Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Histogram equalization using a reduced feature set of background speakers' utterances for speaker recognition*

Myung-jae KIM, Il-ho YANG, Min-seok KIM, Ha-jin YU‡

(*School of Computer Science, University of Seoul, Seoul 02504, Korea*)

E-mail: mj@uos.ac.kr; heisco@hanmail.net; ms@uos.ac.kr; hjyu@uos.ac.kr

Received Nov. 3, 2015; Revision accepted Apr. 18, 2016; Crosschecked Feb. 21, 2017

Abstract: We propose a method for histogram equalization using supplement sets to improve the performance of speaker recognition when the training and test utterances are very short. The supplement sets are derived using outputs of selection or clustering algorithms from the background speakers' utterances. The proposed approach is used as a feature normalization method for building histograms when there are insufficient input utterance samples. In addition, the proposed method is used as an i-vector normalization method in an i-vector-based probabilistic linear discriminant analysis (PLDA) system, which is the current state-of-the-art for speaker verification. The ranks of sample values for histogram equalization are estimated in ascending order from both the input utterances and the supplement set. New ranks are obtained by computing the sum of different kinds of ranks. Subsequently, the proposed method determines the cumulative distribution function of the test utterance using the newly defined ranks. The proposed method is compared with conventional feature normalization methods, such as cepstral mean normalization (CMN), cepstral mean and variance normalization (MVN), histogram equalization (HEQ), and the European Telecommunications Standards Institute (ETSI) advanced front-end methods. In addition, performance is compared for a case in which the greedy selection algorithm is used with fuzzy C-means and K-means algorithms. The YOHO and Electronics and Telecommunications Research Institute (ETRI) databases are used in an evaluation in the feature space. The test sets are simulated by the Opus VoIP codec. We also use the 2008 National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) corpus for the i-vector system. The results of the experimental evaluation demonstrate that the average system performance is improved when the proposed method is used, compared to the conventional feature normalization methods.

Key words: Speaker recognition; Histogram equalization; i-vector
http://dx.doi.org/10.1631/FITEE.1500380 **CLC number:** TN912.34

1 Introduction

Current speaker recognition systems have achieved good performance when the training and test environments are well matched. When the training and test environments do not match, the systems show poor performance. Therefore, feature normalization methods are applied to reduce the channel and acoustic mismatches. Conventional feature normalization methods, such as cepstral mean normalization (CMN) (Atal, 1974) and cepstral mean and variance normalization (MVN) (Viikki and Laurila, 1998), are widely used to remove linear channel effects. However, they are unsuitable for the removal of nonlinear effects. To compensate for nonlinear effects, linear approximation methods such as the vector Taylor series (VTS) method (Moreno *et al.*, 1996) and static linear approximation (SLA) (Kim, 1998) have been applied. Also, the European Telecommunications Standards Institute (ETSI) advanced front-end (AFE) in the ETSI standard (ETSI ES

‡ Corresponding author

* Project supported by the IT R&D Program of MOTIE/KEIT (No. 10041610)

ORCID: Ha-Jin YU, http://orcid.org/0000-0003-3657-0665

202 212) (ETSI, 2005) has been introduced for noise reduction using a Wiener filter approach.

Histogram equalization (HEQ) has also been proposed for feature normalization. Originally, HEQ was used for brightness and contrast adjustment in digital images (Gonzalez and Wintz, 1987). Later, HEQ was applied in speech recognition systems as a feature normalization method (Segura *et al.*, 2004; Torre *et al.*, 2005). The HEQ approach transforms an input test sample to a reference value based on a probability density function (PDF) (Segura *et al.*, 2004). Gaussian density may be appropriate for the reference and test feature distributions of speech utterances, as traditionally most of the models for speech processing are based on Gaussian distributions. This approach can be used as a nonlinear transformation, unlike conventional feature normalization methods. In addition, various applications of HEQ such as feature warping (Pelecanos and Sridharan, 2001) and modified segmental HEQ (Skosan and Mashao, 2006) have been adopted in speaker recognition systems. These approaches divide the samples of an input utterance into small windows, and transform the samples by applying HEQ to each of them. The sizes of the windows should be at least 3 s long. According to Blanco's research, approximately 500 ordered samples are sufficient for an effortless estimation of robust cumulative distribution functions (CDFs) (Blanco *et al.*, 2000). However, when an input test utterance has fewer than 500 samples, the utterance may not be usable for building histograms. The speaker recognition systems that use HEQ for building histograms with fewer samples show poorer performance compared to speaker recognition systems that use conventional feature normalization methods.

In this research, to build robust histograms, we estimate the CDFs of the input utterance by selecting samples from the training set for the universal background model (UBM) that is widely used for modeling general speakers. We apply various algorithms, such as the fuzzy C-means (Cannon *et al.*, 1986), K-means (Duda *et al.*, 2012), and greedy selection (Franc, 2005; Kim *et al.*, 2008) algorithms, to create the supplement sets. Each algorithm is applied to randomly selected speakers' speech from the UBM training set. A supplement set is created by arranging the outputs of the selection algorithms in ascending order.

We also propose a method to apply histogram equalization to the state-of-the-art i-vector-based (Dehak *et al.*, 2011) probabilistic linear discriminant analysis (PLDA) approach (Garcia-Romero and Espy-Wilson, 2011) as an i-vector normalization method. The proposed method is similar to the rank norm method (Stolcke *et al.*, 2008), which is a Gaussian mixture model (GMM) mean supervector normalization method. Originally, rank norm normalizes a GMM mean supervector to follow a uniform distribution. Then, a Gaussianized rank norm is introduced (Jiang *et al.*, 2012) and the rank norm normalizes an i-supervector to follow a Gaussian distribution instead of a uniform distribution. The Gaussianized rank norm is used as a whitening transform of i-supervectors. However, because i-supervectors have high dimensionality and data scarcities, it is difficult to apply a whitening transform. In this study, we apply a Gaussianized rank norm to i-vectors and propose a method using a reduced background set to shorten the processing time for the rank norm. We show experimental results of combining the proposed method and length normalization, such as eigen factor radial (EFR) (Bousquet *et al.*, 2011) and spherical nuisance normalization (SphNorm) (Bousquet *et al.*, 2012), using common data sets. The results show that we can reduce processing time and improve speaker recognition performance.

2 Conventional histogram equalization

The purpose of HEQ is to convert an input test utterance distribution into a reference distribution, e.g., a standard normal distribution. To achieve this goal, HEQ transforms samples of the input test utterance into corresponding values of the reference distribution. Each sample is mapped onto a reference value that has the same CDF. The transform function F for HEQ is given as

$$y = F(x) = C_{\text{ref}}^{-1}(C_X(x)), \tag{1}$$

where X is the observed sequence, x is an element of the observed sequence, y is the transformed value, $C_X(x)$ is the CDF of the element x, and C_{ref}^{-1} is the inverse CDF of the reference distribution. A lookup table is used for the transformation of the sample. Fig. 1 shows an example of HEQ. The input utterance (Fig. 1a) has a multimodal distribution,

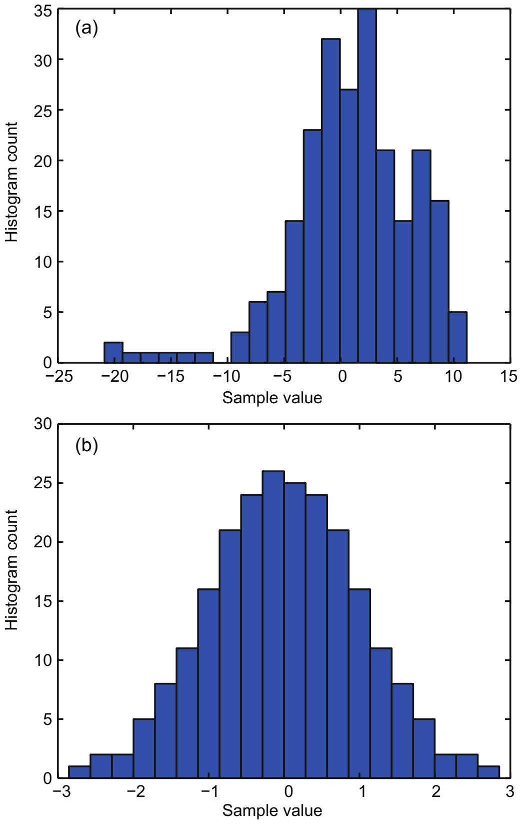

Fig. 1 Example of the transformed distribution using histogram equalization: (a) distribution of an input sequence; (b) distribution of the converted sequence

whereas the transformed samples (Fig. 1b) have a unimodal distribution.

We use a standard normal distribution as the reference distribution:

$$P(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{2}$$

2.1 Cumulative histogram-based histogram equalization

To estimate CDFs, the cumulative histogram-based HEQ (CHEQ) (Skosan and Mashao, 2006) counts the samples in each bin. The details of CHEQ are as follows: define an observed sequence X consisting of T frames of a particular (dth) component as

$$X = O^d = \{o_1^d, o_2^d, \cdots, o_t^d, \cdots, o_T^d\}. \tag{3}$$

We then determine the minimum value, $\min(X)$, and the maximum value, $\max(X)$. The range of values between $\min(X)$ and $\max(X)$ is divided into M equally-spaced non-overlapping bins. The bins, B, are defined such that a bin B_i has the range $[b_i, b_{i+1})$.

$$\min(X) = b_1 < b_2 < \cdots < b_{M+1} = \max(X). \tag{4}$$

These bins are used to construct histograms of the observed sequence X. To build histograms, we count the number of observations belonging to each bin. Each bin is normalized by the total number of observations N_X as follows:

$$p_X(x_t \in B_i) = \frac{n_i}{N_X}, \tag{5}$$

where n_i is the number of observations in B_i. The computation of cumulative histograms is performed by using a normalized histogram:

$$C_X(x_t \in B_i) = \sum_{j=1}^{i} \frac{n_j}{N_X}. \tag{6}$$

A transformed sequence can be obtained as follows:

$$y_t = C_{\text{ref}}^{-1}(C_X(x_t)), \tag{7}$$

which is in direct correspondence to Eq. (1).

2.2 Order statistic-based histogram equalization

To estimate the CDFs, an order statistic-based HEQ (OHEQ) (Segura *et al.*, 2004) uses the ranks of an observed sequence X consisting of T frames of a particular (dth) component:

$$X = O^d = \{o_1^d, o_2^d, \cdots, o_t^d, \cdots, o_T^d\}. \tag{8}$$

Ranks are obtained using a sequence of X that is sorted in ascending order:

$$x_{r(1)} \leq x_{r(2)} \leq \cdots \leq x_{r(t)} \leq \cdots \leq x_{r(T)}, \tag{9}$$

where $r(t)$ is the index of the observed sequence (i.e., $x_{r(t)}$ is the tth element in the sorted sequence). A CDF is estimated as follows:

$$\varPhi_t = (R(x_t) - 0.5)/T, \tag{10}$$

where $R(x_t)$ denotes the rank of x_t in the sorted sequence. A transformed sequence using the CDF estimated by the order statistic can be represented as

$$y = C_{\text{ref}}^{-1}(\varPhi), \tag{11}$$

which directly corresponds to Eq. (1).

3 Feature reduction methods

In this section, we describe the methods to create a supplement set for compensating for an insufficient number of samples for histogram equalization. To make a supplement set, two types of algorithms, clustering and selection algorithms, are considered. We select two clustering algorithms, fuzzy C-means (Cannon *et al.*, 1986) and K-means (Duda *et al.*, 2012). A set of centroids of clusters is used as a supplement set. We also select a greedy selection algorithm. This algorithm solves large-scale problems using training set reduction. The reduced training set is used as a supplement set.

3.1 Fuzzy C-means

Fuzzy C-means is a conventional clustering algorithm that uses fuzzy memberships of clusters. The fuzzy memberships are equivalent to the probability in maximum likelihood estimation. An observed sequence X consisting of T vectors is defined as follows:

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T\}. \tag{12}$$

The C centroids of fuzzy C-means are given as

$$\mu = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_c, \cdots, \boldsymbol{\mu}_C\}. \tag{13}$$

The purpose of fuzzy C-means is to find memberships minimizing the objective function J, which is given as follows:

$$J = \sum_{j=1}^{C} \sum_{i=1}^{T} \omega_{ij}^m \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2, \tag{14}$$

where m is a free parameter (a real number greater than 1) and ω_{ij}^m is the membership of the observed sequence with the clusters. The membership is obtained as

$$\omega_{ij} = \left[\sum_{k=1}^{C} \left(\frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|}{\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|}\right)^{\frac{2}{m-1}}\right]^{-1}. \tag{15}$$

The membership should satisfy the following equation:

$$\sum_{j=1}^{C} \omega_{ij} = 1, \qquad i = 1, 2, \cdots, T. \tag{16}$$

Using Eq. (15), the centroids can be updated as follows:

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{T} \omega_{ij}^m \cdot \boldsymbol{x}_i}{\sum_{i=1}^{T} \omega_{ij}^m}. \tag{17}$$

The centroids are updated using Eqs. (15) and (17) iteratively. The iteration can be stopped when

$$\sum_{j=1}^{C} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j| < \varepsilon, \tag{18}$$

where ε is a real number.

3.2 Greedy selection

A greedy selection algorithm is used to reduce the training set for large-scale problems. The greedy selection algorithm is used for greedy kernel PCA (GKPCA) (Franc, 2005; Kim *et al.*, 2008). The reduced training set is used to find the coordinates that maximize the covariance. We define a training set X consisting of T feature vectors as follows:

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T\}. \tag{19}$$

A reduced set G with L sample vectors is given by

$$G = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_l, \cdots, \boldsymbol{g}_L\}. \tag{20}$$

The purpose of the greedy selection algorithm is to obtain a reduced set G from the training set ($G \subset X$). To obtain the reduced set G, the training set X and the reduced set G are mapped into the feature space, having an infinite dimension such as $X_\phi = \{\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \cdots, \phi(\boldsymbol{x}_t), \cdots, \phi(\boldsymbol{x}_T)\}$ and $G_\phi = \{\phi(\boldsymbol{g}_1), \phi(\boldsymbol{g}_2), \cdots, \phi(\boldsymbol{g}_l), \cdots, \phi(\boldsymbol{g}_L)\}$, respectively. Let $\mathit{\Gamma} = \{1, 2, \cdots, t, \cdots, T\}$ be the set of indices of the training set X, and $\mathcal{L} = \{1, 2, \cdots, l, \cdots, L\}$ be the set of indices of the reduced set G. Using X_ϕ and G_ϕ, we obtain the reconstructed training set $\tilde{X}_\phi$:

$$\tilde{X}_\phi = BG_\phi, \tag{21}$$

where $\tilde{X}_\phi = \{\phi(\tilde{\boldsymbol{x}}_1), \phi(\tilde{\boldsymbol{x}}_2), \cdots, \phi(\tilde{\boldsymbol{x}}_t), \cdots, \phi(\tilde{\boldsymbol{x}}_T)\}$ is a $T \times \infty$ matrix, and $B = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_t, \cdots, \boldsymbol{\beta}_T\}$ is a $T \times L$ matrix. $\boldsymbol{\beta}_t$ is an L-dimensional row vector with coefficients of the reduced set G_ϕ, which is an $L \times \infty$ matrix. The objective function of the reduced set to minimize the reconstruction error is the mean squared error:

$$\varepsilon_{\mathrm{MS}} = \frac{1}{T} \sum_{t \in \mathit{\Gamma}} \left\| \phi(\boldsymbol{x}_t) - \sum_{l \in \mathcal{L}} \boldsymbol{\beta}_{tl} \phi(\boldsymbol{g}_l) \right\|^2. \tag{22}$$

The objective function can be re-expressed using a kernel trick as follows:

$$\begin{aligned} \varepsilon_{\mathrm{MS}} =& \frac{1}{T} \sum_{t \in \mathit{\Gamma}} [k(\boldsymbol{x}_t, \boldsymbol{x}_t) - 2\boldsymbol{K}^g \boldsymbol{k}^g(x_t) \\ &+ \boldsymbol{k}^g(\boldsymbol{x}_t) \cdot \boldsymbol{K}^g \boldsymbol{k}^g(\boldsymbol{x}_t)], \end{aligned} \tag{23}$$

where $\boldsymbol{K}^g$ is an $L \times L$ kernel matrix of the selected set G_ϕ and $\boldsymbol{k}^g(\boldsymbol{x}_t) = \{k(\boldsymbol{g}_1, \boldsymbol{x}_t), k(\boldsymbol{g}_2, \boldsymbol{x}_t), \cdots, k(\boldsymbol{g}_l, \boldsymbol{x}_t)\}$. To select G_ϕ, we use the mean squared error $\varepsilon_{\text{MS}}^{(y)}$ in the yth iteration. The mean squared error $\varepsilon_{\text{MS}}^{(y)}$ can be upper-bounded as the following inequality:

$$\varepsilon_{\text{MS}}^{(y)} \leq \frac{1}{T}(T-y) \max_{t \in \Gamma \backslash \mathcal{L}} \left\| \phi(\boldsymbol{x}_t) - \sum_{l \in \mathcal{L}} \boldsymbol{\beta}_{tl} \phi(\boldsymbol{g}_l) \right\|^2 . \tag{24}$$

In this study, we select a linear kernel $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^{\text{T}} \boldsymbol{x}_2$ instead of a Gaussian radial basis function (RBF) kernel because HEQ transforms samples in the linear space. The details of GKPCA were described by Franc (2005) and Kim *et al.* (2008).

4 Proposed method

The proposed method is based on the order statistic-based HEQ and helps in computing CDFs for a short input test utterance for which it is quite difficult to estimate robust CDFs. The proposed method involves creating a supplement set that represents the entire distribution of the UBM training set. The proposed method uses the centroids obtained by applying clustering algorithms, such as fuzzy C-means (Cannon *et al.*, 1986), K-means, or the greedy selection algorithm, to derive the supplement sets. To describe the proposed method, we let U be the set of utterances of P speakers that is used to train UBM, given as

$$U = \{u(1), u(2), \cdots, u(p), \cdots, u(P)\}, \tag{25}$$

where $u(p)$ is the set of utterances of the pth speaker consisting of T_p frames in the UBM training set. The greedy selection algorithm selects N samples of a particular (dth) component from each $u(p)$. Let the set of the selected sets of samples be

$$G = \{g(1), g(2), \cdots, g(p), \cdots, g(P)\}. \tag{26}$$

The selected set of samples $g(p)$, consisting of N samples, can be presented as

$$M = g^d(p) = \{g_1^d, g_2^d, \cdots, g_n^d, \cdots, g_N^d\}, \tag{27}$$

where N is the number of samples selected from each speaker ($N \ll T_p$). An observed input sequence X consisting of L samples is given by

$$X = O^d = \{o_1^d, o_2^d, \cdots, o_l^d, \cdots, o_L^d\}. \tag{28}$$

We calculate two kinds of ranks of the samples in the observed input sequence. One is composed of the ranks of X in the reduced development set M (denoted as r^M). The other is composed of the ranks of the samples in sequence X (denoted as r^X). We then define a new rank as the sum of the two ranks as follows:

$$r^{\text{new}} = r^M + r^X. \tag{29}$$

Using the new rank, we define CDFs as follows:

$$\varPhi^{\text{new}} = \frac{r^{\text{new}} - 0.5}{(P \cdot N + 1) + L}, \tag{30}$$

where $P \cdot N$ is the number of elements in the supplement set which has N frames of P speakers. The 'plus-one' term of the denominator is used to forbid CDFs from exceeding 1.0, because the values over 1.0 are mapped to infinite by the inverse Gaussian distribution function. Eqs. (29) and (30) can be rewritten using the weight coefficients as

$$\varPhi^{\text{new}} = \frac{2(\alpha r^M + \beta r^X) - (\alpha + \beta)}{2\alpha(P \cdot N + 1) + 2\beta L}, \tag{31}$$

where $\alpha + \beta = 1$. Eq. (31) is equivalent to Eq. (30) when $\alpha = \beta = 0.5$. The samples are transformed by using Eq. (1) with the estimated CDFs. The mapping function is given as

$$y = C_{\text{ref}}^{-1}(\varPhi^{\text{new}}). \tag{32}$$

5 Experimental setups and results

We evaluate the proposed method in the acoustic feature space and i-vector space. In Section 5.1, the experimental setups and results of acoustic feature normalization are shown for a speaker identification system. In Section 5.2, we describe the setups and results of i-vector normalization for a speaker verification system.

5.1 Acoustic feature normalization

In this section, we describe the experimental setups and results for speaker identification in the acoustic feature space. Fig. 2 shows the proposed system flow for acoustic features. The samples selected by the proposed method are used in both offline and online steps. The system performance is evaluated in the Opus codec environment.

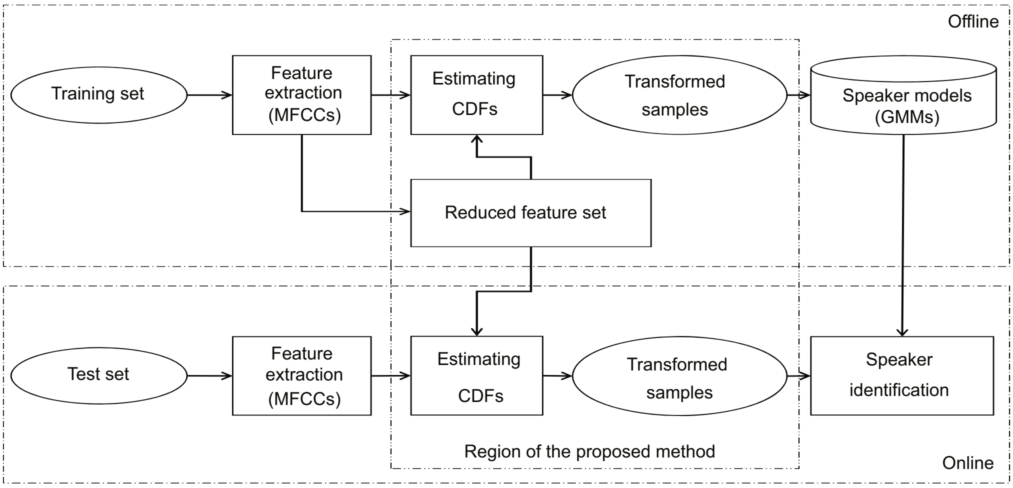

Fig. 2 Speaker identification system flow of the proposed method for acoustic feature normalization (CDFs: cumulative distribution functions; MFCCs: mel-frequency cepstral coefficients; GMMs: Gaussian mixture models)

5.1.1 Databases

To evaluate the proposed system, we use the YOHO database (Linguistic Data Consortium), and the Korean Speaker Recognition Database Recording using Middle Price Microphone (Electronics and Telecommunications Research Institute, ETRI). The data in the YOHO database is recorded at 8 kHz, and consists of enrollment and verification sessions. The enrollment data is used for the adaptation of speaker models, and the verification data is used for testing. The data in the ETRI database is recorded at 16 kHz, and consists of week, month, and season modes that are recorded in intervals of weeks, months, and seasons, respectively. The week, month, and season modes are composed of 100, 100, and 50 speakers, respectively. Each mode consists of four sessions, each of which consists of five trials. We use the sentence parts numbered 10 to 19 of the first trial of the first session in the week mode for the adaptation of the speaker models. The same parts of the first trial of the second, third, and fourth sessions in the week mode are used to test the speaker identification systems. To train UBM, we use the sentence part of the first trial of the first session in month and season modes. To match the sampling rate of YOHO and ETRI databases, the waveforms are downsampled from 16 kHz to 8 kHz. The total number of utterances for UBM training is 1500 (150 speakers × 10 sentences), the total number of utterances for speaker model adaptation is 1000 (100 speakers × 10 sentences), and the total number of utterances for testing the systems is 3000 (3 sessions × 100 speakers × 10 sentences). Each utterance consists of 200 to 300 sample feature vectors. This is reduced to 150 to 250 samples by silence elimination. To simulate the test sets, we apply the Opus codec (Valin *et al.*, 2012) to the waveforms.

5.1.2 Baseline setup

To extract speech features from the waveforms, the filtered waveforms are partitioned into frames that are 25 ms long with 10 ms intervals. A Hamming window is applied to each frame. After partitioning, we apply pre-emphasis with a factor of 0.97. The silence frames are eliminated using the energy threshold method. The frames are then filtered by mel-scaled triangular filters, and the outputs of the mel-scaled triangular filters transformed by discrete cosine transformation into 18-dimensional mel-frequency cepstral coefficients (MFCCs). For an ETSI advanced front-end (AFE), we use the same parameters as those of MFCCs. The statistics library in Scipy (Jones *et al.*, 2001) is used for CDF mapping in Eq. (1).

The Gaussian mixture model, universal background model (GMM-UBM) method (Reynolds *et al.*, 2000), is used for speaker modeling. In this method, the speaker models are obtained from UBM using the maximum a posteriori (MAP) adaptation method ($\tau = 1$). The UBM in this case consists of 128 mixture components.

5.1.3 Experimental results

The proposed system was compared with MFCC (the baseline system without a feature normalization method), CMN, MVN, CHEQ (which has 1000 bins), OHEQ, and AFE.

Fig. 3a shows the error rates in the YOHO database. The proposed method outperforms other feature normalization methods. Among these methods, the proposed method using greedy selection with a supplement set consisting of 350 samples (10 speakers × 35 samples) shows the best performance. The proposed method using K-means also shows better performance than the other feature normalization methods.

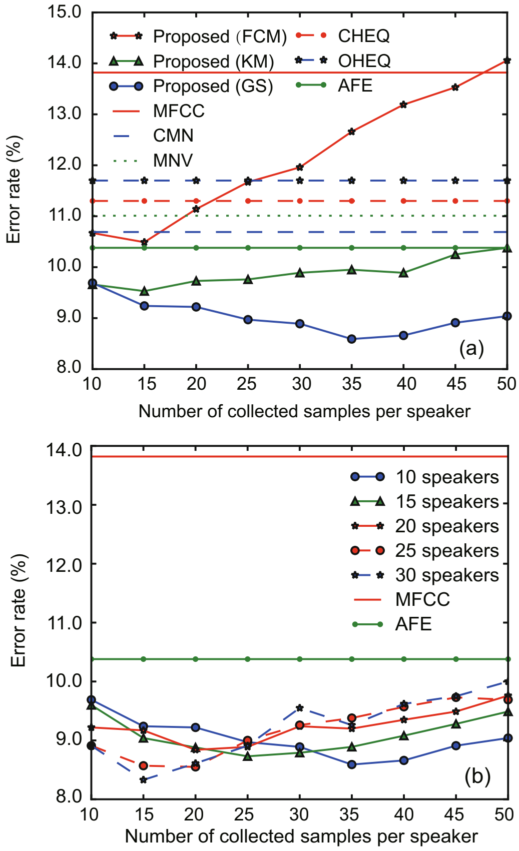

Fig. 3 Speaker identification error rates of the proposed method in the YOHO database: (a) according to the size of the supplement set of 10 speakers; (b) using greedy selection according to the number of speakers

Fig. 3b shows the results according to the number of speakers in the YOHO database. In Fig. 3b, we use greedy selection for the proposed method because the method has shown the best performance in Fig. 3a. For simplification, we show only the results of MFCC and AFE. The results of the other normalization methods are equivalent to those in Fig. 3a. The proposed method is able to use fewer supplement samples per speaker by increasing the number of speakers. When the supplement set has 450 samples (30 speakers × 15 samples), the proposed method has the best performance. There is no significant difference among the performances of the supplement set with 350 samples (10 speakers × 35 samples). These results show that the proposed method is dependent on the total number of supplement samples to a greater degree than it is dependent on the number of speakers.

The results with the ETRI database (Fig. 4a) are similar to those with the YOHO database. Using the ETRI database, the proposed method outperforms other feature normalization methods. The proposed method using greedy selection with a supplement set consisting of 350 samples (10 speakers × 35 samples) gives the best performance.

Fig. 4b shows the results according to the number of speakers for the ETRI database. Greedy selection is also used for the proposed method which shows the best performance in Fig. 4a. Other normalization methods, except MFCC and AFE, are removed in Fig. 4b for simplification. The results of other normalization methods are equivalent to those in Fig. 4a. These results also show better performance than the conventional feature normalization methods when the supplement set has a number of features between 300 and 500.

Table 1 shows the relative error reductions obtained by the proposed methods with their best performance as shown in Figs. 3a and 4a.

Figs. 5a and 5b show the results when the test utterances are limited to 1.5 s in each database.

Table 1 Relative error rate reductions when the proposed methods have the best performance in the YOHO database (Fig. 3a) and the ETRI database (Fig. 4a)

Method	Relative error reduction rate (%)	
	YOHO	ETRI
MFCC	37.84	29.06
CMN	19.64	45.88
MVN	21.98	41.83
CHEQ	23.98	49.98
OHEQ	26.58	45.42
AFE	17.24	21.12

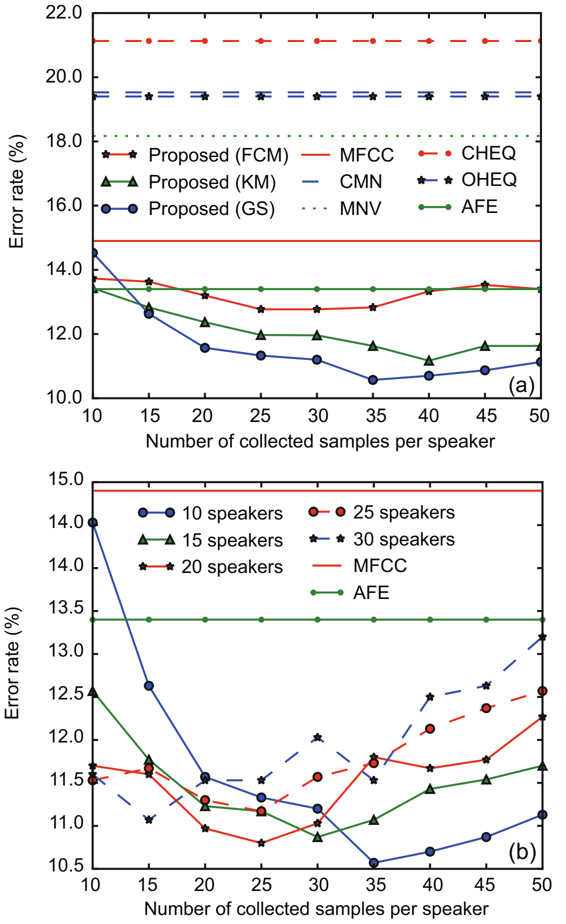

Fig. 4 Speaker identification error rates of the proposed method in the ETRI database: (a) according to the size of the supplement set of 10 speakers; (b) using greedy selection according to the number of speakers

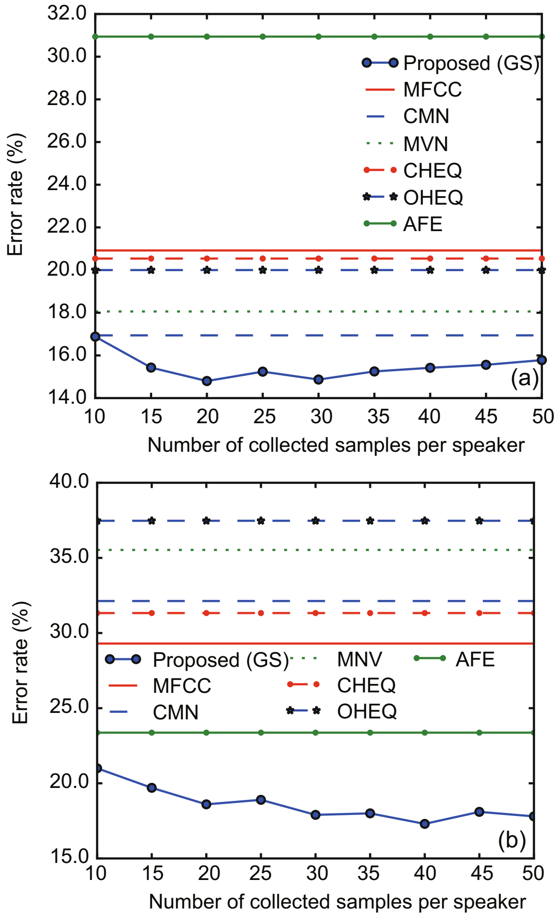

Fig. 5 Speaker identification error rates of the proposed method according to the size of the supplement set with 10 speakers using test utterances with limited duration of 1.5 s in the YOHO database (a) and the ETRI database (b)

A duration of 1.5 s is the shortest duration of the databases. We restrict the duration of test utterances and evaluate the proposed method. In Figs. 5a and 5b, our proposed method using greedy selection has shown the best performances in both databases. In both cases good performances have been obtained when the supplement set has 300–400 samples.

Figs. 6a and 6b show the results of the proposed method with the weight coefficients applied for Eq. (31). We evaluate with three values of weight coefficients (α, β), i.e., (0.75, 0.25), (0.50, 0.50), and (0.25, 0.75). α and β denote the weight coefficients of the supplement set and a test utterance, respectively. Three types of conditions have shown almost identical speaker identification performances when the proposed method using greedy selection is used. However, the difference of coefficients changes the quantity necessary for the supplement set. It shows a slightly better performance when $\alpha = 0.5$ and $\beta = 0.5$.

Now we give an analysis on how the method is affected by the input sample size. Figs. 7a and 7b show the performance plotted against the input sample size to see whether our proposed method breaks down at some sample sizes. We choose the supplement set of 350 samples (10 speakers × 35 samples) which show the best performances in Figs. 3a and 4a. The test utterances are merged and divided into specific durations (2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 s) in the feature space. Our proposed method shows better performances than the other methods when the test utterances have a longer duration than 2.5 s. The results show that the proposed methods will not fail for longer utterances.

Fig. 8 shows the distributions of the first dimension of MFCCs of the supplement set

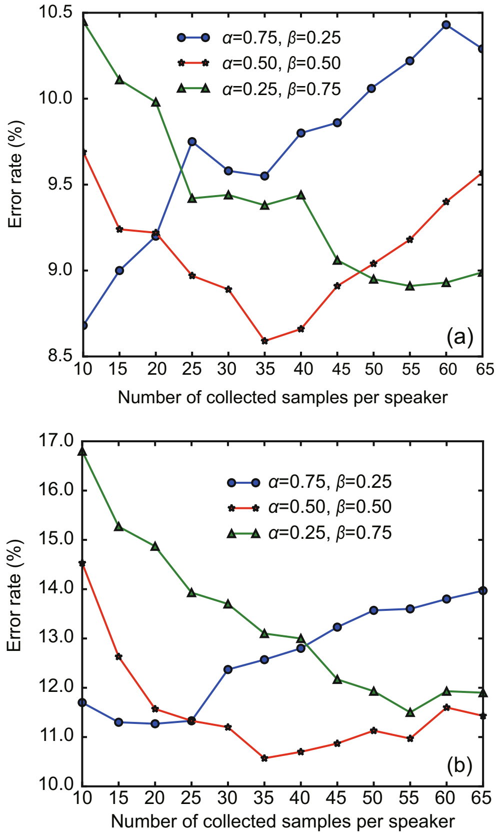

Fig. 6 Speaker identification error rates of the proposed method with the weights applied using greedy selection according to the size of the supplement set of 10 speakers in the YOHO database (a) and the ETRI database (b)

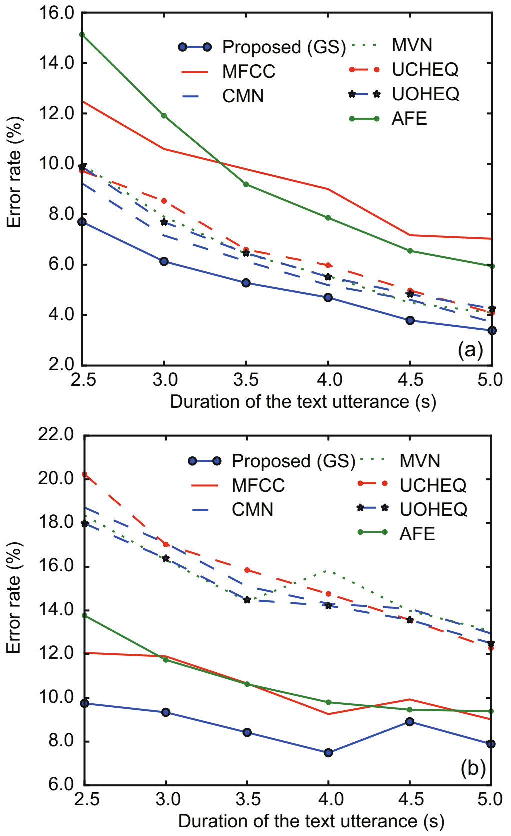

Fig. 7 Speaker identification error rates of the proposed method with the fixed size of the supplement set in the longer utterances of the YOHO database (a) and the ETRI database (b)

consisting of 350 samples (10 speakers × 35 samples). Fig. 8a–8d are the distributions of UBM, the supplement set using greedy selection, the supplement set using K-means, and the supplement set of fuzzy C-means, respectively. The greedy selection and K-means methods properly extract supplement samples according to the density of the distribution of UBM, whereas fuzzy C-means extracts narrowly distributed supplement samples. Greedy selection extracts more widely distributed supplement samples from UBM than K-means. We can interpret the results as indicating that the method explaining the entire distribution of UBM, which represents the entire acoustic feature space, affects the performance of the proposed methods.

5.2 i-vector normalization

Fig. 9 shows the entire system flows using i-vectors, which is the state-of-the-art feature for speaker recognition. The system has online and offline steps. The offline step is composed of rank normalization, length normalization, and PLDA parameter estimation. In the rank normalization step, we perform K-means to collect centroids of the entire background set. In the acoustic feature space, greedy selection shows the best performance. However, the method is limited by data scarcity. Therefore, K-means is performed in the entire background i-vector set, because only one i-vector is extracted from an utterance, whereas many acoustic features are extracted from one utterance. Therefore, instead of greedy selection, K-means is performed in the entire background i-vector set. Then, we use rank normalized i-vectors to estimate parameters of length normalization. Length normalization is performed with estimated parameters and followed by PLDA parameters estimation. In the online step, target and test i-vectors are normalized by rank normalization and length normalization processes consecutively. Then, the i-vector system processes the PLDA scoring and

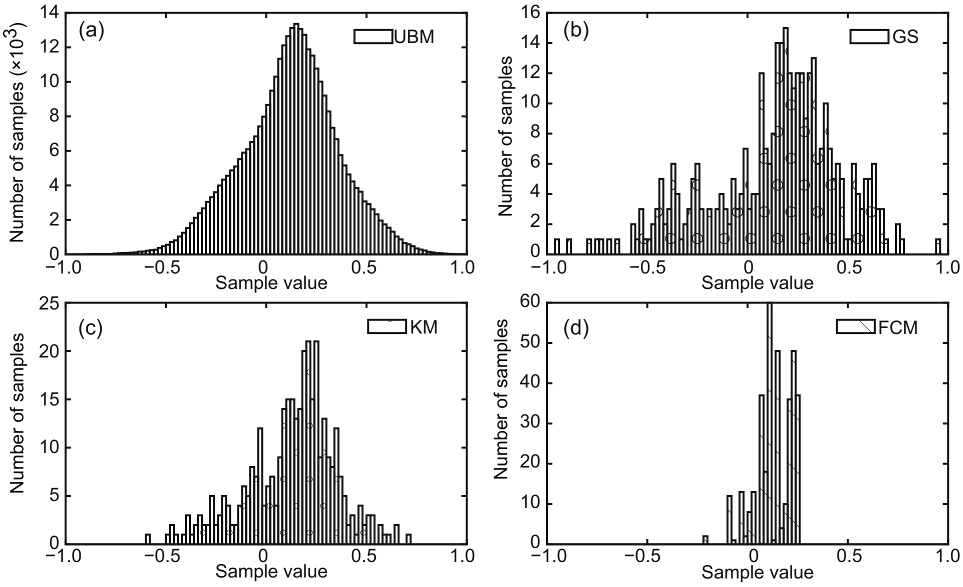

Fig. 8 Distributions of the first dimension of MFCCs of the UBM set (a), the set selected by greedy selection (b), the set generated by *K*-means (c), and the set generated by fuzzy *C*-means (d) from the UBM set which affect the performance of the proposed method

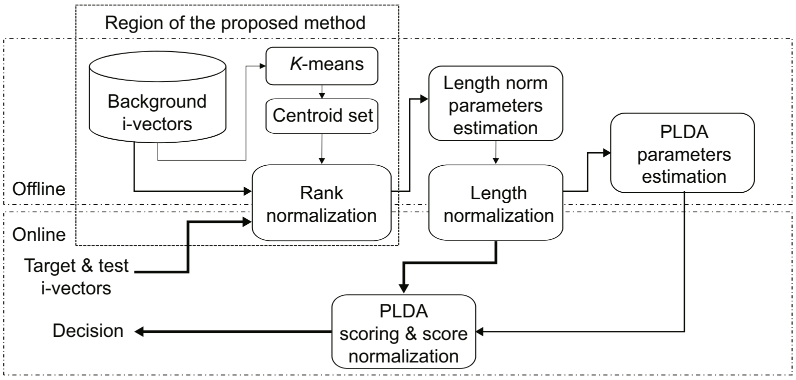

Fig. 9 Speaker verification system flow of the proposed method for i-vector normalization (PLDA: probabilistic linear discriminant analysis)

score normalization steps. Finally, the i-vector system makes a speaker verification decision.

5.2.1 Baseline setup

We use 50-dimensional MFCCs parameters (19 MFCCs, 19Δ, 11$\Delta\Delta$, and ΔE) from a pre-emphasized speech signal every 10 ms using a 25 ms Hamming window. The extracted features are subjected to an energy-based silence removal process and are applied to cepstral mean and variance normalization (MVN). We use a gender-dependent UBM containing 512 Gaussian components. The UBM is trained using the NIST 2004 speaker recognition evaluation (SRE) corpus. We use a gender-dependent total variability matrix and a total variability subspace of 400 dimensions. The total variability matrix is trained using NIST 2005 SRE and Switchboard II phases 1, 2, and 3.

The linear discriminant analysis (LDA) matrix and the probabilistic LDA (PLDA) parameters are trained using the same corpora used in the total variability matrix training. By using a trained 400 × 200 gender-dependent LDA matrix, 400-dimensional raw i-vectors are projected to 200-dimensional i-vectors. A speaker variability subspace of 200 dimensions is used for the PLDA based system.

For evaluation of the proposed method, we use the male part of the telephone conditions (det 6 and det 7) of the short2-short3 condition of NIST 2008 SRE. The i-vectors in the experiments are 200-dimensional LDA-projected i-vectors. We use the PLDA classifier with LDA-projected i-vectors as the baseline. S-norm is used for score normalization (Kenny, 2010). The i-vector platform is based on the ALIZE 3.0 open source platform (Larcher *et al.*, 2013).

5.2.2 Experimental results

We apply length normalization methods such as eigen factor radial (EFR) and spherical nuisance normalization (SphNorm). The parameters of the length normalization methods are estimated using the same corpora as in PLDA training. We perform length normalization with two iterations. Our proposed method is performed with K=500 and K=1000, where K denotes the number of centroids of the K-means.

Table 2 shows the experimental results. 'Baseline' denotes the standard i-vector/LDA/PLDA system. EFR and SphNorm denote the standard i-vector/LDA system, followed by EFR and SphNorm-based PLDA processing, respectively. 'Rank norm' denotes an i-vector performed by LDA followed by rank normalization. 'Proposed' denotes the proposed method instead of rank normalization. We obtain 10.3% and 4.23% improvements in the relative equal error rates (EER) for the det 6 condition compared to the baseline system with EFR and the rank norm with EFR, respectively. We also obtain 17.7% and 0.3% improvements in the relative EER for the det 7 condition compared to the baseline system with EFR and the rank norm with EFR, respectively. In the det 7 condition, we do not obtain significant improvements. However, we obtain performance that is comparable to that of a large background set. The proposed method has a computational advantage when the proposed system estimates the rank of an input i-vector, because the size of the reduced background set is much smaller than the size of the original background set which has 21 391 utterances used in the rank norm. Rank estimation with the large background set of an input i-vector is a time-consuming task. However, our proposed method uses a small background set that can reduce the processing time.

Table 3 shows the processing time for 100 utterances. We obtain 92.6% and 91.7% relative reductions of the processing time with the proposed method with K=500 and K=1000, respectively. The processing time of each method is measured in a single thread with a 2.5 GHz Xeon CPU using Python 2.7.3 on Ubuntu 12.04.

Table 2 Results of the male set evaluated on the dets 6 and 7 conditions of the NIST 2008 SRE short2-short3 condition

Method	EER		DCF	
	det 6	det 7	det 6	det 7
Baseline	5.95%	2.93%	0.0353	0.0180
+ EFR	5.61%	2.79%	0.0329	0.0154
+ SphNorm	5.85%	2.91%	0.0332	0.0158
Rank norm	5.97%	2.95%	0.0351	0.0169
+ EFR	5.25%	2.31%	0.0295	0.0158
+ SphNorm	5.39%	2.31%	0.0301	0.0157
Proposed (K=500)	5.50%	2.91%	0.0326	0.0161
+ EFR	5.14%	2.71%	0.0279	**0.0151**
+ SphNorm	5.11%	**2.30**%	0.0282	0.0154
Proposed (K=1000)	5.68%	2.75%	0.0329	0.0159
+ EFR	**5.03**%	2.55%	**0.0276**	0.0154
+ SphNorm	5.13%	2.48%	0.0281	0.0158

The numbers in bold represent the best performance achieved by the combining methods

Table 3 Processing time of 100 utterances of the rank normalization and the proposed method

Method	Processing time (s)
Rank norm	8.378
Proposed (K=500)	0.621
Proposed (K=1000)	0.691

6 Conclusions

We have proposed a novel approach to HEQ using a greedy selection algorithm. The greedy selection algorithm is generally used to select a set of samples that can represent the entire set with a reduced number of samples. In the proposed method, we used the samples obtained from the UBM training set by applying the greedy selection algorithm to estimate a robust cumulative distribution function. The experimental results demonstrate the variation in the system performance according to the number of samples selected and the number of speakers. The system performance was evaluated in the Opus codec environment. The results showed that the system performance was improved when the proposed method was used with the GS, KM, and FCM algorithms, as compared to the conventional feature normalization methods and AFE. Good performance was achieved with the proposed method when the supplement set consisted of 10 speakers and had 20–40 sample vectors. The number of sample feature vectors of the test utterances ranged from 150 to 250. We observed similar performance when the number of speakers was increased from 10 to 30. We could improve the system with state-of-the-art i-vector features by using the proposed method.

The method we propose suffers from two inherent limitations. On the one hand, it adopts a 'bag-of-frames' representation while equalizing a speech utterance. On the other hand, it is not feasible for real-time or online processing because of the 'look ahead' that would be required. The former is an inherent limitation of speaker recognition using Gaussian mixture models, because the models do not consider the temporal relationships of the frames. The latter can be overcome by using short windows of frames that are long enough to take advantage of the proposed methods, although the response time is not critical because our goal is to focus on short utterances that take short processing time.

In further work, we plan to use a few well-practiced nonparametric density estimation algorithms that can also be used to facilitate accurate estimation of feature distributions. We also plan to find a method to determine the optimal size of the supplement set to decrease the recognition time, and find a correlation with the system performance and the distribution of the supplement set.

References

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Am.*, **55**(6):1304-1312. http://dx.doi.org/10.1121/1.1914702

Blanco, Y., Zazo, S., Principe, J.C., 2000. Alternative statistical Gaussianity measure using the cumulative density function. Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation, p.537-542.

Bousquet, P., Matrouf, D., Bonastre, J., 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. INTERSPEECH, p.485-488.

Bousquet, P., Larcher, A., Matrouf, D., *et al.*, 2012. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. Odyssey: the Speaker and Language Recognition Workshop, p.157-164.

Cannon, R.L., Dave, J.V., Bezdek, J.C., 1986. Efficient implementation of the fuzzy *c*-means clustering algorithms. *IEEE Trans. Patt. Anal. Mach. Intell.*, **8**(2):248-255. http://dx.doi.org/10.1109/TPAMI.1986.4767778

Dehak, N., Kenny, P., Dehak, R., *et al.*, 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.*, **19**(4):788-798. http://dx.doi.org/10.1109/TASL.2010.2064307

de la Torre, A., Peinado, A.M., Segura, J.C., *et al.*, 2005. Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.*, **13**(3):355-366. http://dx.doi.org/10.1109/TSA.2005.845805

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons, Tronto.

ETSI, 2005. Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Extended Advanced Front-End Feature Extraction Algorithm; Compression Algorithms; Back-End Speech Reconstruction Algorithm, ETSI ES 202 212. European Telecommunication Standards Institute, Sophia Antipolis.

Franc, V., 2005. Optimization Algorithms for Kernel Methods. PhD Thesis, Centre for Machine Perception, Czech Technical University, Prague, Czech Republic.

Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. INTERSPEECH, p.249-252.

Gonzalez, R.C., Wintz, P., 1987. Digital Image Processing. Addision-Wesley Publishing Company, Boston.

Jiang, Y., Lee, K., Tang, Z., *et al.*, 2012. PLDA modeling in i-vector and supervector space for speaker verification. INTERSPEECH, p.1680-1683.

Jones, E., Oliphant, T., Peterson, P., 2001. Scipy: Open Source Scientific Tools for Python. http://www.scipy.org/

Kenny, P., 2010. Bayesian speaker verification with heavy-tailed priors. Odyssey: the Speaker and Language Recognition Workshop.

Kim, M., Yang, I., Yu, H., 2008. Robust speaker identification using greedy kernel PCA. 20th IEEE Int. Conf. on Tools with Artificial Intelligence, p.143-146. http://dx.doi.org/10.1109/ICTAI.2008.105

Kim, N., 1998. Statistical linear approximation for environment compensation. *IEEE Signal Process. Lett.*, **5**(1):8-10. http://dx.doi.org/10.1109/97.654866

Larcher, A., Bonastre, J., Fauve, B., *et al.*, 2013. Alize 3.0—open source toolkit for state-of-the-art speaker recognition. INTERSPEECH, p.2768-2772.

Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.733-736. http://dx.doi.org/10.1109/ICASSP.1996.543225

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. Odyssey: the Speaker and Language Recognition Workshop, p.213-218.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. *Dig. Signal Process.*, **10**(1):19-41. http://dx.doi.org/10.1006/dspr.1999.0361

Segura, J.C., Benítez, C., de la Torre, A., *et al.*, 2004. Cepstral domain segmental nonlinear feature transformations for robust speech recognition. *IEEE Signal Process. Lett.*, **11**(5):517-520. http://dx.doi.org/10.1109/LSP.2004.826648

Skosan, M., Mashao, D., 2006. Modified segmental histogram equalization for robust speaker verification. *Patt. Recog. Lett.*, **27**(5):479-486. http://dx.doi.org/10.1016/j.patrec.2005.09.009

Stolcke, A., Kajarekar, S., Ferrer, L., 2008. Nonparametric feature normalization for SVM-based speaker verification. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.1577-1580.

Valin, J.M., Vos, K., Terriberry, T., 2012. Definition of the Opus Audio Codec. http://opus-codec.org/

Viikki, O., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.*, **25**(1):133-147. http://dx.doi.org/10.1016/S0167-6393(98)00033-8