(8):1338- 1353, 2000.

[13] R. Faltlhauser and G. Ruske, "Robust speaker clustering in eigenspace," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

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

[15] J. Ajmera, H., Bourlard, I., Lapidot, and I., McCowan, "Unknown-multiple speaker clustering using HMM," *Proc. International Conference on Spoken Language Processing* (*ICSLP*), 2002.

[16] Y. Moh, P. Nguyen, and J. C. Junqua, "Towards domain independent speaker clustering," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 2003.

[17] D. Liu and F. Kubala, "Online speaker clustering," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 2003.

[18] I., Lapidot, "SOM as likelihood estimator for speaker clustering," *Proc. European Conference on Speech Communication and Technology* (*Eurospeech)*, 2003.

[19] S., Kwon and S., Narayanan, "Unsupervised speaker indexing using generic models," *IEEE Transactions on Speech and Audio Processing*, 13(5):72-83, 2005.

[20] J. P. Campbell, "Speaker recognition: a tutorial," *PROCEEDINGS OF THE IEEE*, 85(9):1437-1462, 1997.

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

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

[23] W. H. Tsai, S. S. Cheng, and H. M. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," *Proc. International Conference on Spoken Language Processing* (*ICSLP*), 2004.

[24] W. H. Tsai, S. S. Cheng, Y. H. Chao, and H. M. Wang, "Clustering speech utterances by speaker using Eigenvoice-motivated vector space models," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), 2005.

[25] W. H. Tsai and H. M. Wang, "Speaker clustering of unknown utterances based on maximum purity estimation," *Proc. European Conference on Speech Communication and Technology* (*Eurospeech*), 2005.

[26] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, 3(1):72-83, 1995.

[27] W. H. Tsai, Y. C. Chu, C. S. Huang, and W. W. Chang, "Background learning of speaker voices for text-independent speaker identification," *Proc. European Conference on Speech Communication and Technology* (*Eurospeech*), 2001.

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

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

[30] L. Kaufman and P. J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.

[31] J. E. Baker, "Adaptive selection methods for genetic algorithm," *Proc. International Conference on Genetic Algorithms and Their Applications*, 1985.

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

[33] LDC, http://www.ldc.upenn.edu/

[34] The VIMAS speech codec, http://www.vimas.com

[35] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proc. European Conference on Speech Communication and Technology* (*Eurospeech*), 1997.

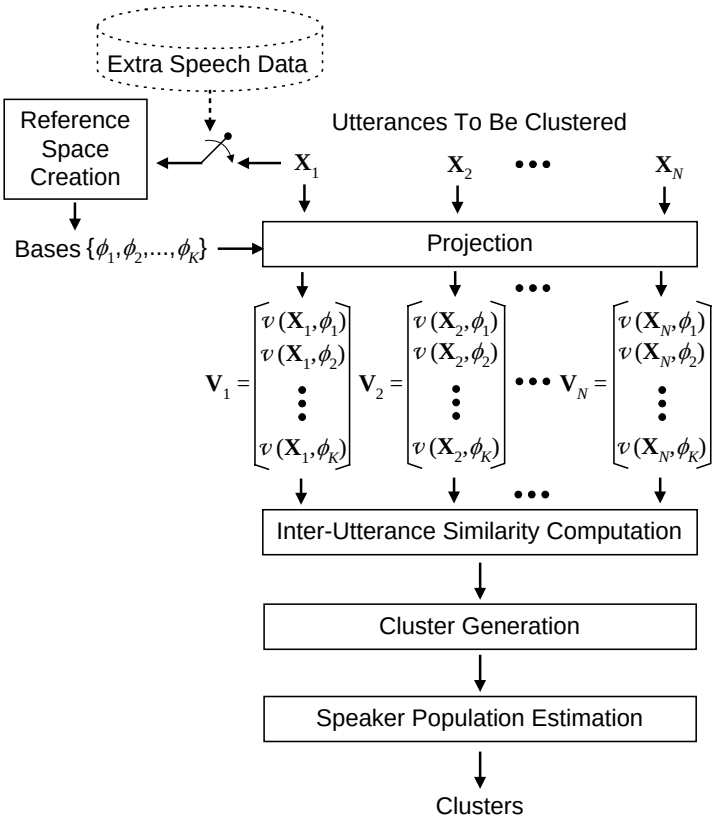


Fig. 1. The proposed speaker-clustering framework.

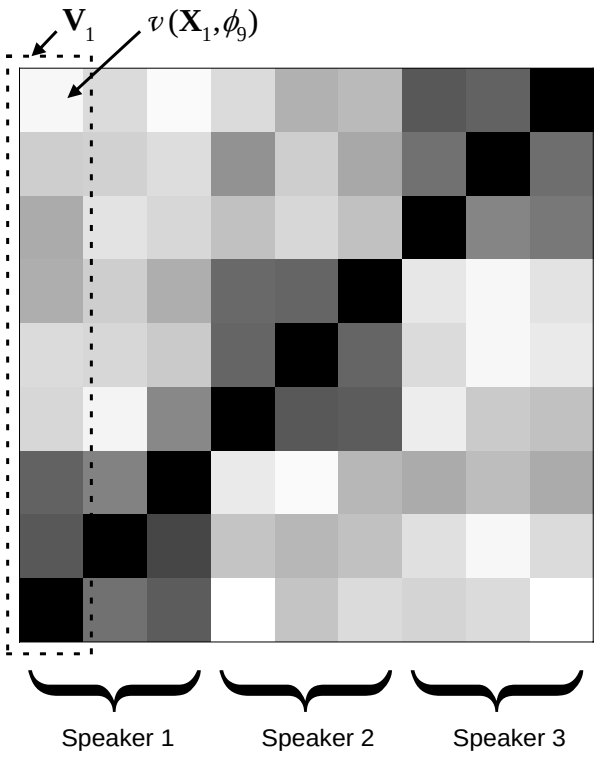


Fig. 2. An example of a projection using a reference space constructed by utterance-dependent Gaussian mixture modeling, in which the projection vectors computed from a collection of nine utterances are shown in gray scale representation.

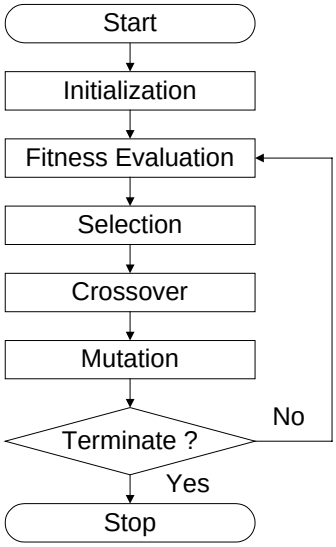
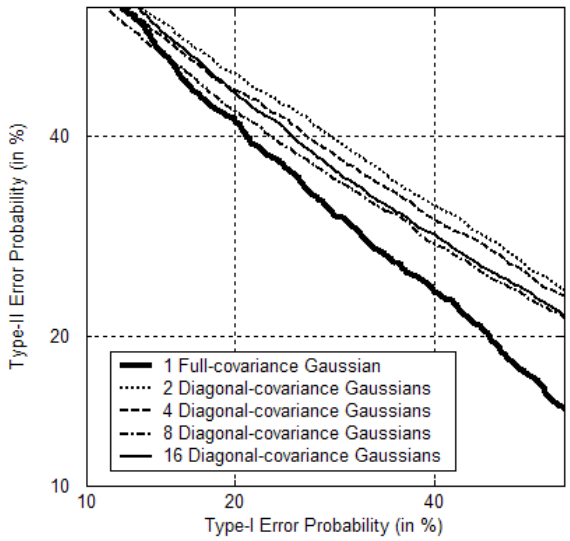
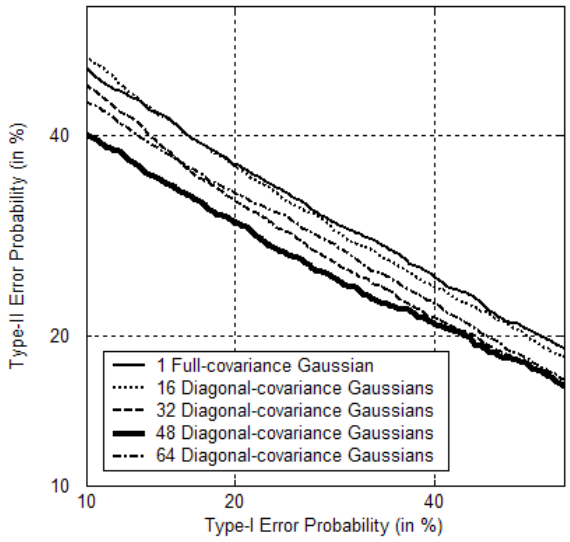


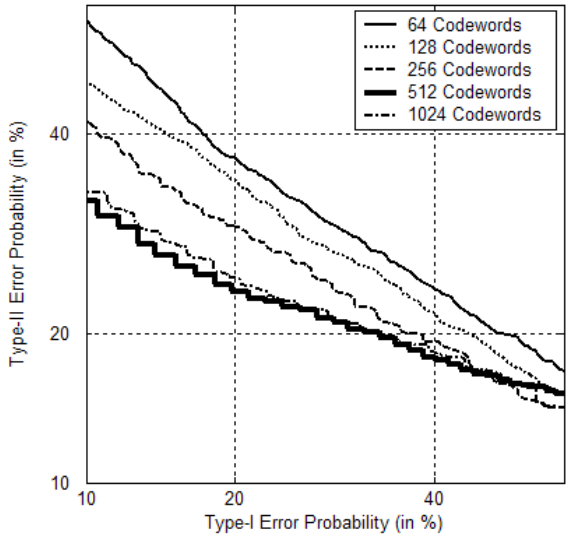
Fig. 3. Flow diagram of the genetic algorithm.



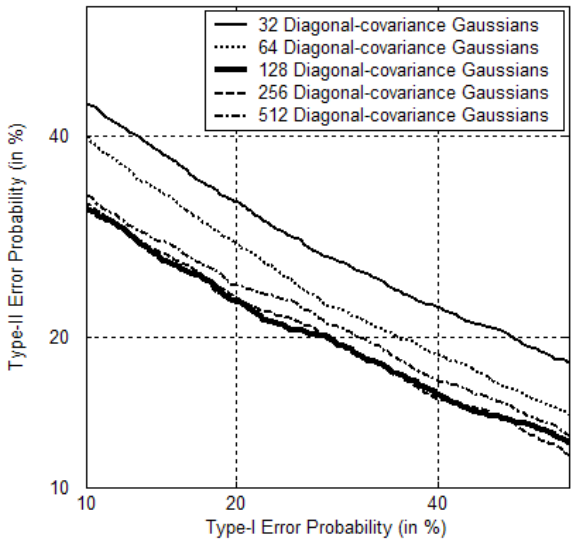
(a) GLR



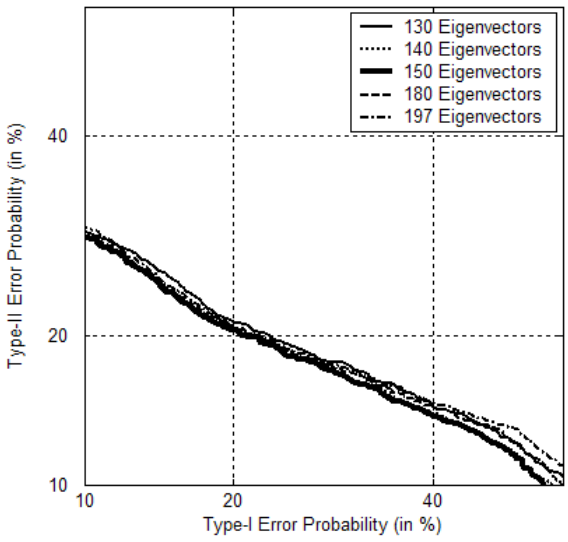
(b) UD-GMM



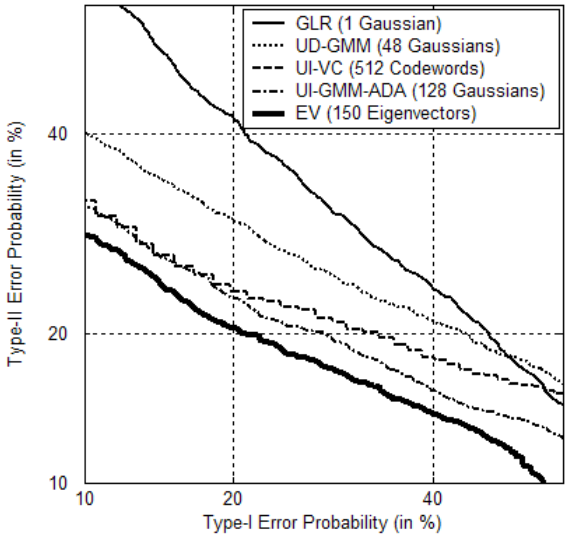
(c) UI-VC



(d) UI-GMM-ADA

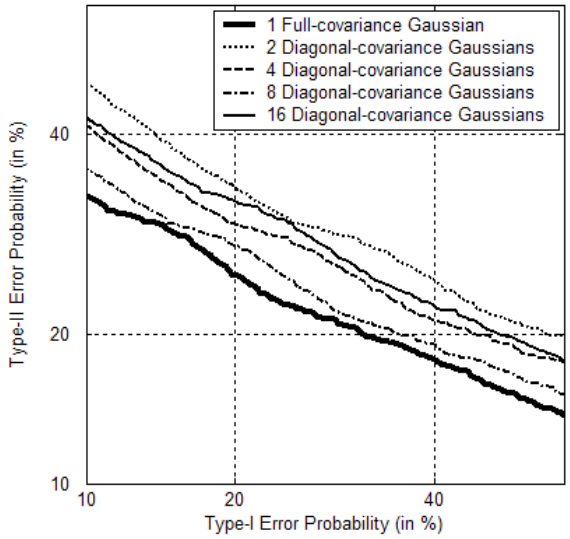


(e) EV

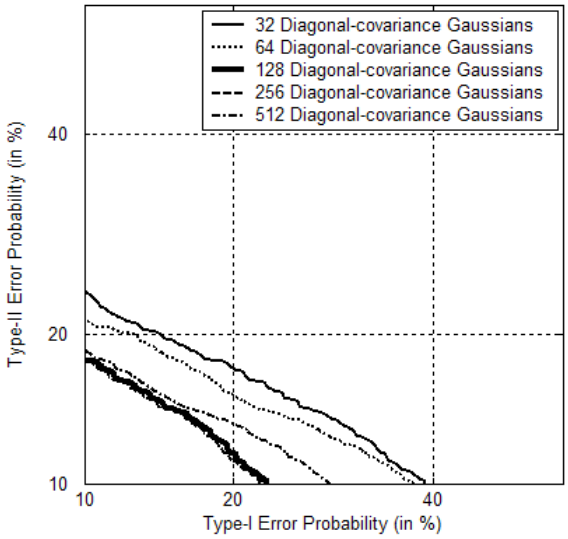


(f) A summarization of the best performances in (a) – (e)

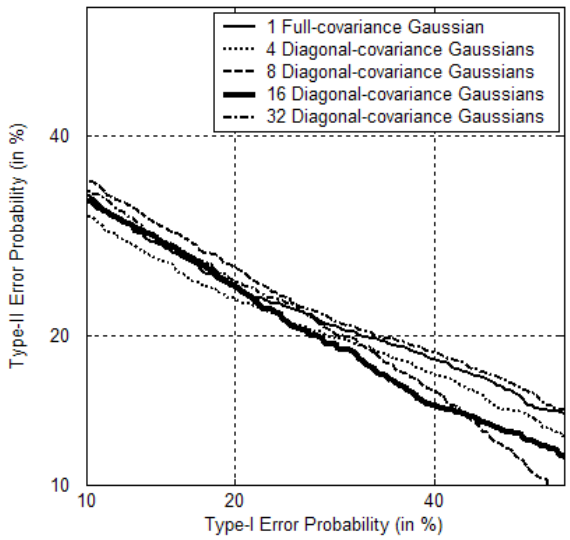
Fig. 4. DET curves obtained with various inter-utterance similarity measures for subset “SRE-01”.



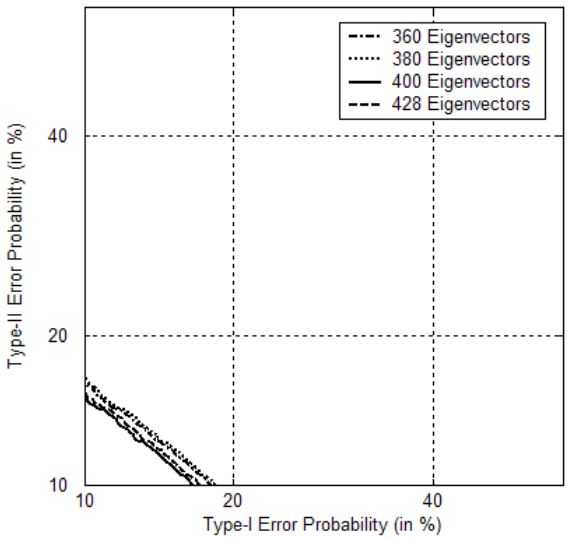
(a) GLR



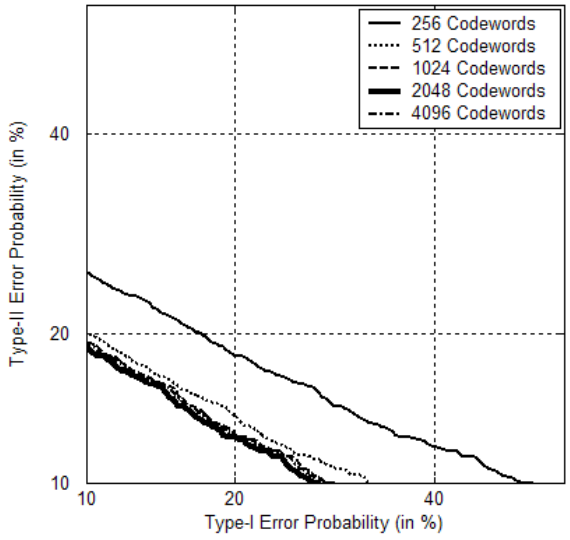
(d) UI-GMM-ADA



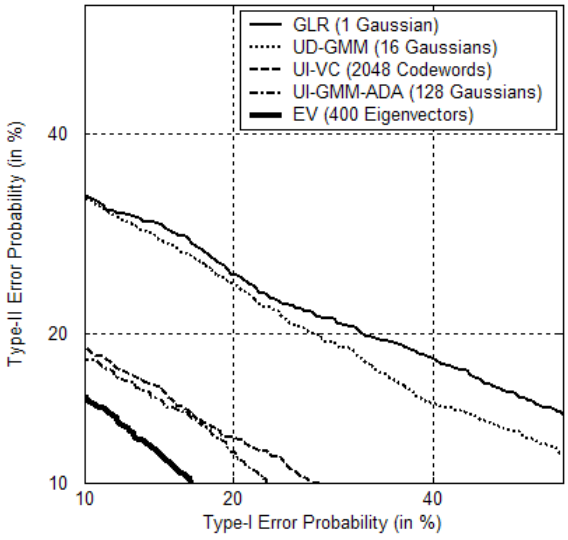
(b) UD-GMM



(e) EV

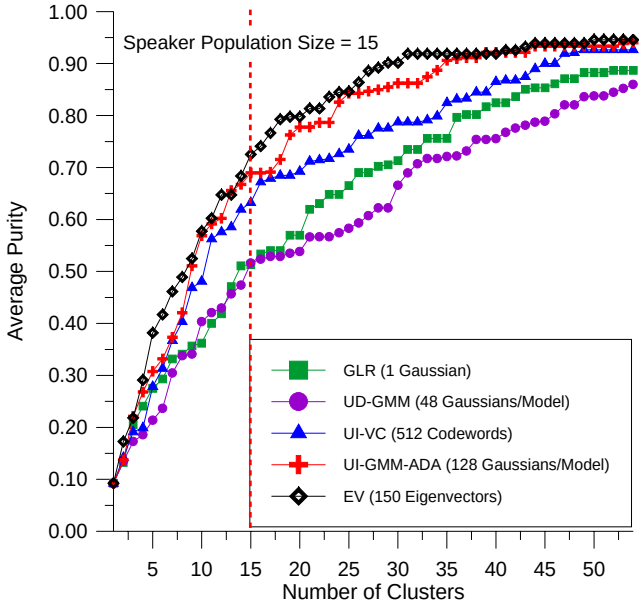


(c) UI-VC

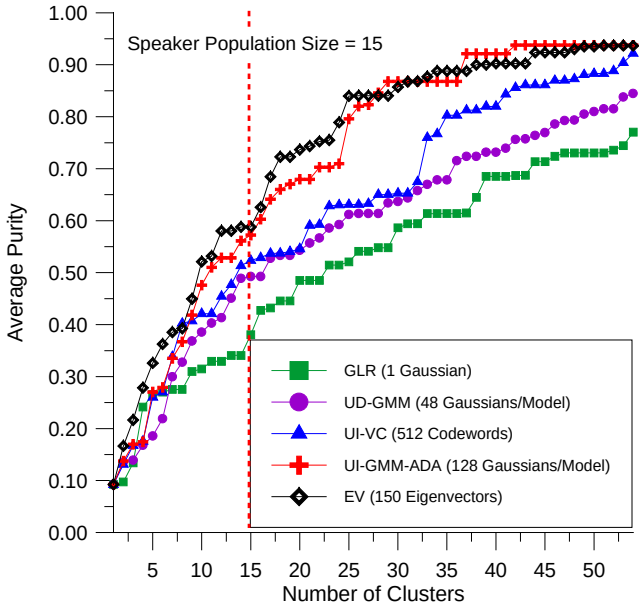


(f) A summarization of the best performances in (a) – (e)

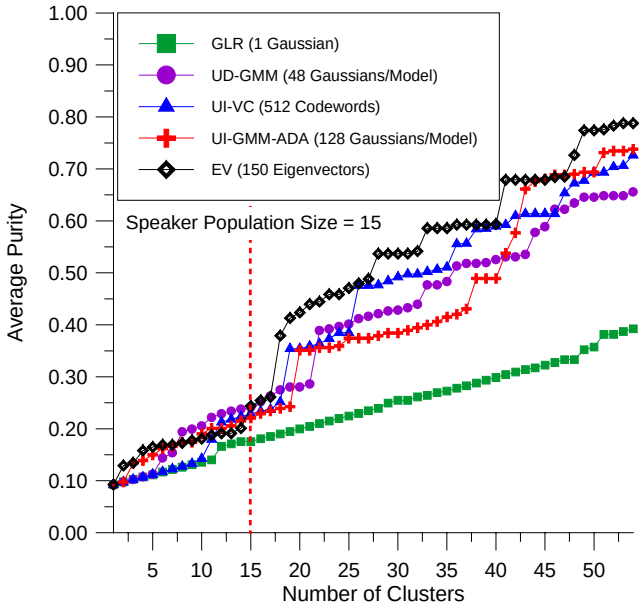
Fig. 5. DET curves obtained with various inter-utterance similarity measures for subset “h4e-98-2”.



(a) Complete linkage agglomerative clustering

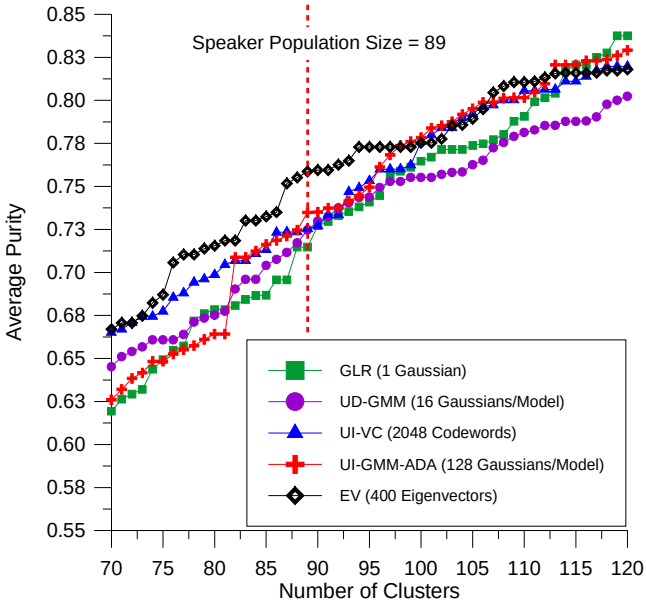


(b) Average linkage agglomerative clustering

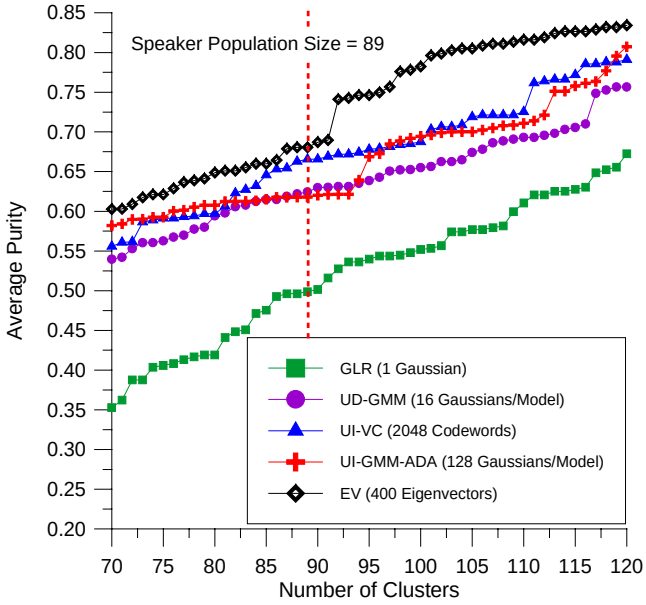


(c) Single linkage agglomerative clustering

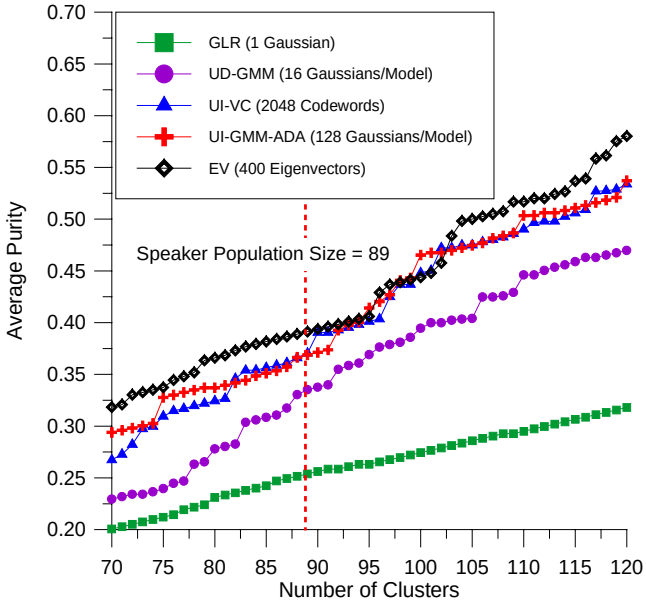
Fig. 6. Performance of speaker clustering obtained with various inter-utterance similarity measures for subset “SRE-01”.



(a) Complete linkage agglomerative clustering

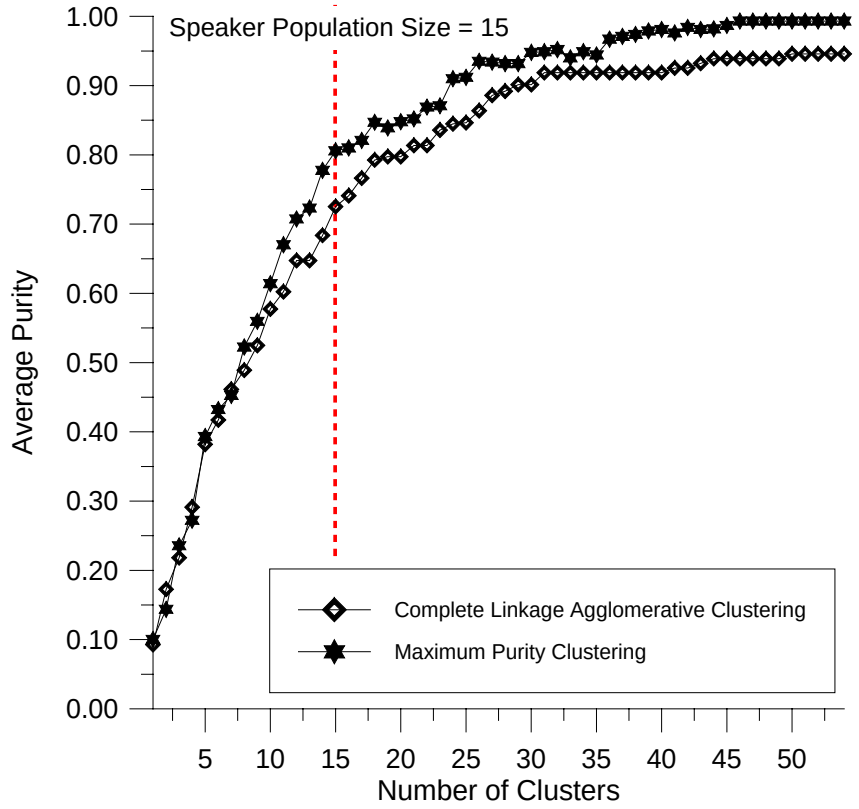


(b) Average linkage agglomerative clustering

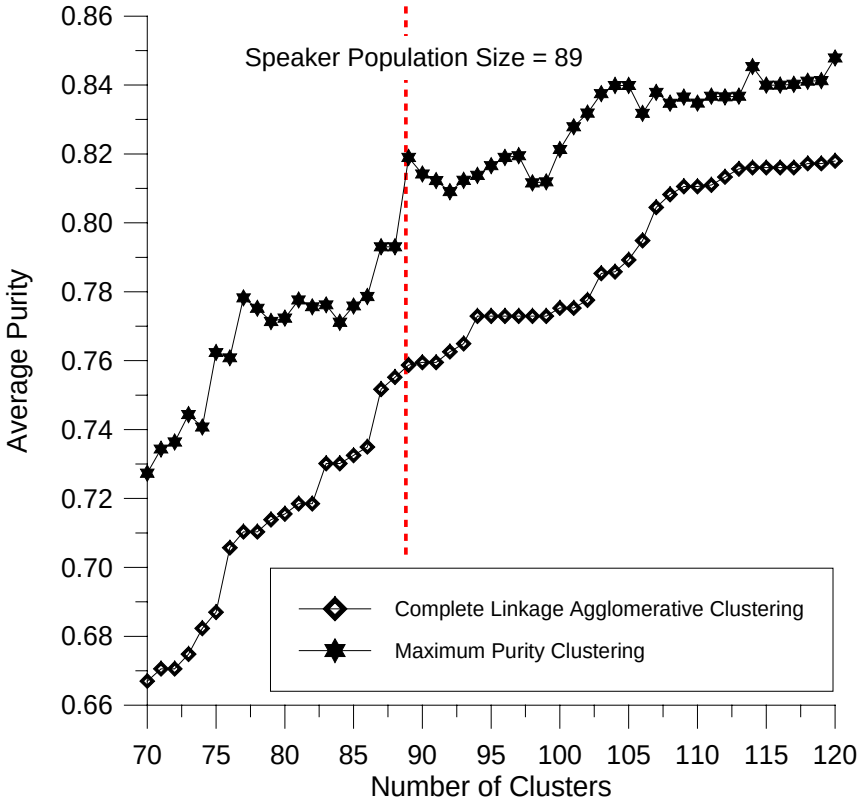


(c) Single linkage agglomerative clustering

Fig. 7. Performance of speaker clustering obtained with various inter-utterance similarity measures for subset “h4e-98-2”.

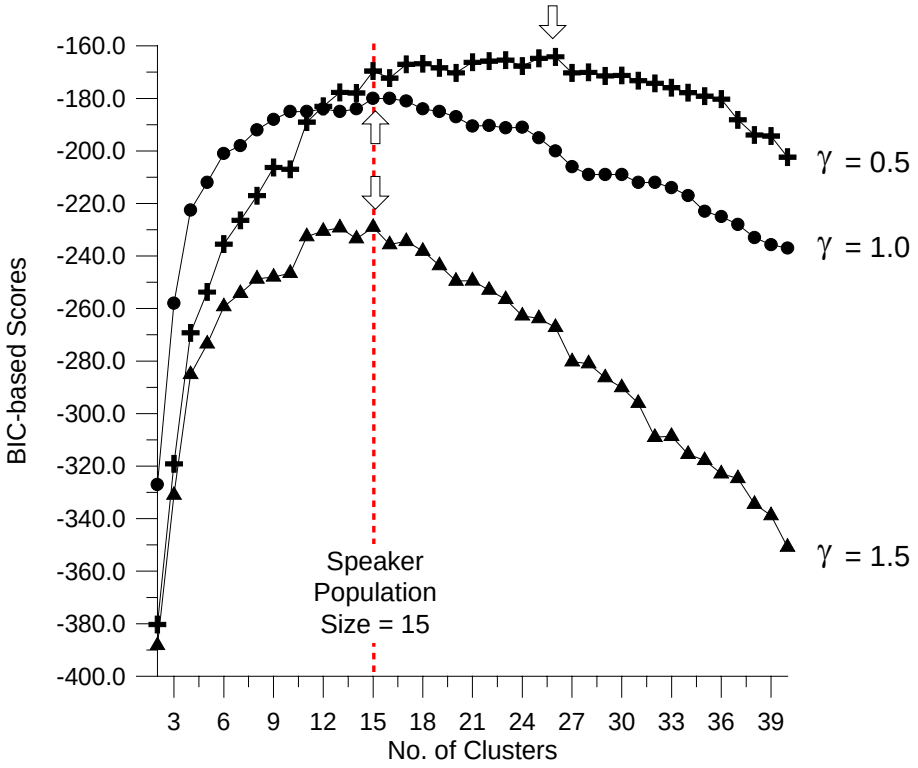


(a) Tests on subset “SRE-01”

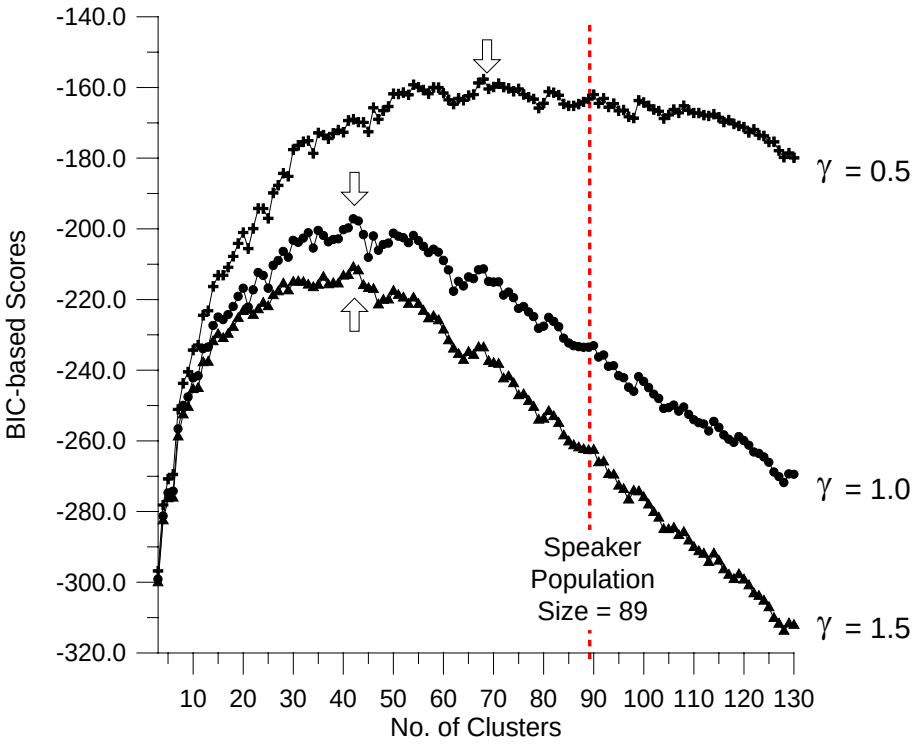


(b) Tests on subset “h4e-98-2”

Fig. 8. Performance of speaker clustering obtained by conventional hierarchical clustering; and the proposed maximum purity clustering, in which the inter-utterance similarities were computed on the basis of an eigenvoice-motivated reference space.



(a) Tests on subset “SRE-01”



(b) Tests on subset “h4e-98-2”

Fig. 9. BIC values as a function of number of clusters.