Clustering-Based I-Vector Formulation for Speaker Recognition

Hung-Shin Lee[1, 2], Yu Tsao[3], Hsin-Min Wang[2], Shyh-Kang Jeng[1]

[1]Department of Electrical Engineering, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

Abstract

In this paper, we first reformulate the derivation of the conventional i-vector scheme, which is the state-of-the-art utterance representation for speaker verification, as a modeling of universal background model (UBM)-based mixtures of factor analyzers (UMFA), and then propose a clustering-based UMFA method called CMFA. In UMFA, each analyzer is characterized by a subspace, and the same projection coordinate of an utterance into individual subspaces is called the i-vector. We relax this assumption by grouping the mixture components of the UBM into clusters according to their acoustic traits. Therefore, in CMFA, each utterance is represented by multiple i-vectors, each of which generated by similar subspaces associated with a same cluster. We also investigate two strategies for merging these i-vectors into a single one to be applied in the classifier of the conventional i-vector framework. The results of experiments conducted on the male portion of the core task in the NIST 2005 Speaker Recognition Evaluation (SRE) in terms of normalized decision cost function (minDCF) and equal error rate (EER) demonstrate the merits of the new i-vector method over the conventional i-vector method.

Index Terms: speaker recognition, i-vector, clustering

1. Introduction

As a branch of audio classification, where sound patterns extracted from raw waveform data are usually discrete symbol sequences or continuous-valued feature vectors, the performance of speaker verification tasks is heavily dependent on the data representation applied [1]. Therefore, to deploy speaker verification systems, many efforts have been devoted to the design of preprocessing pipelines, which result in a representation of speech signals that can effectively and robustly support various backend classifiers. As has been widely practiced, once the speech signal is digitized into a series of acoustic frames, each containing 80-200 ms temporal information, such as MFCCs or LPCCs, by a speech parameterization process [2, 3], two problems immediately arise. The first is how to represent a speech utterance of variable length by a fixed-size vectorial token so that the comparison between any pair of utterances can be straightforwardly effectuated through a suitable metric, such as the cosine similarity or the Euclidean distance. The second is how to eliminate external noise and distortions within an utterance or compensate session/channel variabilities induced by various sources in order that the speaker characteristics can be abundantly preserved and the mismatch between training and test conditions can be ulteriorly reduced.

In recent years, the Gaussian mixture model (GMM) has been found beneficial to serve as a sound tokenizer to represent a speech utterance by a probabilistic distribution with C mixture components [4]. The resulting representation is usually a super-vector $\boldsymbol{M} \in \mathbb{R}^{Cp}$, which is a concatenation of C p-dimensional mean vectors $\{\boldsymbol{M}_c \in \mathbb{R}^p | c = 1, \dots, C\}$ [5], or an i-vector, which is a compact form derived from the super-vector with a much lower dimension, where the aforementioned issues are addressed simultaneously [6]. Unlike the joint factor analysis (JFA) [7], in which the factor-loading matrices are composed of two distinct subspaces for separating the speaker and channel components residing in an utterance, the i-vector formulation only defines a single subspace, characterized by a matrix $\boldsymbol{T} \in \mathbb{R}^{Cp \times d}$ and termed the total variability, in which the speaker discriminatory information is presumably retained as much as possible. Given the subspace, the acoustic characteristics or physical quantities within the utterance are assumed to be embedded in a set of values [8], called the i-vector, which is not only conceived as the coordinate where the utterance locates in the subspace specified by $\boldsymbol{T}$, but essentially the expectation of Gaussian distributed latent variables of a factor analyzer or a total variability model. With the latent vector $\boldsymbol{w} \in \mathbb{R}^d$ ($d < Cp$), $\boldsymbol{M}$ is modeled as $\boldsymbol{T}\boldsymbol{w} + \boldsymbol{m}_0$, where $\boldsymbol{m}_0 \in \mathbb{R}^{Cp}$ is a bias term obtained by vertically stacking the mean vectors $\{\boldsymbol{m}_c \in \mathbb{R}^p | c = 1, \dots, C\}$ of a pre-trained GMM called the universal background model (UBM).

Given an utterance, the i-vector is derived by computing $\mathbb{E}[\boldsymbol{w}]$, i.e., the expectation over $\boldsymbol{w}$ drawn from the conditional probability given the utterance. From the probabilistic modeling perspective, the formulation can be interpreted as an attempt to recover a parsimonious set of latent random variables that describe a distribution over the observed acoustic feature vectors within the utterance, while encouraging careful modeling of the measurement of noise [9]. In the presence of latent variables, the expectation-maximization (EM) algorithm is employed to optimize the parameters with respect to the marginal likelihood, i.e., integrating the joint log likelihood over all values of the latent variables under their posterior probability.

Furthermore, due to the linearity of the total variability model, $\boldsymbol{M}_c$, the c-th block of $\boldsymbol{M}$, can be modeled as $\boldsymbol{T}_c\boldsymbol{w} + \boldsymbol{m}_c$, where $\boldsymbol{T}_c \in \mathbb{R}^{p \times d}$ and $\boldsymbol{m}_c$ is the c-th block of $\boldsymbol{m}_0$. That is, the super-vector based factor analyzer can be taken apart to form C factor analyzers; each corresponds to a mixture component of the UBM and has its own factor-loading matrix $\boldsymbol{T}_c$, but all share the same latent factors $\boldsymbol{w}$. We call this viewpoint of the i-vector formulation as the UBM-based mixture of factor analyzers (UMFA). The UMFA interpretation neither serves to group data points by similarity between features as the mixture of factor analyzers (MFA) defined in [10, 11] or [12] does, nor functions for dimensionality reduction on the p-dimensional acoustic feature vectors as applied in [13]. In contrast, it aims at how to represent an utterance as a single vector, whose size is manipulably *expanded* to d ($d > p$), and how to use the pre-existent UBM, each component of which characterizes some kind of sound characteristics, to locally guide the derivation of the total variability $\boldsymbol{T}$. In this sense, we can briefly conclude that the UMFA method gives consideration to two aspects simultaneously. In regard to feature expansion, the information

needed for discriminating utterances (not speakers) can be fairly preserved in a single vector by flexibly modifying the size of the vector. As for distribution-based clustering, since different acoustic features may be correlated within different mixture components of the UBM, the metric for feature expansion, which is controlled by the factor-loading matrix $\boldsymbol{T}_c$, may need to vary among different mixture components.

In UMFA, each factor-loading matrix $\boldsymbol{T}_c$ can be referred to as a subspace while the latent factors $\boldsymbol{w}$ with respect to an utterance denote the associated coordinate representation that describes the location of the utterance when projected in the subspace. Consequently, a geometrical question arises: is it reasonable to assume that the coordinates of the utterance represented in *different* subspaces utterly coordinate with one another? The answer is obviously no. Is it possible to relax this strong assumption to some extent? In this paper, we attempt to figure out this issue by making a more reasonable assumption that each mixture component has its own subspace $\boldsymbol{T}_c$ and the mixture components belonging to the same cluster share the same coordinate representation. This can be achieved by grouping the C mixture components of the UBM into G different clusters, where $G \leq C$, according to their acoustic similarity. Since similar mixture components are likely to result in similar factor-loading matrices, the respective projection locations of the utterance in these subspaces can be supposed to be nearly identical. Therefore, the proposed proposition not only enables each utterance to be represented by a set of i-vectors $\{\mathbb{E}[\boldsymbol{w}_g]|g = 1, \dots, G\}$ in appearance, but also in essence extends the conventional i-vector scheme to a more general form. In an extreme case when $G = 1$, i.e., all mixture components belong to a single cluster, the proposed approach is apparently reduced to the UMFA method. In another extreme, when $G = C$ and $d < p$, it becomes a special form of acoustic factor analysis (AFA) presented in [13], which is utilized to extract lower-dimensional enhanced features within each mixture component. Moreover, with the purpose of feature expansion, i.e., $d > p$, two more issues will be addressed in this paper. First, since the UMFA model in each cluster can be treated as a system of linear equations, it is necessary to avoid the underdetermined case while keeping the solvability for $\boldsymbol{T}_c$ by controlling the size of each cluster, the dimensionality of acoustic feature vectors, and the number of latent factors. Second, when $G \geq 2$, multiple i-vectors are produced for each utterance; thus, the integration or merging of these i-vectors becomes a new issue.

The remainder of this paper is organized as follows. In Section 2, we briefly re-interpret the conventional i-vector formulation in terms of UMFA via a directed factor graph. Section 3 presents our proposed framework, which is divided into two parts: clustering-based UMFA and the integration of multiple i-vectors. Finally, experiments, conclusions and future work are outlined in Sections 4 and 5, respectively.

2. UBM-based mixture of factor analyzers

Given a pre-trained universal background model (UBM), we can view the conventional i-vector formulation as a special form of mixture of factor analyzers (MFA) called UBM-based MFA (UMFA). From a generative viewpoint of the UMFA method, the observed vector is obtained by first choosing a vector for the latent variables with respect to a mixture component of the UBM, and then sampling the observed variables conditioned on the latent variables.

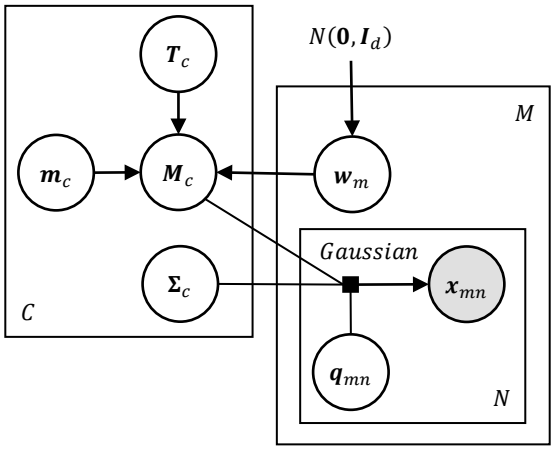

Figure 1. The directed factor graph of the UMFA method.

2.1. Formulation

Let $\mathcal{X} = \{\boldsymbol{x}_{mn} \in \mathbb{R}^p | m = 1, \dots, M; n = 1, \dots, N_m\}$ be the training data collected from M utterances spoken by various speakers in diverse environments or channel conditions, where $\boldsymbol{x}_{mn}$ denotes the n-th acoustic feature vector of the m-th utterance expressed by $\mathcal{Y}_m = \{\boldsymbol{x}_{mn} \in \mathbb{R}^p | n = 1, \dots, N_m\}$. Additionally, let $\mathcal{W} = \{\boldsymbol{w}_m \in \mathbb{R}^d | m = 1, \dots, M\}$ be a set of latent or unobservable random vectors, where each element $\boldsymbol{w}_m$ is associated with the m-th utterance and assumed to identically and independently follow a standard normal distribution, i.e.,

$$p(\boldsymbol{w}_m) = N(\mathbf{0}, \boldsymbol{I}_d), \tag{1}$$

where $\boldsymbol{I}_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix. The expectation of the responsibility that $\boldsymbol{w}_m$ takes for explaining the observations $\mathcal{Y}_m$, i.e., $\mathbb{E}(\boldsymbol{w}_m | \mathcal{Y}_m)$, is the so-called i-vector with respect to the m-th utterance.

Given a UBM with mixture components $\{U_c | c = 1, \dots C\}$, the distribution of each observation $\boldsymbol{x}_{mn}$ can be modeled as

$$\boldsymbol{x}_{mn} = \boldsymbol{m}_c + \boldsymbol{T}_c \boldsymbol{w}_m + \boldsymbol{e}_c \tag{2}$$

with $\boldsymbol{q}_{mn} = [q_{mn1}, \dots, q_{mnC}]^T \in \mathbb{R}^c$, in which $q_{mnc} = p(U_c | m, n)$ denotes the probability that $\boldsymbol{x}_{mn}$ is generated by U_c. In (2), $\boldsymbol{T}_c \in \mathbb{R}^{p \times d}$ and $\boldsymbol{m}_c \in \mathbb{R}^p$ refer to the total variability and the mean vector with respect to U_c, respectively. Moreover, the errors or noise within U_c are modeled by $\boldsymbol{e}_c \in \mathbb{R}^{p \times 1}$ distributed as $N(\mathbf{0}, \boldsymbol{\Sigma}_c)$, where $\boldsymbol{\Sigma}_c$ is a diagonal matrix. As illustrated in the directed factor graph in Figure 1 [14], the mixture component U_c is first picked according to q_{mnc}. Then, the observation $\boldsymbol{x}_{mn}$ is generated by sampling $\boldsymbol{w}_m$ from a Gaussian prior, passing it through the $\boldsymbol{T}_c$ matrix, and adding noise. Since the linear combination of two independent random variables having a Gaussian distribution also has a Gaussian distribution, the conditional density of each $\boldsymbol{x}_{mn} \subseteq \mathcal{Y}_m$ given $\boldsymbol{w}_m$ and $\{\boldsymbol{e}_c | c = 1, .., C\}$ can be expressed by

$$p_{UMFA}(\boldsymbol{x}_{mn} | \boldsymbol{w}_m) \propto \prod_{c=1}^{C} N(\boldsymbol{x}_{mn} | \boldsymbol{M}_c, \boldsymbol{\Sigma}_c)^{q_{mnc}}, \tag{3}$$

where $\boldsymbol{M}_c = \boldsymbol{m}_c + \boldsymbol{T}_c \boldsymbol{w}_m$ [12].

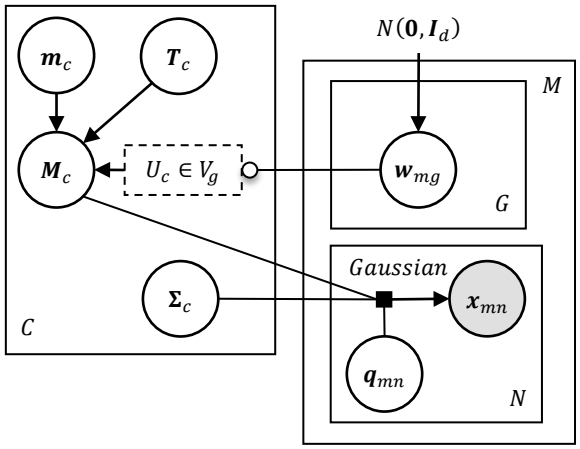

Figure 2. The directed factor graph of the clustering-based UMFA (CMFA) method.

2.2. Inference

We can follow [12] and [15] in applying the expectation maximization (EM) algorithm for estimating the parameters of a UMFA model, i.e., $\boldsymbol{T}_c$ and $\boldsymbol{\Sigma}_c$ for each mixture component. First, we treat each $\boldsymbol{w}_m$ as a missing-data vector, and take $\mathcal{Y}_m$ along with $\boldsymbol{w}_m$ as the complete-data vector. During the E-step, since the complete-data vector is related to a linear Gaussian model [16], the conditional distribution of $\boldsymbol{w}_m$ given $\mathcal{Y}_m$ along with the conditional expectation of $\mathbb{E}[\boldsymbol{w}_m]$ and $\mathbb{E}[\boldsymbol{w}_m \boldsymbol{w}_m^T]$ can be derived by referring to Chapter 2 in [17] or through a similar way of Proposition 1 in [18]. During the M-step, the expected complete-data log-likelihood with $\mathbb{E}[\boldsymbol{w}_m]$ and $\mathbb{E}[\boldsymbol{w}_m \boldsymbol{w}_m^T]$ is maximized with respect to the model parameters $\{\boldsymbol{T}_1, \dots, \boldsymbol{T}_C, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_C\}$. The resulting close forms for updating the parameters can be found consistent with those in [18]. This is one of the reasons why we take the UMFA model as an alternative path to derive the i-vector.

3. The proposed method

As discussed in Section 1, the spirit of the UMFA-based interpretation of the i-vector formulation lies in that, each of the different acoustic traits, characterized by some mixture component of the UBM, has its own linear subspace for feature expansion. However, such a *local* dimension expansion scheme gives a seemingly unreasonable assumption in regard of the geometry of linear transformations that, the coordinates, where an utterance is projected in individual subspaces for different acoustic traits, are exactly the same. In this section, we give a supplementary assumption to make it more advisable: only *similar* acoustic traits (or mixture components) are likely to result in *similar* linear subspaces for feature expansion, and the locations of an utterance projected in these subspaces are supposed to be nearly identical.

3.1. Formulation

Suppose the mixture components $\{U_c | c = 1, \dots, C\}$ of the UBM are grouped into G clusters $\{V_g | g = 1, \dots, G\}$ through some clustering algorithms, such as k-means clustering [22] conducted on the component mean vectors or other density-based clustering methods [24]. Compared with Figure 1, the generative process of the proposed clustering-based UMFA (CMFA) method in Figure 2 adds one gate, which is indicated by a dashed rectangle with a condition, to represent selection. It means that $\boldsymbol{M}_c$ would be affected by $\boldsymbol{w}_{mg}$ if the c-th component U_c belongs to the g-th cluster V_g. Therefore, for each utterance, more than one i-vector will be generated when $G > 1$, and $\boldsymbol{x}_{mn}$ can be modeled by modifying (2) as

$$\boldsymbol{x}_{mn} = \boldsymbol{m}_c + \boldsymbol{T}_c \boldsymbol{w}_{mg} + \boldsymbol{e}_c, \tag{4}$$

subject to $U_c \in V_g$. In the proposed model, the definitions of $\boldsymbol{q}_{mn}$, $\boldsymbol{T}_c$, $\boldsymbol{m}_c$, $\boldsymbol{e}_c$, and the prior distribution of $\boldsymbol{w}_{mg}$ are the same as those depicted in Section 2.1. For simplicity, we also assume that all clusters are statistically independent and uniformly distributed, i.e., $p(V_g) = 1/G$, $g = 1, \dots, G$. The conditional density of each $\boldsymbol{x}_{mn} \subseteq \mathcal{Y}_m$ given $\boldsymbol{w}_{mg}$ and $\{\boldsymbol{e}_c | c = 1, .., C\}$ can be expressed by

$$p_{CMFA}(\boldsymbol{x}_{mn} | \boldsymbol{w}_{mg}) \propto \prod_{c, U_c \in V_g} N(\boldsymbol{x}_{mn} | \boldsymbol{M}_c, \boldsymbol{\Sigma}_c)^{q_{mnc}}, \tag{5}$$

where $\boldsymbol{M}_c = \boldsymbol{m}_c + \boldsymbol{T}_c \boldsymbol{w}_{mg}$, if $U_c \in V_g$.

3.2. Inference

In a similar vein to Section 2.2, the complete data becomes $\{\boldsymbol{x}_{mn}, \boldsymbol{w}_{mg} | m = 1, \dots, M; n = 1, \dots, N_m; g = 1, \dots, G\}$ and its log likelihood is given by

$$\log \mathcal{L}_{CMFA}(\{\boldsymbol{T}_c\}_{c=1}^C, \{\boldsymbol{\Sigma}_c\}_{c=1}^C) = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{g=1}^{G} \sum_{c, U_c \in V_g} q_{mnc} \log N(\boldsymbol{x}_{mn} | \boldsymbol{m}_c + \boldsymbol{T}_c \boldsymbol{w}_{mg}, \boldsymbol{\Sigma}_c). \tag{6}$$

Due to the assumption of cluster independence, the conditional distribution of $\boldsymbol{w}_{mg}$ given $\mathcal{Y}_m$ can be derived one by one. Therefore, at the k-th iteration, the EM algorithm is implemented as follows. In the E-step, given the current fits $\{\boldsymbol{T}_c^{(k-1)}\}$ and $\{\boldsymbol{\Sigma}_c^{(k-1)}\}$, $\forall c, \text{s.t.}\, U_c \in V_g$, and the zero-order and the first-order statistics with respect to the m-th utterance over U_c expressed by $N_{mc} = \sum_{n, \boldsymbol{x}_{mn} \subseteq \mathcal{Y}_m} q_{mnc}$ and $\boldsymbol{F}_{mc} = \sum_{\boldsymbol{n}, \boldsymbol{x}_{mn} \subseteq \mathcal{Y}_m} q_{mnc} (\boldsymbol{x}_{mn} - \boldsymbol{m}_c)$, respectively, the conditional expectations over $\boldsymbol{w}_{mg}$ given $\mathcal{Y}_m$ can be derived through

$$\begin{aligned} p_{CMFA}^{(k)}(\boldsymbol{w}_{mg} | \mathcal{Y}_m) &\propto p^{(k)}(\{\boldsymbol{x}_{mn}\}_{n=1}^{N_m} | \boldsymbol{w}_{mg}) p(\boldsymbol{w}_{mg}) \\ &\propto \exp\left((\boldsymbol{w}_{mg} - \boldsymbol{a}_{mg})^T \boldsymbol{l}_{mg}^{-1} (\boldsymbol{w}_{mg} - \boldsymbol{a}_{mg}) \right), \end{aligned} \tag{7}$$

where

$$\begin{cases} \boldsymbol{l}_{mg}^{-1} = \left(\boldsymbol{I}_d + \sum_{c, U_c \in V_g} N_{mc} \boldsymbol{T}_c^{(k-1)T} \boldsymbol{\Sigma}_c^{(k-1)-1} \boldsymbol{T}_c^{(k-1)} \right)^{-1} \\ \boldsymbol{a}_{mg} = \boldsymbol{l}_{mg}^{-1} \sum_{c, U_c \in V_g} \boldsymbol{T}_c^{(k-1)T} \boldsymbol{\Sigma}_c^{(k-1)-1} \boldsymbol{F}_{mc} \end{cases}.$$

Then, the conditional expectations are given by

$$\begin{cases} \mathbb{E}^{(k)}[\boldsymbol{w}_{mg}] = \boldsymbol{a}_{mg} \\ \mathbb{E}^{(k)}[\boldsymbol{w}_{mg} \boldsymbol{w}_{mg}^T] = \boldsymbol{a}_{mg} \boldsymbol{a}_{mg}^T + \boldsymbol{l}_{mg}^{-1} \end{cases}. \tag{8}$$

The M-step is implemented by maximizing the expected complete-data log-likelihood $\mathbb{E}[\log \mathcal{L}_{CUMFA}]$ over $\mathcal{X}$ with the results derived in the E-step. This yields the updated estimates of $\widetilde{\boldsymbol{T}}_c$ and $\widetilde{\boldsymbol{\Sigma}}_c$, $\forall c, \text{s.t.}\, U_c \in V_g$, as

$$\begin{cases} \widetilde{\boldsymbol{T}}_c^{(k)} = \left(\sum_{m=1}^{M} \boldsymbol{F}_{mc} \mathbb{E}^{(k)}[\boldsymbol{w}_{mg}^T]\right)\left(\sum_{m=1}^{M} N_{mc} \mathbb{E}^{(k)}[\boldsymbol{w}_{mg}\boldsymbol{w}_{mg}^T]\right)^{-1} \\ \widetilde{\boldsymbol{\Sigma}}_c^{(k)} = N_c^{-1}\mathrm{diag}\left(\boldsymbol{S}_c - \left(\sum_{m=1}^{M} \boldsymbol{F}_{mc} \mathbb{E}^{(k)}[\boldsymbol{w}_{mg}^T]\right)\widetilde{\boldsymbol{T}}_c^{(k)T}\right) \end{cases},$$

where N_c and $\boldsymbol{S}_c$ are defined by $\sum_{m,n} q_{mnc}$ and $\sum_{m,n} q_{mnc}\,(\boldsymbol{x}_{mn} - \boldsymbol{m}_c)(\boldsymbol{x}_{mn} - \boldsymbol{m}_c)^T$, respectively.

3.3. Utterance representation

When $G \geq 2$, more than one i-vectors are produced for each utterance. Here we introduce two ways, namely pooling and augmentation, to integrate or merge multiple i-vectors $\{\mathbb{E}[\boldsymbol{w}_{mg}] | g = 1, \ldots, G\}$ into a single vector.

In the case of pooling, the ultimate i-vector $\boldsymbol{z}_m$ is obtained by averaging $\{\mathbb{E}[\boldsymbol{w}_{mg}] | g = 1, \ldots, G\}$ over the posterior distribution of the cluster membership given $\mathcal{Y}_m$; that is,

$$\begin{cases} \boldsymbol{z}_m = \sum_{g=1}^{G} P_{mg}\mathbb{E}[\boldsymbol{w}_{mg}] \\ P_{mg} = \sum_{n,c,U_c \in V_g} q_{mnc} \Big/ \sum_{n,c} q_{mnc} \end{cases}. \quad (9)$$

The pooling strategy is an instinctively simple way to combine several vectors, but it might cause a query that the summation in (9) is problematic since the physical meanings of the i-th components in the vectors $\{\mathbb{E}[\boldsymbol{w}_{mg}] | g = 1, \ldots, G\}$ might differ due to their dissimilar subspaces $\boldsymbol{T}_c$'s.

In contrast, the augmentation strategy seems to be capable of avoiding the above problem by forming the ultimate i-vector with larger dimensionality as

$$\boldsymbol{z}_m = [\mathbb{E}[\boldsymbol{w}_{m1}]^T \quad \cdots \quad \mathbb{E}[\boldsymbol{w}_{mG}]^T]^T. \quad (10)$$

An advantage of the augmentation strategy is that most of metrics defined in the Euclidean space or the inner product space, such as the 2-norm distance and the cosine similarity, are plausible when applied to compare any pair of $\boldsymbol{z}_m$'s. Moreover, although the size of $\boldsymbol{z}_m$ becomes Gd, which is larger than d when the pooling strategy is used, it is still acceptable since G is usually much smaller than C and a smaller d can be set in this case.

4. Experiments

All the experiments in this paper were carried out on the male portion of the core condition (1conv4w-1conv4w) in NIST SRE05, where each target speaker provided only one 5-min conversational utterance for enrollment [19]. The evaluation task contains 1,220 true trials and 11,513 false trials. We used equal error rate (EER) and normalized minimum decision cost function (minDCF) as acknowledged metrics for evaluation. With the frame length of 25 ms and the frame shift of 10 ms, speech parameters were represented by a 60-dimensional feature vector of Mel-frequency cepstral coefficients (MFCC) with first and second derivatives appended using a 2-frame window, followed by data distribution-based feature warping with a 300-frame window in order to compensate for the effects of environmental mismatch [20].

A gender-dependent UBM consisting of 2,048 Gaussian components with diagonal covariance matrices, the UMFA and CMFA models, as well as the back-end classifier, i.e., the PLDA model [25], were trained with the data drawn from SRE04, which contains 1,867 utterances spoken by 122 speakers. For the sake of fair comparison, the dimensionality of the ultimate vector fed into the back-end classifier is fixed to 600, and the ranks of the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$, which pertain to the speaker and session variability in the PLDA model, are set to 300 and 300, respectively.

Table 1. *The average Frobenius norm between any pair of $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$, given that U_i and U_j belong to the same cluster, for various numbers of clusters generated by the k-means method.*

# Clusters	1	2	4	8	16	32
$\overline{d}(\boldsymbol{T}_i, \boldsymbol{T}_j)$	181.20	106.93	70.11	46.83	29.59	18.20

Table 2. *EER (%) and minDCF with the pooling strategy for various d and G, which denote the size of the i-vector generated by each cluster and the number of clusters, respectively, evaluated on SRE05.*

d, G	600, 1	600, 2	600, 4	600, 8
EER (%)	**5.83**	8.29	13.36	20.25
minDCF	**0.30**	0.39	0.54	0.73

Table 3. *EER (%) and minDCF with the augmentation strategy for various d and G evaluated on SRE05.*

d, G	600, 1	300, 2	200, 3	150, 4
EER (%)	5.83	**5.49**	5.98	6.16
minDCF	0.30	**0.28**	0.30	0.29

To measure the difference between two subspaces $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ that belong to the same cluster, we compute their Frobenius norm, $d(\boldsymbol{T}_i, \boldsymbol{T}_j) = \|\boldsymbol{T}_i - \boldsymbol{T}_j\|_F$ [23]. Table 1 shows the average Frobenius norm with respect to various numbers of clusters generated by the k-means method [22]. From the table, it is clear that the average Frobenius norm over a cluster becomes smaller with the increasing number of clusters. The results reveal that the subspaces associated with the mixture components in a cluster are indeed more similar to one another than those of ungrouped ones, and implicitly resolve our concerns discussed in Section 3. The results of the proposed CMFA model with the pooling strategy and the augmentation strategy are, respectively, shown in Tables 2 and 3, where the second column with $d = 600$ and $G = 1$ denotes the baseline results of the conventional UMFA model. From Table 2, it is found that the proposed CMFA model with the pooling strategy does not offer any improvement over the conventional UMFA model. In contrast, from Table 3, we can see that the proposed CMFA model with the augmentation strategy outperforms the baseline when $d = 300$ and $G = 2$. The reason why the CMFA model did not perform well when $G > 2$ could be that, although the number of clusters controls the homogeneity of the mixture components in a cluster, the condition with a larger G might suffer from the problem of data insufficiency. Actually, the CMFA model cannot guarantee that all mixture components are evenly allotted to the clusters, nor can it assure that each training utterance possesses all characteristics dwelling in each cluster.

5. Conclusions

This paper has presented a clustering-based scheme to solve the problem that the conventional i-vector formulation might meet. With a suitable merging mechanism, our proposed method can perform well on the NIST SRE corpus. Although a good start in this direction has been made, much more research is needed. E.g., more merging strategies should be studied. How to keep the clusters balanced is also worth study.

6. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[3] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561-580, 1975.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1, pp. 19-41, 2000.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, May 2006.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788-798, 2011.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980-988, 2008.

[8] P. Kenny, "A small footprint i-vector extractor," presented at *the Odyssey - The Speaker and Language Recognition Workshop*, 2012.

[9] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611-622, 1999.

[10] G. McLachlan and D. Peel, "Mixtures of factor analyzers," presented at *the International Conference on Machine Learning (ICML)*, 2000.

[11] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2004.

[12] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," *CRG-TR-96-1*, 1997.

[13] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 842-853, 2013.

[14] L. Dietz, "Directed factor graph notation for generative models," *Max Planck Institute for Informatics*, 2010.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1-38, 1997.

[16] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305-345, 1999.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345-354, May 2005.

[19] The NIST Year 2005 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2005/index.htm.

[20] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," presented at *the Odyssey - The Speaker and Language Recognition Workshop*, 2011.

[21] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," presented at *Interspeech*, 2011.

[22] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society, Series C*, vol. 28, no. 1, pp. 100-108, 1979.

[23] E. W. Weisstein, "Frobenius norm," from *MathWorld--A Wolfram Web Resource*. http://mathworld.wolfram.com/FrobeniusNorm.html

[24] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," presented at *the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[25] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144-157, 2012.