www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Archives of Acoustics Vol. 49, No. 1, pp. 37–48 (2024), doi: 10.24425/aoa.2024.148768

Research Paper

# Short Utterance Speaker Recognition Based on Speech High Frequency Information Compensation and Dynamic Feature Enhancement Methods

Yunfei ZI*, Shengwu XIONG

*School of Computer and Artificial Intelligence, Wuhan University of Technology*
Wuhan, China

*Corresponding Author e-mail: yfzi@whut.edu.cn

(*received April 5, 2022; accepted September 20, 2023; published online January 8, 2024*)

This work aims to further compensate for the weaknesses of feature sparsity and insufficient discriminative acoustic features in existing short-duration speaker recognition. To address this issue, we propose the Bark-scaled Gauss and the linear filter bank superposition cepstral coefficients (BGLCC), and the multi-dimensional central difference (MDCD) acoustic feature extracted method. The Bark-scaled Gauss filter bank focuses on low-frequency information, while linear filtering is uniformly distributed, therefore, the filter superposition can obtain more discriminative and richer acoustic features of short-duration audio signals. In addition, the multi-dimensional central difference method captures better dynamics features of speakers for improving the performance of short utterance speaker verification. Extensive experiments are conducted on short-duration text-independent speaker verification datasets generated from the VoxCeleb, SITW, and NIST SRE corpora, respectively, which contain speech samples of diverse lengths, and different scenarios. The results demonstrate that the proposed method outperforms the existing acoustic feature extraction approach by at least 10% in the test set. The ablation experiments further illustrate that our proposed approaches can achieve substantial improvement over prior methods.

**Keywords:** Bark-scaled Gauss; linear filter; filter bank superposition; multi-dimensional central difference; speaker recognition.



## 1. Introduction

Speaker recognition, as one of the most popular biometric technologies (Wu *et al.*, 2016) today has been widely used in many fields such as access control, forensic evidence provision, security, and telephone banking user authentication (Vogt *et al.*, 2010). The purpose of speaker recognition is to recognize the claimed identity of the speaker, which includes speaker verification and speaker identification (Campbell, 1997). One of its main purposes is to determine whether the test sound from the speaker is acceptable. After decades of development, the technology of speaker verification has been extensively studied, and the recognition system has achieved relatively satisfactory performance, provided that the enrollment and test voices are long enough and the signal-to-noise ratio (SNR) is large enough (Zinchenko *et al.*, 2017; Greenberg *et al.*, 2013; Kinnunen, Li, 2010).

However, in some application scenarios, it is not easy to collect a suitable speech. The current speaker verification system has a significant decrease of the recognition rate in a short utterance environment (Nosratighods *et al.*, 2010). A short-duration speech means that the speech contains insufficient acoustic characteristics. Obtaining enough speech data is difficult for many real-world applications and users are reluctant to provide sufficient voice data, especially during the testing phase asking the user to speak for a long time, for instance in phone banking. In other cases, it is very difficult to collect enough data, e.g., in forensic applications, in the security field. The performance degradation caused by insufficient data is called the short-duration issue.

Current speaker recognition systems have achieved great success and performed well when the enrollment and test data are sufficiently long; hence, the traditional acoustic feature extraction methods are designed

based on long-duration speech, and the long-duration speech feature extraction filter arrangement method mainly focuses on the low-frequency domain, this makes high-frequency domain features more sparse in the short duration speech, and high-frequency domain information best represents timbre and detail (Huang, Pun, 2020). At the same time, the traditional acoustics features include fewer dynamic features of speakers, as a result, fewer acoustic features are extracted that can be discriminated for speaker recognition. Research on the more challenging short-duration text-independent speaker recognition of discriminative feature compensation has been more in demand lately, which is also our focus in this work.

Although the traditional speaker model has obvious feature specificity, because the number of features is too few, it is still susceptible to noise interference, and awful recognition performance. The acoustic feature extraction design should address how to extract the high discriminative embeddings more effectively in short-duration audio speaker recognition. Therefore, how improving the effectiveness of discriminative acoustic feature extraction, in short utterance speaker environment, is an urgent problem to be solved.

To address the problems, the solution is proposed in this paper. In the Bark-scaled Gauss filter bank acoustic feature extraction method the filter bank distribution puts more emphasis on the low-frequency bands, which portray the low-frequency spectrum of speech in great detail. In comparison, the Bark-scaled Gauss filter distribution less emphasizes the high-frequency bands, so some helpful information is easily lost from the high-frequency domain. However, the details of the high frequency can enhance the information of one's timbre. To enhance the valuable information on the high-frequency, the Bark-scaled Gauss and linear filter bank superposition cepstrum coefficients (BGLCC) are proposed to portray more precise high-frequency details. The filter bank of the conventional acoustic feature extraction method puts more emphasis on the low-frequency band. In contrast, the linear triangle filter is uniformly distributed, which can remedy the weakness of the sparse high-frequency information and insufficient acoustic feature extraction brought by the uneven distribution of a single filter, thus, integrating the advantages of both and constructing new hybrid feature parameters is a way to enhance the feature sparsity problem.

Moreover, aiming to capture better dynamics features of speakers, we propose multi-dimensional central difference (MDCD) features based on the BGLCC features matrix, simultaneously, to improve the performance of short utterance speaker recognition. The MDCD are multi-dimensional central difference features in the time-frequency plane. Different speakers speak the same word or sentence in different ways. The proposed MDCD feature concatenate information about the speaker from four different dimensions, this can explain why it performs significantly better than traditionally used speech features in speaker recognition tasks under various conditions. Therefore, the MDCD features can further compensate for the limited and sparse dynamic acoustic characteristics of short-duration audio signals based on extracting dynamic speaker features.

### *1.1. Related works*

To enhance the efficiency of performance of short-duration audio speaker recognition algorithms, some approaches have been presented by previous research studies. In terms of front-end acoustic feature extraction, the vast majority of existing acoustic feature extraction is based on some form of the short-term frequency spectrum to implement short utterance speaker recognition algorithms like Mel-frequency cepstral coefficients (MFCCs) (Herrera-Camacho *et al.*, 2019; Paseddula, Gangashetty, 2018) linear prediction cepstral coefficients (LPCCs) (Yang *et al.*, 2019; Atal, 1974) and constant Q cepstral coefficients (CQCC) (Todisco *et al.*, 2017), acoustic features. For instance, by judiciously combining MFCC and LPCC for short-duration audio signal speaker recognition (Chowdhury, Ross, 2020), the hypothesis is that MFCC and LPC capture two different aspects of speech, namely, speech perception and speech production. By using the model method, there is speaker recognition based on GMM-UBM from MFCC features in the limited enrollment and test data (Omar, Pelecanos, 2010). Another work is the I-vector approach and factor analysis subspace estimation introduced by (Kenny *et al.*, 2005; Dehak *et al.*, 2010) to reduce the number of redundant model parameters, resulting in more accurate speaker models. Some approaches attempt to increase performance by selecting segments with better discriminability based on speaker features (Nosratighods *et al.*, 2010) GMM and the CNN hybrid method (Liu *et al.*, 2018), the method is an initial alignment method for short utterance feature, which can improve the effect of short utterance speaker recognition. In their work, front-end feature extraction methods are based on Fourier transform Mel-triangle filtering and linear prediction cepstral coefficients for model training and testing as well as model inference.

With further developments in deep learning, various methods for speaker recognition or short utterance speaker recognition have been proposed, by Povey *et al.* (2018), the factorized time delay neural network (F-TDNN) has been proposed which divides the parameter matrix of TDNN into smaller matrices to increase the training effectiveness and the extended time delay neural networks (E-TDNN) was proposed in (Snyder *et al.*, 2019), E-TDNN is based on its broader and deeper network structure, thus allow-

ing more information to be learned, they both improve speaker recognition performance significantly. In (Villalba *et al.*, 2020), based on F-TDNN and E-TDNN, the best results were obtained for speaker evaluation in SRE18 and in the field. In addition, a focus on aggregation information, channel attention, and propagation method were proposed (Desplanques *et al.*, 2020), called TDNN-based speaker verification (ECAPA-TDNN), which further improves the robustness of speaker recognition. After years of development, the performance of short utterance speaker recognition has improved considerably, but it is still unsatisfactory in some complex scenarios.

Most of the aforementioned methods would benefit from the optimization model, enhance data characteristics and extract more discriminative features for speaker recognition. With 5~10 seconds of speech duration, they all improve speaker recognition performance when audio speech becomes shorter, but they still face significant challenges.

Generally speaking, there are two types of speech recognition features, namely linear prediction cepstral coefficients (LPCCs) and Mel-frequency cepstral coefficients (MFCCs), but when used in a short-duration environment, they suffer from a drop in performance. As we know, there is no reasonably good short-duration speaker verification model. Unfortunately, there is no better feature extraction method to obtain sufficient and discriminative speaker information models from short-duration speech signals, there are no better training methods.

### *1.2. Contribution*

To compensate for the problems of difficult short-utterance discriminative feature capture and insufficient discriminative acoustic features, we propose a filter superposition-based multi-dimensional central difference discriminative acoustic feature extraction method for feature compensation and enhancement of short-duration speech speaker recognition. The proposed method can significantly improve the performance and accuracy of the the short-duration speech speaker recognition system.

The contributions of this paper:

- we propose the Bark-scaled Gauss and linear filter bank superposition acoustic feature extraction method, which compensates for the weakness of the sparse filter and the sparse feature in the high-frequency information for a short utterance feature, this can improve the performance of short utterance speaker recognition by providing rich timbre information;
- we propose the multi-dimensional central difference method for capturing the dynamic features of speakers, which is used to simulate real speech and enhance the diversity of acoustic features with limited speech data.

### *1.3. Organization*

This paper is organized as follows. Section 2 details the proposed filter superposition-based multi-dimensional central difference discriminative acoustic feature extraction method. Then we analyze the experiments and results of the proposed method in Sec. 3. Finally, the conclusion is given in Sec. 4.

## 2. Proposed method

In this section, which mainly includes the discriminative acoustic feature extraction algorithm, we elaborate on the proposed feature extraction technique, which the design based on the Bark-scaled Gauss and linear filter banks superposition algorithm and then the multi-dimensional central difference dynamic features extraction method based on the BGLCC features matrix. In addition, the effect of the introduced feature extraction of BGLCC and MDCD was achieved through mathematical analysis.

### *2.1. BGLCC feature extraction method*

The speech signal is performed by a high-pass filter as pre-emphasis, this filter is equivalent to:

$$H(z) = 1 - az^{-1}, \tag{1}$$

where $a$ is a pre-emphasis coefficient, the value is chosen in the interval [0.95, 0.97] and it can increase the energy of higher frequencies.

The following Hamming window $w$ is used for smoothing the edge of framed speech signals:

$$w(k) = \left[0.54 - 0.46\cos\left(\frac{2\pi k}{K-1}\right)\right] R_K(k), \tag{2}$$

where $K-1$ is the window length, $K-1$ equals 256, $0 \le k \le K-1$, $R_K(k)$ equals rectangular window.

In speech processing, the Bark-frequency cepstrum (BFC) affects the speech short-term power spectrum, which is transformed on the Bark-scale of frequency. The BFC can be obtained as:

$$F_{\text{Bark}}(f) = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f}{7500}\right)^2. \tag{3}$$

In contrast to the well-known Mel-scaled triangular filter, the proposed Bark-scaled Gauss filter structure has a smoother response and enhances the correlation between adjacent sub-bands. The coefficients are derived from a type of cepstral representation of the speech clip. The frequency response of the Bark-scaled Gauss filter bank can be obtained as:

$$H_{\text{Bark}_b}(k) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{\frac{-[k-f(b)]^2}{2\sigma_b^2}}, \tag{4}$$

where $\sigma_b$ is the standard deviation, and $f(b)$ is the $b$-th filter boundary point (Bark-scaled center frequency), as defined:

$$\sigma_b = \frac{f(b+1) - f(b)}{\alpha}, \tag{5}$$

where $\alpha$ is equal to 2.0.

The signal presents 24 critical bands in the band, which is also the Bark center frequency, and this is the Bark domain.

Next, the linear triangle filter bank processing details. The power spectrum is then processed, on the frequency, by a linear uniform filter bank. In these linear filter banks, each filter is a triangle filter. The filter can be defined as:

$$H_{\text{Linear}_l}(k) = \begin{cases} 0 & k < f(l-1), \\ \dfrac{k - f(l-1)}{f(l) - f(l-1)} & f(l-1) \le k < f(l), \\ 1 & k = f(l), \\ \dfrac{f(l+1) - k}{f(l+1) - f(l)} & f(l) < k \le f(l+1), \\ 0 & k > f(l+1), \end{cases} \tag{6}$$

where $f(l)$ is the center frequency, $0 \le l < L$, and $L$ is the number of filter banks, and the value of $L$ is 24. We use more filter bands than usual on account that the resolution of high-frequency domains is essential for the timbre. Finally, we get the linear filter features.

The raw speech signal $x(n)$ is preprocessed to obtain $x_w(n)$. Subsequently, the fast Fourier transform of the framed speech signal to transform the speech data from the time domain to the frequency domain, the mathematical calculation can be written as:

$$X(i,n) = \text{FFT}[x_w(i,n)], \tag{7}$$

where $x_w(n)$ indicates that after adding the window function $i$ is the number of speech frames.

The power spectrum is calculated as:

$$E(i,n) = |X(i,n)|^2. \tag{8}$$

Therefore, the Bark-scaled Gauss and linear filter banks superposition feature extraction is made based on the power spectrum of the output from the fast Fourier transform. Thus, the BGLCC power calculation procedure can be given by:

$$S(i,t) = \sum_{k=0}^{N-1} E(i,n)[H_{\text{Bark}_b}(k) + H_{\text{Linear}_l}(k)], \qquad 0 \le b \le u, \qquad 0 \le l \le v, \tag{9}$$

where $t$ denotes the $t$-th superposition filter, $b$ denotes the $b$-th Bark-scaled Gauss filter, and $l$ denotes the $l$-th linear triangle filter, respectively, $u$ is the number of the Bark-scaled Gauss filter, $v$ is the number of the linear triangle filter, $t$, $u$, $v$ all are 48-channel filter banks; $S(i,t)$ is equivalent to multiplying the power spectrum $E(i,n)$ and the superposition of $H_{\text{Bark}_b}(k)$, the Bark-scaled Gauss filter and $H_{\text{Linear}_l}(k)$ the linear triangle filter on the frequency domain.

$$\text{BGLCC}(i,t) = \sum_{t=0}^{T-1} \log[S(i,t)] \cos\left[\frac{\pi r(2t-1)}{2T}\right], \tag{10}$$

where $S(i,t)$ is the BGLCC power, $i$ denotes the $i$-th frame, $r$ is the spectral line after discrete cosine transformation, $t$ denotes the $t$-th superposition filter, $T$ is the number of superposition filters, and the value of $T$ is 48.

The Bark-scaled Gauss and linear filter bank superposition features (BGLCC) are processed as shown in Fig. 1.

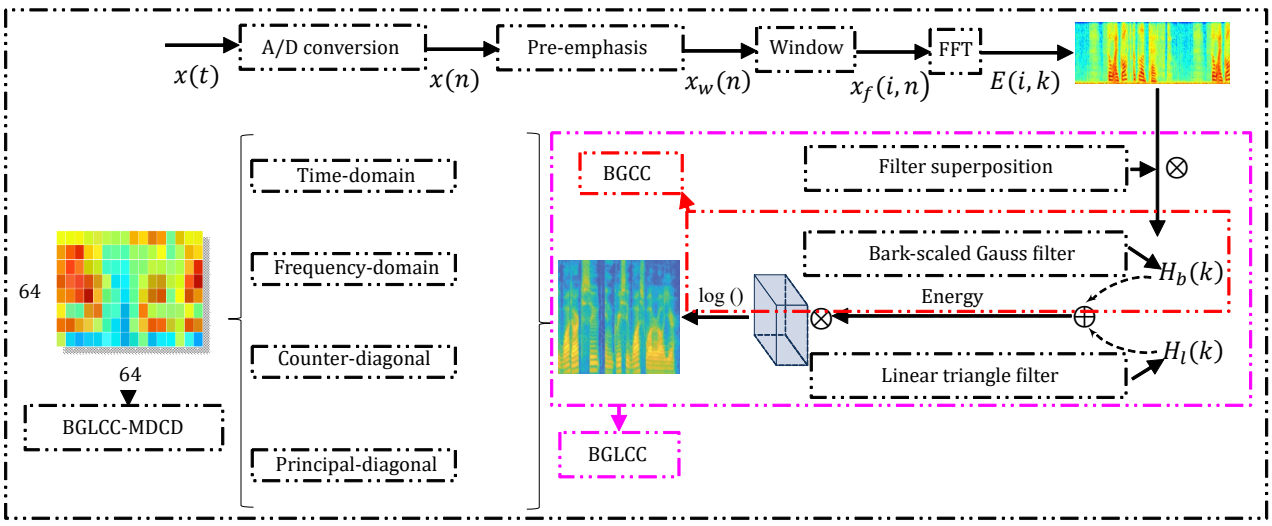


Fig. 1. Structure of the proposed acoustic features extraction method.

### *2.2. MDCD dynamic feature extraction method*

The proposed multi-dimensional central difference dynamic feature extraction method was applied to the different dimensions of the BGLCC time-frequency matrix, where the horizontal dimension is the time domain axis and the vertical is the frequency domain axis dimension and it captures speech time-domain relevance and speech high-low-frequency correlation of the speaker. Similarly, the central difference of linear regression is applied to the time-frequency matrix principal diagonal and counter diagonal, therefore it can capture the voiceprint of the speaker.

The process of the proposed method is shown in Fig. 1; MDCD dynamic feature extraction of different dimensions on the BGLCC time-frequency matrix. First, a series of pre-processing is performed on a frame of the speech signal, which converts the input signal from a time-domain speech signal to a frequency-domain speech signal. Next, the proposed Bark-scaled Gauss and linear filter bank features superposition is applied to divide the spectrum into certain frequency bands, and the log compression is applied. Then, multi-dimensional central difference obtains four different types of features based on the BGLCC time-frequency matrix, which are calculated as in Eqs. (11)–(14):
time-domain:

$$T_h = M^t_{t,f} = \frac{M_{t+1,f} - 2M_{t,f} + M_{t-1,f}}{h^2}, \tag{11}$$

frequency-domain:

$$F_h = M^f_{t,f} = \frac{M_{t,f+1} - 2M_{t,f} + M_{t,f-1}}{h^2}, \tag{12}$$

counter-diagonal domain:

$$P_h = M^P_{t,f} = \frac{M_{t+1,f+1} - 2M_{t,f} + M_{t-1,f-1}}{h^2}, \tag{13}$$

principal-diagonal domain:

$$C_h = M^C_{t,f} = \frac{M_{t+1,f-1} - 2M_{t,f} + M_{t-1,f+1}}{h^2}. \tag{14}$$

In these equations, the value of $h$ is 2, as the central difference of linear regression has been applied. Here, $t$ stands for the time domain axis and $f$ stands for the frequency domain axis. $M$ is the point along which different dimensions of the axis have been applied.

The time domain's central difference and the frequency domain's central difference can better capture the contour of the speaker formants. By doing the matrix principal diagonal's central difference and matrix counter diagonal's central difference, speaker information about the uttering text phoneme of each speaker can be captured. Thus, the different dimensions of the time-frequency spectrum central difference can be regarded as multi-dimensional dynamic speaker information of each speaker and this explains the excellent results of the proposed MDCD features. To reduce the computationally derived high-dimensional MDCD features, we compress and decorrelate them by DCT.

It was our goal to perform speaker verification through the proposed BGLCC-MDCD as acoustic features, and use 34-layer ResNet as the backbone model, to perform the short-duration speaker verification. The detailed configuration is listed in Table 1.

Table 1. Detailed configuration of the backbone model of 34-layer ResNet. The input size is $T \times 64$.

| Layer | Structure | Output shape |
|---|---|---|
| Conv0 | CNN $(7 \times 7, 32)$, stride 2 | $T \times 64 \times 32$ |
| Conv1 | $\begin{pmatrix}(3 \times 3, 32)\\(3 \times 3, 32)\end{pmatrix} \times 3$, stride 2 | $T/2 \times 32 \times 32$ |
| Conv2 | $\begin{pmatrix}(3 \times 3, 64)\\(3 \times 3, 64)\end{pmatrix} \times 4$, stride 2 | $T/2 \times 16 \times 64$ |
| Conv3 | $\begin{pmatrix}(3 \times 3, 128)\\(3 \times 3, 128)\end{pmatrix} \times 6$, stride 2 | $T/2 \times 8 \times 128$ |
| Conv4 | $\begin{pmatrix}(3 \times 3, 256)\\(3 \times 3, 256)\end{pmatrix} \times 3$, stride 2 | $T/2 \times 4 \times 256$ |

## 3. Experiments and analysis

### *3.1. Experiments*

The short-duration speaker verification experiments presented in this paper are conducted using the three well-known speaker recognition datasets with different scenarios: VoxCeleb (Nagrani *et al.*, 2017; Chung *et al.*, 2018), Speaker in the Wild (SITW) (McLaren *et al.*, 2016), and the NIST SRE 2010 (Martin, Greenberg, 2010) to evaluate our proposed algorithm.

The short-duration text-independent dataset is generated from the VoxCeleb, SITW, and NIST SRE corpus, respectively. After removing silence frames using an energy-based VAD, the speech utterances are chopped into short segments (ranging from 0.25 to 10 seconds). This is to illustrate the efficiency of the work of our proposed method under short-duration audio conditions.

The three different scenarios of speech datasets: VoxCeleb, SITW, and NIST SRE corpus aim to evaluate the generalizability of the methods across a range of different audio lengths of scenarios. We focus on conducting speaker verification trials on voice samples of different speech lengths, which are used to investigate the effect of testing speech sample length changes and to validate the efficiency of the presented method on the performance of the speaker verification method. One thing to keep in mind is that in all of our tests, we assume that there is only one speaker in each voice sample and that there is no overlapping voice from several speakers in any of the training or testing speeches.

### 3.2. Corpus description

#### 3.2.1. VoxCeleb and SITW corpus

VoxCeleb is a large open-source speaker recognition dataset with over a million utterances, 7000 speakers, and 2000 hours of audio. The average duration of utterances in the VoxCeleb dataset is 8 seconds, and the majority of utterances have a duration of fewer than 10 seconds. The audio sampling rate is 16kHz. VoxCeleb includes two sub-datasets, VoxCeleb-1 and VoxCeleb-2. The SITW dataset contains open-source media recordings of 299 public celebrities. The SITW dataset is used to generate the short-duration text-independent dataset. SITW speech segments range in length from 6 seconds to 180 seconds, where the majority are long utterances. As a result, the two datasets can be used to assess the performance of our proposed architectures on utterances of varying lengths as well as the model's generalizability.

Each of the three datasets, VoxCeleb-1, VoxCeleb-2, and SITW, is divided into two parts: development and testing (evaluation). The training set consists of 1 092 009 utterances and 5994 speakers from the VoxCeleb-2 development part (VoxCeleb2-Dev). The remaining datasets were treated as test sets, with two parts: the VoxCeleb-1 dataset and the SITW evaluation (SITW-Eval) set. There are 4706 utterances and 37 611 trials in the VoxCeleb-1. There are 1202 utterances and 721 788 trials in the SITW evaluation (SITW-Eval).

#### 3.2.2. NIST SRE corpus

The NIST SRE corpus was used to generate the short-duration text-independent dataset. The SRE04-08, Switchboard II phase 2, 3, and Switchboard Cellular Part 1, Part 2 comprise the training set. The final training set includes 4000 speakers with 40 short utterances each. Similarly, the enrollment and test sets are derived from NIST SRE 2010. The enrollment speech includes 150 male and 150 female speakers, each of whom is enrolled by five utterances. The 4500 utterances in the enrollment speech data are used to test from the same 300 speakers. The trial list that was generated contains 392 660 trials. The website GitHub provides access to the trial list and the comprehensive segmentation files.

### 3.3. Feature extraction

All experiments use a 64-dimensional input feature from a 25 ms window with a 10 ms frameshift. The experiments evaluate using features: LPCC, MFCC, MFCC-LPCC, the proposed BGCC, BGLCC, and BGLCC-MDCD. The 64-dimensional features were extracted for LPCCs, with 32 for linear regression along the time axis and 32 along the frequency axis. The MFCCs used 64-dimensional features, and the 64-dimensional MFCC-LPCC features contain 32-dimensional MFCC and LPCC features, respectively. The use of delta 1/2 inputs is also a 64-dimensional feature. For the proposed acoustic feature, BGCC, BGLCC, the 64-dimensional feature vector has been extracted, BGCC-MDCD, BGLCC-MDCD, which contain 16 time-domain features, 16 frequency-domain features, 16 counter-diagonal domain features, 16 principal-diagonal domain features, respectively.

### 3.4. Loss function

In (Schroff *et al.*, 2015), the triplet loss was initially proposed to learn discriminatory image embedding. The embeddings need to satisfy the following relationship for model training to be successful. The cosine triplet embedded *Loss* (Zhang *et al.*, 2018) for training the model is:

$$\left\|f(s_i^a) - f(s_i^p)\right\|_2^2 + \alpha_{\text{margin}} < \left\|f(s_i^a) - f(s_i^n)\right\|_2^2, \quad \forall (f(s_i^a), f(s_i^p), f(s_i^n)) \in \tau, \tag{15}$$

$$L = \sum_i^N \left[\left\|f(s_i^a) - f(s_i^p)\right\|_2^2 - \left\|f(s_i^a) - f(s_i^n)\right\|_2^2 + \alpha_{\text{margin}}\right]. \tag{16}$$

The cosine triplet embedding the loss function $L$ is used here, where $\tau$ is the batch of triplet, with $(s_i^a, s_i^p, s_i^n)$ is a triplet. $N$ is the batch size. Samples of speech from a specific "$a$" are $s_i^a$, the anchor sample, and $s_i^p$, the positive sample with the same person. The negative sample, $s_i^n$, is a sample of speech from another person "$b$", so that $a \neq b$. The $\alpha_{\text{margin}}$ is a user-tunable hyper-parameter at the value of 0.25 that determines the minimum distance between negative and positive speech samples.

### 3.5. Implementation and reproducibility

The proposed discriminative acoustic feature method uses the PyTorch (Paszke *et al.*, 2017) toolkit to conduct the experiment, and training using the Triplet-loss (Schroff *et al.*, 2015). The initial learning rate is 0.001 and lasts for 200 epochs. The experiment embeds the cosine triplet loss, and the value of the $\alpha_{\text{margin}}$ hyper-parameter is 0.25, which is the best trade-off. The network is optimized using the Adam optimizer with a minibatch size of 32 and softmax as a classifier. The fully connected layers after the statistic pooling layer have 512 nodes. The training was done on a single Nvidia A100 GPU.

### 3.6. Evaluation metrics

We use the following metrics to evaluate the model performance: the Equal Error Rate (EER, in %), and

the minimum detection cost function at the prior probability of specifying the targeted speaker of (MinDCF*100), which is a standard-setting (Nagrani *et al.*, 2017), and partial AUC (pAUC) with $\alpha = 0$ and $\beta = 0.05$, the pAUC represents the partial area under the ROC curve, it meets the evaluation requirement of real-world applications that work on different parts of ROC curves. It is a supplement evaluation metric to the existing metrics. The pAUC is defined by two false positive rate (FPR) parameters: $\alpha$ and $\beta$, which is a detailed calculation (Bai *et al.*, 2020). The pAUCMetric evaluates the similarity between two speaker features by the squared Mahalanobis distance.

### *3.7. Results and analysis*

#### *3.7.1. Overall performance*

*Performance comparison of different acoustic features.* Table 2 and Fig. 2 show the performance of our proposed acoustic features and the compared acoustic features on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, respectively. Table 2 lists the results in terms of EER, Min-DCF, and pAUC, Fig. 2 plots the detection error trade-off (DET) curves of different acoustic features under 10 s speech length that include no dynamic features, using delta 1/2 dynamic features and using MDCD dynamic features. The acoustic feature extraction level for the short-duration audio signal, contains three conventional baseline features, which are MFCC, LPCC, MFCC-LPCC, and our proposed BGCC and BGLCC acoustic features. The speech length ranges from 0.25 to 10 seconds, including 3 segments.

From Table 2, on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, it can be observed that BGLCC-MDCD acoustic feature significantly outperforms MFCC, LPCC, and MFCC-LPCC in terms of EER, Min-DCF, and pAUC, and BGLCC-MDCD acoustic feature achieves better performance in short-duration speaker verification.

Across the LPCC experiment in Table 2, on the VoxCeleb-1 dataset, compared to LPCC features, the proposed BGLCC features improve by 15.0%, compared to LPCC-delta1/2 features, BGLCC-MDCD features improve 19.0%, under 2 s duration speech length in terms of EER.

Across the MFCC experiment in Table 2, on the VoxCeleb-1 dataset, compared to MFCC features, the proposed BGLCC features improve by 10.6%, compared to MFCC-delta1/2 features, BGLCC-MDCD features improve 15.0%, under 2 s duration speech length in terms of EER.

Across the MFCC-LPCC experiment in Table 2, on the VoxCeleb-1 dataset, compared to MFCC-LPCC features, the proposed BGLCC features improve by 9.1%, compared to MFCC-LPCC-delta1/2 features, BGLCC-MDCD features improve 13.3%, under 2 s duration speech length in terms of EER.

At the same time, on the other speech with different lengths from VoxCeleb-1, SITW, and NIST SRE 2010 datasets, the proposed BGLCC-MDCD acoustic features for short-duration speaker verification achieve better performance, compared with conventional MFCC, LPCC, and MFCC-LPCC fusion acoustic features. The comparison of the performance of the baseline is shown in Table 2.

In order to visualize the effectiveness of our proposed acoustic features on the different length speech, we plot detection error trade-off (DET) curves for all comparable features, as illustrated in Fig. 2. The performance advantage of proposed BGLCC and MDCD can also be seen from the DET curves in Fig. 2. For example, the results of experiment 1 present the DET curves of the LPCC acoustic feature under three conditions: no dynamic features, using delta 1/2 dynamic features, and using our MDCD dynamic features, under 10 s speech length on the VoxCeleb-1 dataset; the results of experiment 2 present the DET curves of the LPCC acoustic feature under three conditions: no dynamic features, using delta 1/2 dynamic features, and using our MDCD dynamic features, under 10 s speech length on the SITW dataset; the results of experiment 3 present the DET curves of the LPCC acoustic feature under three conditions: no dynamic features, using delta 1/2 dynamic features, and using our MDCD dynamic features, under 10 s speech length on the NIST SRE 2010 dataset.

Similarly, experiments 4–6 represent the DET curves of the MFCC acoustic feature under three conditions, on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, respectively; experiments 7–9 represent the DET curves of the MFCC-LPCC acoustic feature under three conditions, on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, respectively; experiments 10–12 represent the DET curves of the BGCC acoustic feature under three conditions, on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, respectively; and experiments 13–15 represent the DET curves of the BGLCC acoustic feature under three conditions, on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, respectively.

The experimental results also show the lower DET curves achieved using our proposed MDCD dynamic features, compared to no dynamic features, and using delta 1/2 dynamic features on VoxCeleb-1, SITW, and NIST SRE 2010 datasets.

The proposed MDCD dynamic acoustic feature achieves lower EER, Min-DCF, and highest pAUC than delta 1/2, thus demonstrating that the proposed multi-dimensional central difference dynamic features perform better and are more effective than single-dimensional dynamic features. The results of that comparison are listed in Table 2.

Table 2. Comparison results of different acoustic features and proposed acoustic features under varying audio lengths using the ResNet-34 network on VoxCeleb-1, SITW, and NIST SRE 2010 datasets.

| Features | Delta2 | MDCD | Duration [s] | VoxCeleb-1 | | | SITW | | | NIST SRE 2010 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | EER [%] | MinDCF | pAUC [%] | EER [%] | MinDCF | pAUC [%] | EER [%] | MinDCF | pAUC [%] |
| LPCC | – | – | 0.25 | 11.19 | 32.94 | 75.38 | 13.22 | 36.31 | 70.52 | 12.01 | 34.17 | 74.43 |
| | √ | – | | 11.18 | 32.92 | 75.39 | 13.20 | 36.29 | 70.54 | 12.00 | 34.15 | 74.44 |
| | – | √ | | **11.13** | **32.83** | **75.46** | **13.15** | **36.24** | **70.66** | **11.95** | **34.09** | **74.54** |
| | – | – | 2 | 3.17 | 17.99 | 95.38 | 5.53 | 23.41 | 92.37 | 4.48 | 23.03 | 93.45 |
| | √ | – | | 3.16 | 17.98 | 95.39 | 5.52 | 23.40 | 92.40 | 4.46 | 23.01 | 93.47 |
| | – | √ | | **3.11** | **17.92** | **95.46** | **5.47** | **23.35** | **92.46** | **4.40** | **22.95** | **93.58** |
| | – | – | 10 | 1.61 | 10.33 | 98.01 | 3.60 | 19.17 | 94.96 | 2.54 | 12.10 | 96.73 |
| | √ | – | | 1.60 | 10.32 | 98.03 | 3.58 | 19.16 | 94.98 | 2.52 | 12.09 | 96.75 |
| | – | √ | | **1.54** | **10.27** | **98.10** | **3.53** | **19.10** | **95.04** | **2.46** | **12.02** | **96.84** |
| MFCC | – | – | 0.25 | 11.04 | 32.52 | 75.74 | 12.52 | 35.83 | 72.23 | 11.50 | 32.93 | 75.01 |
| | √ | – | | 11.02 | 32.51 | 75.75 | 12.51 | 35.82 | 72.26 | 11.48 | 32.92 | 75.03 |
| | – | √ | | **10.98** | **32.47** | **75.79** | **12.45** | **35.74** | **72.44** | **11.44** | **32.86** | **75.12** |
| | – | – | 2 | 3.01 | 17.90 | 95.40 | 4.46 | 23.01 | 93.53 | 3.33 | 18.07 | 95.27 |
| | √ | – | | 3.00 | 17.88 | 95.41 | 4.45 | 23.00 | 93.54 | 3.32 | 18.05 | 95.28 |
| | – | √ | | **2.95** | **17.84** | **95.60** | **4.40** | **22.95** | **93.58** | **3.27** | **17.99** | **95.31** |
| | – | – | 10 | 1.37 | 10.12 | 98.34 | 3.24 | 17.96 | 95.32 | 2.11 | 10.83 | 97.62 |
| | √ | – | | 1.37 | 10.11 | 98.35 | 3.23 | 17.95 | 95.33 | 2.10 | 10.81 | 97.64 |
| | – | √ | | **1.36** | **10.04** | **98.40** | **3.19** | **17.66** | **95.51** | **2.05** | **10.75** | **97.74** |
| MFCC-LPCC | – | – | 0.25 | 10.97 | 32.47 | 75.82 | 12.42 | 35.73 | 72.47 | 11.41 | 32.82 | 75.17 |
| | √ | – | | 10.96 | 32.46 | 75.83 | 12.41 | 35.72 | 72.49 | 11.40 | 32.81 | 75.19 |
| | – | √ | | **10.90** | **32.42** | **75.86** | **12.34** | **35.70** | **72.58** | **11.35** | **32.75** | **75.30** |
| | – | – | 2 | 2.96 | 17.79 | 95.44 | 4.37 | 21.98 | 93.73 | 3.28 | 18.01 | 95.30 |
| | √ | – | | 2.94 | 17.78 | 95.45 | 4.35 | 21.97 | 93.74 | 3.27 | 17.99 | 95.31 |
| | – | √ | | **2.88** | **17.71** | **95.66** | **4.30** | **21.92** | **93.76** | **3.21** | **17.92** | **95.35** |
| | – | – | 10 | 1.36 | 9.92 | 98.36 | 3.17 | 17.99 | 95.38 | 2.05 | 10.75 | 97.72 |
| | √ | – | | 1.35 | 9.91 | 98.38 | 3.16 | 17.98 | 95.39 | 2.04 | 10.73 | 97.74 |
| | – | √ | | **1.34** | **9.82** | **98.42** | **3.11** | **17.92** | **95.46** | **1.99** | **10.70** | **97.89** |
| BGCC | – | – | 0.25 | 10.98 | 32.47 | 75.79 | 12.43 | 35.74 | 72.46 | 11.42 | 32.84 | 75.14 |
| | √ | – | | 10.97 | 32.46 | 75.80 | 12.42 | 35.73 | 72.48 | 11.40 | 32.83 | 75.15 |
| | – | √ | | **10.91** | **32.42** | **75.84** | **12.35** | **35.71** | **72.57** | **11.36** | **32.77** | **75.28** |
| | – | – | 2 | 2.96 | 17.79 | 95.42 | 4.38 | 22.00 | 93.64 | 3.28 | 18.01 | 95.30 |
| | √ | – | | 2.95 | 17.78 | 95.44 | 4.37 | 21.98 | 93.65 | 3.27 | 17.99 | 95.31 |
| | – | √ | | **2.90** | **17.72** | **95.64** | **4.31** | **21.92** | **93.69** | **3.23** | **17.94** | **95.34** |
| | – | – | 10 | 1.36 | 9.93 | 98.35 | 3.18 | 18.01 | 94.36 | 2.06 | 10.77 | 97.72 |
| | √ | – | | 1.36 | 9.92 | 98.36 | 3.17 | 17.99 | 94.38 | 2.05 | 10.75 | 97.74 |
| | – | √ | | **1.35** | **9.84** | **98.41** | **3.13** | **17.94** | **95.44** | **2.00** | **10.72** | **97.84** |
| BGLCC | – | – | 0.25 | 10.71 | 31.95 | 75.84 | 12.26 | 34.91 | 72.57 | 11.11 | 32.64 | 75.34 |
| | √ | – | | 10.70 | 31.94 | 75.85 | 12.25 | 34.90 | 72.58 | 11.10 | 32.62 | 75.36 |
| | – | √ | | **10.58** | **31.42** | **75.89** | **12.05** | **34.62** | **72.64** | **10.95** | **32.22** | **75.84** |
| | – | – | 2 | 2.69 | 17.03 | 95.63 | 4.16 | 21.88 | 93.73 | 3.06 | 17.91 | 95.35 |
| | √ | – | | 2.67 | 17.02 | 95.64 | 4.15 | 21.87 | 93.74 | 3.05 | 17.91 | 95.36 |
| | – | √ | | **2.55** | **16.95** | **95.72** | **3.99** | **21.02** | **93.78** | **2.86** | **17.57** | **95.72** |
| | – | – | 10 | 1.34 | 9.82 | 98.42 | 2.96 | 17.79 | 95.42 | 1.87 | 10.66 | 97.85 |
| | √ | – | | 1.34 | 9.82 | 98.43 | 2.95 | 17.78 | 95.44 | 1.85 | 10.64 | 97.87 |
| | – | √ | | **1.32** | **9.37** | **98.48** | **2.66** | **16.82** | **95.94** | **1.63** | **10.34** | **98.00** |

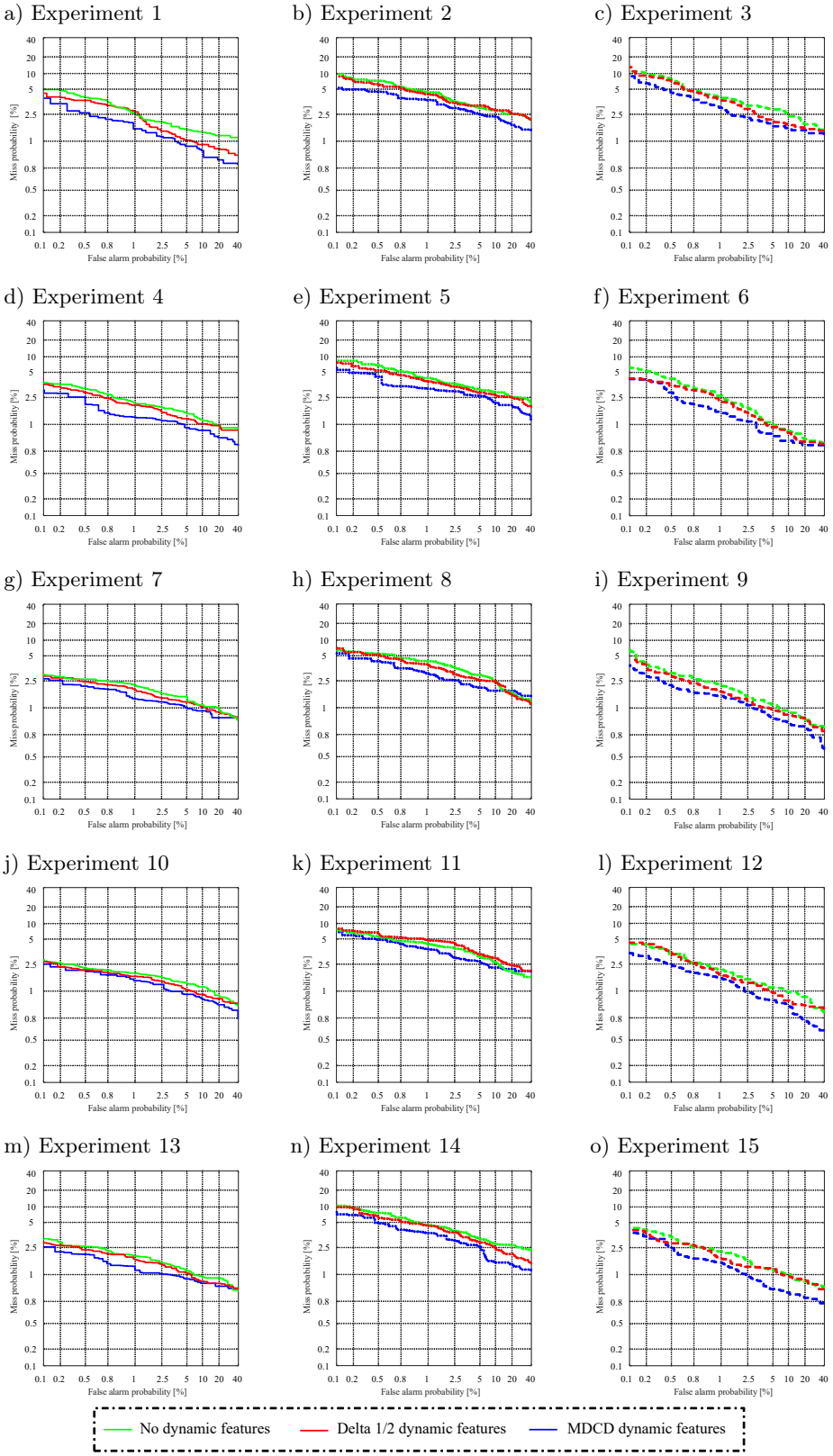


Fig. 2. The DET curves of different acoustic features and different dynamic features for speaker verification under varying audio lengths using the ResNet-34 model on VoxCeleb-1, SITW, and NIST SRE 2010 datasets. The experiments 1 to 3, the DET curves indicate, that on VoxCeleb-1, SITW, and NIST SRE 2010, under 10 s speech length, the LPCC uses no dynamic features, delta 1/2 dynamic features, and MDCD dynamic features, respectively. Similarly, experiments: 4–6 represent the MFCC method, 7–9 represent the MFCC-LPCC method, 10–12 represent the BGCC method, and 13–15 represent the BGLCC method.

At the same time, in the experiments comparing the different attributes of source information combination for short-duration speaker recognition (Das *et al.*, 2016), the proposed multi-source discriminative acoustic feature achieves consistent performance benefits across short-duration speech dataset experiments.

### *3.7.2. Ablation experiments*

To evaluate each component of the BGLCC-MDCD feature, we conducted several ablation experiments on VoxCeleb-1, SITW, and NIST SRE 2010 datasets, where the results are shown in Tables 2 and 3, and Figs. 2 and 3.

First, we evaluate the effectiveness of our proposed enhancement of discriminative acoustic features. Table 2 lists the EER, Min-DCF, and pAUC results of different features on VoxCeleb-1, SITW, and NIST SRE 2010 datasets. From Table 2, it can be observed that the proposed acoustic feature vastly outperforms the baseline feature, and it is seen from Fig. 2 that the DET curve of using MDCD dynamic features is lower than that without dynamic features, and using delta 1/2 dynamic features. The main reason for the performance improvement is our proposed BGLCC feature which employs the Bark-scaled Gauss and the linear filter bank superposition methods, it can remedy the weakness of the sparse high-frequency information and insufficient acoustic feature extraction by enhancing more high-frequency domain information. Similarly, MDCD through four different dimension differences captures better dynamics features of voiceprints, and it can further compensate for the limited and sparse dynamic acoustic features of short-duration audio signals. The experimental results also prove this.

To verify that different multi-dimensional central differences can capture dynamic features of the voiceprint, we conducted several ablation experiments,

Table 3. Ablation study for different multi-dimensional dynamic features based on BGLCC under varying audio lengths using the ResNet-34 network on VoxCeleb-1, SITW, and NIST SRE 2010 datasets.

| Methods | Duration [s] | VoxCeleb-1 | | | SITW | | | NIST SRE 2010 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | EER [%] | MinDCF | pAUC [%] | EER [%] | MinDCF | pAUC [%] | EER [%] | MinDCF | pAUC [%] |
| MDCD-$T_h$ | 0.25 | 10.67 | 31.90 | 75.88 | 12.19 | 34.80 | 72.62 | 11.04 | 32.29 | 75.43 |
| | 2 | 2.65 | 16.99 | 95.68 | 4.08 | 21.72 | 93.81 | 3.01 | 17.89 | 95.42 |
| | 10 | 1.32 | 9.71 | 98.46 | 2.91 | 15.69 | 95.96 | 1.83 | 10.62 | 97.89 |
| MDCD-$F_h$ | 0.25 | 10.68 | 31.91 | 75.87 | 12.21 | 34.82 | 72.61 | 11.05 | 32.31 | 75.42 |
| | 2 | 2.66 | 17.00 | 95.67 | 4.10 | 21.81 | 93.79 | 3.02 | 17.88 | 95.40 |
| | 10 | 1.33 | 9.72 | 98.44 | 2.92 | 17.70 | 95.94 | 1.84 | 10.63 | 97.88 |
| MDCD-$P_h$ | 0.25 | 10.70 | 31.94 | 75.85 | 12.25 | 34.85 | 72.59 | 11.09 | 32.34 | 75.38 |
| | 2 | 2.68 | 17.02 | 95.65 | 4.14 | 21.84 | 93.76 | 3.05 | 17.90 | 95.37 |
| | 10 | 1.35 | 9.82 | 98.42 | 2.95 | 17.73 | 95.37 | 1.86 | 10.65 | 97.86 |
| MDCD-$C_h$ | 0.25 | 10.69 | 31.93 | 75.86 | 12.23 | 34.84 | 72.60 | 11.07 | 32.32 | 75.39 |
| | 2 | 2.67 | 17.01 | 95.66 | 4.12 | 21.83 | 93.77 | 3.04 | 17.89 | 95.38 |
| | 10 | 1.34 | 9.81 | 98.43 | 2.94 | 17.72 | 95.38 | 1.85 | 10.64 | 97.87 |
| MDCD | 0.25 | **10.58** | **31.42** | **75.89** | **12.05** | **34.62** | **72.64** | **10.95** | **32.22** | **75.84** |
| | 2 | **2.55** | **16.95** | **95.72** | **3.99** | **21.02** | **93.78** | **2.86** | **17.57** | **95.72** |
| | 10 | **1.32** | **9.37** | **98.48** | **2.66** | **16.82** | **95.94** | **1.63** | **10.34** | **98.00** |

a) Experiment (EER, VoxCeleb-1) b) Experiment (EER, SITW) c) Experiment (EER, NIST SRE 2010)

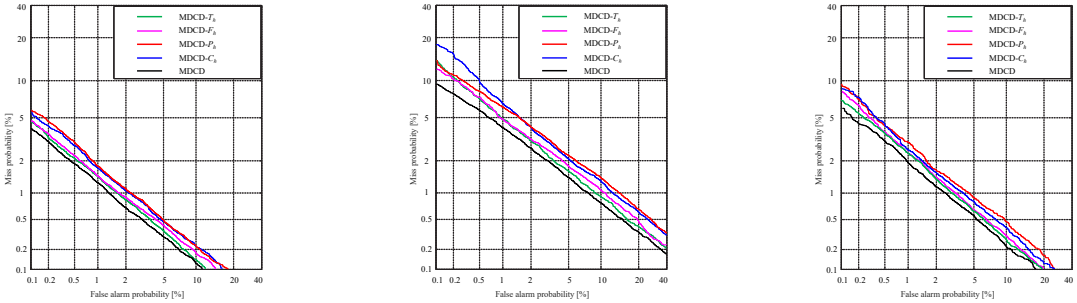


Fig. 3. DET curves of different dimensional dynamic features on VoxCeleb-1 (a), SITW (b), and NIST SRE 2010 (c) datasets under 10 s duration speech using the ResNet-34 model.

where the results are shown in Table 3 and Fig. 3. Compared to the diagonal domain, the time-frequency domain central difference captures better dynamic features, and the MDCD achieves the lower EER and Min-DCF. Figure 3 visualizes the DET curve of each dimension branch under the 10 s length utterance. The time-frequency domain performs better than the diagonal domain which is since the signal is mainly analyzed in the time-frequency domain.

Hence, the proposed BGLCC-MDCD discriminative acoustic features are the key reasons for the performance improvement in short utterance speaker verification, which: (a) extracts speaker-reliant characteristics successfully, from the BGLCC features to remedy the weakness of insufficient acoustic features to solve the problem of less emphasizes high-frequency information from the conventional acoustic feature extraction filter design; (b) then, the MDCD method can capture better dynamics features of voiceprints from short-duration audio signals.

## 4. Conclusion

In this paper, we propose the Bark-scaled Gauss and the linear filter bank superposition acoustic features extraction methods to enhance high-frequency domain information of short-duration audio, to deal with the problem of the high-frequency band feature sparsity. Compared with traditional acoustic features such as MFCC, LPCC, etc., our proposed BGLCC feature extraction method emphasizes a focus on both the low-high frequency band of speech, which is more helpful in extracting more discriminative acoustic features to compensate the sparsity of the effective information. Furthermore, a multi-dimensional central difference dynamic acoustic feature is proposed following the BGLCC spectrum characteristics, aiming to capture more diverse dynamic information. The MDCD feature concatenate information of the speaker from four different dimensions, this can explain why it performs significantly better than traditionally used speech features in short utterance speaker verification tasks under various conditions.

The proposed methods are evaluated on well-known datasets, VoxCeleb-1, SITW, and NIST SRE 2010 corpus. From the experimental results, the proposed method achieves continuous improvement over traditional acoustic features in all test sets. The ablation experiments further indicate that the proposed approaches substantially improve the enhanced discriminant features for speaker verification tasks. Future work involves the combination of acoustic feature-based and model-based compensations for short-duration speech speaker verification, and to improve the performance, accuracy, and richness of acoustic feature extraction in short-duration audio signals.

## Acknowledgments

This work was in part supported by NSFC (grants nos. 62176194, 62101393), the Major project of IoV (grant no. 2020AAA001), Sanya Science and Education Innovation Park of the Wuhan University of Technology (grant no. 2021KF0031), CSTC (grant no. cstc2021jcyj-msxmX1148), and the Open Project of the Wuhan University of Technology Chongqing Research Institute (ZL2021-6).

## References

1. Atal B.S. (1974), Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *The Journal of the Acoustical Society of America*, **55**(6): 1304–1312, doi: 10.1121/1.1914702.

2. Bai Z., Zhang X.-L., Chen J. (2020), Speaker verification by partial AUC optimization with Mahalanobis distance metric learning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**: 1533–1548, doi: 10.1109/TASLP.2020.2990275.

3. Campbell J.P. (1997), Speaker recognition: A tutorial, *Proceedings of the IEEE*, **85**(9): 1437–1462, doi: 10.1109/5.628714.

4. Chowdhury A., Ross A. (2020), Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals, *IEEE Transactions on Information Forensics and Security*, **15**: 1616–1629, doi: 10.1109/TIFS.2019.2941773.

5. Chung J.S., Nagrani A., Zisserman A. (2018), Voxceleb2: Deep speaker recognition, [in:] *Proceedings of Interspeech 2018*, pp. 1086–1090, doi: 10.21437/Interspeech.2018-1929.

6. Das R.K., Mahadeva Prasanna S.R. (2016), Exploring different attributes of source information for speaker verification with limited test data, *The Journal of the Acoustical Society of America*, **140**(1): 184, doi: 10.1121/1.4954653.

7. Dehak N., Dehak R., Glass J., Reynolds D., Kenny P. (2010), Cosine similarity scoring without score normalization techniques, [in:] *Proceedings of Odyssey 2010 – The Speaker and Language Recognition Workshop*.

8. Desplanques B., Thienpondt J., Demuynck K. (2020), ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, [in:] *Proceedings of Annual conference of the International Speech Communication Association 2020*, pp. 3830–3834, doi: 10.21437/Interspeech.2020-2650.

9. Greenberg C.S. *et al.* (2013), The 2012 NIST speaker recognition evaluation, [in:] *Proceedings of Interspeech 2013*, pp. 1971–1975, doi: 10.21437/Interspeech.2013-469.

10. Herrera-Camacho A., Zúñiga-Sainos A., Sierra-Martínez G., Tramgol-Curipe J., Mota-Montoya M., Jarquín-Casas A. (2019), Design and testing of a corpus for forensic speaker recognition using MFCC, GMM and MLE, [in:] *Proceedings of International Conference on Video, Signal and Image Processing 2019*, pp. 105–110, doi: 10.1145/3369318.3369330.
11. Huang L., Pun C.-M. (2020), Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**: 1813–1825, doi: 10.1109/TASLP.2020.2998870.
12. Kenny P., Boulianne G., Dumouchel P. (2005), Eigenvoice modeling with sparse training data, *IEEE Transactions on Speech and Audio Processing*, **13**(3): 345–354, doi: 10.1109/TSA.2004.840940.
13. Kinnunen T., Li H. (2010), An overview of text-independent speaker recognition: From features to supervectors, *Speech Communication*, **52**(1): 12–40, doi: 10.1016/j.specom.2009.08.009.
14. Liu Z., Wu Z., Li T., Li J., Shen C. (2018), GMM and CNN hybrid method for short utterance speaker recognition, *IEEE Transactions on Industrial Informatics*, **14**(7): 3244–3252, doi: 10.1109/TII.2018.2799928.
15. Martin A.F., Greenberg C.S. (2010), The NIST 2010 speaker recognition evaluation, [in:] *Proceedings of Interspeech 2010*, pp. 2726–2729, doi: 10.21437/Interspeech.2010-722.
16. McLaren M., Ferrer L., Castan D., Lawson A. (2016), The speakers in the wild (SITW) speaker recognition database, [in:] *Proceedings of Interspeech 2016*, pp. 818–822, doi: 10.21437/Interspeech.2016-1129.
17. Nagrani A., Chung J.S., Zisserman A. (2017), VoxCeleb: A large-scale speaker identification dataset, [in:] *Proceedings of Interspeech 2017*, pp. 2616–2620, doi: 10.21437/Interspeech.2017-950.
18. Nosratighods M., Ambikairajah E., Epps J., Carey M.J. (2010), A segment selection technique for speaker verification, *Speech Communication*, **52**(9): 753–761, doi: 10.1016/j.specom.2010.04.007.
19. Omar M.K., Pelecanos J.W. (2010), Training universal background models for speaker recognition, [in:] *Proceedings of Odyssey 2010 – The Speaker and Language Recognition Workshop*, pp. 52–57.
20. Paseddula C., Gangashetty S.V. (2018), DNN based acoustic scene classification using score fusion of MFCC and inverse MFCC, [in:] *Proceedings of International Conference on Industrial and Information Systems 2018*, pp. 18–21, doi: 10.1109/ICIINFS.2018.8721379.
21. Paszke A. *et al.* (2017), Automatic differentiation in PyTorch, [in:] *Proceedings of NIPS 2017 Workshop*, pp. 1–4.
22. Povey D. *et al.* (2018), Semi-orthogonal low-rank matrix factorization for deep neural networks, [in:] *Proceedings of Interspeech 2018*, pp. 3743–3747, doi: 10.21437/Interspeech.2018-1417.
23. Schroff F., Kalenichenko D., Philbin J. (2015), FaceNet: A unified embedding for face recognition and clustering, [in:] *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2015*, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.
24. Snyder D., Garcia-Romero D., Sell G., McCree A., Povey D., Khudanpur S. (2019), Speaker recognition for multi-speaker conversations using x-vectors, [in:] *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2019*, pp. 5796–5800, doi: 10.1109/ICASSP.2019.8683760.
25. Todisco M., Delgado H., Evans N. (2017), Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification, *Computer Speech & Language*, **45**: 516–535, doi: 10.1016/j.csl.2017.01.001.
26. Villalba J. *et al.* (2020), State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations, *Computer Speech & Language*, **60**: 101026, doi: 10.1016/j.csl.2019.101026.
27. Vogt R., Sridharan S., Mason M. (2010), Making confident speaker verification decisions with minimal speech, *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(6): 1182–1192, doi: 10.1109/TASL.2009.2031505.
28. Wu Z., Yu Z., Yuan J., Zhang J. (2016), A twice face recognition algorithm, *Soft Computing – A Fusion of Foundations, Methodologies and Applications*, **20**(3): 1007–1019, doi: 10.1007/s00500-014-1561-9.
29. Yang H., Deng Y., Zhao H.-A. (2019), A comparison of MFCC and LPCC with deep learning for speaker recognition, [in:] *Proceedings of International Conference on Big Data and Computing 2019*, pp. 160–164, doi: 10.1145/3335484.3335528.
30. Zhang C., Koishida K., Hansen J.H.L. (2018), Text-independent speaker verification based on triplet convolutional neural network embeddings, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(9): 1633–1644, doi: 10.1109/TASLP.2018.2831456.
31. Zinchenko K., Wu C.-Y., Song K.-T. (2017), A study on speech recognition control for a surgical robot, *IEEE Transactions on Industrial Informatics*, **13**(2): 607–615, doi: 10.1109/TII.2016.2625818.