Hindawi
Scientific Programming
Volume 2020, Article ID 4748606, 13 pages
https://doi.org/10.1155/2020/4748606

Research Article

Data-Driven Decision-Support System for Speaker Identification Using E-Vector System

He Ma,[1] Yi Zuo ,[1,2,3,4] Tieshan Li,[1,2] and C. L. Philip Chen[1,2]

[1]*Navigation College, Dalian Maritime University, Dalian 116026, China*
[2]*Maritime Big Data & Artificial Intelligent Application Center, Dalian Maritime University, Dalian 116026, China*
[3]*Collaborative Innovation Center for Transport Studies, Dalian Maritime University, Dalian 116026, China*
[4]*The Research Institute for Socionetwork Strategies, Kansai University, Osaka 5648680, Japan*

Correspondence should be addressed to Yi Zuo; zuo@dlmu.edu.cn

Received 12 November 2019; Revised 24 February 2020; Accepted 12 June 2020; Published 29 June 2020

Academic Editor: Rahman Ali

Recently, biometric authorizations using fingerprint, voiceprint, and facial features have garnered considerable attention from the public with the development of recognition techniques and popularization of the smartphone. Among such biometrics, voiceprint has a personal identity as high as that of fingerprint and also uses a noncontact mode to recognize similar faces. Speech signal-processing is one of the keys to accuracy in voice recognition. Most voice-identification systems still employ the mel-scale frequency cepstrum coefficient (MFCC) as the key vocal feature. The quality and accuracy of the MFCC are dependent on the prepared phrase, which belongs to text-dependent speaker identification. In contrast, several new features, such as d-vector, provide a black-box process in vocal feature learning. To address these aspects, a novel data-driven approach for vocal feature extraction based on a decision-support system (DSS) is proposed in this study. Each speech signal can be transformed into a vector representing the vocal features using this DSS. The establishment of this DSS involves three steps: (i) voice data preprocessing, (ii) hierarchical cluster analysis for the inverse discrete cosine transform cepstrum coefficient, and (iii) learning the E-vector through minimization of the Euclidean metric. We compare experiments to verify the E-vectors extracted by this DSS with other vocal features measures and apply them to both text-dependent and text-independent datasets. In the experiments containing one utterance of each speaker, the average accuracy of the E-vector is improved by approximately 1.5% over the MFCC. In the experiments containing multiple utterances of each speaker, the average micro-F1 score of the E-vector is also improved by approximately 2.1% over the MFCC. The results of the E-vector show remarkable advantages when applied to both the Texas Instruments/Massachusetts Institute of Technology corpus and LibriSpeech corpus. These improvements of the E-vector contribute to the capabilities of speaker identification and also enhance its usability for more real-world identification tasks.

1. Introduction

Over the last decades, recognition technologies based on biometrics such as fingerprint, facial features, voiceprint, and iris scans have been widely used in target identification for access security, system identity, private confirmation, etc. In terms of technical and practical usage, recognition through iris scans is the most secured and accurate and is applied to meet the requirements of military standards [1]. For mass requirements, fingerprint recognition is one of the most popular and mature identity-recognition technologies [2]. As fingerprint acquisition and recognition need specific devices [3], it has been increasingly replaced by face recognition in recent years, often preferred for its noncontact mode [4]. Facial data can be collected more easily than iris and fingerprint data, as most smartphones already have an inbuilt camera. However, the accuracy of face recognition is dependent on the recognition conditions, such as environmental brightness and camera angle [5]. Similar to face recognition, voice recognition is also a noncontact mode technique. The voiceprint can be easily collected using a microphone and other voice receivers, and its quality

requirements are less dependent on environmental factors than those of face recognition [6]. Similar to fingerprints, voiceprints also contain unique biometric features and are superior to facial features on recognition accuracy. Nevertheless, voice recognition has many merits compared with other recognition technologies. Feature extraction from voiceprint data is the main technical bottleneck, and its real-world applications are fewer than those of fingerprint and face recognition in daily life. In contrast to face recognition, which employs image processing methods, a voiceprint is composed of classic mechanical waves that need signal processing methods to transform voice signals from time-domain representation into frequency-domain representation. Such vocal features are difficult to implement, but quite efficient for speaker identification owing to the biometric variances and personal characteristics between different voiceprints. In most methods of speaker identification (SI), there are two main processes: one is to extract the vocal features, and the other is to learn the identification model based on these features. The vocal features not only adequately represent the common properties of the same speaker, but also separate the different speakers as far apart as possible. Therefore, an effective extraction of vocal features can determine the performance of SI models, such that these models can definitely identify the target speaker from multiple speakers' utterances.

The general process of SI can be regarded as a decision-making support process that decides the identity of the corresponding speakers by their utterances. In the field of automatic speech recognition (ASR), most methods of SI are constructed by extracting the vocal features, which is also one of the most important applications of the decision-support system (DSS) for SI tasks. The linear prediction coefficient (LPC) was extracted by a linear combination of the exiting speech, which was the first proposed speech feature in 1967 [7]. Since the vocal feature named mel-scale frequency cepstrum coefficient (MFCC) was proposed in 1980 [8], it has been extensively applied in SI systems. The perceptive linear prediction coefficient was extracted by putting speech signals into the auditory model based on LPC in 2011 [9]. In 2012, a histogram frequency-domain transformation on the discrete cosine transform (DCT) cepstrum coefficient (HDCC) was carried out based on the idea of the MFCC feature extraction [10]. Subsequently, Kim and Stern applied a power-law nonlinear transformation instead of the traditional log nonlinear transformation of MFCC by auditory processing, and they proposed a new feature called power-normalized cepstral coefficients (PNCC) in 2016 [11]. In comparison with the traditional vocal features, an SI model based on the identity vector (i-vector) was also proposed in a data-driven approach [12], which is a popular topic in the field of ASR. Furthermore, Variani et al. applied a deep neural network to generate the d-vector, which is a similar feature to the i-vector [13]. Based on this d-vector, an end-to-end SI approach was also proposed [14].

The existing methods for the extraction of vocal features mostly used model-based approaches, such as MFCC, PNCC, and LPC. In contrast, several new vocal features such as d-vector are based on data-driven approaches. However, these approaches are "black box" in vocal feature learning [14]. Consequently, in this study, a novel method using the data-driven approach of hierarchical cluster analysis for SI is proposed. There are three main contributions: (1) a novel vocal feature extraction method is proposed based on a data-driven approach of hierarchical cluster analysis; (2) the Euclidean metric is used as a measure to generate an adaptive feature vector called the "E-vector"; (3) DSS is established based on the E-vector to provide decision-making support services for SI tasks. In the data-driven hierarchical clustering approach, various personal phonetic features are considered to learn and extract the vocal feature vector. The distances between different cepstral coefficients of the same speaker are measured using the Euclidean metric, and the E-vector is generated through a hierarchical clustering approach by minimizing the Euclidean metric. In the comparative experiments of single utterance SI, the E-vector method improves the identification accuracy by approximately 3% over the MFCC and 5% over the HDCC, where the DSS of SI is based on the Gaussian mixture model (GMM). In the comparative experiments of multiple utterances SI, the micro-F1 score of the E-vector is better than the MFCC and HDCC, where the DSS of SI is based on both the GMM and hidden Markov model (HMM).

The remainder of this paper is organized as follows: the problem statement and E-vector are introduced in Section 2. The conducted comparative experiments to evaluate the performance of the E-vector are described in Section 3. Finally, a conclusion is provided in Section 4.

2. Materials and Methods

2.1. Problem Statement

2.1.1. Model-Based Extraction of Vocal Features. The existing vocal features used in SI are mostly based on model-based approaches, such as MFCC, HDCC, and PNCC. MFCC is a widely used speech feature first proposed in the 1980s. MFCC applies a discrete Fourier transform method to transform the time-domain signal into a frequency-domain signal. In an MFCC transformation, we use the following equation to translate the frequency-domain signal into mel-frequency:

$$\mathrm{mel}(f) = 2595 * \ln\left(1 + \frac{f}{700}\right), \tag{1}$$

where f is the original frequency, and $\mathrm{mel}(f)$ represents mel-frequency. Subsequently, the amplitude based on mel-frequency is calculated by a series of triangular filters, as in Figure 1(a). Finally, MFCC is obtained by making a cepstrum analysis of the signal using the mel-frequency and triangular filters [8]. HDCC is a new feature proposed with the influence of MFCC. The HDCC creates a two-term span of histogram bins: 50–500 Hz with a span of 50 Hz each and 600–1000 Hz with a 100 Hz span of each as shown in Figure 1(b). After DCT cepstrum coefficients of each bin are obtained from histogram analysis, we can extract the HDCC for each bin [10]. PNCC has similar parts of the first two steps of MFCC in its initial process. Next, PNCC obtains the

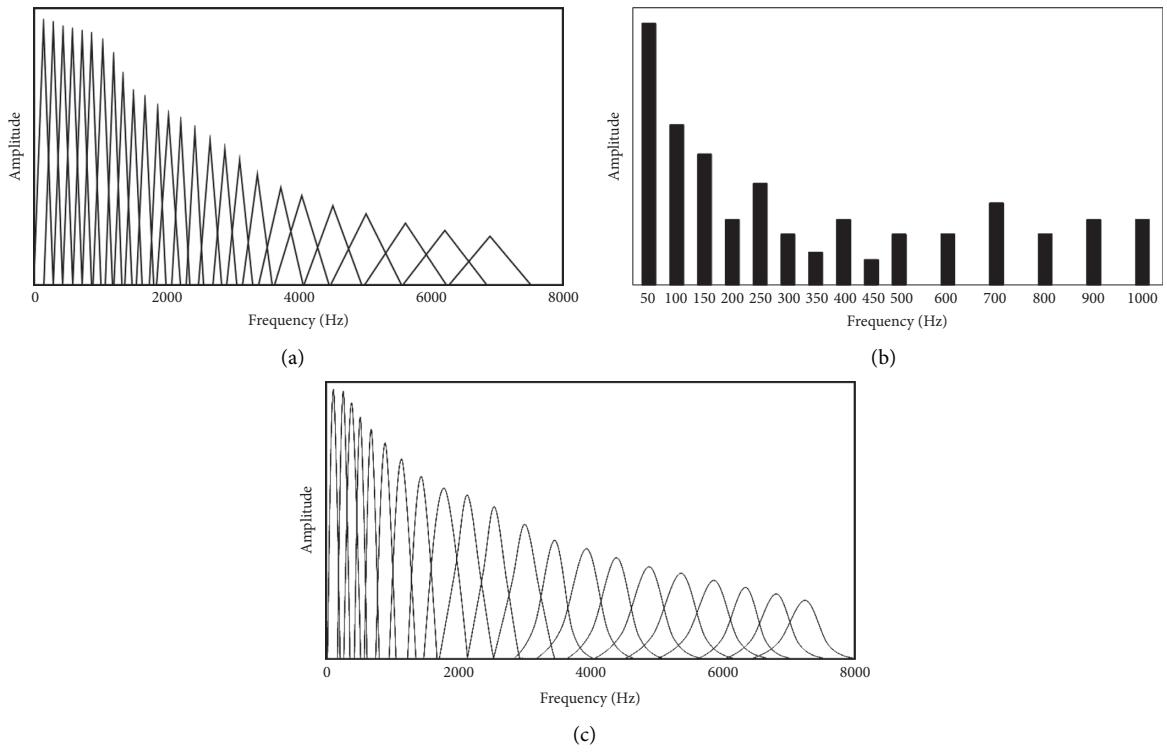


Figure 1: (a) MFCC filters. (b) HDCC filters. (c) PNCC filters.

short-time spectral power using the squared gammatone summation. As shown in Figure 1(c), gammatone filters are power-law nonlinear transformations, different from the traditional log nonlinearity used in the MFCC. Finally, smoothing-weight processing is used on each frame, and spectral subtraction is applied to realize noise suppression [11].

2.1.2. Data-Driven Extraction of Vocal Features. The existing research using data-driven approaches for SI mostly focused on clustering different speakers via their feature similarity. For instance, the Alibaba group proposed a speech recognition method based on a clustering method in 2017 [15]. They obtained the feature vector based on cluster analysis of training data. Then, the feature vector model was established for speech recognition [15]. Nevertheless, there are few researches on applying data-driven approach to extract vocal feature. The i-vector, d-vector, and end-to-end SI approach were proposed based on data-driven approach; however, they are black box (method without a transparent working process) [12–14]. Actually, vocal data has relevant regularities accounting for the speaker's personal phonetic features. Therefore, in this paper, a novel vocal feature, E-vector, extracted using a data-driven approach is proposed. The method learns the SI models based on the E-vector realized as a DSS.

2.2. Determining the Decision Objective. In decision-making process, there are generally four steps, as shown in Figure 2 (a). At the beginning, decision objective (DO) should be determined after finding out the problem. Then, the scheme will be designed based on the decision environment. Next, the scheme will be evaluated in order to carry the scheme. In an SI task, the SI process can be regarded as a multilabel classification task. The number of speakers is the number of classes; the labels are the utterance of each speaker. The DO is achieving the classification of all speakers by identifying all speakers' identities based on vocal features, as shown in Figure 2 (b).

2.3. E-Vector System for Speaker Identification. In this section, we introduce the E-vector system for SI-DSS. It is shown in Figure 3(a) that the SI-DSS based on E-vector system is established in three steps: (i) data preprocessing, (ii) cluster analysis, and (iii) learning models. When a continuous speech signal is put into the E-vector system, data preprocessing is applied to obtain the inverse discrete cosine transform (IDCT) cepstrum coefficient; then the clustering method is used to analyze the IDCT cepstrum coefficient, and finally, GMM and HMM are applied to classify the speakers. The following charts show the detailed introduction of the E-vector system.

2.3.1. Step 1: Data Preprocessing. The competency of data preprocessing is storing speech data in the form of the IDCT cepstrum coefficient. The original speech data is in the form of a continual signal wave, and the spectrogram is generally used to describe the continual signal wave. In this study, the spectrogram is extracted by the following three steps:

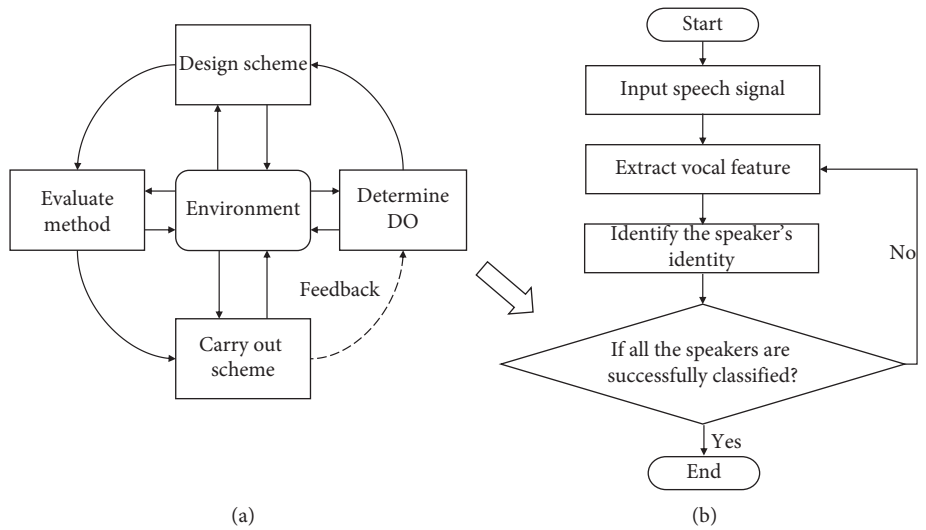


Figure 2: The relationship between DSS and SI. (a) DSS. (b) The SI system based on DSS.

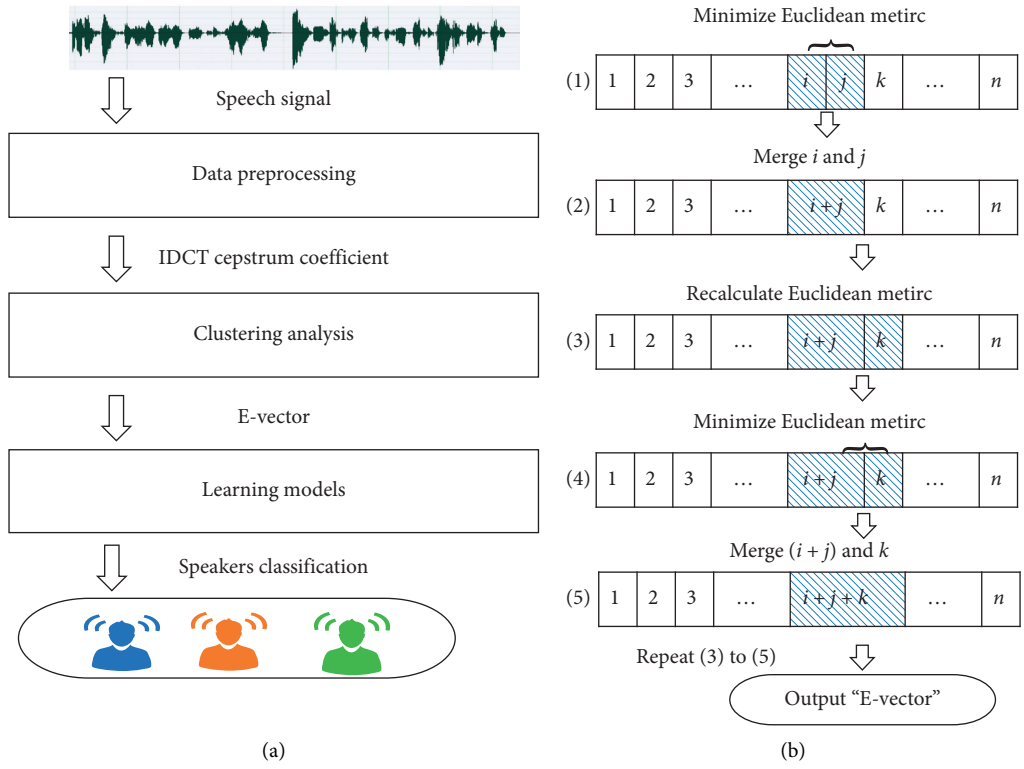
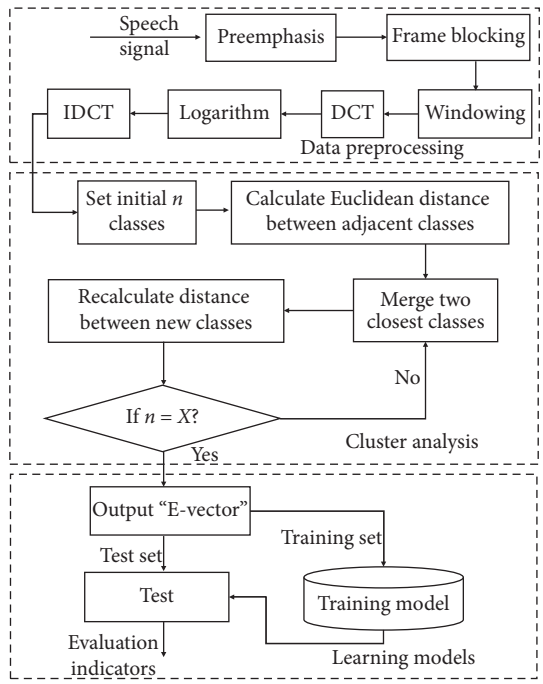


Figure 3: Continued.



(c)

Figure 3: Overview of E-vector system for SI-DSS. (a) Flowchart of SI-DSS. (b) Flowchart of E-vector procedure. (c) Flowchart of SI system.

(1) The first step aims at making the speech signal wave more significant. High-pass filtering process shown in the following equation is used to preemphasize the input signal wave [7]:

$$H(z) = 1 - \mu z^{-1}. \tag{2}$$

Here, z is the input speech signal, H (z) is the output preemphasis speech signal, and the value of μ is 0.97 in this study.

(2) In the second step, the preweighted speech signal is segmented into small blocks to get a frame signal (frames of 20 ms in this study).

(3) The third step is adding a Hamming window, W, to the framed signal. The Hamming window function is defined as

$$W(n, 0.46) = 0.54 - a \times \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \ll N-1, \tag{3}$$

where N is the number of each frame. It can make the voice data more periodic for analyzing each frame signal.

Then, homomorphic signal processing is applied to obtain the IDCT cepstrum coefficient. The processing involves three steps:

(1) In the first step, the DCT is applied to obtain the multiplicative signal as equation (4) from all frames of the speech signal. The process of DCT is defined as

$$C(a) = \sum_{m=0}^{N-1} S(b)\cos\left(\frac{\pi n(b-0.5)}{M}\right), \quad a = 1, 2, \ldots, L. \tag{4}$$

Here, $S(b)$ is the input speech signal and M is its points; $C(a)$ is the output signal, and a is the points of transformation:

$$C(a) = S_h(b) \times S_l(b). \tag{5}$$

Here, $S_h(b)$ is the high-frequency signal; $S_l(b)$ represents the low-frequency signal.

(2) The second step is calculating the logarithmic energy of the output signal to convert the multiplicative signal into an additive signal as follows:

$$\log C(a) = \log\left(S_h(b)\right) + \log\left(S_l(b)\right). \tag{6}$$

(3) The third step is applying the IDCT to obtain the cepstrum coefficient as follows:

$$c(a) = s_h(b) \times s_l(b). \tag{7}$$

Here, $c(a)$ is the IDCT cepstrum coefficient, $s_h(b)$ is the output high-frequency signal, and $s_l(b)$ is the output low-frequency signal.

2.3.2. Step 2: Cluster Analysis for IDCT Cepstrum Coefficient. The obtained IDCT cepstrum coefficient is a data matrix, and the length of a row is proportional to the time duration of the input vocal signal, and the length of a column is proportional to the number of the speakers in the input signal. The analysis process of the IDCT cepstrum coefficients consists of five steps:

(1) If the input vocal signal contains m speakers' speech, the IDCT cepstrum coefficient can be described by set A. The speech of speaker p ($p = 1, 2, \ldots, m$) can be described as in (9), where n is the number of A's columns. For such a data matrix, the cluster method can be applied to analyzing using the data-driven approach. An improved hierarchical cluster method is proposed to analyze the IDCT cepstrum coefficient as only adjacent columns can be grouped:

$$A = \{A_1; A_2; \ldots; A_m\}, \tag{8}$$

$$A_p = \{a_1^P, a_2^P, \ldots, a_n^P\}. \tag{9}$$

(2) Each column of set A is regarded as a class, so there are n classes, as shown in Figure 3(b).

(3) Calculate the distances of adjacent classes and define the similarity values as a set. Here, the Euclidean distance measure [16] is used to calculate the distance. The smaller the value of the distance, the greater the similarity. The Euclidean distance l_i ($l_i \in S$) of a_i and a_{i+1} can be described as in (10). Thus, the set S can be composed of $n - 1$ number of distance values. It can be described as in (11):

$$l_i = \mathrm{Dis}(a_i, a_{i+1}) = \sqrt{\sum_{j=1}^{mn} (a_{i,j} - a_{i+1,j})}, \tag{10}$$

$$S = \{l_1, l_2, \ldots, l_{n-1}\}. \tag{11}$$

(4) Compare all values in S; if the Euclidean distance l_i of a_i and a_{i+1} is the minimum one, group a_i and a_{i+1} into a class. Update the classes of set A.

(5) Iterate step (2) to step (4) until the number n of classes in set A is equal to X (X is determined by identification accuracy) as shown as Figure 3(c). Then, the classes in set A constitute the E-vector.

2.3.3. Step 3: Learning SI Models. The identification process matches the input feature with the model feature set by the degree of similarity. In this study, the model feature set is established using the GMM and HMM based on the E-vector feature Algorithm 1. The HMM achieves the identification task by searching for the sequence most likely to produce a particular output sequence in the implicit state; the process consists of six steps:

(1) Define a vocal feature set $A = \{a_1, a_2, \ldots, a_X\}$ for the model, where a_X is the number X class of the vocal feature set A.

(2) Accumulate certain vocal features with their labels in each class a_i in the vocal feature set A to establish the training feature set.

(3) Obtain the best model λ_i for a_i, based on the training set, as shown in Figure 3(c).

(4) Produce an unknown observation sequence O for input feature.

(5) Estimate the probability of the input feature $\Pr(O/\lambda_i)(i = 1, 2, \ldots, X)$. The input feature belongs to the class with the maximal probability.

(6) Calculate evaluation indicators, and the value of X is determined by the state with the highest accuracy.

3. Results and Discussion

3.1. Experimental Design

3.1.1. Datasets of Vocal Corpus. In the experiments, we used two vocal datasets. One was the Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus, and the other one was the LibriSpeech corpus. The TIMIT corpus was applied as the representative of a text-dependent experiment, which contained 6300 sentences spoken by 630 persons [17]. The LibriSpeech corpus is a variety of audio datasets that consists of text and voice. Thus, the LibriSpeech corpus was used as the representative of a text-independent experiment [18]. Two groups of experiments were conducted with the TIMIT corpus and LibriSpeech corpus. Table 1 shows the number of speakers used in the experiments with the TIMIT and LibriSpeech corpus.

3.1.2. Evaluation Indicators. In the research of identity identification, several evaluation indicators were applied to evaluate the algorithm's performance. The false rejection rate (FRR) is the proportion of cases mistaking the matched voiceprint as the unmatched voiceprint. It refers to the proportion of cases in which the same voiceprint is mistakenly considered as a different voiceprint when testing the voiceprint recognition on the standard voiceprint database:

$$\mathrm{FRR} = 1 - \mathrm{recall}. \tag{12}$$

In this study, we applied the measures accuracy, precision, recall, and micro-F1 score to evaluate the performance of the E-vector against other features. If the number of speakers is m, TP_i is the true positive number of "i" ($0 \le i \le m$) speakers, FP_i is the false positive number of "i," TN_i is the true negative number of "i," and FN_i is the false negative number of "i."

The accuracy is calculated as follows:

$$\mathrm{accuracy} = \frac{\sum_{i=0}^{m} (\mathrm{TP}_i + \mathrm{TN}_i)}{\sum_{i=0}^{m} (\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{TN}_i + \mathrm{FN}_i)}. \tag{13}$$

Input: z (continuous speech signal); frame 20 ms; Step 10 ms; $n = X$
Output: E-vector $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_\mathbf{X}\}$; accuracy; micro-F1
(1) Initialization: $c\ (a) \longrightarrow$IDCT cepstrum coefficient
(2) Data preprocessing: $c\ (a) =$ data preprocessing (z)
(3) Cluster: Set S (S$\leftarrow\varnothing$), Name $c\ (a) \longrightarrow \mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_\mathbf{i}, \ldots, \mathbf{a}_\mathbf{n})$.
(4) **For** each $\mathbf{a}_\mathbf{i}$ in A do
(5) $\quad \mathbf{l}_\mathbf{i} = \mathbf{Dis}(\mathbf{a}_\mathbf{i}, \mathbf{a}_{\mathbf{i+1}})$
(6) $\quad$ Put all $\mathbf{l}_\mathbf{i}$ into S
(7) $\quad$ **If** $\mathbf{l}_\mathbf{i} = \mathbf{arcmin}(\mathbf{S})$
(8) $\quad \mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_\mathbf{i} + \mathbf{a}_{\mathbf{i+1}}, \mathbf{a}_{\mathbf{i+2}} \ldots, \mathbf{a}_\mathbf{n})$.
(9) $\quad\quad$ **If** $n = X$
(10) $\quad\quad\quad$ break;
(11) $\quad\quad$ **else**
(12) $\quad\quad\quad$ continue;
(13) $\quad\quad$ **end**
(14) $\quad$ **end**
(15) **end**
(16) Learn models: put A into GMM, HMM $\longrightarrow$ accuracy, micro-F1

Algorithm 1: E-vector system for speaker identification.

Table 1: Datasets of TIMIT and LibriSpeech corpuses.

	TIMIT				LibriSpeech			
Set no.	T1	T2	T3	T4	L1	L2	L3	L4
Number of speakers	100	300	500	630	10	20	30	40

The precision is calculated as follows:

$$\text{precision} = \frac{\sum_{i=0}^{m} \text{TP}}{\sum_{i=0}^{m} (\text{TP}_i + \text{FP})}. \tag{14}$$

The recall represents the percentage actually true in the positive set, and it is described as follows:

$$\text{recall} = \frac{\sum_{i=0}^{m} \text{TP}_\text{i}}{\sum_{i=0}^{m} (\text{TP}_i + \text{FN}_i)}. \tag{15}$$

The formula of micro-F1 is described as follows:

$$\text{micro} - F1 = 2 \times \frac{recall \times precision}{\text{recall} + \text{precision}}. \tag{16}$$

3.2. Comparison Experiments

3.2.1. Optimization of E-Vector Dimension. In order to decide the optimal dimension of the E-vector based on hierarchical cluster analysis, we selected 630 people of the TIMIT corpus (i.e., T4) and 40 people of the LibriSpeech corpus (i.e., L4) and measured the training accuracy. The proposed E-vector features with different dimensions of 15, 25, and 35 were used with GMM and HMM for the SI task. The results in Table 2 show that 15 dimensions of the E-vector obtain the highest training accuracy. In the experiments with the TIMIT corpus, the voice signals of 630 people were selected for the experiments. The voice signals of 40 people were selected for the experiments with the LibriSpeech corpus. It is shown in Table 2 that the E-vector obtains the same highest accuracy when it consists of 15 and 35 dimensions. We chose the smaller dimension, 15, as the dimension for the E-vector.

Table 2: Identification accuracy comparison.

Model	Set no.	E-vector (15)	E-vector (25)	E-vector (35)
GMM	T4	1.000	0.970	1.000
	L4	1.000	1.000	1.000
HMM	T4	0.930	0.930	0.930
	L4	0.950	0.950	0.950

3.2.2. Single-Utterance Comparison Experiments. We first tested the 15-dimensional E-vector, 13-dimensional MFCC, and 15-dimensional HDCC for SI with an input speech signal containing one utterance of each speaker. The different numbers of speakers in the TIMIT corpus and LibriSpeech corpus are identified using the GMM and HMM. The accuracy results are shown in Table 3. The best performances for each test on each corpus are shown in boldface type.

The following is shown: (1) In the TIMIT corpus, the E-vector performs best with accuracy 1.000 when using the GMM, and the results of MFCC are relatively worse than those of the E-vector; HDCC is inferior to MFCC and E-vector with approximately 10% gap, as shown in Figure 4(a). When the recognition model is HMM, as in Figure 4(c), the results of MFCC and E-vector are both approximately 0.850. It can be found that all these characteristic parameters have good performance with a recognition accuracy of over 0.75. (2) In the LibriSpeech

Table 3: Single-utterance experiment identification accuracy comparison.

Model	Feature	T1	T2	T3	T4	L1	L2	L3	L4
GMM	MFCC	**1.000**	0.990	1.000	0.995	1.000	0.950	0.970	0.975
	HDCC	0.940	0.920	0.926	0.924	**1.000**	**1.000**	0.970	0.950
	E-vector	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HMM	MFCC	0.930	0.866	**0.874**	**0.860**	**1.000**	0.900	**0.930**	**0.950**
	HDCC	0.833	0.830	0.776	0.760	**1.000**	**0.950**	**0.930**	0.925
	E-vector	**0.940**	**0.870**	0.780	**0.860**	**1.000**	**0.950**	**0.930**	**0.950**

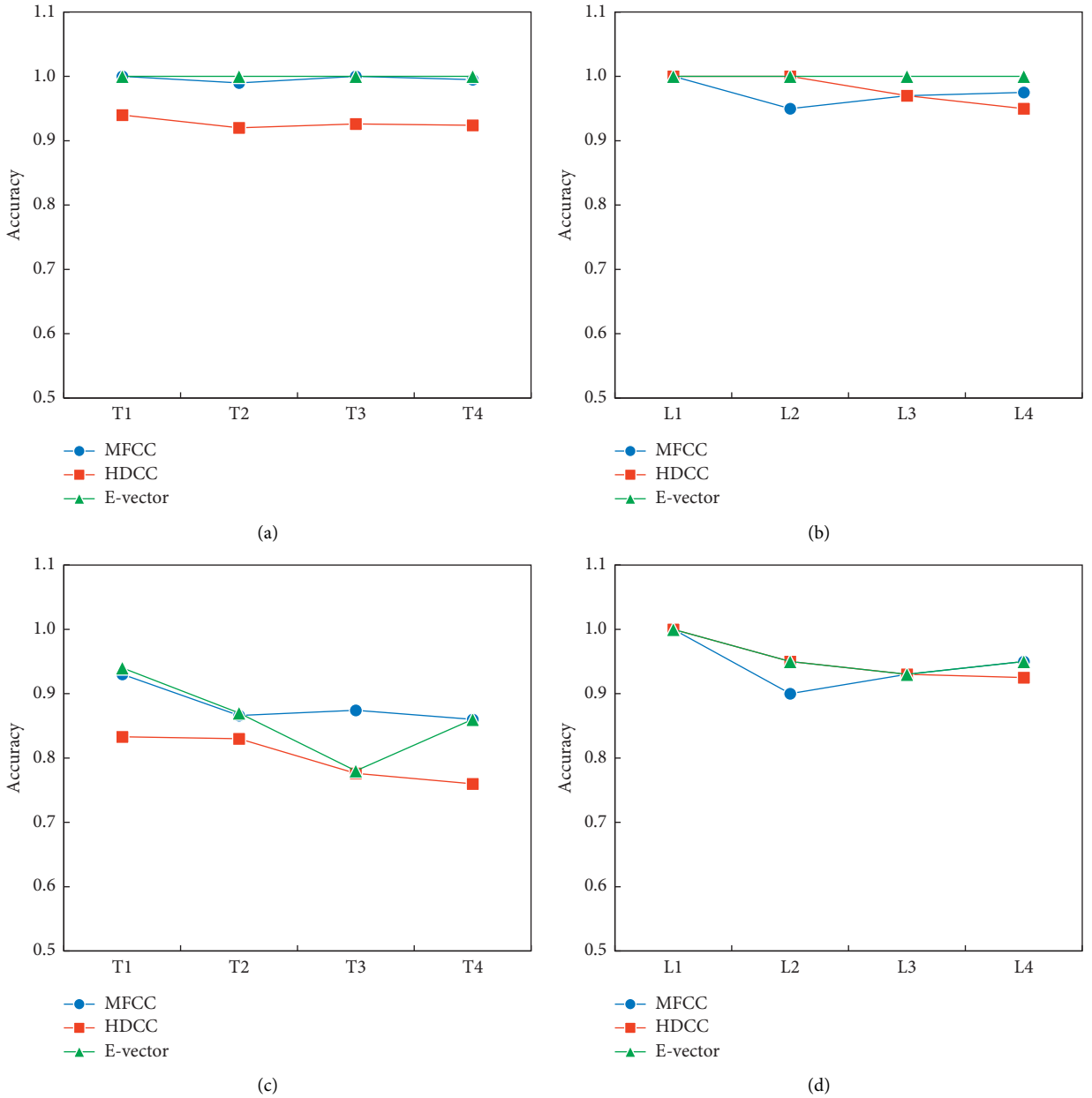


Figure 4: Single-utterance experiment accuracy comparison. (a) TIMIT corpus using GMM. (b) LibriSpeech corpus using GMM. (c) TIMIT corpus using HMM. (d) LibriSpeech corpus using HMM.

Table 4: Three-utterance experiment micro-F1 score comparison.

Model	Feature	T1	T2	T3	T4	L1	L2	L3	L4
GMM	MFCC	**0.990**	0.971	0.965	0.973	**1.000**	**1.000**	**1.000**	**1.000**
	HDCC	0.897	0.793	0.778	0.765	**1.000**	**1.000**	0.978	0.992
	E-vector	0.977	**0.978**	**0.975**	**0.976**	**1.000**	**1.000**	**1.000**	**1.000**
HMM	MFCC	0.940	0.903	0.902	0.895	1.000	0.933	0.925	0.900
	HDCC	0.877	0.827	0.815	0.807	**1.000**	**0.950**	0.956	0.958
	E-vector	**0.943**	**0.920**	**0.905**	**0.899**	**1.000**	**0.950**	**0.967**	**0.975**

Table 5: Five-utterance experiment micro-F1 score comparison.

Model	Feature	T1	T2	T3	T4	L1	L2	L3	L4
GMM	MFCC	0.962	0.948	0.939	0.928	**1.000**	0.990	0.987	0.990
	HDCC	0.706	0.699	0.659	0.630	0.970	**1.000**	0.960	0.970
	E-vector	**0.968**	**0.959**	**0.954**	**0.948**	**1.000**	**1.000**	**0.987**	**0.990**
HMM	MFCC	0.916	0.889	0.886	0.860	**1.000**	0.940	**0.953**	0.940
	HDCC	0.816	0.772	0.753	0.735	**1.000**	0.890	**0.953**	0.955
	E-vector	**0.912**	**0.895**	**0.887**	**0.876**	**1.000**	**0.950**	0.920	**0.965**

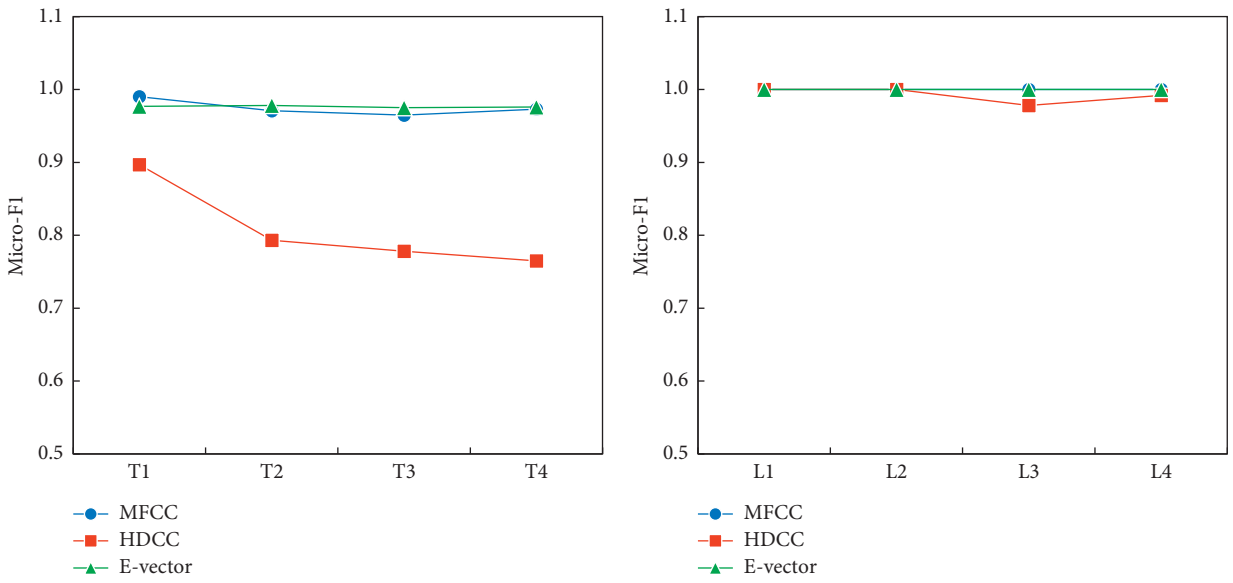


Figure 5: Continued.

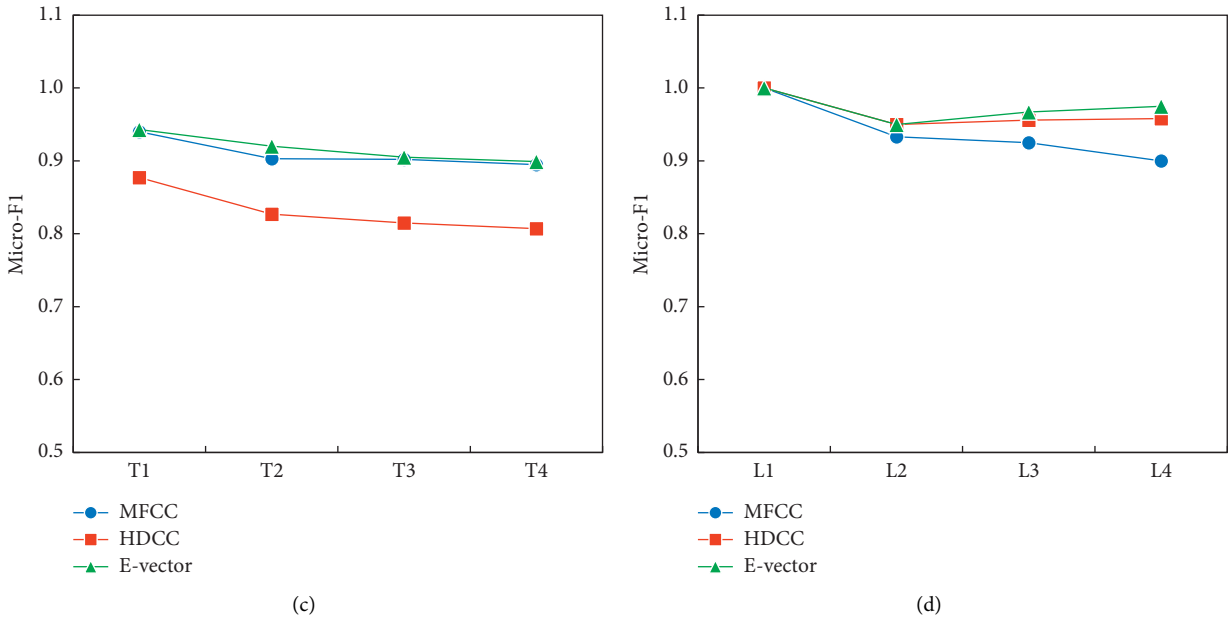


Figure 5: Three-utterance experiment micro-F1 score comparison. (a) TIMIT corpus using GMM. (b) LibriSpeech corpus using GMM. (c) TIMIT corpus using HMM. (d) LibriSpeech corpus using HMM.

corpus, the E-vector also performs best with accuracy 1 when using the GMM; MFCC and HDCC are inferior to E-vector as shown in Figure 4(b). The MFCC is almost similar to the HDCC as shown in Figure 4(d), where the accuracy is approximately 0.93 using HMM. The identification result of E-vector shows remarkable advantages in single-utterance comparison experiments.

3.2.3. Multiple-Utterance Comparison Experiments. In order to further verify the effectiveness of E-vector, we conducted the experiments with the input signal speech containing multiple utterances with GMM and HMM. We firstly used signals containing three utterances from each speaker, and the experiment results are shown in Table 4. Subsequently, we added the utterance of each speaker and used the signals with five utterances of each speaker, and the results are shown in Table 5.

In Figure 5, the identification results with the input signal containing three utterances are shown. (1) When GMM is used as the SI model, the micro-F1 scores of E-vector are a little higher than MFCC, approximately 1%, in both TIMIT and LibriSpeech corpuses. The micro-F1 score of HDCC is less than MFCC and E-vector by approximately 20% (see Figure 5(a)). (2) When HMM is the SI model, the micro-F1 score of E-vector is almost equal to MFCC. HDCC is inferior to others by approximately 10% (see Figure 5(c)) and it is almost equal to MFCC and E-vector in the LibriSpeech (see Figures 5(b) and 5(d)).

In Figure 6, the identification results with the input signal containing five utterances are described. When using GMM as the SI model, the micro-F1 scores of E-vector and MFCC are almost equal, as shown in Figures 6(a) and 6(b). When using HMM as the SI model, the results of the MFCC and E-vector are almost the same level, as shown in Figures 6(c) and 6(d). The micro-F1 score of HDCC is less than MFCC and E-vector by approximately 20% (see Figure 6(a)), and it is a little inferior to others (see Figure 6(b)). In the LibriSpeech corpus, we can find in Figures 6(b) and 6(d) that both MFCC and E-vector show good performances with the score of micro-F1 over 0.96.

Since mirco-F1 score is a collaborative measure of precision and recall, it can better denote the identification performance and stability. Therefore, we calculated the average value of micro-F1 score (Avg. micro-F1) and standard deviation of the micro-F1 score (Std. Dev. micro-F1) in the experiments of multiple utterances (three utterances and five utterances) and compared the results of E-vector with MFCC and HDCC based on different models and corpus databases (see Table 6 and Figure 7).

In Figure 7(a), we can found that the both Avg. micro-F1 and Std. Dev. micro-F1 of E-vector were superior to MFCC and much better than HDCC. However, Figure 7(b) shows that E-vector obtained a similar level of Std. Dev. micro-F1 as MFCC. E-vector still outperformed MFCC and HDCC in comparison of Avg. micro-F1. Subsequently, we can also obtain the same results in Figures 7(c) and 7(d). Particularly, in case of TIMIT corpus (see Figure 7(c)), the Avg. micro-F1 of E-vector can be improved by 0.65% and 21.40% against MFCC and HDCC, and Std. Dev. micro-F1 of E-vector can be improved by 5.41% and 21.40% against the other two vocal features. In a word, the above investigations revealed that the average proportion of mistaking the matched utterance as the unmatched utterance of E-vector is less than MFCC and HDCC; namely, FRR of E-vector is lower in multiple-utterance SI tasks.

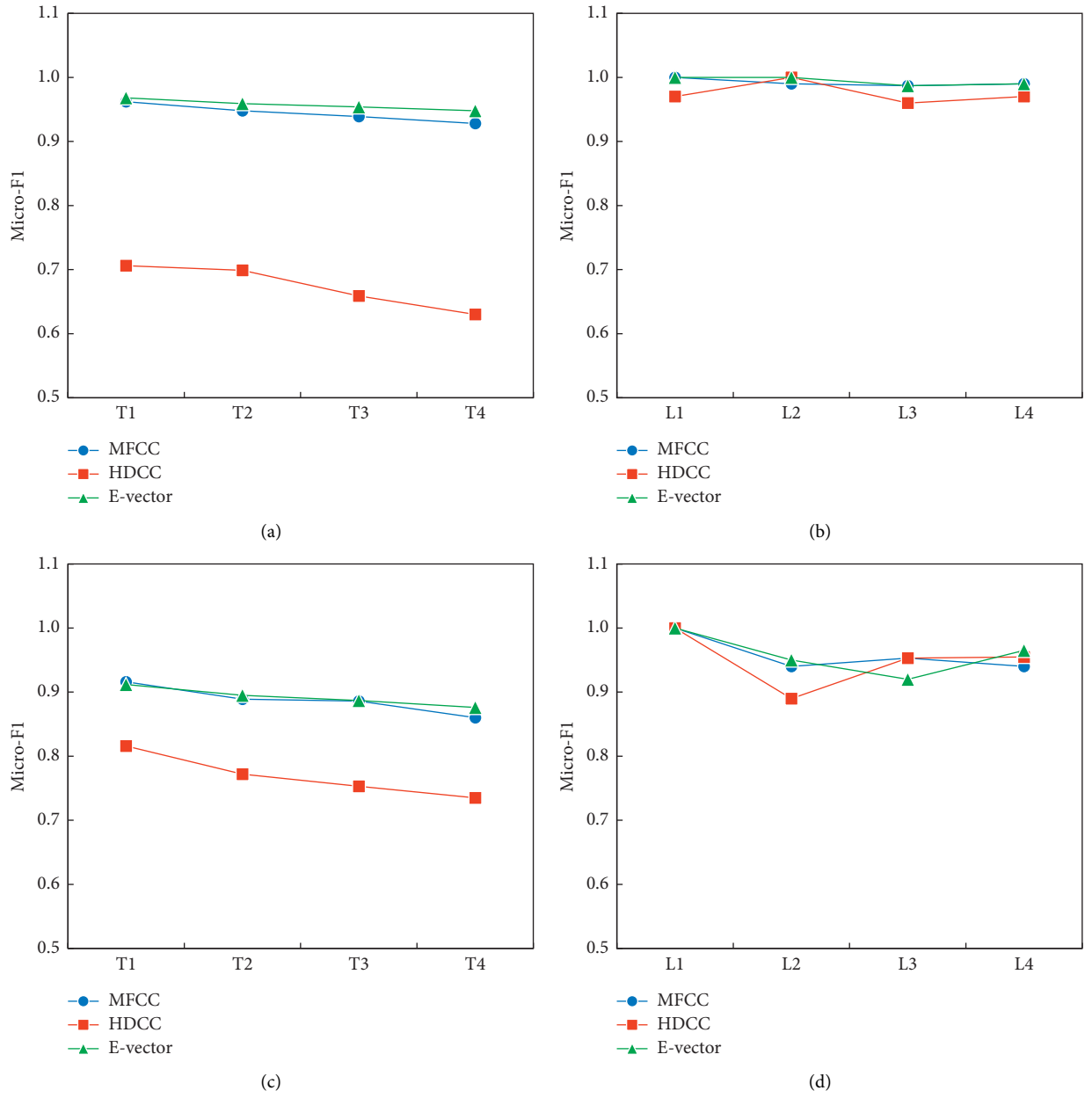


Figure 6: Five-utterance experiment micro-F1 score comparison. (a) TIMIT corpus using GMM. (b) LibriSpeech corpus using GMM. (c) TIMIT corpus using HMM. (d) LibriSpeech corpus using HMM.

Table 6: Comparison of Avg. micro-F1and Std. Dev. micro-F1 SD of multiple-utterance identification.

Micro-F1	Different models					
	GMM			HMM		
	E-vector	MFCC	HDCC	E-vector	MFCC	HDCC
Avg.	↑0.982	0.978	0.862	↑0.935	0.924	0.879
Std. Dev.	↓0.018	0.023	0.139	↓0.039	↓0.039	0.090
	Different corpus databases					
	TIMIT			LibriSpeech		
	E-vector	MFCC	HDCC	E-vector	MFCC	HDCC
Avg.	↑0.936	0.930	0.771	↑0.982	0.972	0.971
Std. Dev.	↓0.035	0.037	0.071	↓0.025	0.034	0.029

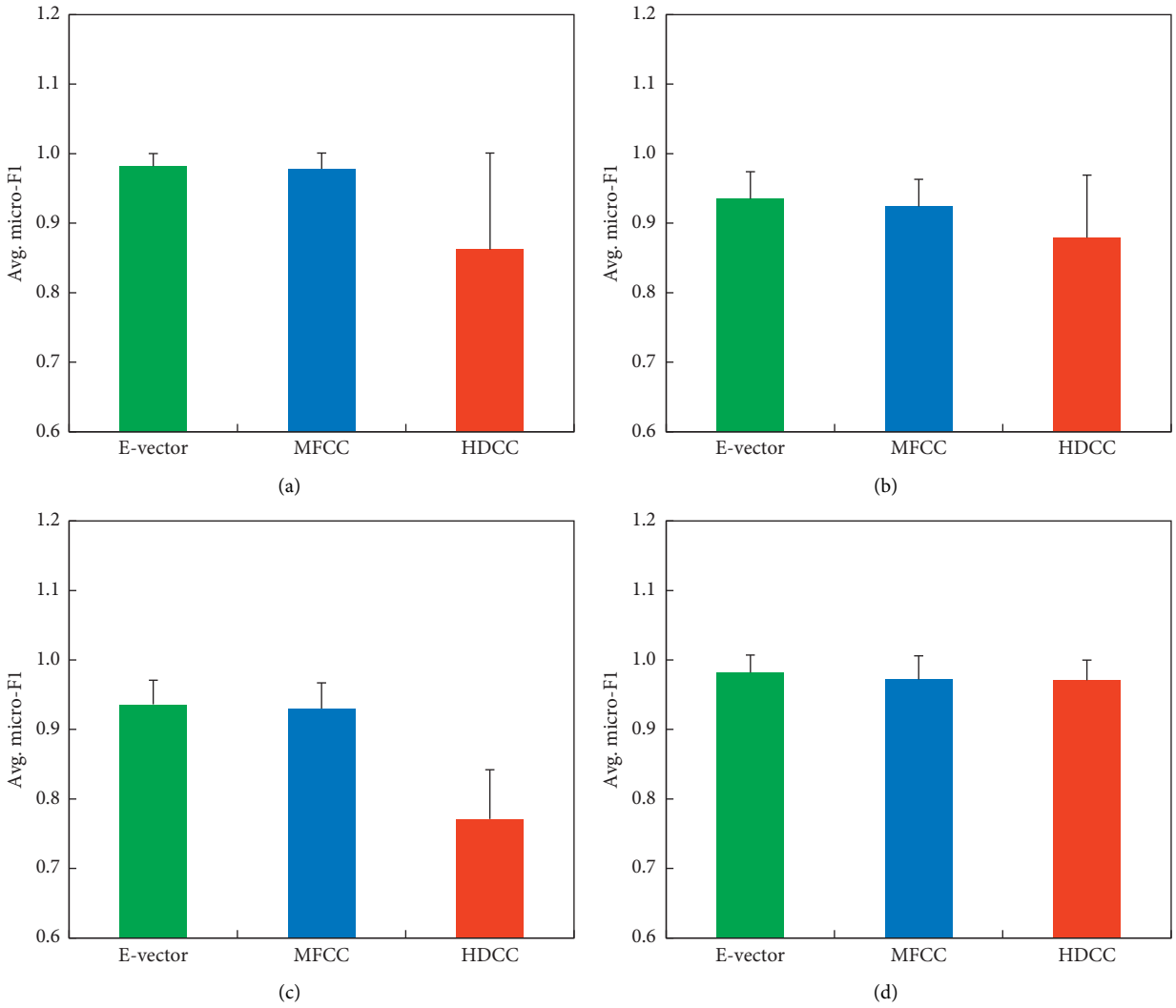


Figure 7: Stability comparison of E-vector, MFCC, and HDCC in different situations. (a) GMM-based SI using TIMIT and LibriSpeech corpus. (b) HMM-based SI using TIMIT and LibriSpeech corpus. (c) SI of TIMIT corpus using GMM and HMM. (d) SI of LibriSpeech corpus using GMM and HMM.

4. Conclusions

In this paper, we proposed a novel data-driven approach for vocal feature extraction based on DSS. Our method learns the E-vector with the minimization of the Euclidean metric using hierarchical analysis for the IDCT cepstrum coefficient, which is obtained by voice data preprocessing. Several different graphs of the experiments illustrate the effectiveness of our method in challenging SI tasks.

As a vocal feature extraction method, the generalization of E-vector is also significant. Our results show that E-vector has perfect identification performances in both one- and multiple-utterance experiments in different corpus databases, with an approximately 1.5% superiority to MFCC at best. It is also shown that our method is suitable to both GMM and HMM, with an approximately 2.1% average micro-F1 score superiority to MFCC at best. These advantages of the proposed method contribute to the capabilities of voice-feature extraction and enhance its usability for more real-world identification tasks. In our future work, we plan to investigate cosine similarity and correlation coefficient calculation methods to extract more optimized feature vectors.

Data Availability

The data and codes used in this article are given as follows: TIMIT corpus: http://academictorrents.com/details/34e2b78745138186976cbc27939b1b34d18bd5b3; LibriSpeech corpus: http://www.openslr.org/12/; E-vector source code: https://github.com/XiaoHe68/voice2vector.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

He Ma and Yi Zuo contributed equally to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grant nos. 61751202, U1813203, 61803064, 51939001, and 61976033), the Science and Technology Innovation Funds of Dalian (under Grant no. 2018J11CY022), the Liaoning Revitalization Talents Program (under Grant nos. XLYC1807046 and XLYC1908018), the Natural Science Foundation of Liaoning Province (under Grant nos. 2019-ZD-0151 and 2020-HYLH-26), and the Fundamental Research Funds for the Central Universities (under Grant no. 3132019345).

References

[1] D. Zhao, W. Luo, L. Ran, and L. Yue, "Negative iris recognition," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 1, pp. 112–125, 2015.

[2] T.-Y. Jea and V. Govindaraju, "A minutia-based partial fingerprint recognition system," *Pattern Recognition*, vol. 38, no. 10, pp. 1672–1684, 2005.

[3] R. Cappelli, M. Ferrara, and D. Maltoni, "Minutia cylinder-code: a new representation and matching technique for fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, 2010.

[4] K. Cao and A. K. Jain, "Automated latent fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 788–800, 2019.

[5] V. Bruce and A. Young, "Understanding face recognition," *British Journal of Psychology*, vol. 77, no. 3, pp. 305–327, 2011.

[6] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3244–3252, 2018.

[7] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 2005.

[8] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Computing*, vol. 2, 2010.

[9] S. Cai, X. Li, X. Zou et al., "Power normalized perceptional linear predictive feature for robust automatic speech recognition," in *Proceedings of INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, Osaka Japan, pp. 3022–3027, 2011.

[10] S. Al-Rawahy, A. Hossen, and U. Heute, "Text-independent speaker identification system based on the histogram of DCT-cepstrum coefficients," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 16, no. 3, pp. 141–161, 2012.

[11] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[12] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[13] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056, IEEE, Florence, Italy, May 2014.

[14] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5115–5119, IEEE, Shanghai, China, May 2016.

[15] Alibaba Group Holding Limited and Grand Cayman (KY), "Method of speech recognition and device thereof: CN 107564513 A," CN Patent 107564513 A, 2018.

[16] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[17] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, IEEE, Brisbane, Australia, April 2015.