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

Latent discriminative representation learning for speaker recognition*

Duolin HUANG[1], Qirong MAO[‡1,2], Zhongchen MA[1], Zhishen ZHENG[1],
Sidheswar ROUTRYAR[1], Elias-Nii-Noi OCQUAYE[1]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*
[2]*Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Zhenjiang 212013, China*
E-mail: 2211708034@stmail.ujs.edu.cn; mao_qr@ujs.edu.cn; zhongchen_ma@ujs.edu.cn;
1209103822@qq.com; sidheswar69@gmail.com; eocquaye@ujs.edu.cn

Received Dec. 10, 2019; Revision accepted July 12, 2020; Crosschecked Nov. 18, 2020; Published online Jan. 29, 2021

Abstract: Extracting discriminative speaker-specific representations from speech signals and transforming them into fixed length vectors are key steps in speaker identification and verification systems. In this study, we propose a latent discriminative representation learning method for speaker recognition. We mean that the learned representations in this study are not only discriminative but also relevant. Specifically, we introduce an additional speaker embedded lookup table to explore the relevance between different utterances from the same speaker. Moreover, a reconstruction constraint intended to learn a linear mapping matrix is introduced to make representation discriminative. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods based on the Apollo dataset used in the Fearless Steps Challenge in INTERSPEECH2019 and the TIMIT dataset.

Key words: Speaker recognition; Latent discriminative representation learning; Speaker embedding lookup table; Linear mapping matrix
https://doi.org/10.1631/FITEE.1900690 **CLC number:** TP391.4

1 Introduction

Speaker recognition (SR) is one of the most widely used biometric recognition technologies, and provides unique advantages in remote authentication. It has been extensively implemented in daily life tasks. Speaker-specific representations play essential roles in many speech recognition applications, such as speaker identification, verification, and clustering. By comparing speaker representations, a system can recognize the identity of a speaker and verify whether the current speaker matches an enrolled target speaker or not (Togneri and Pullella, 2011). Depending on the restrictions on utterances, we can classify SR models into two categories, text-dependent and text-independent (Fisusi and Yesufu, 2007). When the transcript of the utterances is lexically constrained, a task is considered text-dependent; otherwise, it is considered text-independent.

Since the 1980s, numerous SR methods have been proposed and achieved state-of-the-art results. Starting with conventional methods and progressing toward current deep learning ones, SR methods can be categorized into three groups: non-parametric models, parametric models, and artificial neural networks. The non-parametric models can be modeled based on features of a speaker using particular operations. Conventional non-parametric models

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. U1836220 and 61672267), the Qing Lan Talent Program of Jiangsu Province, China, and the Jiangsu Province Key Research and Development Plan (Industry Foresight and Key Core Technology) (No. BE2020036)

ORCID: Duolin HUANG, https://orcid.org/0000-0002-3149-2605; Qirong MAO, https://orcid.org/0000-0002-0616-4431

include dynamic time warping (DTW) (Yu et al., 1995; Dey et al., 2017) and vector quantization (VQ) (Singh and Rajan, 2011). In DTW, an identification template and a reference template are compared, and then similarities between these two templates are found. The VQ method extracts multi-dimensional time-series vectors from a speech signal and establishes a speaker model by selecting particular representative vectors.

The parametric models include the hidden Markov model (HMM) (Rabiner, 1989) and Gaussian mixture model (GMM) (Reynolds and Rose, 1995). GMM is the superposition of several Gaussian functions, and the linear combination can simulate the continuous probability distribution of the speaker vector feature, that is, describe the speaker's characteristics. Furthermore, a GMM universal background model (GMM-UBM) (Reynolds et al., 2000) and a support vector machine (Wan and Campbell, 2000) were proposed to describe a target speaker. GMM-UBM aims to avoid overfitting caused by insufficient training data. As a result of introducing join factor analysis (Kenny et al., 2007) and identity vector (i-vector) (Dehak et al., 2011) modeling methods, the cooperation between i-vector and probabilistic linear discriminant analysis (PLDA) (van Leeuwen and Saeidi, 2013) has shown excellent performance, which has remained unsurpassed for a long time.

Artificial neural networks have proved their applicability to SR with the success of deep learning in representation learning. Many researchers have developed deep neural architecture (DNA) methods to generate speaker-specific representations (Variani et al., 2014; Heigold et al., 2016; Li et al., 2017). Deep learning frameworks are deemed powerful tools for complex data analysis, and many studies have proposed training deep nonlinear extractors as solutions to SR. These frameworks have demonstrated perspective results for text-dependent and text-independent SR tasks.

Several studies based on deep neural networks (DNNs) suggested replacing GMM with an i-vector framework to obtain statistics (Lei et al., 2014), until Google d-vector was proposed (Variani et al., 2014). The method based on d-vector is the first SR system entirely based on a DNN framework. Furthermore, an end-to-end SR system based on triplet loss subsequently emerged, and has been extensively studied (Li et al., 2017; Zhang C and Koishida, 2017). Besides the loss function, various speaker embedding models have been introduced, including the d-vector model for text-dependent tasks (Variani et al., 2014) and the x-vector system for text-independent tasks (Snyder et al., 2016, 2018).

In SR, learning discriminative features based on an original input is one of the most important research issues, including facial expression recognition (Zhang FF et al., 2018), speech recognition (Luo et al., 2018), speech emotion recognition (Mao et al., 2014), and visual recognition of zero-shot learning (Chen XB et al., 2015; Jiang et al., 2017). In recent years, many feature learning methods employed in other recognition tasks have been applied in SR. For example, discriminative methods used in face recognition (Schroff et al., 2015; Wen et al., 2016) have been successfully applied to SR tasks (Li et al., 2017; Yadav and Rai, 2018). The basic idea implemented in the above mentioned studies is that in a latent feature space, the distance between samples from the same class is reduced or limited, while that between samples from different classes is increased.

Inspired by the idea mentioned above, in this study, we propose a latent discriminative representation learning (LDRL) method for SR. In this method, we first use a dictionary learning framework to model an original latent representation space, so that the space can be constructed by dictionary items. Then, a speaker embedding lookup table is constructed to learn the relevance of samples from the same speaker. Furthermore, to make latent representations discriminative, we introduce a reconstruction constraint to learn a linear mapping as an SR classifier. This constraint maps training samples to the corresponding speaker labels. Finally, the latent representations learned by LDRL are used to maintain the correlation between samples from the same speaker and provide sufficient discriminative capability. Experimental results based on the Apollo dataset (Hansen et al., 2018) used in the Fearless Steps Challenge in INTERSPEECH2019 and the TIMIT dataset (Garofolo et al., 1993) demonstrate that the proposed approach can achieve stable and robust recognition performance in complex scenes (for example, with speaker variation and noise).

The major contributions of this paper can be formulated as follows:

1. Using a novel objective function, we can learn latent representations that are relevant and discriminative for SR. Specifically, LDRL from supervised learning is divided into three blocks: basic latent representation learning, latent relevance learning, and latent discriminative learning. The latent representations are discriminative and robust, providing significant performance improvement of SR.

2. The proposed embedding lookup table can be used to learn correlations between different utterances of the same speaker. Using this lookup table, correlations between utterances from the same speaker can be preserved, enabling latent representations clustered in the latent representation space.

3. The proposed LDRL method achieves state-of-the-art SR performance on the TIMIT dataset, and performs better than the baselines of the Fearless Steps Challenge in INTERSPEECH2019 based on the Apollo dataset.

2 Related works

A typical SR system consists of two main components: a front-end processing unit which extracts appropriate features from speech data and a classifier which identifies the target speaker of a speech utterance. In this section, we briefly review the methods widely used in SR.

2.1 Feature extraction for speaker recognition

A speech signal comprises multiple features, including the important and unimportant ones. In the past few decades, many speaker-specific features have been introduced, including conventional manual features and the current deep learning ones, such as Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980), linear predictive cepstral coefficients (LPCCs) (Huang et al., 2001), line spectral frequencies (LSFs) (Huang et al., 2001), and perceptual linear prediction (PLP) coefficients (Hermansky, 1990). In the early 1980s, MFCCs were first introduced into speech recognition and then applied to SR. They were computed using a group of filter banks followed by a logarithmic compression and the discrete cosine transform. LPCCs, LSFs, and PLP use predictor coefficients to transform themselves into robust and less correlated features. In recent years, with the development of deep learning in various fields, d-vector (Variani et al., 2014), x-vector (Snyder et al., 2017), and j-vector (Chen NX et al., 2015) extracted from speech signals using DNNs have been introduced into SR. Although these studies have investigated the problem of feature learning in SR using various techniques, their focus has been emphasized mainly on learning discriminative features using input data. However, it is still a challenging task to identify an appropriate mode to learn discriminative and relevant features in SR. In this study, we introduce an LDRL method based on a dictionary learning framework. Using a novel object function, it allows to integrate latent relevance learning and latent discriminative learning.

2.2 Classifiers for speaker recognition

SR refers to recognizing people from their voices. Two individuals cannot have the same voice because their vocal tract shape, throat size, and other vocal organs are different. Besides these physical differences, everyone has a unique way to speak, including specific accents, rhythms, intonation styles, pronunciation patterns, and vocabulary choices. State-of-the-art SR systems rely on the aforementioned features in parallel, attempting to cover these unique aspects and use them to achieve accurate recognition. Various methods have been applied for SR, including VQ (Singh and Rajan, 2011), HMM (Rabiner, 1989), GMM (Reynolds and Rose, 1995; Kim et al., 2017), DTW (Yu et al., 1995; Dey et al., 2017), i-vector/PLDA (van Leeuwen and Saeidi, 2013), and DNN (Variani et al., 2014; Heigold et al., 2016; Li et al., 2017). However, experimental results indicated that each method has its own advantages and limitations. To combine the merits of different methods, an aggregation method has been proposed (Desai and Joshi, 2013).

3 Latent relevant and discriminative representation learning for speaker recognition

The architecture of the proposed method is illustrated in Fig. 1. It has three stages: (1) latent representation by dictionary learning; (2) latent relevance learning; (3) latent discriminative learning. In this section, we will describe each learning stage in detail.

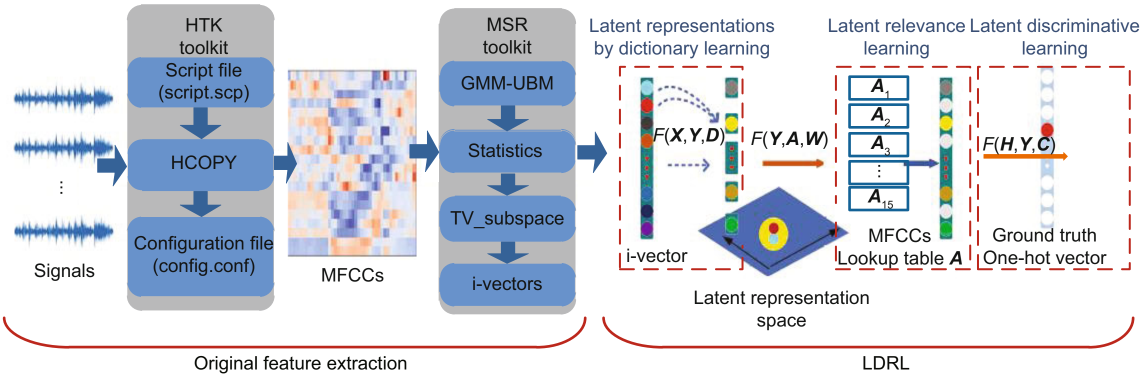

Fig. 1 Architecture of the latent discriminative representation learning method for speaker recognition based on dictionary learning

After extracting the original features, the implementation of the proposed approach is framed using a red rectangle. Three steps of the implementation method are denoted by three red rectangles. The first rectangle refers to the input of the i-vector to learn latent representations through dictionary learning. The second one represents the latent representations of different utterances of the same speaker keeping relevance in the lookup table. The final one means using the reconstruction constraint to learn a linear mapping matrix to make latent representations to classify different speakers, making the latent representations discriminative. References to color refer to the online version of this figure

3.1 Latent representation by dictionary learning

One of the key issues associated with SR tasks in the proposed approach is to find a latent representation space that can be used to specify relationships between utterances. The latent representation space is constructed based on a dictionary learning framework. The problem can be formulated as follows:

$$l_1\{\boldsymbol{D},\boldsymbol{Y}\} = \underset{\boldsymbol{D},\boldsymbol{Y}}{\arg\min}\,\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{Y}\|_{\mathrm{F}}^2$$
$$\text{s.t.}\quad \|\boldsymbol{d}_i\|_2^2 \le 1\ \ \forall_i, \tag{1}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $\boldsymbol{d}_i$ is the i^{th} column of the learned dictionary $\boldsymbol{D}\in\mathbb{R}^{m\times h}$. $\boldsymbol{X}\in\mathbb{R}^{m\times n}$ represents the the set of original input features (i-vectors) of n labeled training samples, $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]$ (where $\boldsymbol{x}_i$ represents the i^{th} i-vector), and m represents the dimension of the i-vector. $\boldsymbol{Y}\in\mathbb{R}^{h\times n}$ is the reconstruction coefficient that corresponds to the target latent representation extracted from the utterance of a speaker.

3.2 Latent relevance learning among utterances of the same speaker

Although $\boldsymbol{Y}$ can be viewed as latent representations in a recognition task, there are still several questions that need to be clarified. First, there are correlations among the latent representations corresponding to the same speaker. Therefore, it is deemed unsuitable to learn representations of each utterance independently. To address this problem, a linear transformation matrix $\boldsymbol{W}$ is used to define the relationship between a lookup table and latent representations, providing relevant information for different latent representations corresponding to the same speaker. The latent relevance learning is defined as follows:

$$l_2\{\boldsymbol{Y},\boldsymbol{W}\} = \underset{\boldsymbol{Y},\boldsymbol{W}}{\arg\min}\,\|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{A}\|_{\mathrm{F}}^2$$
$$\text{s.t.}\quad \|\boldsymbol{w}_i\|_2^2 \le 1\ \ \forall_i, \tag{2}$$

where $\boldsymbol{w}_i$ is the i^{th} column of $\boldsymbol{W}\in\mathbb{R}^{h\times k}$. It can be inferred from Eq. (2) that the latent representations can be represented as the linear combination of a lookup table. $\boldsymbol{A}\in\mathbb{R}^{k\times n}$ denotes the speaker embedding lookup table composed of MFCCs, where k denotes the dimension of the MFCCs. Speaker embedding is identical for different latent representations for the same speaker category, and the construction of lookup table $\boldsymbol{A}$ will be described in the original feature extraction section (i.e., Section 4.2). $\boldsymbol{W}\boldsymbol{A}$ is encouraged to be similar to latent representation $\boldsymbol{Y}$ in Eq. (2), thereby ensuring that the learned bases depict dictionary items. $\boldsymbol{Y}$ can be viewed as a linear combination of speaker embeddings that implicitly combine the strongly correlated ones. The lookup table is aimed at guiding the latent relevance learning of different latent representations corresponding to the same speaker. This is similar to building a cluster point for each speaker in a latent representation space, which can together pull different utterances from the same speaker. In this way, relevant information of the same speaker can be passed between different latent representations.

3.3 Latent discriminative learning among different speakers

Eq. (1) is to ensure that sparse coding for a corresponding utterance provides the minimum reconstruction error. However, it does not guarantee the best classification performance. Utterances share some common information between different speakers, such as voice and linguistic information in human speech. To make a recognition task effective, the latent representations must be discriminative. In this case, we use a speaker classifier to make latent representations discriminative. Specifically, linear mapping $\boldsymbol{C}$ is learned by latent representations for final speaker categories:

$$\begin{aligned} &l_3\{\boldsymbol{C},\boldsymbol{Y}\} = \underset{\boldsymbol{C},\boldsymbol{Y}}{\arg\min}\,\|\boldsymbol{H}-\boldsymbol{C}\boldsymbol{Y}\|_{\mathrm{F}}^2 \\ &\text{s.t.}\quad \|\boldsymbol{c}_i\|_2^2 \le 1\ \ \forall_i, \end{aligned} \tag{3}$$

where $\boldsymbol{c}_i$ is the i^{th} column of $\boldsymbol{C} \in \mathbb{R}^{s\times h}$. $\boldsymbol{C}$ can be considered a speaker classifier in a latent representation space, and $\boldsymbol{H} \in \mathbb{R}^{s\times n}$ represents the speaker labels of the samples, where $\boldsymbol{h}_i = [0,...,0,1,0,...,0]^{\mathrm{T}}$ (i.e., i^{th} column of $\boldsymbol{H}$) is a one-hot vector with dimension s. Dimension s represents the total number of speaker categories. Eq. (3) is aimed to make latent representations sufficiently discriminative to classify different speakers. It implicitly pulls together utterances from the same speaker and pushes them away from different speakers. In summary, the overall objective function of SR is defined as follows:

$$l = l_1 + \alpha l_2 + \beta l_3, \tag{4}$$

where α and β are super parameters. By adjusting α and β, the constraint strength can be modified.

The latent representations learned by the proposed method are not only discriminative, but also relevant. First, in Eq. (2), the lookup table and latent representations are connected using a linear transformation matrix $\boldsymbol{W}$. Using matrix $\boldsymbol{W}$, we can recover the speaker embeddings by the latent features from the same speaker. Thus, latent representations share relevant information. Second, linear mapping $\boldsymbol{C}$ in a latent representation space can be viewed as a speaker classifier, and is used to map latent representations to the corresponding speaker category. This constraint enables latent representations to be discriminative.

3.4 Optimization

An obvious challenge is represented in Eq. (4), which is not simultaneously convex for $\boldsymbol{D}$, $\boldsymbol{Y}$, $\boldsymbol{W}$, and $\boldsymbol{C}$. An alternative optimization method can optimize the overall objective function as follows:

1. Update latent representation $\boldsymbol{Y}$ by fixing $\boldsymbol{D}$, $\boldsymbol{W}$, and $\boldsymbol{C}$. The subproblem can be formulated as follows:

$$\underset{\boldsymbol{Y}}{\arg\min}\,\|\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{D}}\boldsymbol{Y}\|_{\mathrm{F}}^2, \tag{5}$$

where

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \alpha\boldsymbol{W}\boldsymbol{A} \\ \beta\boldsymbol{H} \end{bmatrix},\quad \tilde{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{D} \\ \alpha\boldsymbol{I} \\ \beta\boldsymbol{C} \end{bmatrix}, \tag{6}$$

and $\boldsymbol{I}$ is the identity matrix. Let the derivative of Eq. (5) be 0. We can obtain the closed solution of $\boldsymbol{Y}$ as

$$\boldsymbol{Y} = \left(\tilde{\boldsymbol{D}}^{\mathrm{T}}\tilde{\boldsymbol{D}}\right)^{-1}\tilde{\boldsymbol{D}}^{\mathrm{T}}\tilde{\boldsymbol{X}}. \tag{7}$$

2. Update latent representation dictionary $\boldsymbol{D}$ in Eq. (1) by fixing $\boldsymbol{Y}$, $\boldsymbol{W}$, and $\boldsymbol{C}$. It can be optimized by the Lagrangian dual. Therefore, the analytical solution to $\boldsymbol{D}$ can be formulated as follows:

$$\boldsymbol{D} = \left(\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}\right)\left(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} + \boldsymbol{\Lambda}\right)^{-1}, \tag{8}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix that comprises Lagrangian dual variables.

3. Update matrix $\boldsymbol{W}$ in Eq. (2) by fixing $\boldsymbol{D}$, $\boldsymbol{Y}$, and $\boldsymbol{C}$. It can also be optimized in the same way as defined in Eq. (1). Thus, the analytical solution to $\boldsymbol{W}$ is defined as follows:

$$\boldsymbol{W} = \left(\boldsymbol{Y}\boldsymbol{A}^{\mathrm{T}}\right)\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T} + \boldsymbol{\Lambda}\right)^{-1}. \tag{9}$$

4. Update linear mapping $\boldsymbol{C}$ in Eq. (3) by fixing $\boldsymbol{D}$, $\boldsymbol{Y}$, and $\boldsymbol{W}$. It can also be optimized in the same way as defined in Eq. (1). The analytical solution to $\boldsymbol{C}$ is formulated as follows:

$$\boldsymbol{C} = \left(\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}}\right)\left(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} + \boldsymbol{\Lambda}\right)^{-1}. \tag{10}$$

3.5 Recognition with latent discriminative representations

As the latent discriminative representation is relevant and discriminative, we can perform the recognition task in a latent representation space. Given a test sample with its input feature $\boldsymbol{x}^{\mathrm{t}}$, we can calculate its latent discriminative representation $\boldsymbol{y}^{\mathrm{t}}$ using the following formula:

$$\boldsymbol{y}^{\mathrm{t}} = \min_{\boldsymbol{y}^{\mathrm{t}}} \|\boldsymbol{x}^{\mathrm{t}} - \boldsymbol{D}\boldsymbol{y}^{\mathrm{t}}\| + \lambda \|\boldsymbol{y}^{\mathrm{t}}\|_2^2, \tag{11}$$

where λ is the weight of the regularization term. In this study, we concatenate different vector representations of an utterance to form the final representation. Then, we apply the nearest neighbor algorithm to perform the SR task using the cosine distance as follows:

$$\mathrm{Score}(\boldsymbol{x}^{\mathrm{t}}) = \begin{cases} \cos < \begin{bmatrix} \boldsymbol{y}^{\mathrm{t}} \\ \boldsymbol{C}\boldsymbol{y}^{\mathrm{t}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{WA} \\ \boldsymbol{H} \end{bmatrix} >, & \boldsymbol{W} \neq \boldsymbol{0},\ \boldsymbol{C} \neq \boldsymbol{0}, \\ \cos < \begin{bmatrix} \boldsymbol{C}\boldsymbol{y}^{\mathrm{t}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{H} \end{bmatrix} >, & \boldsymbol{W} = \boldsymbol{0}, \\ \cos < \begin{bmatrix} \boldsymbol{y}^{\mathrm{t}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{WA} \end{bmatrix} >, & \boldsymbol{C} = \boldsymbol{0}, \end{cases} \tag{12}$$

where Score$(\cdot)$ denotes the score for the test sample and all speaker categories. The speaker with the highest score is the target speaker. $\boldsymbol{W} = \boldsymbol{0}$ denotes removing the latent relevance learning in LDR. $\boldsymbol{C} = \boldsymbol{0}$ denotes removing the latent relevance learning (LRL). However, LDRL contains both of them.

4 Experiments

4.1 Datasets

To investigate the performance of the proposed approach, we conducted experiments based on the TIMIT and Apollo datasets.

1. TIMIT

The TIMIT dataset contains the broadband recordings corresponding to 630 speakers of eight major dialect regions (DRs) of American English, and each reading is composed of 10 phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions and a 16-bit, 16-kHz speech waveform file for each utterance. Details are provided in Table 1.

2. Apollo

The Apollo dataset was supplied by the organizer of the Fearless Steps Challenge in INTERSPEECH2019. It comprises 19 000 h of naturalistic multi-channel data. However, the organizer disclosed only more than 10 000 speech files, including the development set and test set. Details are presented in Table 2. Data in this dataset is characterized by multiple classes of noise, degradation, and overlapping instances over most channels. Most audio channels are degraded due to high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, and tag aging. These pose great challenges, and the performance of the SR system depends on the speech length. Contiguous speech by a single speaker with the length of 0.4 to 50 s is provided in this dataset, and a significant portion of short utterances exists in the corpus.

4.2 Original feature extraction

In the experiments, speech signals were sampled at a rate of 16 kHz and divided by a 25-ms Hamming window with a 10-ms shift. An HMM toolkit (HTK) (Young, 1993) was used to extract the MFCC feature normalized by the cepstrum mean. The Microsoft Research (MSR) identity toolbox (Sadjadi et al., 2013) was used to develop a GMM-UBM system and extract the i-vector as the input into the proposed model. Details are presented in Fig. 1.

After extracting the MFCC features, we applied two techniques. First, the MFCC features were used to construct a speaker embedding lookup table. In the experiments, we randomly selected an exemplar from multiple speeches of a speaker, accumulated

Table 1 Dialect distribution of speakers

DR	Male (%)	Female (%)	Total (%)
1	31 (63.3)	18 (36.7)	49 (7.8)
2	71 (69.6)	31 (30.4)	102 (16.2)
3	79 (77.5)	23 (22.5)	102 (16.2)
4	69 (69.0)	31 (31.0)	100 (15.9)
5	62 (63.3)	36 (36.7)	98 (15.5)
6	30 (65.2)	16 (34.8)	46 (7.3)
7	74 (74.0)	26 (26.0)	100 (15.9)
8	22 (66.7)	11 (33.3)	33 (5.2)
	438 (69.5)	192 (30.5)	630 (100)

Table 2 Statistics of the development set (Dev) and evaluation set (Eval) on the Apollo dataset

Set	Number of speakers	Average duration per utterance (s)	Number of utterances
Dev	183	5.35	8394
Eval	183	4.69	3600

the MFCC features over all frames, and averaged and normalized them. The processed feature was used to represent a particular speaker in a lookup table. This procedure was inspired by the process of extracting d-vectors. Second, the total variability matrix and UBM were used to model i-vector based on the MFCC features. The dimension of the extracted i-vector was fixed between 100 and 400 with an interval of 100. Thereafter, the i-vector was used as the input feature in LDRL. We conducted several experiments by altering the i-vector dimension and GMM mixture size to find the optimal condition of the proposed model. At the UBM training step, different GMM mixture sizes (16, 32, 64, 128, 256, 512, and 1024) were used.

4.3 Baselines

The LDRL method was compared with the following baseline approaches:

1. GMM-UBM

Following the conventional GMM-UBM framework, a single speaker-independent universal background model (UBM) was constructed with the mixture sizes of 256 and 1024. It performed training during 10 iterations based on the training data.

2. i-vector/PLDA

Gender-independent i-vector extractors were trained based on the TIMIT dataset. The probabilistic LDA (PLDA) (Cumani et al., 2013) scoring was then applied to perform recognition. As the i-vector contained both speaker and channel information, we focused only on the former. Therefore, implementing a channel compensation algorithm called PLDA was required to enable channel compensation. To compare the results reported by Yoshimura et al. (2018), the i-vector was extracted based on the total variability model with rank 200, and then modeled using PLDA comprising 100 dimensions of the speaker subspace.

4.4 Model ablation

Model ablation was executed on the TIMIT dataset and was not performed on the Apollo dataset due to the elapsed deadline of the Fearless Steps Challenge in INTERSPEECH2019. To obtain the best parameter set, we performed several experiments using different i-vector dimensions and mixture sizes. Data division was the same as in Al-Kaltakchi et al. (2016). Table 3 presents the speaker recognition accuracy (SRA) based on different GMM mixture sizes and i-vector dimensions realized in the proposed method.

Table 3 Speaker recognition accuracy (SRA) based on different mixture sizes and i-vector dimensions for 120 speakers

Mixture size	SRA (%)			
	i-vector dimension			
	100	200	300	400
16	91.67	90.21	85.00	83.54
32	97.50	96.04	91.00	78.75
64	98.54	98.33	98.13	92.08
128	99.38	98.33	98.54	97.08
256	99.79	99.58	98.96	98.13
512	100	99.58	99.17	98.54
1024	100	99.58	99.17	97.50

It was observed that SRA augments with an increase in the mixture size corresponding to the same i-vector dimension when the i-vector dimension is smaller than 400. However, at the same mixture size, SRA decreases with an increase in the i-vector dimension. Experimental results demonstrated that in the proposed method, the latent representation obtained using the low-dimensional i-vector has greater discriminative power in the recognition tasks. As shown in Table 3, when the i-vector dimension is 100 and mixture size larger than 512, accuracy could reach 100%.

The proposed method is associated with latent relevance learning and latent discriminative learning, which makes the latent representations relevant and discriminative. To verify the effectiveness of each component, as described in Section 3.5, we compared three different approaches. The results on the TIMIT dataset with two data divisions are provided in Table 4: (1) recognition with latent discriminative learning (LDL) by removing the second term in Eq. (4); (2) recognition with LRL by removing the third term in Eq. (4); (3) recognition with all constraint terms in Eq. (4), i.e., LDRL. To verify how different feature extraction methods affect the performance of the proposed method, we compared the input i-vector with the x-vector features.

Comparing the performances of LDL, LRL, and LDRL, we observed that the latent representations are deemed the most applicable to SR tasks. Moreover, by imposing the discriminative constraint on the basis of latent relevance learning in Eq. (4), the recognition accuracy was improved, as shown through comparing LRL with LDRL in Table 4. In addition, adding latent relevant learning to LDL has indeed improved the speaker recognition performance. Analyzing the experimental results of the inputs x-vector and i-vector, we could see that the effect of i-vector was better. The reason could be that x-vector is based on the features extracted based on the discriminative loss obtained through training a deep learning model. Therefore, compared with i-vector, there is a lack of relevant information between different x-vectors from the same speaker. As shown in Table 5, experiments based on the TIMIT and Apollo datasets were conducted using Ubuntu 16.04 LST with Matlab R2014b on NVIDIA TITAN X GPU. A simple grid search method was used to select the values of α, β, and λ in [0, 1] with the i-vector dimension of 100 and mixture size of 512.

4.5 LDRL performance evaluation

To evaluate the LDRL in terms of the classification accuracy, we compared it with several other well-established SR methods: GMM-UBM, i-vector/PLDA, extreme learning machine (EML) (Al-Kaltakchi et al., 2017), and variational autoencoder (VAE) (Yoshimura et al., 2018). We considered the deep learning methods called convolutional long short-term memory (CLSTM) (Kumar et al., 2018) and unsupervised adversarial invariance (UAI) (Peri et al., 2019). Specifically, EML was composed of an i-vector by three fusion methods using an ELM classifier. VAE was defined as the extended version of the variational autoencoder for sequence modeling. The proposed approach could directly process variable-length observation sequences. UAI employed an unsupervised adversarial invariance disentangled method to obtain various robust speaker embeddings, which are learned by separating the speaker-related information from all other factors. In the CLSTM and UAI methods, the TIMIT dataset was not used as a test dataset. Therefore, in this study, we reproduce the code of the CLSTM and UAI methods to verify the TIMIT dataset. In the process of reproduction, we cut out the keyword detection branch in CLSTM and adopted the supervised training policy to train UAI.

Table 4 Speaker recognition accuracy (SRA) comparison of 120 and 630 speakers on the TIMIT dataset under different i-vector dimensions and different mixture sizes

Method	Number of speakers	Dim_i	Mixture size	SRA (%)
LDL	120	100	256	98.13
LRL	120	100	256	97.12
LDRL_x				99.47
LDRL_i	120	100	256	**99.79**
LDL	630	200	1024	98.89
LRL	630	200	1024	97.20
LDRL_x				99.15
LDRL_i	630	200	1024	**99.21**

Dim_i denotes the i-vector dimension; subscripts x and i denote the input features of x-vector and i-vector, respectively. Best results are in bold

Table 5 Parameter setting based on the TIMIT and Apollo datasets

Dataset	α	β	λ	$\boldsymbol{C}$	Number of iterations
TIMIT	0.8	0.60	0.85	[120, 1800]	300
Apollo	0.8	0.65	0.90	[183, 800]	400

1. Comparison with the baseline and state-of-the-art methods based on the TIMIT dataset

We evaluated the performances of the proposed method, baseline approach GMM-UBM (Al-Kaltakchi et al., 2016), and state-of-the-art one EML (Al-Kaltakchi et al., 2017) based on the TIMIT dataset. To ensure the consistency with the experimental setup used in EML (Al-Kaltakchi et al., 2017), we considered 120 speakers corresponding to dialects from one to four, and then extracted six sentences per speaker for training on all five sentences (SX, phonetically compact sentences). One sentence (SA, dialect sentence) was employed for training, and other sentences (SI, phonetically diverse sentences) were used for testing the last sentence (SA). Detailed information is presented in Table 6. Table 7 provides the results of testing based on the TIMIT dataset obtained by comparing the proposed method with other methods. Using the same mixture size (256) and i-vector dimension (100) as in Al-Kaltakchi et al. (2016, 2017), the results clearly indicated that LDRL outperformed the existing methods, improving SRA by 3.12% (state-of-the-art method, EML) to 4.79% (baseline method, GMM-UBM). Compared with CLSTM and UAI (Peri et al., 2019), LDRL improved SRA by 3.48% and 2.37%, respectively.

To compare the performance of the proposed method with that of VAE (Yoshimura et al., 2018), we used the data of all speakers (630 speakers) for model training. Nine sentences per speaker were selected for training, and the remaining one sentence for testing (Table 8). Table 9 indicated that the LDRL generally outperformed the baseline approaches GMM-UBM and i-vector/PLDA with an increase of SRA by 8.42% and 8.26%, respectively. Compared with state-of-the-art results of VAE in Yoshimura et al. (2018), LDRL yielded a significant improvement of 2.7%. Table 9 showed that the results of the deep learning methods CLSTM and UAI based on the TIMIT dataset were 4.08% and 2.11% lower than that of LDRL, respectively. The results in Tables 7 and 9 demonstrated that the proposed method is effective for SR tasks.

Table 6 Development set and test set based on the TIMIT dataset for 120 speakers

Set	Number of speakers	Number of utterances per speaker	Number of utterances
Dev	120	6 (5/SX, 1/SA)	720
Test	120	4 (3/SI, 1/SA)	480
Total		10 (5/SX, 3/SI, 2/SA)	1200

SX means phonetically compact sentences, SI means dialect sentences, and SA means phonetically diverse sentences

Table 7 Speaker recognition accuracy (SRA) comparison based on the TIMIT dataset for 120 speakers

Method	SRA (%)
GMM-UBM (Al-Kaltakchi et al., 2016)	95.00
CLSTM$^+$ (Kumar et al., 2018)	96.31
ELM (Al-Kaltakchi et al., 2017)	96.67
UAI$^+$ (Peri et al., 2019)	97.42
LDRL	**99.79**

+ denotes that the TIMIT dataset was not used as a test set in the literature. We reimplemented the code of this method and tested it on the TIMIT dataset. Best result is in bold

Table 8 Development set and test set based on the TIMIT dataset for 630 speakers

Set	Number of speakers	Number of utterances per speaker	Number of utterances
Dev	630	9	5670
Test	630	1	630
Total		10	6300

Table 9 Speaker recognition accuracy (SRA) comparison based on the TIMIT dataset for 630 speakers

Method	SRA (%)
GMM-UBM (Yoshimura et al., 2018)	90.79
i-vector/PLDA (Yoshimura et al., 2018)	90.95
CLSTM$^+$ (Kumar et al., 2018)	95.13
VAE model (Yoshimura et al., 2018)	96.51
UAI$^+$ (Peri et al., 2019)	97.10
LDRL	**99.21**

+ denotes that the TIMIT dataset was not used as a test set in the literature. We reimplemented the code of this method and tested it on the TIMIT dataset. Best result is in bold

2. Comparison based on the Apollo dataset used in the Fearless Steps Challenge in INTERSPEECH2019

In SR tasks, there are many cases in which recognition accuracy is affected significantly. Environmental factors such as ambient noise, reverberation, microphone type, multiple speakers, and capture devices are deemed common sources of such effect. We provided the results of performance evaluation by applying the proposed method to a real environment dataset, i.e., the Apollo dataset. The experimental setup of the Apollo dataset provided by the organizer of the Fearless Steps Challenge in INTERSPEECH2019 has been presented in Table 2. More than 350 known speakers contributed with varying degrees of content; however, the data of only 183 speakers was provided for the experiments. The total voice content had a length of at least 10 s. These speakers were distributed in the development set (Dev, training set) and evaluation set (Eval, test set).

Table 10 shows the results of the baseline approaches obtained during the Fearless Steps Challenge in INTERSPEECH2019 and those of the proposed method. The baselines Dev and Eval denote the baseline results provided by the organizer of the Fearless Steps Challenge in the development set and evaluation set, respectively. LDRL (Dev) and LDRL (Eval) represent the results of the proposed method in the development set and evaluation set, respectively. The ground truth of each utterance appeared in the top five labels of a prediction (Table 10). It was assumed that the prediction result of the utterance was correct. SRA obtained in the development set by the proposed method was based on the known speaker labels, while the sample labels in the evaluation set were unknown. Therefore, we fixed the top five speaker labels of each utterance prediction into a file and sent the file to the organizer of the Fearless Steps Challenge who provided us with the SRA obtained by the proposed method.

As shown in Table 10, the results of baseline approaches Dev and Eval on the development and evaluation sets provided by the Fearless Steps Challenge were 58.17% and 47.00%, respectively. These low SRA results were caused by the presence of multiple instances with rapid switching of speakers, short duration, and environmental noise. The results demonstrated that LDRL (Dev) outperformed the baseline (Dev) by 32.98% and that LDRL (Eval) surpassed the baseline (Eval) by 36.33%. These results proved that the proposed model is robust when applied to a speech with complex background noise. No model ablation experiment was conducted on the Apollo dataset owing to the elapsed deadline of the Fearless Steps Challenge in INTERSPEECH2019.

3. Feature distribution visualization before/after using LDRL

To provide an intuitive understanding of the learned latent discriminative representations by LDRL, we first visualized the i-vector for seven people and transformed the corresponding latent representation into a two-dimensional space of the same scale using the proposed method. Specifically, Fig. 2 shows the distribution visualization of the i-vector and latent discriminative representation. As shown in Fig. 2a, the i-vector features corresponding to different speakers were separated from each other. Therefore, the i-vector was deemed sufficiently discriminative to recognize speakers. However, different i-vector features corresponding to the same speaker scattered relatively. Fig. 2b shows that latent discriminative representation outperformed the i-vector in the two-dimensional space of the same scale. Not only the latent discriminative representations of different speakers were separated, but more importantly, the different latent discriminative representations of the same speaker were close to each other. The latent representations obtained by the proposed method were sufficiently discriminative for different speakers and had strong correlations concerning the same speaker. Finally, these results indicated that we have achieved the goal of learning latent discriminative representations with relevance from the original feature input.

Table 10 Top five speaker recognition accuracy (SRA) comparison based on the Apollo dataset for 183 speakers

Method	SRA (%)
Baseline (Dev)	58.17
LDRL (Dev)	91.15
Baseline (Eval)	47.00
LDRL (Eval)	83.33

5 Conclusions

In this study, we have proposed a novel speaker recognition approach based on latent discriminative representation learning (LDRL). We have demonstrated its effectiveness by comparison with other methods during testing on the TIMIT dataset. Moreover, we have conducted experiments on the Apollo dataset introduced in the Fearless Steps Challenge in INTERSPEECH2019. The experimental results confirmed that the latent representations learned by the proposed method are sufficiently discriminative for SR. Finally, we have experimentally proved that performing a latent representation depends on the following aspects: (1) i-vector features used in dictionary learning to derive a basic latent representation; (2) a lookup table connected to the latent representation space; (3) linear transformation matrix $\boldsymbol{C}$ regarded as a speaker clas-

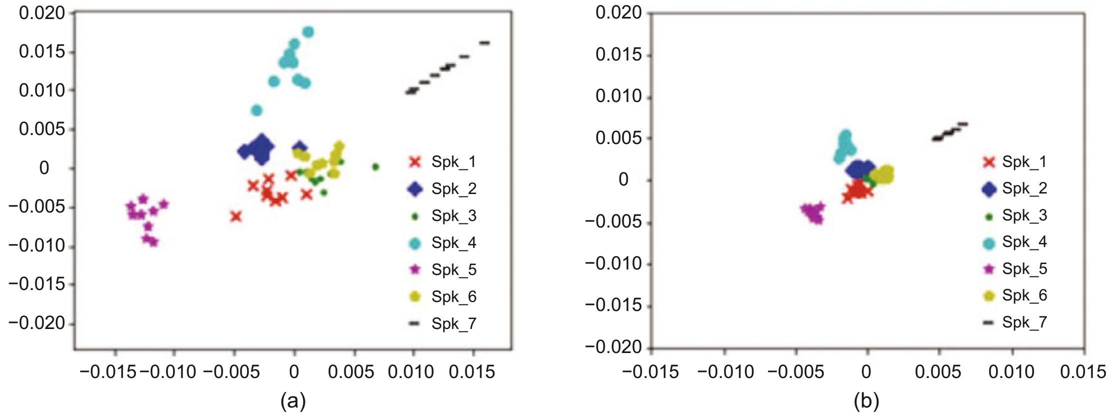

Fig. 2 Distribution visualization of the i-vector (a) and latent discriminative representation (b) in a two-dimensional space (References to color refer to the online version of this figure)

sifier that makes the latent representation discriminative. The latent representation obtained is not only discriminative, but also relevant. In addition, the visualization results supported these conclusions. In the future, we plan to extend the proposed method using deep networks and to evaluate its performance on large-scale complicated speech datasets.

Contributors

Duolin HUANG and Qirong MAO designed the research. Duolin HUANG processed the data. Duolin HUANG and Qirong MAO drafted the manuscript. Zhongchen MA, Zhishen ZHENG, Sidheswar ROUTRAY, and Elias-Nii-Noi OCQUAYE helped organize the manuscript. Duolin HUANG and Qirong MAO revised and finalized the paper.

Compliance with ethics guidelines

Duolin HUANG, Qirong MAO, Zhongchen MA, Zhishen ZHENG, Sidheswar ROUTRAY, and Elias-Nii-Noi OCQUAYE declare that they have no conflict of interest.

References

Al-Kaltakchi MTS, Woo WL, Dlay SS, et al., 2016. Study of statistical robust closed set speaker identification with feature and score-based fusion. IEEE Statistical Signal Processing Workshop, p.1-5. https://doi.org/10.1109/SSP.2016.7551807

Al-Kaltakchi MTS, Woo WL, Dlay SS, et al., 2017. Speaker identification evaluation based on the speech biometric and i-vector model using the TIMIT and NTIMIT databases. Proc 5th Int Workshop on Biometrics and Forensics, p.1-6. https://doi.org/10.1109/IWBF.2017.7935102

Chen NX, Qian YM, Yu K, 2015. Multi-task learning for text-dependent speaker verification. Proc 16th Annual Conf of the Int Speech Communication Association, p.185-189.

Chen XB, Cai YF, Chen L, et al., 2015. Discriminant feature extraction for image recognition using complete robust maximum margin criterion. *Mach Vis Appl*, 26(7-8):857-870. https://doi.org/10.1007/s00138-015-0709-7

Cumani S, Plchot O, Laface P, 2013. Probabilistic linear discriminant analysis of i-vector posterior distributions. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.7644-7648. https://doi.org/10.1109/ICASSP.2013.6639150

Davis S, Mermelstein P, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Process*, 28(4):357-366. https://doi.org/10.1109/TASSP.1980.1163420

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

Desai D, Joshi M, 2013. Speaker recognition using MFCC and hybrid model of VQ and GMM. Proc 2nd Int Symp on Intelligent Informatics, p.53-63. https://doi.org/10.1007/978-3-319-01778-5_6

Dey S, Motlicek P, Madikeri S, et al., 2017. Template-matching for text-dependent speaker verification. *Speech Commun*, 88:96-105. https://doi.org/10.1016/j.specom.2017.01.009

Fisusi A, Yesufu T, 2007. Speaker recognition systems: a tutorial. *Afr J Inform Commun Technol*, 3(2):42-52. https://doi.org/10.5130/ajict.v3i2.508

Garofolo JS, Lamel LF, Fisher WM, et al., 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report N, 93:27403, NASA, USA.

Hansen JHL, Sangwan A, Joglekar A, et al., 2018. Fearless steps: Apollo-11 corpus advancements for speech technologies from Earth to the Moon. Proc 19th Annual Conf of the Int Speech Communication Association, p.2758-2762. https://doi.org/10.21437/Interspeech.2018-1942

Heigold G, Moreno I, Bengio S, et al., 2016. End-to-end text-dependent speaker verification. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.5115-5119. https://doi.org/10.1109/ICASSP.2016.7472652

Hermansky H, 1990. Perceptual linear predictive (PLP) analysis of speech. *J Acoust Soc Am*, 87(4):1738-1752. https://doi.org/10.1121/1.399423

Huang XD, Acero A, Hon HW, 2001. Spoken Language Processing: a Guide to Theory, Algorithm and System Development. Upper Saddle River, Prentice Hall PTR, USA.

Jiang HJ, Wang RP, Shan SG, et al., 2017. Learning discriminative latent attributes for zero-shot classification. IEEE Int Conf on Computer Vision, p.4233-4242. https://doi.org/10.1109/ICCV.2017.453

Kenny P, Boulianne G, Ouellet P, et al., 2007. Speaker and session variability in GMM-based speaker verification. *IEEE Trans Audio Speech Lang Process*, 15(4):1448-1460. https://doi.org/10.1109/TASL.2007.894527

Kim MJ, Yang IH, Kim MS, et al., 2017. Histogram equalization using a reduced feature set of background speakers' utterances for speaker recognition. *Front Inform Technol Electron Eng*, 18(5):738-750. https://doi.org/10.1631/FITEE.1500380

Kumar R, Yeruva V, Ganapathy S, 2018. On convolutional LSTM modeling for joint wake-word detection and text dependent speaker verification. Proc 19th Annual Conf of the Int Speech Communication Association, p.1121-1125. https://doi.org/10.21437/Interspeech.2018-1759

Lei Y, Scheffer N, Ferrer L, et al., 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1695-1699. https://doi.org/10.1109/ICASSP.2014.6853887

Li C, Ma XK, Jiang B, et al., 2017. Deep speaker: an end-to-end neural speaker embedding system. https://arxiv.org/abs/1705.02304

Luo Y, Liu Y, Zhang Y, et al., 2018. Speech bottleneck feature extraction method based on overlapping group lasso sparse deep neural network. *Speech Commun*, 99:56-61. https://doi.org/10.1016/j.specom.2018.02.005

Mao QR, Dong M, Huang ZW, et al., 2014. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans Multim*, 16(8):2203-2213.
https://doi.org/10.1109/TMM.2014.2360798

Peri R, Pal M, Jati A, et al., 2019. Robust speaker recognition using unsupervised adversarial invariance.
https://arxiv.org/abs/1911.00940

Rabiner LR, 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77(2):257-286.

Reynolds DA, Rose RC, 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans Speech Audio Process*, 3(1):72-83.
https://doi.org/10.1109/89.365379

Reynolds DA, Quatieri TF, Dunn RB, 2000. Speaker verification using adapted Gaussian mixture models. *Dig Signal Process*, 10(1-3):19-41.
https://doi.org/10.1006/dspr.1999.0361

Sadjadi SO, Slaney M, Heck L, et al., 2013. MSR Identity Toolbox v1.0: a MATLAB Toolbox for Speaker Recognition Research. Microsoft Research Technical Report, Piscataway, NJ, USA.

Schroff F, Kalenichenko D, Philbin J, 2015. FaceNet: a unified embedding for face recognition and clustering. IEEE Conf on Computer Vision and Pattern Recognition, p.815-823.
https://doi.org/10.1109/CVPR.2015.7298682

Singh S, Rajan EG, 2011. Vector quantization approach for speaker recognition using MFCC and inverted MFCC. *Int J Comput Appl*, 17(1):1-7.

Snyder D, Ghahremani P, Povey D, et al., 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. IEEE Spoken Language Technology Workshop, p.165-170.
https://doi.org/10.1109/SLT.2016.7846260

Snyder D, Garcia-Romero D, Povey D, et al., 2017. Deep neural network embeddings for text-independent speaker verification. Proc 18th Annual Conf of the Int Speech Communication Association, p.999-1003.
https://doi.org/10.21437/Interspeech.2017-620

Snyder D, Garcia-Romero D, Sell G, et al., 2018. X-vectors: robust DNN embeddings for speaker recognition. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.5329-5333.
https://doi.org/10.1109/ICASSP.2018.8461375

Togneri R, Pullella D, 2011. An overview of speaker identification: accuracy and robustness issues. *IEEE Circ Syst Mag*, 11(2):23-61.
https://doi.org/10.1109/MCAS.2011.941079

van Leeuwen DA, Saeidi R, 2013. Knowing the non-target speakers: the effect of the i-vector population for PLDA training in speaker recognition. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.6778-6782.
https://doi.org/10.1109/ICASSP.2013.6638974

Variani E, Lei X, McDermott E, et al., 2014. Deep neural networks for small footprint text-dependent speaker verification. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.4052-4056.
https://doi.org/10.1109/ICASSP.2014.6854363

Wan V, Campbell WM, 2000. Support vector machines for speaker verification and identification. Neural Networks for Signal Processing X. Proc IEEE Signal Processing Society Workshop, p.775-784.
https://doi.org/10.1109/NNSP.2000.890157

Wen YD, Zhang KP, Li ZF, et al., 2016. A discriminative feature learning approach for deep face recognition. Proc 14th European Conf on Computer Vision, p.499-515.
https://doi.org/10.1007/978-3-319-46478-7_31

Yadav S, Rai A, 2018. Learning discriminative features for speaker identification and verification. Proc 19th Annual Conf of the Int Speech Communication Association, p.2237-2241.
https://doi.org/10.21437/Interspeech.2018-1015

Yoshimura T, Koike N, Hashimoto K, et al., 2018. Discriminative feature extraction based on sequential variational autoencoder for speaker recognition. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf, p.1742-1746.
https://doi.org/10.23919/APSIPA.2018.8659722

Young S, 1993. The HTK Hidden Markov Model Toolkit: Design and Philosophy. Department of Engineering, Cambridge University, Cambridge.

Yu K, Mason J, Oglesby J, 1995. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE Proc Vis Image Signal Process*, 142(5):313-318.
https://doi.org/10.1049/ip-vis:19952144

Zhang C, Koishida K, 2017. End-to-end text-independent speaker verification with triplet loss on short utterances. Proc 18th Annual Conf of the Int Speech Communication Association, p.1487-1491.
https://doi.org/10.21437/Interspeech.2017-1608

Zhang FF, Zhang TZ, Mao QR, et al., 2018. Joint pose and expression modeling for facial expression recognition. IEEE Conf on Computer Vision and Pattern Recognition, p.3359-3368.
https://doi.org/10.1109/CVPR.2018.00354