

RESEARCH ARTICLE

Robust Feature Extraction Using Temporal Context Averaging for Speaker Identification in Diverse Acoustic Environments

YASSIN TERRAF^{1,2}, (Student Member, IEEE), AND YOUSSEF IRAQI¹, (Senior Member, IEEE)

¹College of Computing, University Mohammed VI Polytechnic, Ben Guerir 43150, Morocco

²Henceforth, Rabat 10000, Morocco

Corresponding author: Yassin Terraf (yassin.terraf@um6p.ma)

This work was supported by Henceforth.

ABSTRACT Speaker identification in challenging acoustic environments, influenced by noise, reverberation, and emotional fluctuations, requires improved feature extraction techniques. Although existing methods effectively extract distinct acoustic features, they show limitations in these adverse settings. To overcome these limitations, we propose the Temporal Context-Enhanced Features (TCEF) approach, which provides a consistent audio representation for better performance under various acoustic conditions. TCEF leverages a context window to average features in adjacent frames, effectively reducing short-term variations caused by noise, reverberation, fluctuations in emotional speech, and those in neutral recordings. This approach improves the distinctive features of a speaker voice, improving speaker identification in challenging and neutral acoustic environments. To evaluate the performance of TCEF against conventional features, One-Dimensional Convolutional Neural Network (1D-CNN) was used for a detailed frame-level analysis and Long Short-Term Memory (LSTM) for a comprehensive sequence-level analysis. We used four datasets to assess the effectiveness of the TCEF approach. The GRID and RAVDESS datasets represent neutral and emotional speech, respectively. To test the robustness of our system under adverse acoustic conditions, we created two additional datasets: GRID-NR and RAVDESS-NR. These are modified versions of the original GRID and RAVDESS, incorporating added noise and reverberation. Performance evaluation results showed that TCEF significantly outperformed existing feature extraction methods in identifying speakers in diverse acoustic environments.


INDEX TERMS Speaker identification, feature extraction, challenging acoustic environments, temporal context-enhanced features, convolutional neural networks, long short-term memory.

I. INTRODUCTION

Automatic speaker identification (ASI), which is the process of extracting a speaker identity based on their vocal characteristics [12], [15], has become a significant focus in research and real-world applications. This technology is essential in various sectors, including user authentication [6], voice-controlled devices [14], smart home personalization, and forensic analysis [27], [45]. The growing need for these applications highlights the importance of creating accurate speaker identification systems. To achieve this accuracy, a crucial component is feature extraction, where

voice recordings are transformed into distinct characteristics representing individual unique vocal features.

Several feature extraction techniques have been used to extract these vocal features. Among the techniques that are used most frequently are Mel-Frequency Cepstral Coefficients (MFCC) [28], [33], [43], Gammatone Frequency Cepstral Coefficients (GTCC) [3], [32], and Power-Normalized Cepstral Coefficients (PNCC) [17], [30]. MFCC is a widely used technique in speech processing that captures essential auditory patterns. GTCC is biologically inspired and designed to mirror the human auditory system closely. On the other hand, PNCC incorporates additional noise reduction techniques to enhance their ability to extract effective features. Although these techniques perform

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson .

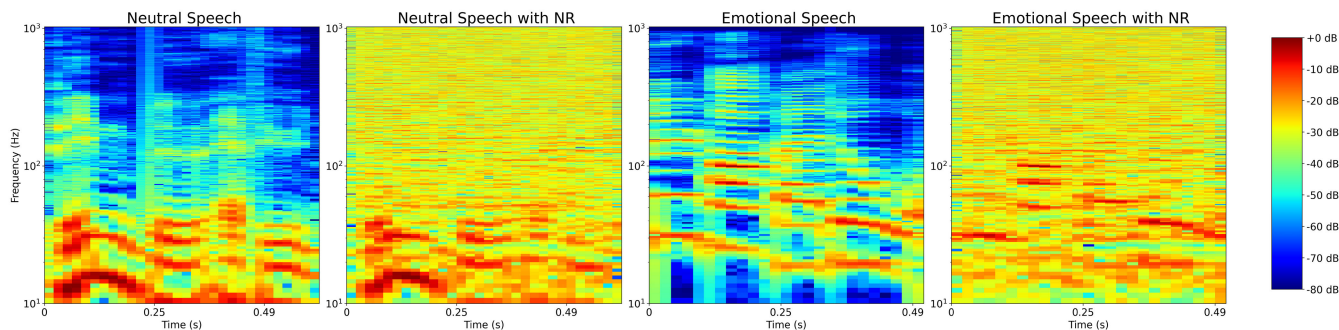


FIGURE 1. Comparative spectrograms of a sample speech under varying acoustic conditions (NR: Noise and Reverberation).

well in various situations, they often struggle to extract effective features when dealing with recordings affected by noise, reverberation, and emotions, making it difficult to identify speakers accurately [16], [45]. Figure 1 illustrates the visual representation of the spectrograms of a sample speech audio under different conditions: neutral speech, neutral speech with added noise and reverberation, emotional speech, and emotional speech with noise and reverberation. This illustration highlights the challenges posed by noise, reverberation, and emotional states in obscuring distinct vocal characteristics, complicating feature extraction, and, subsequently, the speaker identification process.

To address these challenges, researchers have investigated several approaches. Multi-condition training is a strategy that has been implemented, involving the addition of noise and reverberation to training data to improve the robustness of speaker identification [36], [38]. Although this approach is effective, it often demands extensive labeled datasets. Other research efforts have focused on the fusion of different feature extraction methods. In the study by Chowdhury et al. [5], MFCC were combined with Linear Predictive Coding to enhance speaker recognition performance under challenging conditions. Similarly, Salvati et al. [32] integrated frequency-domain features, specifically GTCC, with time-domain raw features to improve the accuracy rate under adverse noise and reverberation conditions. However, while expanding the set of features can enhance the discriminative ability of the features extracted from speech, it also increases the dimensionality of the feature vectors. This expanded dimensionality adds complexity to the computational tasks involved in feature extraction and model training, making them more resource-intensive. Additionally, there have been efforts to adapt conventional feature extraction methods to improve their robustness against environmental noise and reverberation. Modifications have been proposed for MFCC [13], GTCC [37], and PNCC [22], [47]. These enhanced feature variants introduce additional complexity to the feature extraction process by incorporating extra computational operations, such as adaptive noise compensation or non-linear transformation techniques, resulting in increased computational demands compared to the original feature computation.

To address the limitations of speaker identification in diverse acoustic environments, this study introduces

Temporal Context-Enhanced Features (TCEF). A method that averages features across adjacent frames within a context window. This approach effectively reduces short-time variations caused by noise, reverberation, and fluctuations in emotional speech, as well as those in neutral recordings. Consequently, TCEF provides more robust features, enhancing speaker identification accuracy in challenging and neutral conditions.

The main contributions of this paper can be summarized as follows.

- TCEF for robust feature extraction under diverse recording conditions for speaker identification.
- Integration of TCEF into conventional feature extraction methods including MFCC, GTCC, and PNCC.
- Application of One-Dimensional Convolutional Neural Network (1D-CNN) for frame-level analysis and Long Short-Term Memory (LSTM) for sequence-level analysis to assess performance differences between TCEF and conventional features.
- Creation of GRID-NR and RAVDESS-NR datasets, incorporating noise and reverberation, to test the robustness of the TCEF approach under varied acoustic conditions.
- Extensive experiments were carried out on GRID, RAVDESS, GRID-NR, and RAVDESS-NR to evaluate the performance of TCEF compared to conventional features for speaker identification.

The remainder of this paper is structured as follows. Section II presents the TCEF. Section III discusses the integration of TCEF with conventional feature extraction methods and the inclusion of dynamic features. Section IV describes the experimental setup, which includes the neural networks employed, the descriptions of the datasets, and a detailed analysis of the experimental results. Section V discusses our findings. Section VI covers the reproducibility of the evaluation, and, finally, Section VII concludes the paper.

II. PROPOSED APPROACH

A. PROBLEM FORMULATION

Speaker identification is the task of determining the identity of a speaker from a given audio recording. In our investigation, we focus on a set of n speakers denoted

$S = \{s_1, s_2, \dots, s_n\}$. Each speaker s_i in the set S is associated with their own set of audio recordings $A_i = \{a_{i1}, a_{i2}, \dots, a_{io}\}$, where o represents the number of recordings for speaker s_i , and these are captured under diverse acoustic conditions that introduce different challenges to the identification process. These conditions include neutral speech, emotional speech, neutral speech with noise and reverberation, and emotional speech with noise and reverberation, resulting in four main conditions:

- C_1 : Neutral speech
- C_2 : Neutral speech with noise and reverberation
- C_3 : Emotional speech
- C_4 : Emotional speech with noise and reverberation

Extracting distinct features representing each speaker is necessary to achieve effective speaker identification. However, human speech is inherently non-stationary over long periods. To overcome this challenge, a framing operation is used to segment each audio recording into shorter frames, where the speech features are assumed to be stationary. This operation results in a set of frames $F = \{f_1, f_2, \dots, f_M\}$ for each audio recording, where M represents the total number of frames in the audio recording. When applying the feature extraction function E to this frame set, we obtain

$$E(F) = \Phi = \{\Phi_1, \Phi_2, \dots, \Phi_M\} \quad (1)$$

where each Φ_i represents the extracted features from the corresponding frame f_i in neutral speech C_1 . To effectively handle real-world scenarios, we also consider the effects of conditions C_2 , C_3 , and C_4 on these extracted features Φ . These conditions introduce significant variations, posing different acoustic challenges in speaker identification. To model these variations, we define a series of transformations on the frame set F . The function N models the influence of noise, while R models the effects of reverberation, and EM represents the modulation of features due to emotional states in speech. For neutral speech impacted by noise and reverberation, we apply a composite function $N(R(\cdot))$. In the case of emotional speech with noise and reverberation, $N(R(EM(\cdot)))$ is used. The following equations represent these transformations applied to F :

$$F_{\text{noise}} = \{N(f_1), \dots, N(f_M)\} \quad (2)$$

$$F_{\text{reverb}} = \{R(f_1), \dots, R(f_M)\} \quad (3)$$

$$F_{\text{emotion}} = \{EM(f_1), \dots, EM(f_M)\} \quad (4)$$

$$F_{\text{nnr}} = \{N(R(f_1)), \dots, N(R(f_M))\} \quad (5)$$

$$F_{\text{enr}} = \{N(R(EM(f_1))), \dots, N(R(EM(f_M)))\} \quad (6)$$

Processing these modified frame sets with the feature extraction process E , we obtain the corresponding feature sets:

$$\Phi_{\text{noise}} = E(F_{\text{noise}}), \quad (7)$$

$$\Phi_{\text{reverb}} = E(F_{\text{reverb}}), \quad (8)$$

$$\Phi_{\text{emotion}} = E(F_{\text{emotion}}), \quad (9)$$

$$\Phi_{\text{nnr}} = E(F_{\text{nnr}}), \quad (10)$$

$$\Phi_{\text{enr}} = E(F_{\text{enr}}). \quad (11)$$

Given an audio recording a from a set A , under a condition C_j , our objective is to match the audio recording a to the correct speaker identity from the set S . Conventional feature extraction methods, denoted by E , struggle to extract speaker-specific features in diverse acoustic environments, leading to challenges in accurate speaker identification. To overcome this limitation, we introduce an advanced feature extraction approach, denoted TCEF. This approach is designed to mitigate the impacts of noise, reverberation, and emotional states on the feature extraction process. By improving the extraction process to be more robust in diverse environments, TCEF aims to significantly improve speaker identification performance.

B. TEMPORAL CONTEXT-ENHANCED FEATURES

Feature extraction based on individual frames can produce features susceptible to short-time variations in acoustic features caused by factors such as noise, reverberation, emotional fluctuations, speaking styles, and microphone quality. These variations may result in inconsistent representations of the distinctive vocal features of a speaker, especially in challenging recording conditions. To address this challenge, we introduce TCEF. This approach integrates the current frame with its neighboring ones, employing a sliding-window technique to smooth out short-term variations from noise, reverberation, emotional fluctuations, and those inherent in neutral speech, thereby emphasizing the distinct features of a speaker voice. Consider a frame f_i with its corresponding feature set $\Phi_i = \{\phi_{i1}, \phi_{i2}, \dots, \phi_{ip}\}$, where each ϕ_{ij} is a feature at index j , and p represents the total number of features extracted from each frame. The feature set can reflect speaker-specific features under various conditions such as neutral Φ , neutral with noise and reverberation Φ_{nnr} , emotional Φ_{emotion} , or emotional with noise and reverberation Φ_{enr} . The TCEF for frame f_i is calculated by averaging the features of the current frame f_i and its next $N - 1$ frames. Here, N represents the size of the context window, a key hyper-parameter that includes the total number of frames to consider for the averaging process. As the context window slides over the sequence of audio frames, it covers $M - N + 1$ frames, where M is the total number of frames. To ensure that all frames are included in the feature extraction process, the context window size N_i for the i^{th} frame is adaptively computed as follows:

$$N_i = \min(N, M - i + 1) \quad (12)$$

For any particular feature indexed by j within Φ_i , the temporal context-enhanced feature $TCEF_{ij}$ is subsequently calculated using the context window size N_i as follows:

$$TCEF_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i-1} \phi_{i+k,j} \quad \text{for } j = 1, \dots, p \quad (13)$$

The TCEF algorithm is comprehensively detailed in the following pseudocode, which outlines the sequence of operations performed to derive the enhanced features.

Algorithm 1 Temporal Context-Enhanced Features (TCEF)

Require: Set of speakers S , set of audio recordings per speaker A_i , context window size N

Ensure: Enhanced features $TCEF$ for each recording

- 1: **for** each speaker s_i in S **do**
- 2: **for** each recording a_{ij} in A_i **do**
- 3: $TCEF_{a_{ij}} = []$
- 4: Divide a_{ij} into frames $F = \{f_1, f_2, \dots, f_M\}$
- 5: **for** $c = 1$ to M **do**
- 6: $TCEF_c = []$
- 7: $N_c = \min(N, M - c + 1)$
- 8: Extract $\Phi_c = \{\phi_{c1}, \phi_{c2}, \dots, \phi_{cp}\}$ from f_c
- 9: **for** $d = 1$ to p **do**
- 10: $TCEF_{cd} = \frac{1}{N_c} \sum_{k=0}^{N_c-1} \phi_{c+k,d}$
- 11: $TCEF_c.Add(TCEF_{cd})$
- 12: **end for**
- 13: $TCEF_{a_{ij}}.Add(TCEF_c)$
- 14: **end for**
- 15: **end for**
- 16: **end for**

III. ENHANCED FEATURE EXTRACTION TECHNIQUES

Feature extraction is one of the most critical aspects of ASI. It converts raw audio signals into features representing the distinct characteristics of speakers' voices. These features are then used as the primary input for modeling processes, improving speaker identification accuracy. This section offers an overview of widely used methods in this field. Subsequently, we describe how our Temporal Context-Enhanced Features technique enhances these conventional methods.

A. CONVENTIONAL FEATURE EXTRACTION TECHNIQUES

1) MFCC

MFCC [24], is a technique frequently used in speech and audio research for its efficacy in capturing significant characteristics of speech signals [2], [11], [46]. This technique, inspired by the human ear frequency response, is beneficial in tasks such as speaker identification. The process of obtaining these coefficients comprises the following steps:

Pre-emphasis: A filter processes the speech signal, amplifying its higher frequencies. This step ensures a balanced representation of both low- and high-frequency components.

Framing: Given the non-stationary nature of speech signals, the pre-emphasised audio is segmented into overlapping intervals, usually of 10-50 ms [41]. In these frames, the signal can be assumed to be stationary for such a short interval, facilitating the subsequent feature extraction process.

Windowing: A window function is applied to each frame to minimize edge discontinuities. This ensures that the frames are more suitable for subsequent frequency analysis, ensuring

an accurate representation of the frequency components in the signal.

Fast Fourier Transform (FFT): The windowed frame is processed using the FFT, converting from time to frequency domain. This provides a spectrum that indicates the frequency components of the frame.

Mel Filter Bank: The spectrum is processed using triangular filters to map it to the Mel scale, consistent with the human ear frequency response. This produces a Mel spectrum that closely matches human auditory perception.

Log Compression: To align with human perception of sound intensity, the energy-representing values in the Mel spectrum are subjected to logarithmic compression, producing a log Mel spectrum.

Discrete Cosine Transform (DCT): The log Mel spectrum is then transformed using DCT. This transformation results in MFCC, which are represented as a series of decorrelated coefficients, ensuring that each coefficient offers unique information for subsequent modeling. Mathematically, the j th MFCC coefficient of the i th frame is computed as [1]:

$$MFCC_{ij} = \sum_{k=1}^K \log(S_{mi}(k)) \cos\left(\frac{\pi j(2k-1)}{2K}\right) \quad (14)$$

where K is the total number of Mel filter banks used, $S_{mi}(k)$ represents the power spectrum of the i th frame as processed by the k th Mel filter bank, and j ranges from 1 to the desired number of cepstral coefficients p .

2) GTCC

In response to the challenges of speaker identification under noisy conditions and extending from the concept of MFCC, Valero and Alias [41] improved the extraction of audio features by introducing GTCC. Following this development, GTCC has been adopted in a variety of studies focusing on robust speech and speaker recognition systems [19], [39]. The computation process of GTCC is similar to that of MFCC. The audio signal is segmented into overlapping frames, usually 10-50 ms. Then, a windowing function is applied to each frame. Subsequently, the FFT is used on the windowed signal to derive the frequency spectrum. In contrast to MFCC, which uses Mel filter banks, GTCC employs the Gammatone filter bank, which is inspired by the cochlea auditory processing comprising a series of overlapping filters, each focused on a specific frequency. When passed through these filters, the frequency spectrum from the FFT produces a cochleagram. After the cochleagram is obtained, a logarithmic compression is applied, reflecting human sound intensity perception. Following this, a DCT is performed to decorrelate the coefficients obtained from the GTCC processing, providing unique coefficients for subsequent analysis. Mathematically, the j th GTCC coefficient of the i th frame is computed as [41]:

$$GTCC_{ij} = \sum_{k=1}^K \log(G_{mi}(k)) \cos\left(\frac{\pi j(2k-1)}{2K}\right) \quad (15)$$

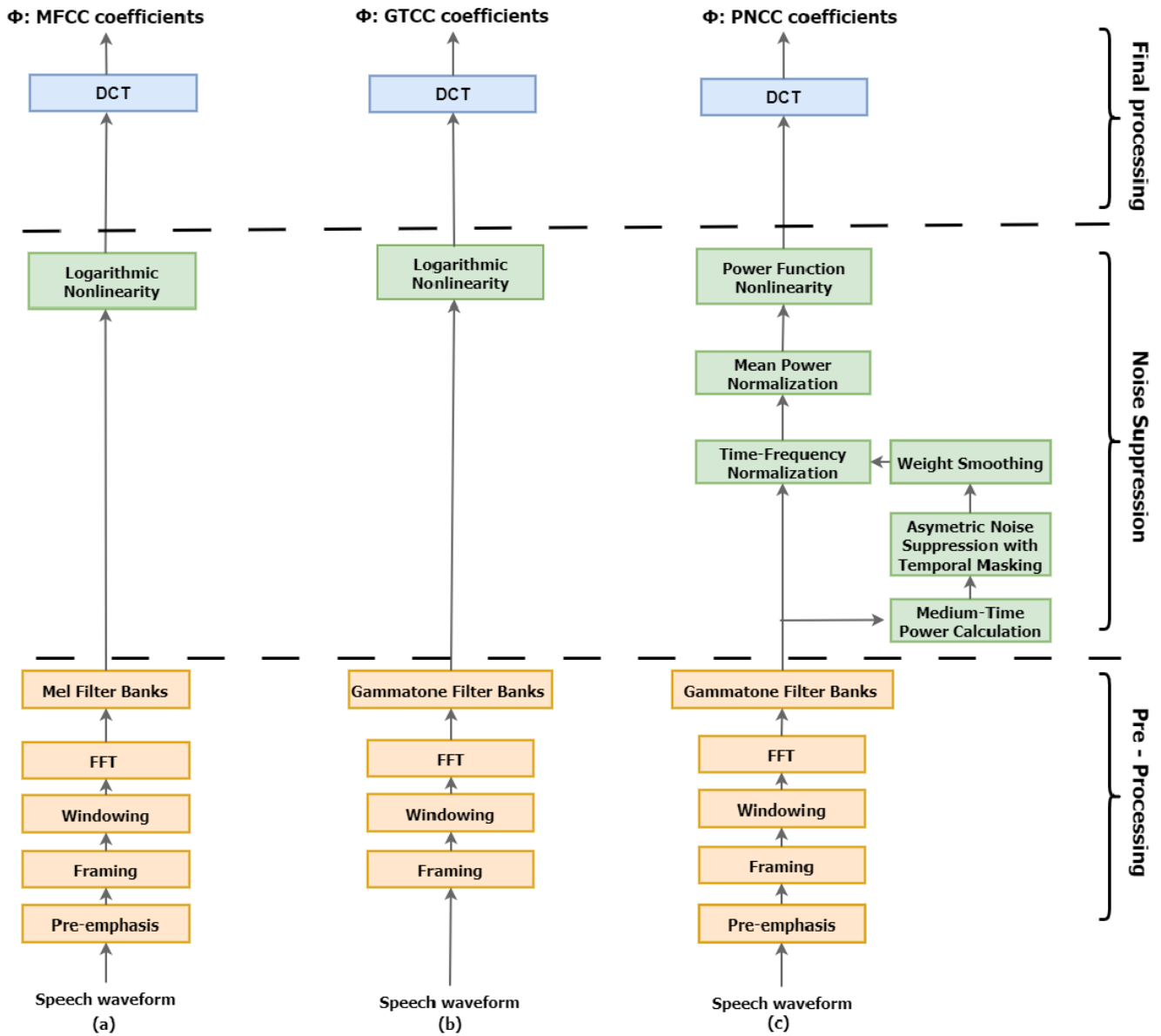


FIGURE 2. Diagrams of the computation of conventional feature extraction techniques: (a) MFCC, (b) GTCC, and (c) PNCC.

where K is the total number of Gammatone filter banks used, $G_{mi}(k)$ represents the output energy of the k th Gammatone filter bank for the i th frame, and j ranges from 1 to the desired number of cepstral coefficients p .

3) PNCC

PNCC, developed by Kim and Stern [18], is a technique designed to improve speech recognition in varying acoustic environments. The effectiveness of PNCC has been recognized and further explored in various research works [21], [31], demonstrating its robustness in diverse acoustic settings. The computation process for PNCC starts with the audio signal that undergoes pre-emphasis to amplify its high-frequency components. Subsequently, the enhanced

signal is segmented into multiple frames, which are then windowed using a window function to minimize edge discontinuities. The windowed frame is transformed into a frequency spectrum using FFT. Subsequent enhancements are achieved by introducing gammatone filter banks designed to model human auditory perception. The asymmetric noise suppression system is used with temporal masking to handle the varied noise levels within the signal. This approach distinguishes the noise spectrum, and adjusted the audio intensity, ensuring that essential audio elements are differentiated from background disturbances. After applying gammatone filter banks and medium-time analysis, spectral weight smoothing ensures uniform power distribution across all frequencies, producing a balanced spectral representation. Following this,

mean power normalization modifies the signal amplitude, providing uniform intensity levels and reducing significant fluctuations in the audio. In addition, a power law function with an exponent of 1/15 is introduced. This function captures the relationship between perceived sound levels and the human auditory system response. Finally, DCT is used to produce the PNCC coefficients. Mathematically, the j th PNCC coefficient of the i th frame is computed as [18]:

$$PNCC_{ij} = \sum_{k=1}^K (PowerNorm_i(k))^\alpha \cos\left(\frac{\pi j(2k-1)}{2K}\right) \quad (16)$$

where K is the total number of critical-band filters used, $PowerNorm_i(k)$ is the normalized energy output of the k th gammatone filter bank for the i th frame after power normalization, α is the exponent in the power-law non-linearity, typically set to 1/15, and j ranges from 1 to the desired number of cepstral coefficients p .

Fig. 2 provides a visual representation of the conventional feature extraction processes. (a) represents the block diagram of the MFCC process, (b) illustrates the GTCC process, and (c) demonstrates the PNCC process, each highlighting the distinct steps involved in these techniques.

B. TEMPORAL CONTEXT-ENHANCED FEATURE EXTRACTION TECHNIQUES

Conventional feature extraction techniques MFCC, GTCC, and PNCC are crucial for speaker identification but often encounter challenges in diverse acoustic conditions, including background noise, reverberation, and variations in emotional states. It is critical to refine these techniques to ensure robust speaker identification. In response to this challenge, conventional MFCC, GTCC, and PNCC features have been enhanced using TCEF. As illustrated in Fig. 3, an input speech signal is analyzed using MFCC, GTCC, or PNCC techniques to produce conventional features. The feature windowing then applies a sliding window operation across these feature sequences with a context size of N_i frames. For parts of the signal where the number of remaining frames is less than the set context window size N , the context window size N_i is adjusted, as specified in (12), to ensure that all frames are included in the analysis. Following this adjustment, the Feature Averaging process computes the average for each coefficient by combining the corresponding coefficients from all frames within the window, as detailed in (13). This averaging operation results in TCEF vectors that provide a richer and more robust representation of the conventional features. Extending the basic MFCC features as defined in (14), the TCEF for MFCC denoted M_TCEF is defined as follows:

$$M_TCEF_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i-1} MFCC_{i+k,j} \quad \text{for } j = 1, \dots, p \quad (17)$$

where M_TCEF_{ij} represents the TCEF of MFCC feature for the j th coefficient at the i th frame, N_i is context window size, and $MFCC_{i+k,j}$ is the conventional MFCC feature for the

j th coefficient at the $(i+k)$ th frame. Similarly, the TCEF representations of GTCC and PNCC features, as delineated in (15) and (16) respectively, are defined as follows:

$$G_TCEF_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i-1} GTCC_{i+k,j} \quad \text{for } j = 1, \dots, p \quad (18)$$

$$P_TCEF_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i-1} PNCC_{i+k,j} \quad \text{for } j = 1, \dots, p \quad (19)$$

where G_TCEF_{ij} represents the TCEF of the GTCC feature for the j th coefficient at the i th frame, $GTCC_{i+k,j}$ is the conventional GTCC feature for the j th coefficient at the $(i+k)$ th frame, P_TCEF_{ij} denotes the TCEF of the PNCC feature for the j th coefficient at the i th frame, and $PNCC_{i+k,j}$ represents the conventional PNCC feature for the j th coefficient at the $(i+k)$ th frame.

C. INCORPORATING DYNAMIC FEATURES IN TEMPORAL CONTEXT-ENHANCED FEATURES

Dynamic features, delta Δ and delta-delta $\Delta\Delta$, representing the first- and second-order derivatives of the extracted feature vectors from speaker speech, were introduced by Furui [8] and expanded by Lawrence [20] for speaker recognition tasks. These features are crucial in speech processing, as they capture the temporal dynamics inherent in speech features over time. Specifically, Δ reflects the speed of change in spectral features, while $\Delta\Delta$ indicates the acceleration of this rate of change. This method effectively models the dynamic and non-stationary nature of speech characteristic of human articulation. The incorporation of delta Δ and delta-delta $\Delta\Delta$ derivatives into conventional acoustic features, namely *MFCC*, *GTCC*, and *PNCC*, denoted Φ , significantly enhances speaker identification. Various studies have validated this enhancement [4], [35], [47]. The calculation of these dynamic features is described by the following equations for the first-order derivative of the conventional feature vector Φ :

$$\Delta\Phi_i = \frac{\sum_{\tau=1}^L (\Phi_{i+\tau} - \Phi_{i-\tau})}{2 \sum_{\tau=1}^L \tau^2} \quad (20)$$

where i indexes the current frame for which the derivative is computed, L denotes the number of frames considered for computing first- and second-order derivatives, and $\Phi_{i+\tau}$ and $\Phi_{i-\tau}$ represent the feature vectors of subsequent and preceding frames relative to frame i . $2 \sum_{\tau=1}^L \tau^2$ normalizes the values to ensure consistent weighting of features. For the second-order derivative of the conventional feature vector Φ :

$$\Delta\Delta\Phi_i = \frac{\sum_{\tau=1}^L (\Delta\Phi_{i+\tau} - \Delta\Phi_{i-\tau})}{2 \sum_{\tau=1}^L \tau^2} \quad (21)$$

where i indexes the current frame for which we compute the second-order derivative, and the terms $\Delta\Phi_{i+\tau}$ and $\Delta\Phi_{i-\tau}$ are the first-order derivative values of the feature vectors for the frames following and preceding frame i . To include first- and second-order derivatives with the enhanced features,

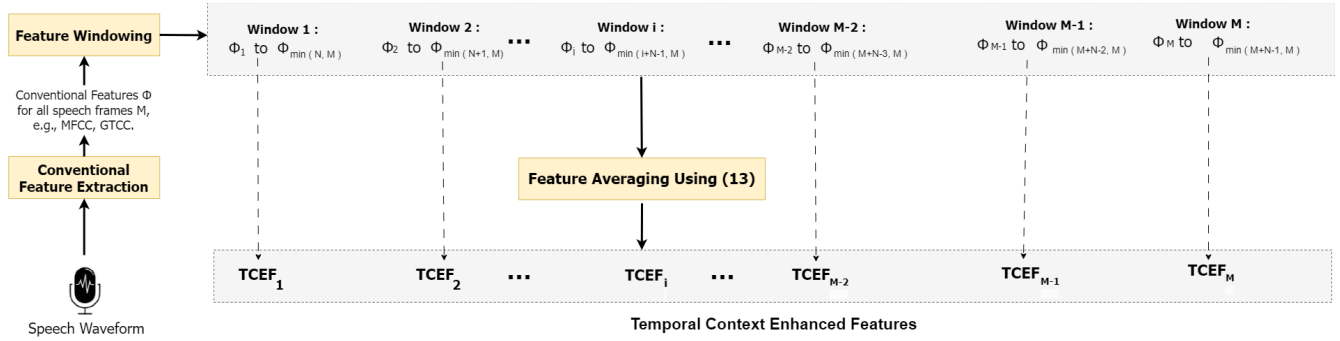


FIGURE 3. Process for generating temporal context enhanced features: Context window size (N), Frame count (M), and Conventional features (Φ).

M_TCEF , G_TCEF , and P_TCEF , referred to as TCEF, we adopt a similar method as used for conventional features but adapted for TCEF. This approach produces the delta $\Delta TCEF$ and delta-delta $\Delta\Delta TCEF$ derivatives, calculated for frame i using the following equations:

$$\Delta TCEF_i = \frac{\sum_{\tau=1}^L (TCEF_{i+\tau} - TCEF_{i-\tau})}{2 \sum_{\tau=1}^L \tau^2} \quad (22)$$

$$\Delta\Delta TCEF_i = \frac{\sum_{\tau=1}^L (\Delta TCEF_{i+\tau} - \Delta TCEF_{i-\tau})}{2 \sum_{\tau=1}^L \tau^2} \quad (23)$$

IV. EXPERIMENTAL SETUP

This section presents a comprehensive analysis of the performance of TCEF compared to conventional features using neural network models. This comparison includes evaluations at both the frame level and sequence level. Following this, we describe the datasets used for speaker identification and the baseline methods used for comparison. We detail the experimental results after defining the evaluation metrics used in our study.

A. PERFORMANCE EVALUATION OF TCEF AND CONVENTIONAL FEATURES USING NEURAL NETWORKS

To validate the effectiveness of TCEF compared to conventional methods in speaker identification, our study used two distinct neural network models. 1D-CNN was explicitly used for detailed analysis at the frame level, while an LSTM network was used for a comprehensive evaluation of sequence-level features.

1) FRAME-LEVEL ANALYSIS USING 1D-CNN

1D-CNN are widely recognized in speaker identification for their ability to process and analyze complex audio data, extracting distinct features essential to differentiate individual speakers [5], [25], [42]. In our study, a 1D-CNN, as shown in Fig. 4 (a), is used specifically for frame-level feature analysis. Frame-level analysis using 1D-CNN plays a crucial role in our comparative evaluation of TCEF and conventional features in speaker identification. It enables us to assess the effectiveness of both TCEF and conventional features in distinguishing unique vocal characteristics within

individual frames in different acoustic environments. The network starts with a convolutional layer that filters and extracts relevant features from audio frames, using multiple filters designed to target specific audio patterns crucial for speaker differentiation. Following this, the ReLU activation function is applied, introducing non-linearity that is vital for recognizing complex audio patterns unique to different speakers. Subsequent layers are organized into two distinct blocks: Block 1 and Block 2. Each block includes convolutional stages paired with batch normalization, followed by a ReLU activation and then MaxPooling. Batch normalization within these blocks accelerates and stabilizes the training by adjusting and scaling activations, significantly contributing to more effective model training. The MaxPooling layers in each block reduce the dimensions of the processed features, effectively downsampling the features and directing the network analysis toward the most relevant features for speaker identification. Following these blocks' convolutional and batch normalization layers, the network includes a flattening layer. This layer transforms the multidimensional feature maps into a single vector, a crucial step for transitioning the data into the dense layers. These dense layers are integral to associating the extracted features with specific output classes. The network concludes with a dense layer using a softmax activation function, which converts the learned features into a probabilistic distribution over the identities of the potential speakers.

2) SEQUENCE-LEVEL ANALYSIS USING LSTM

LSTM networks are widely used in speaker identification due to their ability to handle temporal dependencies effectively [26], [44]. These networks excel at capturing and preserving information through sequences of data, which is essential for recognizing speaker identity by analyzing speech patterns that change over time. In our study, as shown in Fig. 4 (b), we used LSTM networks for the sequence-level analysis. This method is vital to our study, as it analyzes sequences of audio frame features, rather than individual frames. This sequence-level analysis is the key to an in-depth comparison between TCEF and conventional features, allowing us to assess how they capture and differentiate unique

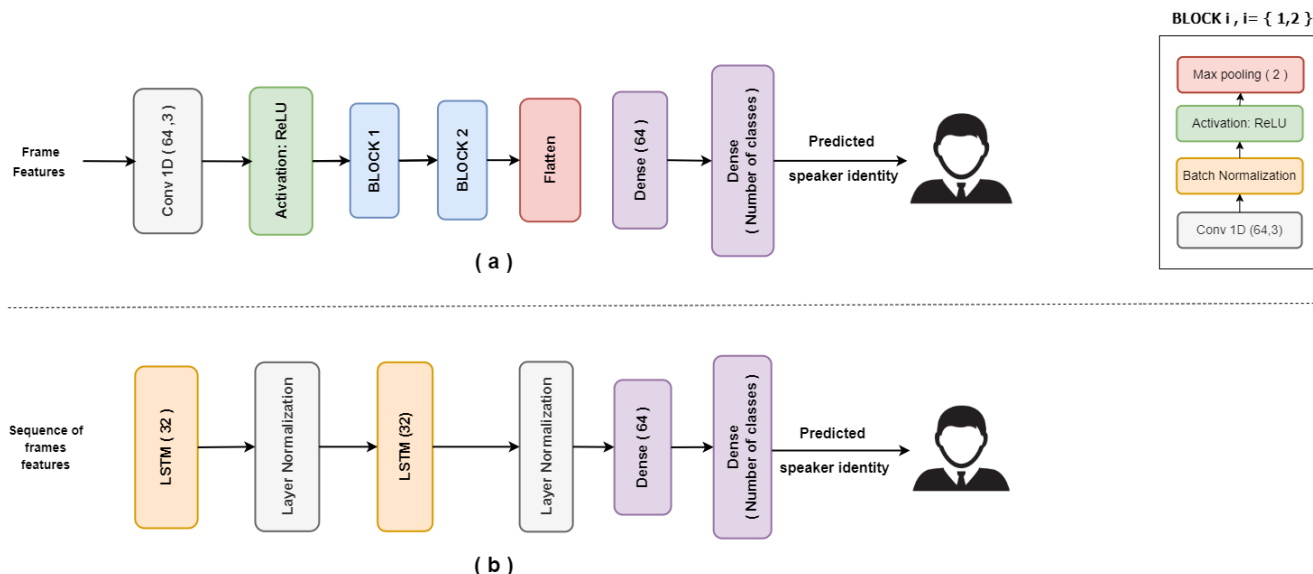


FIGURE 4. Neural network architectures to assess TCEF compared to conventional features in speaker identification. (a) 1D-CNN architecture for frame-level analysis. (b) LSTM architecture for sequence-level analysis.

vocal characteristics in a sequence of frames. The LSTM layer in our neural network consists of 32 units each, chosen for its effectiveness in processing temporal data. Following these layers, Layer Normalization is applied to stabilize and expedite the training process. Subsequently, a second LSTM layer, comprising 32 units, is used to refine the temporal analysis further. Another Layer Normalization follows this step to maintain the consistency and stability of the learning process, ensuring accurate modeling of temporal dynamics in speech. The network then incorporates a series of dense layers, further processing the temporal data extracted by the LSTM. The final stage of the network is the softmax activation function in a dense layer. This layer is critical as it converts the complex temporal features identified by the LSTM network into a probability distribution across the potential speaker identities.

B. DATASETS

1) GRID DATASET

The GRID audiovisual corpus [7] is a valuable resource for text-independent speaker identification. It contains high-quality audio and video recordings from 34 speakers (18 male and 16 female), totaling 34,000 sentences. Each speaker contributed 1,000 sentences recorded under neutral conditions at a standard speech rate. The speakers were instructed to deliver each sentence in 1 to 2 seconds. All utterances were recorded at a 25 kHz sampling rate and are available in WAV file format. The GRID dataset is used in our study to evaluate the performance of TCEF compared to conventional features in neutral speech.

2) RAVDESS DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [23] is widely used in the field

of speaker identification [9], [10], known for its diverse emotional speech content in a North American accent. This dataset includes 16-bit audio files sampled at 48 kHz, available in WAV format, with 1,440 utterances. It contains recordings from 24 professional actors (12 female and 12 male), each contributing 60 trials. These actors expressed the two sentences “Dogs are sitting by the door” and “Kids are talking by the door” across various emotional states. RAVDESS encompasses various emotional expressions such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Apart from neutral, each emotional state is represented at two intensity levels: normal and strong, with each emotion having 192 utterances and the neutral category comprising 96 utterances. In our study, the RAVDESS dataset is crucial for evaluating the effectiveness of TCEF compared to conventional features in emotional speech conditions.

3) GRID-NR AND RAVDESS-NR SIMULATED DATASETS

To evaluate the robustness of TCEF against conventional features in challenging acoustic environments, we developed two simulated datasets: GRID-NR and RAVDESS-NR. These datasets are derived from the original GRID and RAVDESS datasets, respectively, specifically engineered to simulate neutral and emotional speech conditions influenced by added noise and reverberation. This approach allowed us to validate the performance of our feature extraction techniques in more complex acoustic scenarios, reflecting the challenges often encountered in real-world environments.

a: ADDING NOISE

Gaussian noise was added to the original audio samples to simulate environmental noise conditions for the GRID-NR and RAVDESS-NR datasets. This modification was intended to generate versions of the GRID and RAVDESS datasets

with incorporated noise characteristics, thus representing neutral and emotional speech in more challenging acoustic settings. The process begins with the original input signal, represented as $S(t)$ from either the neutral GRID or emotional RAVDESS dataset. The initial step involves calculating the power of this signal, denoted P_{signal} , which is determined by averaging the square of the signal amplitude over time:

$$P_{\text{signal}} = \frac{1}{T} \sum_{t=1}^T S(t)^2 \quad (24)$$

where, T is the total duration of the signal in samples, and $S(t)$ is the signal amplitude at time t . Next, the noise power P_{noise} is determined based on a desired Signal-to-Noise Ratio (SNR), set to vary uniformly between 5 and 20 dB:

$$P_{\text{noise}} = P_{\text{signal}} \times 10^{-\frac{\text{SNR}_{\text{dB}}}{10}} \quad (25)$$

To create the noisy signal, Gaussian noise is generated with a mean of zero and a variance equal to P_{noise} . This noise is then added to the original signal, resulting in the noisy signal $S_{\text{noisy}}(t)$, which is computed as follows:

$$S_{\text{noisy}}(t) = S(t) + N_s \quad (26)$$

where $N_s \sim \mathcal{N}(0, \sqrt{P_{\text{noise}}})$ denotes Gaussian white noise.

b: ADDING REVERBERATION

To simulate reverberation effects in the audio samples of the GRID and RAVDESS datasets, we used room simulation techniques based on the Pyroomacoustics package [34]. This process included creating a virtual room with specific dimensions and sound-absorbing properties, setting a source and microphone position within this virtual space, and then computing the Room Impulse Response (RIR). The RIR, denoted $R(t)$, reflects how the sound propagates and interacts with the surfaces of the room. The original audio signal $S(t)$ was then convolved with this RIR to add reverberation, simulating the effect of a real-life reverberant environment. This convolution is mathematically represented as follows:

$$S_{\text{reverb}}(t) = S(t) * R(t) \quad (27)$$

where, $S_{\text{reverb}}(t)$ is the resulting reverberant signal, and $*$ is the convolution operator.

c: COMBINING NOISE AND REVERBERATION

We sequentially combined noise and reverberation effects to simulate complex acoustic environments in our datasets. After introducing Gaussian noise to the original signal, resulting in $S_{\text{noisy}}(t)$, we then applied reverberation to this noisy signal. The final signal, which incorporates both noise and reverberation and denoted $S_{\text{noisy_reverb}}(t)$ is obtained by convolving the noisy signal with RIR $R(t)$:

$$S_{\text{noisy_reverb}}(t) = S_{\text{noisy}}(t) * R(t) \quad (28)$$

In this equation, $S_{\text{noisy_reverb}}(t)$ is the signal that has been processed through both acoustic transformations, first adding

noise and then applying reverberation. This approach ensures that the final simulated signal accurately mimics the audio characteristics of real-world environments where noise and reverberation are present. These modified datasets GRID-NR and RAVDESS-NR with added noise and reverberation provided a comprehensive platform to compare the effectiveness of TCEF against conventional features under more realistic conditions.

C. BASELINES

For the baseline comparison in our study, we used three conventional feature extraction techniques commonly used in speaker identification: MFCC, GTCC, and PNCC, along with their respective first- and second-order derivatives.

1) CONVENTIONAL FEATURE EXTRACTION TECHNIQUES

The parameters selected for conventional feature extraction techniques are commonly used in speech processing research [40], [46], ensuring alignment with established practices in the field.

MFCC: The extraction involved pre-emphasizing the speech signal, segmenting into 25ms frames with a 10ms shift, applying a Hamming window, transforming using FFT of 1024, filtering through 40 Mel-scaled triangular filters, and deriving 12 coefficients with a DCT.

GTCC: The procedure included dividing the signal into 25 ms frames with a 10 ms shift, windowing using Hamming, applying FFT of 1024, filtering through a 40-filter Gammatone bank, logarithmic compression, and generating 12 coefficients with a DCT.

PNCC: This method started with pre-emphasizing the signal, framing into 25 ms with 10 ms overlap, using a Hamming window, applying FFT of 1024, filtering through a 40-filter Gammatone bank, applying a power-law function with an exponent of 1/15, and concluding with a DCT to obtain 12 coefficients.

2) COMPUTATION OF FIRST AND SECOND-ORDER DERIVATIVES

In our analysis, first- and second-order derivatives were calculated for each feature extraction technique: MFCC, GTCC, and PNCC, using a span of five frames. This method considers the target frame and two frames before and after, providing a broader temporal context for each frame. Such a five-frame span ensures a thorough examination of the temporal dynamics in the speech signal. Applying this uniform approach in MFCC, GTCC, and PNCC is crucial for an accurate evaluation of temporal variations in speech, which contributes significantly to the robustness of our feature analysis.

D. OVERVIEW OF FEATURE SETS FOR EVALUATION

We investigated various feature extraction techniques, including conventional features, TCEF, and an examination of the effects of combining these two types of features. In the feature

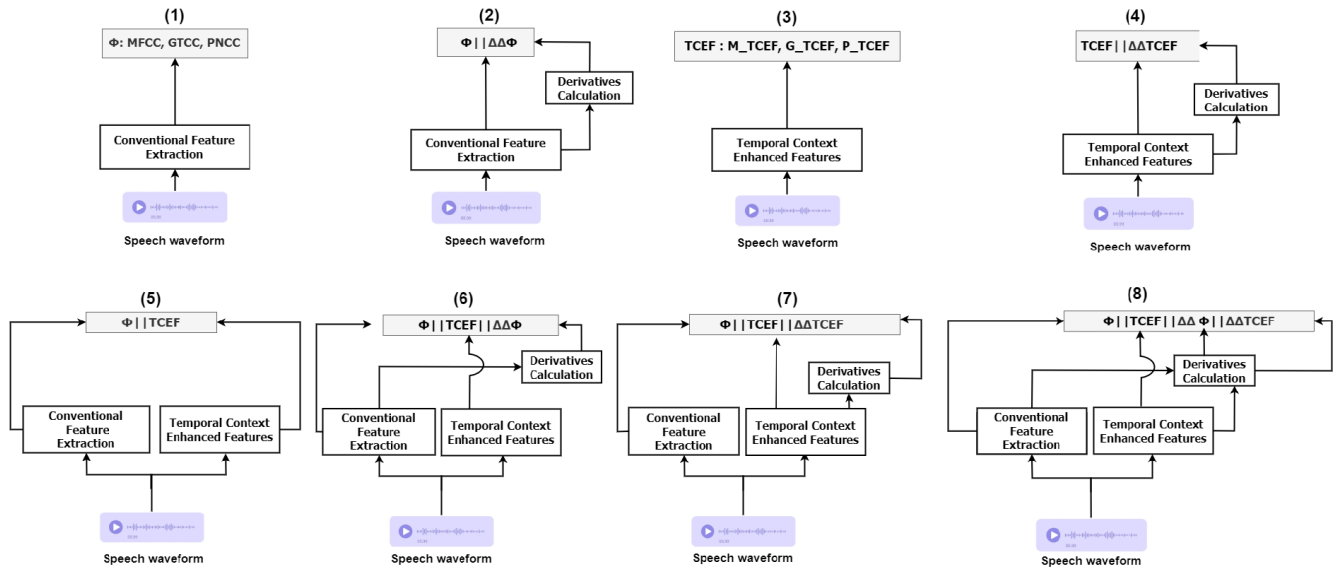


FIGURE 5. Combinations of feature sets for enhanced speaker identification analysis. (1) conventional features, (2) conventional features with derivatives, (3) temporal context enhanced features, (4) temporal context enhanced features with derivatives, (5) combination of TCEF features with conventional features, (6) combination of TCEF features with conventional features and their derivatives, (7) combination of TCEF features with conventional features and derivatives of TCEF features, and (8) combination of TCEF features with conventional features including all derivatives. The vertical stacking of feature sets is denoted by the ‘||’ operator.

sets detailed in this section, the ‘||’ operator signifies the vertical stacking of feature vectors. This method combines different feature sets, leading to an integrated feature vector that encapsulates the characteristics of each set. The various feature sets used in this investigation are depicted in Fig. 5 and are:

- 1) Conventional features:
 - MFCC
 - GTCC
 - PNCC
- 2) Conventional features with derivatives:
 - MFCC||ΔΔMFCC
 - GTCC||ΔΔGTCC
 - PNCC||ΔΔPNCC
- 3) Temporal context enhanced features:
 - M_TCEF
 - G_TCEF
 - P_TCEF
- 4) Temporal context enhanced features with derivatives:
 - M_TCEF||ΔΔM_TCEF
 - G_TCEF||ΔΔG_TCEF
 - P_TCEF||ΔΔP_TCEF
- 5) Combination of TCEF features with conventional features:
 - MFCC||M_TCEF
 - GTCC||G_TCEF
 - PNCC||P_TCEF
- 6) Combination of TCEF features with conventional features and their derivatives:
 - MFCC||M_TCEF||ΔΔMFCC
 - GTCC||G_TCEF||ΔΔGTCC
 - PNCC||P_TCEF||ΔΔPNCC

- 7) Combination of TCEF features with conventional features and derivatives of TCEF features:
 - MFCC||M_TCEF||ΔΔM_TCEF
 - GTCC||G_TCEF||ΔΔG_TCEF
 - PNCC||P_TCEF||ΔΔP_TCEF
- 8) Combination of TCEF features with conventional features including all derivatives:
 - MFCC||M_TCEF||ΔΔMFCC||ΔΔM_TCEF
 - GTCC||G_TCEF||ΔΔGTCC||ΔΔG_TCEF
 - PNCC||P_TCEF||ΔΔPNCC||ΔΔP_TCEF

E. EXPERIMENTAL SETTINGS

In our study, which employed 1D-CNN for frame-level analysis and LSTM networks for sequence-level analysis, we maintained uniform experimental settings to ensure a fair comparison between conventional features and TCEF. An essential aspect of our experimental setup was the alignment of the context window size in TCEF with the sequence length in the LSTM network. This alignment is crucial for a fair and accurate comparison between the performance of TCEF and conventional feature extraction methods. For TCEF, the size of the context window was varied from 1 to 10 frames to assess the effectiveness with smaller and larger window sizes, providing information on the impact of temporal context on feature extraction. In particular, at a context window size of 1, TCEF generates conventional features, serving as a baseline for comparison. Concurrently, the LSTM network sequence length was adjusted to match the TCEF context window size, ensuring that both LSTM and conventional features were evaluated under equivalent temporal conditions. For example, if the context window size is 5, the LSTM processes sequences

of 5 frames, thus maintaining consistency in the analysis. To further maintain this consistency, the network models were optimized using the Adam optimizer, with a learning rate set to 0.001. The categorical cross-entropy loss function was employed for its suitability in multi-class classification tasks. To avoid overfitting, a dropout rate of 0.5 was implemented. Training was conducted over 100 epochs with a batch size of 32. The dataset was consistently split with 70% allocated for training, 20% for testing, and 10% for validation, uniformly applied for each considered dataset, including GRID, RAVDESS, GRID-NR, and RAVDESS-NR. All experiments were conducted in a high-performance computing environment equipped with an Intel Xeon(R) CPU at 2.20 GHz, 83.48 GB of memory, and an NVIDIA A100-SXM4-40GB GPU with 40 GB of memory.

F. EVALUATION METRICS

To effectively measure the performance of TCEF and conventional features in our speaker identification models, we used four widely recognized metrics for classification problems: Accuracy, Precision, Recall, and F1 Score. Each metric contributes to a comprehensive understanding of the models' performance.

1) ACCURACY

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (29)$$

Accuracy reflects the model overall rate of correct predictions in correctly identifying speakers. *TP* represents the number of correct positive identifications by the model, and *TN* represents the number of correct negative identifications. *FP* refers to the number of incorrect positive identifications, where the model incorrectly identifies a non-speaker as a speaker, and *FN* denotes the number of incorrect negative identifications, where the model fails to identify the actual speaker.

2) PRECISION

$$Precision = \frac{TP}{TP + FP} \quad (30)$$

Precision indicates the proportion of correctly identified speakers *TP* out of all instances where the model predicted the speaker identity.

3) RECALL

$$Recall = \frac{TP}{TP + FN} \quad (31)$$

Recall assesses the model ability to correctly identify actual speakers, measuring how many actual speakers *TP* were identified correctly out of all actual speaker instances.

4) F1 SCORE

$$F1 \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (32)$$

F1 score combines *Precision* and *Recall* into a single metric, balancing the accurate identification of speakers *Precision* and the ability to identify as many actual speakers as possible *Recall*. The *F1 Score* is used to validate the model effectiveness in correctly identifying speakers and minimizing false identifications, providing a comprehensive measure of performance accuracy.

5) STATISTICAL ANALYSIS

This study used Approximate Randomization (AR) [29] to evaluate the statistical significance of performance differences between speaker identification models using different feature extraction techniques. AR was selected for its adaptability to datasets that may not conform to normal distribution assumptions. This method involves shuffling data between groups multiple times and recalculating the performance metrics. The resulting *p* value indicates the probability that the observed performance differences could occur randomly without an actual distinction between the models. This approach provides a robust means to determine whether differences in model performance are statistically significant.

G. EXPERIMENTAL RESULTS

This section presents the experimental results for speaker identification obtained from the GRID, GRID-NR, RAVDESS, and RAVDESS-NR datasets. These results were achieved using both TCEF and conventional features, applied in frame- and sequence-level analyses.

1) EVALUATION ON GRID DATASET - NEUTRAL SPEECH CONDITIONS

In the frame-level analysis using 1D-CNN with the GRID dataset, representative of neutral speech conditions, TCEF consistently shows superior performance compared to conventional features. Table 1 presents the detailed performance results with the GRID dataset comparing TCEF to conventional features for both frame-level and sequence-level with context window size 10. Table 1 reveals that M_TCEF achieves an accuracy of 77.92%, compared to the 63.66% accuracy with conventional MFCC features, resulting in a difference of 14.26%. Similarly, in terms of the F1 score, M_TCEF records 77.86%, surpassing the 63.41% achieved by conventional MFCC features by a margin of 14.45%. Comparable performance enhancements are observed with GTCC and PNCC, where G_TCEF and P_TCEF show a difference in accuracy of 11.7% and 15.24%, respectively, compared to conventional features. Similarly, the F1 score differences for G_TCEF and P_TCEF are 12.31% and 14.64%, respectively, over their respective conventional feature extraction methods.

Integrating first- and second-order derivatives into the feature set significantly improves the performance of both TCEF and conventional features. This enhancement is clearly shown in Table 1, where M_TCEF $\|\Delta\Delta$ M_TCEF outperforms conventional features with derivatives. For instance,

TABLE 1. Speaker identification performance on the GRID dataset: comparative analysis of frame-level analysis with 1D-CNN and sequence-level analysis with LSTM using TCEF and conventional features under neutral conditions for context window size 10.

Feature Extraction	Technique	1D-CNN				LSTM			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MFCC	MFCC	63.66	63.73	63.10	63.41	91.05	91.06	90.40	90.73
	M_TCEF	77.92	78.04	77.68	77.86	91.91	92.08	91.55	91.81
	MFCC $\Delta\Delta$ MFCC	70.08	70.17	69.82	69.99	92.14	92.16	91.61	91.88
	M_TCEF $\Delta\Delta$ M_TCEF	84.04	84.08	83.98	84.03	94.17	94.26	93.92	94.09
	MFCC M_TCEF	82.07	82.22	81.65	81.93	93.26	93.40	92.92	93.16
	MFCC M_TCEF $\Delta\Delta$ MFCC	84.30	84.49	83.86	84.17	94.25	94.50	93.58	94.04
	MFCC M_TCEF $\Delta\Delta$ M_TCEF	84.54	84.56	84.07	84.31	94.10	94.12	93.33	93.72
MFCC M_TCEF $\Delta\Delta$ MFCC $\Delta\Delta$ M_TCEF	84.49	84.51	83.89	84.20	94.13	94.17	93.29	93.72	
GTCC	GTCC	67.71	67.97	67.11	67.54	92.23	92.35	91.88	92.11
	G_TCEF	79.41	79.48	79.04	79.26	92.53	92.84	92.03	92.43
	GTCC $\Delta\Delta$ GTCC	73.55	73.86	73.15	73.51	93.69	93.92	93.06	93.49
	G_TCEF $\Delta\Delta$ G_TCEF	85.62	85.87	85.07	85.46	95.88	96.16	95.12	95.64
	GTCC G_TCEF	83.94	84.11	83.38	83.75	94.27	94.77	93.46	94.11
	GTCC G_TCEF $\Delta\Delta$ GTCC	85.64	85.72	85.12	85.42	95.24	95.76	94.71	95.23
	GTCC G_TCEF $\Delta\Delta$ G_TCEF	86.06	86.38	85.60	85.99	95.65	96.15	95.02	95.58
GTCC G_TCEF $\Delta\Delta$ GTCC $\Delta\Delta$ G_TCEF	86.04	86.20	85.48	85.84	95.40	95.46	95.21	95.33	
PNCC	PNCC	56.04	56.11	55.89	56.00	88.84	89.00	88.21	88.61
	P_TCEF	71.28	71.67	70.74	71.20	88.97	89.08	88.34	88.71
	PNCC $\Delta\Delta$ PNCC	61.22	61.45	60.82	61.14	88.85	89.21	88.43	88.82
	P_TCEF $\Delta\Delta$ P_TCEF	77.73	78.08	77.35	77.71	91.35	91.64	90.94	91.29
	PNCC P_TCEF	74.74	75.06	74.17	74.61	90.44	90.94	89.70	90.32
	PNCC P_TCEF $\Delta\Delta$ PNCC	77.49	77.79	76.91	77.35	91.69	91.95	91.30	91.62
	PNCC P_TCEF $\Delta\Delta$ P_TCEF	77.42	77.46	77.30	77.38	91.02	91.15	90.59	90.87
PNCC P_TCEF $\Delta\Delta$ PNCC $\Delta\Delta$ P_TCEF	77.10	77.35	76.63	76.99	91.26	91.36	90.66	91.01	

M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 84.04%, which reflects a difference of 13.96% when compared to the 70.08% accuracy of MFCC|| $\Delta\Delta$ MFCC. This trend of improved performance after integrating derivatives is consistently observable for both G_TCEF|| $\Delta\Delta$ G_TCEF and P_TCEF|| $\Delta\Delta$ P_TCEF compared to the conventional GTCC|| $\Delta\Delta$ GTCC and PNCC|| $\Delta\Delta$ PNCC.

In the sequence-level analysis using LSTM with the GRID dataset, TCEF continues to exhibit superior performance over conventional features. As mentioned in Table 1, M_TCEF achieves an accuracy of 91.91% compared to 91.05% for conventional MFCC features, a difference of 0.86%. Similar performance enhancements are seen with G_TCEF and P_TCEF compared to the conventional GTCC and PNCC.

Incorporating first- and second-order derivatives into the feature set further enhances the performance for both TCEF and conventional features. As Table 1 shows, M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 94.17%, significantly exceeding the 92.14% accuracy of MFCC|| $\Delta\Delta$ MFCC. Comparable results are observed for G_TCEF|| $\Delta\Delta$ G_TCEF and P_TCEF|| $\Delta\Delta$ P_TCEF compared to the conventional GTCC|| $\Delta\Delta$ GTCC and PNCC|| $\Delta\Delta$ PNCC.

When examining the combinations of conventional features with TCEF in both frame- and sequence-level analyses on the GRID dataset, it is clear that the integration of conventional features with TCEF leads to enhanced performance compared to the use of conventional features alone and with their derivatives. As indicated in Table 1 in the frame-level analysis, MFCC||M_TCEF achieves an accuracy of 82.07%. Incorporating derivatives

of conventional features MFCC||M_TCEF|| $\Delta\Delta$ MFCC, the accuracy improves to 84.30%, and with the addition of M_TCEF derivatives, MFCC||M_TCEF|| $\Delta\Delta$ M_TCEF reaches 84.54%, while combining both types of derivatives MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF results in an accuracy of 84.49%. Similarly, in sequence-level analysis using LSTM, MFCC||M_TCEF achieves an accuracy of 93.26%. This accuracy increases to 94.25% for MFCC||M_TCEF|| $\Delta\Delta$ MFCC. It reaches 94.10% for MFCC||M_TCEF|| $\Delta\Delta$ M_TCEF. For the configuration MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF, the accuracy is 94.13%. Despite these enhancements, the AR analysis indicates that the performance differences between combinations of TCEF with conventional features and derivatives are not statistically significant. Furthermore, these combinations do not demonstrate a statistically significant improvement over the exclusive use of TCEF|| $\Delta\Delta$ TCEF.

Detailed performance results with the GRID dataset comparing TCEF with conventional features for both frame-level and sequence-level across different context window sizes, ranging from 1 to 10, are presented in Table 5. The TCEF-based approaches perform better than conventional feature extraction approaches for all context window sizes.

2) EVALUATION ON GRID-NR DATASET - NEUTRAL SPEECH WITH NOISE AND REVERBERATION

In the frame-level analysis using 1D-CNN on the GRID-NR dataset, which simulates neutral speech conditions with added noise and reverberation, TCEF consistently outperforms conventional feature extraction methods. Table 2 presents the detailed performance results with the GRID-NR

TABLE 2. Speaker identification performance on the GRID-NR dataset: comparative analysis of frame-level analysis with 1D-CNN and sequence-level analysis with LSTM using TCEF and conventional features under neutral conditions with noise and reverberation for context window size 10.

Feature Extraction	Technique	ID-CNN				LSTM			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MFCC	MFCC	37.02	37.23	36.72	36.97	71.82	72.01	71.20	71.60
	M_TCEF	52.40	52.68	51.96	52.32	73.10	73.27	72.57	72.92
	MFCC $\Delta\Delta$ MFCC	43.10	43.25	42.90	43.07	74.85	75.21	74.21	74.71
	M_TCEF $\Delta\Delta$ M_TCEF	59.39	59.67	58.98	59.32	81.19	81.49	80.61	81.05
	MFCC M_TCEF	56.14	56.35	55.65	56.00	78.96	79.26	78.65	78.95
	MFCC M_TCEF $\Delta\Delta$ MFCC	59.90	60.00	59.66	59.83	80.88	81.01	80.55	80.78
	MFCC M_TCEF $\Delta\Delta$ M_TCEF	<u>59.69</u>	<u>59.70</u>	<u>59.44</u>	<u>59.57</u>	81.14	81.19	80.71	80.95
MFCC M_TCEF $\Delta\Delta$ MFCC $\Delta\Delta$ M_TCEF	59.60	59.61	59.48	59.55	80.72	80.96	80.25	80.60	
GTCC	GTCC	50.04	50.29	49.60	49.94	79.19	79.59	78.53	79.06
	G_TCEF	62.27	62.35	62.15	62.25	81.68	81.68	81.44	81.56
	GTCC $\Delta\Delta$ GTCC	56.69	56.93	56.22	56.57	81.99	82.00	81.64	81.82
	G_TCEF $\Delta\Delta$ G_TCEF	71.24	71.38	70.70	71.04	86.68	86.89	86.23	86.56
	GTCC G_TCEF	67.41	67.56	66.97	67.26	84.85	85.08	84.45	84.77
	GTCC G_TCEF $\Delta\Delta$ GTCC	<u>71.33</u>	<u>71.47</u>	<u>71.07</u>	<u>71.27</u>	86.04	86.12	85.28	85.70
	GTCC G_TCEF $\Delta\Delta$ G_TCEF	70.89	71.10	70.67	70.88	87.02	87.03	86.95	86.99
	GTCC G_TCEF $\Delta\Delta$ GTCC $\Delta\Delta$ G_TCEF	71.52	71.65	71.32	71.48	86.61	86.82	85.96	86.39
PNCC	PNCC	39.10	39.25	38.81	39.03	74.10	74.28	73.46	73.87
	P_TCEF	53.69	53.87	53.47	53.67	76.83	76.91	76.51	76.71
	PNCC $\Delta\Delta$ PNCC	44.97	45.16	44.74	44.95	<u>76.47</u>	<u>76.62</u>	<u>76.16</u>	<u>76.39</u>
	P_TCEF $\Delta\Delta$ P_TCEF	61.09	61.27	60.72	60.99	81.55	81.70	81.27	81.48
	PNCC P_TCEF	57.72	58.04	57.36	57.70	79.70	80.11	79.29	79.70
	PNCC P_TCEF $\Delta\Delta$ PNCC	61.72	61.96	61.43	61.70	80.91	80.96	80.77	80.87
	PNCC P_TCEF $\Delta\Delta$ P_TCEF	61.82	62.09	61.45	61.77	<u>81.21</u>	<u>81.27</u>	<u>80.84</u>	<u>81.06</u>
	PNCC P_TCEF $\Delta\Delta$ PNCC $\Delta\Delta$ P_TCEF	61.22	61.45	60.74	61.09	81.38	81.57	80.98	81.27

dataset comparing TCEF to conventional features for both frame-level and sequence-level with context window size 10. As shown in Table 2, M_TCEF attains an accuracy of 52.40%, showing a difference of 15.38% compared to the 37.02% accuracy achieved by conventional MFCC features. In the case of GTCC, G_TCEF records an accuracy of 62.27%, which is 12.23% higher than the 50.04% accuracy of conventional GTCC features. Similarly, with PNCC, P_TCEF achieves an accuracy of 53.69%, exceeding the 39.10% accuracy of conventional features by 14.59%. Comparable improvements are noted for other performance evaluation metrics.

Integrating first- and second-order derivatives into the feature set significantly improves the performance of both TCEF and conventional features. As indicated in Table 2, M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 59.39%, which is notably higher by +16.29% compared to the 43.10% accuracy of MFCC|| $\Delta\Delta$ MFCC. Similar trends of enhanced performance are observed with the inclusion of derivatives for GTCC and PNCC.

In the sequence-level analysis performed with LSTM and the GRID-NR dataset, TCEF consistently demonstrates superior performance over conventional features. The results, as presented in Table 2, show M_TCEF achieving an accuracy of 73.10%, which is a difference of 1.28% compared to the 71.82% accuracy of conventional MFCC features. Similar improvements are observed with GTCC and PNCC.

Incorporation of first- and second-order derivatives into the feature set improves the performance of both TCEF and conventional features. As indicated in Table 2, M_TCEF|| $\Delta\Delta$ M_TCEF achieves the higher accuracy of

81.19%, reflecting a difference of 6.34% compared to the 74.85% accuracy of MFCC|| $\Delta\Delta$ MFCC. Similar performance enhancements are observed for GTCC and PNCC.

Investigation of combining conventional features with TCEF in both frame- and sequence-level analyses on the GRID-NR dataset indicates that combining conventional features with TCEF leads to better performance compared to using conventional features and conventional features with their derivatives. Specifically, Table 2 shows that for frame-level analysis using MFCC||M_TCEF achieves an accuracy of 56.14%. The addition of derivatives of conventional features MFCC||M_TCEF|| $\Delta\Delta$ MFCC increases this accuracy to 59.90%. Furthermore, incorporating TCEF derivatives, M_TCEF|| $\Delta\Delta$ M_TCEF results in an accuracy of 59.69%, while combining both types of derivatives MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF yields an accuracy of 59.60%. In the sequence-level analysis with LSTM, the inclusion of TCEF shows improved performance. MFCC||M_TCEF achieves an accuracy of 78.96%, which is further enhanced to 80.88% for MFCC||M_TCEF|| $\Delta\Delta$ MFCC, 81.14% for MFCC||M_TCEF|| $\Delta\Delta$ M_TCEF, while achieving 80.72% for MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF. However, according to AR, the performance differences between combinations of TCEF with conventional features and derivatives do not show statistical significance. Furthermore, these combinations do not show a statistically significant improvement over the TCEF|| $\Delta\Delta$ TCEF.

Detailed performance results on the GRID-NR dataset that compare TCEF to conventional features in frame-level and sequence-level analysis for different context window

TABLE 3. Speaker identification performance on the RAVDESS dataset: comparative analysis of frame-level analysis with 1D-CNN and sequence-level analysis with LSTM using TCEF and conventional features in emotional conditions for context window size 10.

Feature Extraction	Technique	1D-CNN				LSTM			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MFCC	MFCC	53.47	53.67	53.25	53.46	84.80	85.16	84.33	84.74
	M_TCEF	75.81	76.06	75.14	75.60	89.76	90.25	89.22	89.73
	MFCC $\Delta\Delta$ MFCC	57.69	57.89	57.40	57.64	87.14	87.24	86.63	86.93
	M_TCEF $\Delta\Delta$ M_TCEF	79.41	79.74	78.88	79.31	92.76	92.84	91.93	92.38
	MFCC M_TCEF	76.13	76.17	75.95	76.06	90.62	90.75	90.45	90.60
	MFCC M_TCEF $\Delta\Delta$ MFCC	78.85	79.23	78.40	78.81	93.00	93.01	92.35	92.68
	MFCC M_TCEF $\Delta\Delta$ M_TCEF	<u>78.51</u>	<u>78.71</u>	<u>78.27</u>	<u>78.49</u>	92.17	92.67	91.37	92.01
MFCC M_TCEF $\Delta\Delta$ MFCC $\Delta\Delta$ M_TCEF	78.32	78.61	77.63	78.12	92.21	92.69	91.61	92.15	
GTCC	GTCC	56.13	56.47	55.73	56.10	85.12	85.22	84.81	85.02
	G_TCEF	75.13	75.38	74.85	75.11	89.31	89.74	88.52	89.13
	GTCC $\Delta\Delta$ GTCC	61.03	61.20	60.56	60.87	87.48	87.81	87.09	87.45
	G_TCEF $\Delta\Delta$ G_TCEF	78.78	78.78	78.66	78.72	92.53	93.03	91.76	92.39
	GTCC G_TCEF	76.22	76.46	75.91	76.19	89.88	89.95	89.63	89.79
	GTCC G_TCEF $\Delta\Delta$ GTCC	77.87	78.21	77.37	77.79	92.97	93.33	92.53	92.93
	GTCC G_TCEF $\Delta\Delta$ G_TCEF	79.77	80.06	79.30	79.68	92.25	92.46	91.44	91.95
GTCC G_TCEF $\Delta\Delta$ GTCC $\Delta\Delta$ G_TCEF	78.99	79.17	78.45	78.81	91.85	91.86	91.52	91.69	
PNCC	PNCC	48.56	48.61	48.19	48.40	84.22	84.25	83.82	84.03
	P_TCEF	70.80	71.03	70.47	70.75	88.12	88.26	87.92	88.09
	PNCC $\Delta\Delta$ PNCC	52.16	52.32	51.90	52.11	84.84	85.21	84.23	84.71
	P_TCEF $\Delta\Delta$ P_TCEF	75.53	75.75	75.17	75.46	90.86	91.15	90.42	90.78
	PNCC P_TCEF	71.35	71.69	70.90	71.29	87.11	87.12	86.33	86.72
	PNCC P_TCEF $\Delta\Delta$ PNCC	73.67	73.81	73.04	73.42	90.30	90.43	90.14	90.28
	PNCC P_TCEF $\Delta\Delta$ P_TCEF	74.43	74.60	74.24	74.42	91.47	91.84	91.09	91.46
PNCC P_TCEF $\Delta\Delta$ PNCC $\Delta\Delta$ P_TCEF	73.36	73.63	72.70	73.16	<u>90.57</u>	<u>90.69</u>	<u>90.15</u>	<u>90.42</u>	

sizes, ranging from 1 to 10, are presented in Table 6. The TCEF-based approaches perform better than conventional feature extraction approaches for all context window sizes.

3) EVALUATION ON RAVDESS DATASET - EMOTIONAL SPEECH

In the frame-level analysis on the RAVDESS dataset for emotional speech using 1D-CNN, TCEF features consistently demonstrate superior performance compared to conventional features for all feature extraction techniques. Table 3 presents the detailed performance results with the RAVDESS dataset comparing TCEF to conventional features for both frame-level and sequence-level with context window size 10. As indicated in Table 3, M_TCEF records an accuracy of 75.81%, showing a difference of 20.34% compared to the 53.47% accuracy of conventional MFCC features. Similar performance enhancements with G_TCEF and P_TCEF are observed, where TCEF shows an increase of 19% and 22.24% in accuracy over conventional GTCC and PNCC features, respectively.

The integration of first- and second-order derivatives into the TCEF features further increases their effectiveness. For MFCC with derivatives, M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 79.41%, reflecting a difference of 21.72% compared to the 57.69% accuracy of MFCC|| $\Delta\Delta$ MFCC. This pattern of enhanced performance with the addition of derivatives is consistent for both GTCC and PNCC.

In the sequence-level analysis with LSTM, TCEF consistently outperforms conventional features. As shown in Table 3 M_TCEF achieves an accuracy of 89.76%, showing a difference of 4.96% compared to the 84.80% accuracy of

conventional MFCC features. These performance improvements are also observed with GTCC and PNCC.

Adding first- and second-order derivatives to both TCEF and conventional features significantly enhances their performance. As reported in Table 3, for instance, M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 92.76%, reflecting a difference of 5.62% compared to the 87.14% accuracy of MFCC|| $\Delta\Delta$ MFCC. This consistent improvement with the addition of derivatives is also observed for both GTCC and PNCC.

The combination of conventional features with TCEF was explored at the frame- and sequence-levels. As indicated in Table 3, integrating conventional features with TCEF leads to better performance compared to using conventional features and conventional features with their derivatives. In the frame-level analysis using 1D-CNN with MFCC, MFCC||M_TCEF achieves an accuracy of 76.13%. The inclusion of derivatives of conventional features, MFCC||M_TCEF|| $\Delta\Delta$ MFCC, increases the accuracy to 78.85%, and with the addition of TCEF derivatives, MFCC||M_TCEF|| $\Delta\Delta$ M_TCEF, the accuracy reaches 78.51%. The combination of both types of derivatives, MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF, results in an accuracy of 78.32%. Similar observations can be made for the sequence-level analysis using LSTM. However, AR analysis indicates that the performance differences between these combinations that incorporate derivatives are not statistically significant. Furthermore, these combinations do not show a statistically significant improvement over the exclusive use of TCEF with derivatives TCEF|| $\Delta\Delta$ TCEF.

Detailed performance results with the RAVDESS dataset that compare TCEF to conventional features in frame-level

TABLE 4. Speaker identification performance on the RAVDESS-NR dataset: comparative analysis of frame-level analysis with 1D-CNN and sequence-level analysis with LSTM using TCEF and conventional features in emotional speech with noise and reverberation for context window size 10.

Feature Extraction	Technique	1D-CNN				LSTM			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MFCC	MFCC	36.94	37.03	36.61	36.82	74.23	74.25	73.80	74.02
	M_TCEF	61.36	61.51	60.82	61.16	80.15	80.34	79.87	80.10
	MFCC $\Delta\Delta$ MFCC	41.01	41.17	40.80	40.98	78.04	78.38	77.59	77.98
	M_TCEF $\Delta\Delta$ M_TCEF	66.93	67.23	66.37	66.80	83.24	83.54	82.93	83.23
	MFCC M_TCEF	59.62	59.64	59.45	59.54	81.35	81.73	80.82	81.27
	MFCC M_TCEF $\Delta\Delta$ MFCC	63.91	64.23	63.43	63.83	84.82	84.92	84.30	84.61
	MFCC M_TCEF $\Delta\Delta$ M_TCEF	62.56	62.77	62.21	62.49	83.56	83.61	83.44	83.52
MFCC M_TCEF $\Delta\Delta$ MFCC $\Delta\Delta$ M_TCEF	62.18	62.43	61.91	62.17	84.85	85.15	84.37	84.76	
GTCC	GTCC	44.04	44.24	43.83	44.03	75.72	76.13	75.29	75.71
	G_TCEF	62.63	62.69	62.39	62.54	81.76	82.00	81.26	81.63
	GTCC $\Delta\Delta$ GTCC	48.09	48.13	47.97	48.05	78.52	78.52	78.31	78.42
	G_TCEF $\Delta\Delta$ G_TCEF	68.01	68.02	67.59	67.80	85.88	86.05	85.63	85.84
	GTCC G_TCEF	64.97	65.29	64.40	64.84	83.55	83.64	83.36	83.50
	GTCC G_TCEF $\Delta\Delta$ GTCC	68.09	68.22	67.95	68.08	85.35	85.48	84.78	85.13
	GTCC G_TCEF $\Delta\Delta$ G_TCEF	69.22	69.31	68.93	69.12	86.29	86.54	85.88	86.21
GTCC G_TCEF $\Delta\Delta$ GTCC $\Delta\Delta$ G_TCEF	68.00	68.19	67.66	67.93	85.89	85.90	85.54	85.72	
PNCC	PNCC	37.37	37.47	37.06	37.26	68.58	68.78	68.26	68.52
	P_TCEF	59.86	60.16	59.38	59.77	77.48	77.57	77.20	77.39
	PNCC $\Delta\Delta$ PNCC	40.44	40.46	40.33	40.40	74.37	74.76	73.82	74.29
	P_TCEF $\Delta\Delta$ P_TCEF	63.33	63.57	63.01	63.29	81.33	81.38	81.01	81.20
	PNCC P_TCEF	58.50	58.51	58.20	58.36	78.18	78.60	77.52	78.06
	PNCC P_TCEF $\Delta\Delta$ PNCC	60.65	61.00	60.18	60.58	82.08	82.43	81.57	82.00
	PNCC P_TCEF $\Delta\Delta$ P_TCEF	61.49	61.57	60.98	61.27	82.66	82.79	82.36	82.57
PNCC P_TCEF $\Delta\Delta$ PNCC $\Delta\Delta$ P_TCEF	60.17	60.17	59.70	59.94	81.85	81.91	81.39	81.65	

and sequence-level analysis for different context window sizes, ranging from 1 to 10, are presented in Table 7. The TCEF-based approaches perform better than conventional feature extraction approaches for all context window sizes.

4) EVALUATION ON RAVDESS-NR DATASET - EMOTIONAL SPEECH WITH NOISE AND REVERBERATION

In the frame-level analysis using 1D-CNN with the RAVDESS-NR dataset for emotional speech with added noise and reverberation, TCEF consistently demonstrates superior performance over all conventional feature extraction techniques. As indicated in Table 4, M_TCEF achieves an accuracy of 61.36%, a difference of 24.42% compared to the 36.94% accuracy of conventional MFCC features. Similar levels of enhanced performance are observed with G_TCEF and P_TCEF, where TCEF shows an increase of 18.59% and 22.49% in accuracy over conventional GTCC and PNCC features, respectively.

Including first- and second-order derivatives in both TCEF and conventional features further improves their effectiveness. Table 4 demonstrates that M_TCEF|| $\Delta\Delta$ M_TCEF reaches an accuracy of 66.93%, reflecting a difference of 25.92% compared to the 41.01% accuracy achieved by conventional MFCC features with derivatives, MFCC|| $\Delta\Delta$ MFCC. This pattern of consistent improvement with integrating derivatives is also observed for GTCC and PNCC.

In the sequence-level analysis using LSTM on the RAVDESS-NR dataset, TCEF consistently outperforms conventional features across all evaluated feature extraction techniques. As detailed in Table 4, M_TCEF achieves an accuracy of 80.15%, showing a difference of 5.92%

compared to the 74.23% accuracy of conventional MFCC features. These improvements in performance are also observed with GTCC and PNCC.

Integrating first- and second-order derivatives into TCEF further enhances its performance, as demonstrated in Table 4. M_TCEF|| $\Delta\Delta$ M_TCEF achieves an accuracy of 83.24%, reflecting a difference of 5.2% compared to the 78.04% accuracy of MFCC|| $\Delta\Delta$ MFCC.

Combining conventional features with TCEF in the RAVDESS-NR dataset improves performance in both frame- and sequence-level analyses compared to using conventional features and conventional features with their first and second derivatives. However, as shown in Table 4, the effectiveness of these combinations is more noticeable when derivatives are included. In the frame-level analysis using 1D-CNN with MFCC, MFCC||M_TCEF achieves an accuracy of 59.62%. With the introduction of derivatives of conventional features, MFCC||M_TCEF|| $\Delta\Delta$ MFCC, the accuracy increases to 63.91%, and with the addition of TCEF derivatives, MFCC||M_TCEF|| $\Delta\Delta$ M_TCEF, the accuracy reaches 62.56%. Combining both types of derivatives, MFCC||M_TCEF|| $\Delta\Delta$ MFCC|| $\Delta\Delta$ M_TCEF, results in an accuracy of 62.18%. Similar observations can be made for the sequence-level analysis using LSTM. Despite these enhancements, AR analysis indicates that the performance differences between these combinations with derivatives are not statistically significant. Furthermore, these combinations do not offer a significant advantage over using TCEF with derivatives TCEF|| $\Delta\Delta$ TCEF, suggesting that the inclusion of derivatives in TCEF captures the essential information for speaker identification in these challenging conditions.

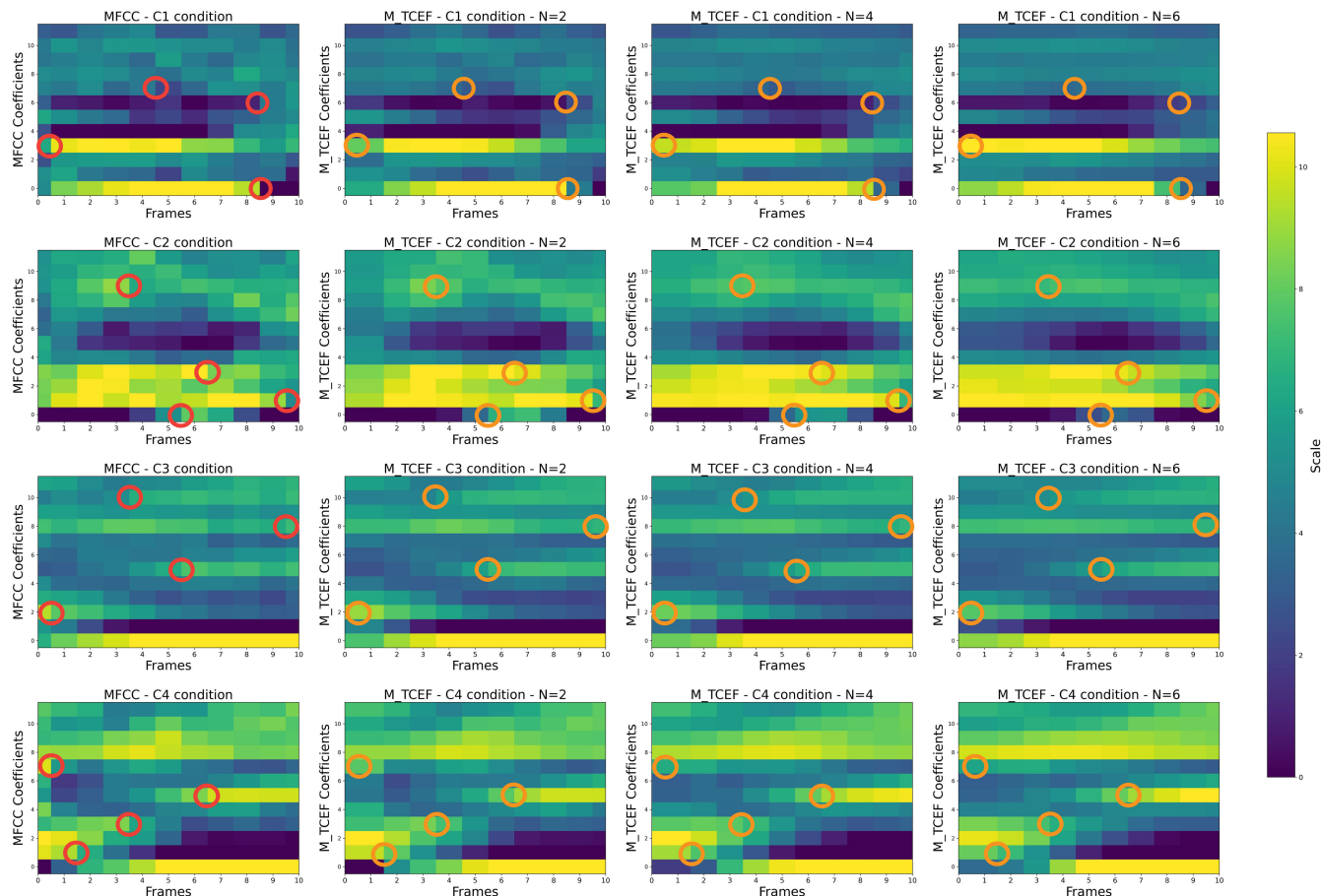


FIGURE 6. MFCC and M_TCEF heatmaps for various recording conditions across various context window sizes in four different scenarios: C1 - Neutral speech, C2 - Neutral speech with noise and reverberation, C3 - Emotional speech, C4 - Emotional speech with noise and reverberation. Red circles indicate regions of short-time variations present in the speech, while the orange circles represent the same regions after smoothing out short-time variations with TCEF.

Detailed performance results with the RAVDESS-NR dataset that compare TCEF to conventional features in frame-level and sequence-level analysis for different context window sizes, ranging from 1 to 10, are presented in Table 8. The TCEF-based approaches perform better than conventional feature extraction approaches for all context window sizes.

V. DISCUSSION

Performance evaluations in the various considered datasets, GRID, GRID-NR, RAVDESS, and RAVDESS-NR, show that TCEF consistently outperforms conventional features in speaker identification. Figure 6 presents heatmaps for MFCC and M_TCEF coefficients in diverse recording conditions. On the x-axis, frame indices from 1 to 10 are shown, each representing a 25 ms time slice, with the context window size varying from 2 to 4 to 6 frames. The y-axis represents the coefficients. In the MFCC heatmap, sharp transient features between consecutive frames within a short period are highlighted with red circles. These transients pose challenges in speaker identification, as they introduce

short-time variations in speaker features that conventional methods struggle to handle effectively. However, in the M_TCEF heatmap, a significant reduction in these short-time variations in acoustic features is noticed. To emphasize the comparative difference, the same regions are highlighted with orange circles in the M_TCEF heatmap. This visualization demonstrates how M_TCEF mitigates the effects of these transients on acoustic features, particularly when increasing the context window size. The impact of this increase leads to a decrease in transient variations and a more detailed representation of individual speaker characteristics, resulting in a more consistent representation of vocal features, essential for accurate speaker identification. Although this figure focuses on the impact of TCEF on MFCC, the effectiveness of TCEF extends to GTCC and PNCC.

Further investigation of the performance of TCEF in sequence-level analysis using LSTM networks under various acoustic conditions has revealed substantial insights. The analysis, as highlighted in Fig. 7, shows that in the GRID dataset, representative of neutral speech conditions, the performance difference between TCEF (M_TCEF, G_TCEF, and P_TCEF) and conventional features (MFCC, GTCC, and

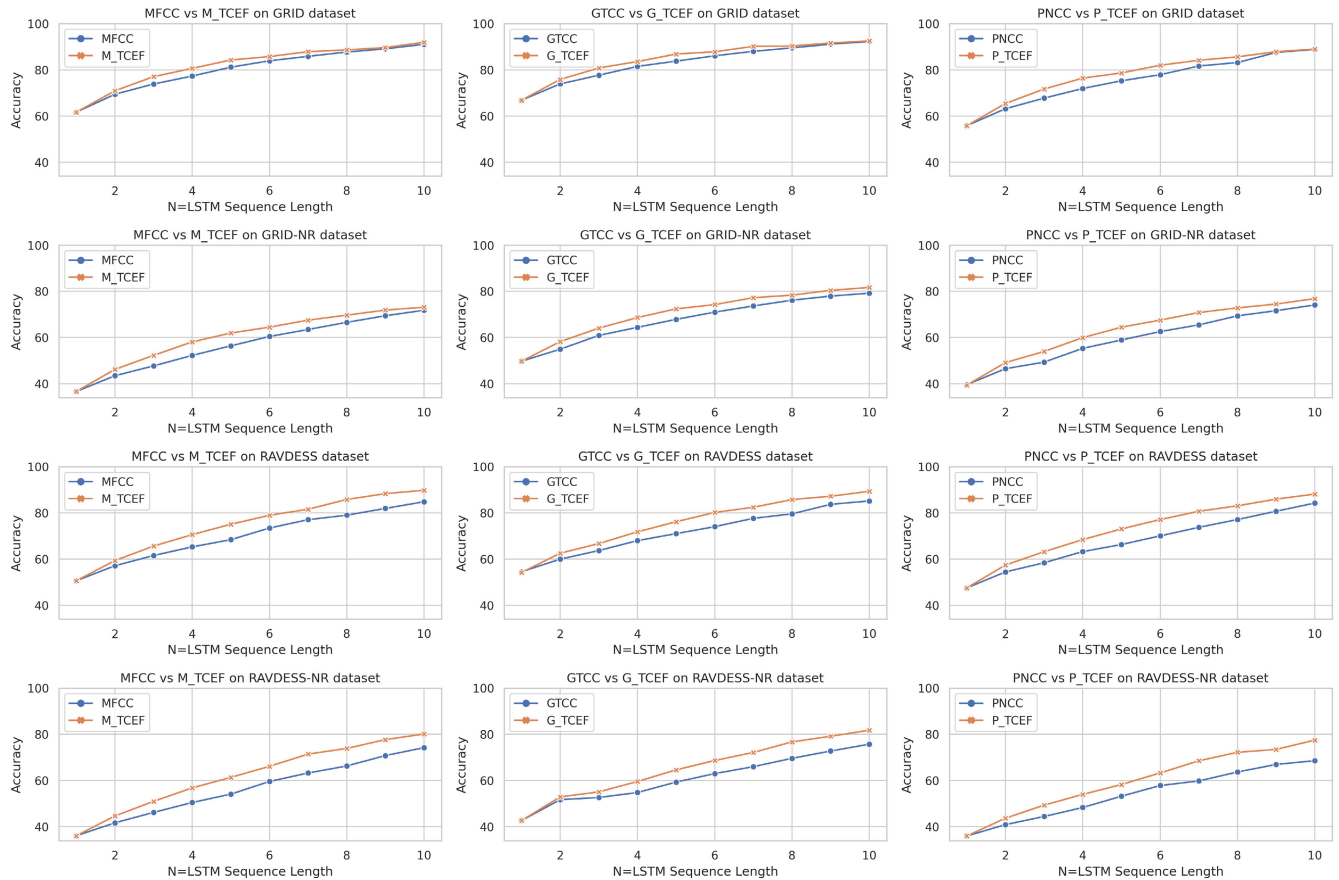


FIGURE 7. Evaluating speaker identification accuracy of MFCC and M_TCEF using LSTM network: impact of context window size and sequence length across datasets.

PNCC) is slight with increasing context window sizes and LSTM sequence lengths. This finding suggests that in simpler acoustic environments, LSTM networks improve the efficacy of conventional features, thus reducing the performance gap with TCEF. The GRID-NR dataset, characterized by noise and reverberation, shows an advantage of TCEF over conventional features. This advantage becomes more apparent in the RAVDESS and RAVDESS-NR datasets, which include emotional speech and emotional speech with noise and reverberation, respectively. In these complex acoustic scenarios, TCEF consistently surpasses conventional features, effectively extracting speaker-specific characteristics across all context window sizes and LSTM sequence lengths. While LSTM networks show a greater ability to capture temporal and speaker-specific features in neutral acoustic settings, TCEF consistently exhibits superior performance in more complex environments. This demonstrates the effectiveness of TCEF for speaker identification in a wide range of challenging acoustic conditions.

The insights into the effectiveness of TCEF in diverse acoustic environments prompt further examination of its computational efficiency and the impact of its integration with conventional features. This exploration extends our

analysis to the results of combining conventional features with TCEF. Analysis of the datasets indicates that the combination of TCEF and conventional features with derivatives does not lead to statistically significant performance improvements when evaluated through AR analysis, particularly when compared to TCEF|| $\Delta\Delta$ TCEF. In our comparative analysis of convergence time, as illustrated in Fig. 8, TCEF|| $\Delta\Delta$ TCEF achieves faster convergence than when conventional features are used alongside TCEF. This observation suggests that TCEF|| $\Delta\Delta$ TCEF offers a better balance between performance and efficiency. This efficiency, combined with its high level of performance, makes TCEF|| $\Delta\Delta$ TCEF a more effective practical approach for speaker identification in diverse acoustic environments.

Following the identification of TCEF|| $\Delta\Delta$ TCEF as a balanced approach between efficiency and performance, our study explored the influence of different context window sizes on its performance. As detailed in Fig. 9, the investigation examines TCEF|| $\Delta\Delta$ TCEF for frame-level analysis using 1D-CNN across context window sizes ranging from 1 to 50 frames. The 50-frame limit is chosen, as it is the maximum number of frames in audio samples across all datasets,

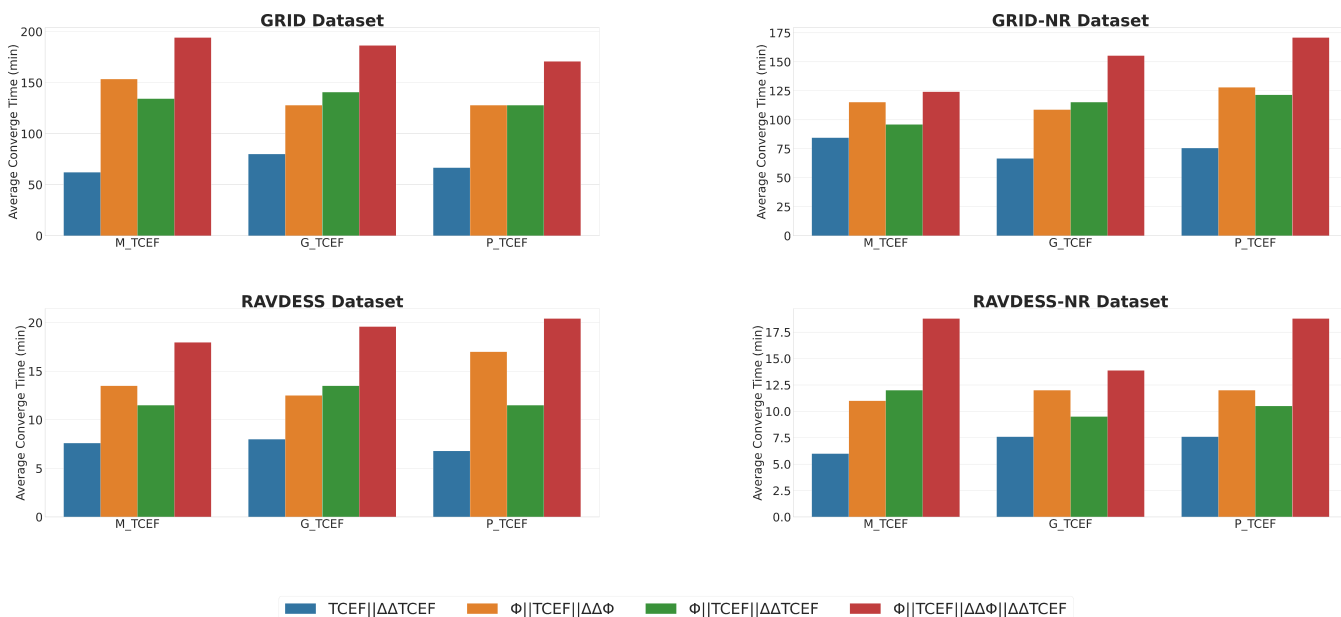


FIGURE 8. Average convergence time for the best performing techniques in all feature extraction techniques and datasets.

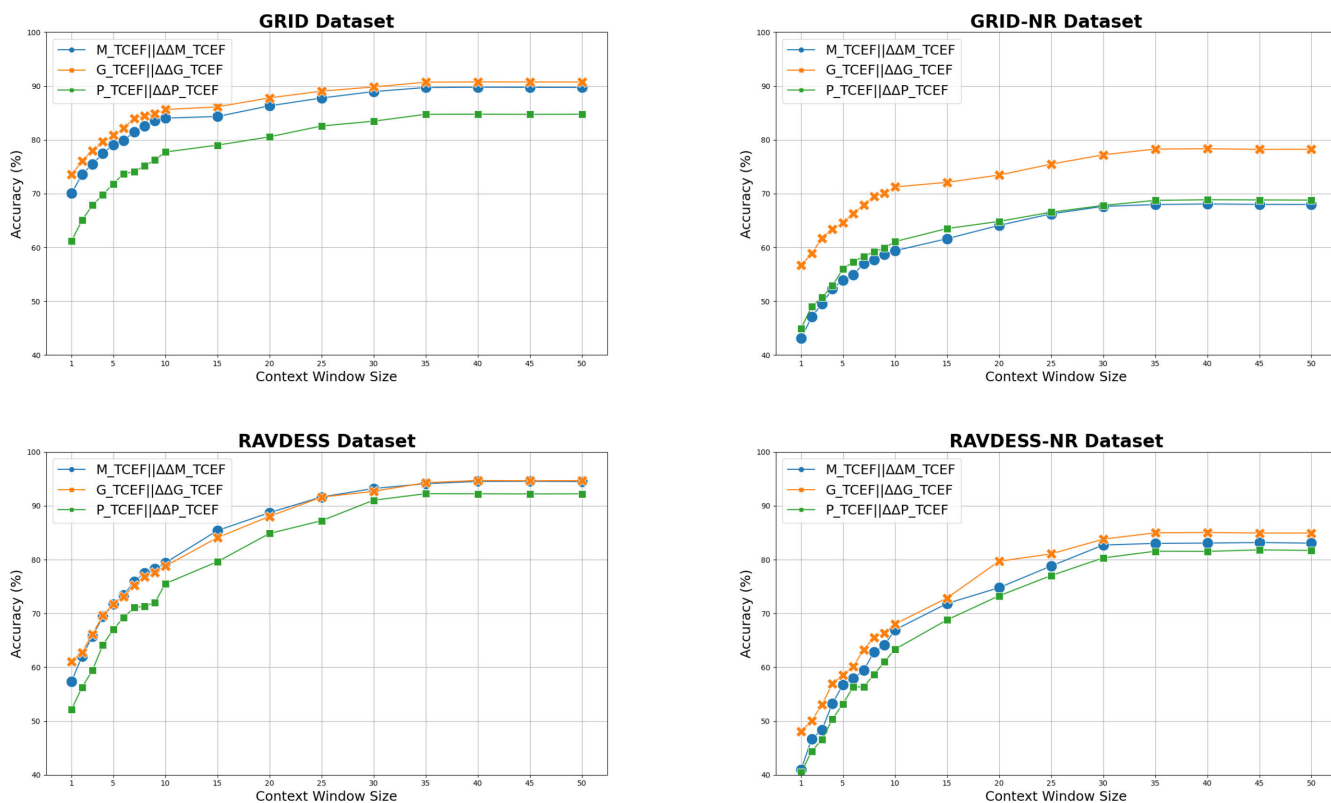


FIGURE 9. Performance evaluation of TCEF||ΔΔTCEF in diverse acoustic conditions: Speaker identification accuracy across context window sizes from 1 to 50 in frame-level analysis using 1D-CNN.

thereby setting the upper bound for the context window size in our analysis. Fig. 9 shows a consistent increase in accuracy for TCEF||ΔΔTCEF as the context window size expands.

However, the performance accuracy begins to stabilize at a context window size of 35 frames, and this pattern remains consistent up to the 50-frame limit. This finding is significant,

TABLE 5. Speaker identification performance on the GRID dataset for context window sizes 1 to 10: comparative analysis of frame-level analysis with 1D CNN and sequence-level analysis with LSTM using TCEF and conventional features under neutral conditions - Accuracy (%).

Feature Extraction	Technique	CW1		CW2		CW3		CW4		CW5		CW6		CW7		CW8		CW9		CW10	
		ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM	ID CNN	LSTM
MFCC	MFCC	63.66	61.65	63.66	69.43	63.66	73.85	63.66	77.30	63.66	81.15	63.66	83.90	63.66	85.85	63.66	87.68	63.66	89.10	63.66	91.05
	M_TCEF	63.66	61.65	66.24	70.92	68.77	77.04	70.79	80.63	72.08	84.25	73.37	85.74	74.88	87.84	75.90	88.62	76.67	89.63	77.92	91.91
	MFCC Δ MFCC	70.08	74.00	70.08	78.82	70.08	81.81	70.08	83.15	70.08	85.00	70.08	87.16	70.08	88.71	70.08	90.18	70.08	91.75	70.08	92.14
	M_TCEF Δ M_TCEF	70.08	74.00	73.54	79.68	75.52	83.17	77.47	86.51	79.08	88.01	79.90	89.73	81.47	91.77	82.59	92.66	83.53	94.17	84.04	94.17
	MFCCIM_TCEF	63.94	64.3	66.57	72.9	70.14	78.74	72.18	83.48	75.31	86.76	77.0	89.29	78.46	90.62	80.35	91.19	81.82	92.59	82.07	93.26
	MFCCIM_TCEF Δ MFCC	69.15	74.33	72.69	79.78	74.85	82.68	76.81	86.66	79.2	88.24	80.23	90.46	81.48	92.03	82.84	92.57	83.7	93.17	84.3	94.25
	MFCCIM_TCEF Δ M_TCEF	69.89	73.73	72.26	79.7	0.74	82.87	76.99	86.29	79.11	87.98	80.67	90.01	81.54	91.22	83.09	92.35	83.68	93.82	84.54	94.1
	MFCCIM_TCEF Δ MFCC Δ M_TCEF	70.08	75.17	72.73	80.01	75.01	82.29	77.45	86.27	79.49	88.52	80.64	90.03	81.67	91.11	82.84	92.57	83.62	93.34	84.49	94.13
	GTCC	67.71	66.85	67.71	73.89	67.71	77.66	67.71	81.48	67.71	83.75	67.71	86.07	67.71	88.00	67.71	89.50	67.71	91.12	67.71	92.23
	G_TCEF	67.71	66.85	69.63	75.82	71.16	80.80	72.42	83.58	74.16	86.83	75.01	87.81	76.61	90.15	77.28	90.31	78.36	91.69	79.41	92.53
GTCC	GTCC Δ GTCC	73.55	79.21	73.55	81.89	73.55	84.37	73.55	85.83	73.55	87.23	73.55	88.80	73.55	90.78	73.55	91.33	73.55	92.79	73.55	93.69
	G_TCEF Δ G_TCEF	73.55	79.21	76.04	82.37	78.00	85.90	79.62	88.27	80.88	89.93	82.15	91.39	83.96	93.93	84.47	94.58	84.82	95.02	85.62	95.88
	GTCCIG_TCEF	67.87	70.31	69.55	77.15	72.82	82.38	74.79	86.15	77.25	88.77	78.94	90.32	80.67	91.96	81.61	92.42	82.87	93.16	83.94	94.27
	GTCCIG_TCEF Δ GTCC	73.99	78.72	76.15	82.46	77.39	85.40	79.33	88.58	81.07	90.83	82.28	91.84	83.74	93.15	84.53	94.11	85.07	94.71	85.64	95.24
	GTCCIG_TCEF Δ G_TCEF	74.05	78.53	75.2	83.09	77.3	86.04	79.86	87.64	81.24	89.95	82.69	91.22	83.73	93.09	84.64	93.49	85.5	94.45	86.06	95.65
	GTCCIG_TCEF Δ GTCC Δ G_TCEF	72.55	79.12	75.65	83.57	77.8	85.91	79.95	88.41	81.55	90.01	82.44	91.90	84.08	92.86	84.73	93.51	85.41	94.71	86.04	95.40
	PNCC	56.04	55.82	56.04	63.15	56.04	67.73	56.04	71.90	56.04	75.24	56.04	77.88	56.04	81.65	56.04	83.15	56.04	87.51	56.04	88.84
	P_TCEF	56.04	55.80	59.05	65.41	61.49	71.68	63.67	76.40	65.28	78.63	66.62	81.99	67.64	84.13	69.02	85.63	70.04	87.83	71.28	88.97
	PNCC Δ PNCC	61.22	68.33	61.22	72.81	61.22	76.24	61.22	78.50	61.22	80.99	61.22	82.59	61.22	84.88	61.22	85.97	61.22	87.04	61.22	88.85
	P_TCEF Δ P_TCEF	61.22	68.33	65.05	75.34	67.85	78.49	69.79	82.17	71.82	85.33	73.71	86.08	74.07	88.39	75.14	89.22	76.26	89.86	77.73	91.35
PNCC	PNCCIP_TCEF	55.72	59.67	58.41	66.71	61.48	73.66	64.88	78.19	67.33	82.19	69.37	85.36	70.68	86.52	72.01	87.91	73.79	89.21	74.74	90.44
	PNCCIP_TCEF Δ PNCC	60.88	67.71	65.11	73.99	66.93	77.2	69.61	82.3	71.1	84.43	73.4	86.73	74.9	88.53	75.35	89.21	77.37	90.73	77.49	91.69
	PNCCIP_TCEF Δ P_TCEF	60.95	68.0	64.56	74.31	67.1	77.74	69.62	82.53	71.82	83.27	73.52	86.54	74.7	88.37	75.82	89.26	76.81	90.89	77.42	91.02
	PNCCIP_TCEF Δ PNCC Δ P_TCEF	60.4	68.54	64.06	75.69	67.55	79.64	69.16	81.74	71.66	84.81	73.2	85.65	75.46	87.65	76.01	89.54	76.95	89.93	77.1	91.26

as it indicates that extending the context window size beyond a specific threshold does not lead to additional significant performance improvements.

This study further investigated the performance of different feature extraction techniques under various acoustic conditions. In the GRID dataset, G_TCEF consistently outperforms M_TCEF and P_TCEF. This can be attributed to G_TCEF comprehensive spectral representation, which accurately captures the fine details of a speaker voice. In the GRID-NR dataset, representing scenarios with added noise and reverberation, G_TCEF significantly outperforms M_TCEF and P_TCEF due to its advanced auditory-simulating design. P_TCEF shows a slight improvement over M_TCEF, primarily because of its focus on noise reduction and speech enhancement. When we focused on the RAVDESS dataset, both M_TCEF and G_TCEF demonstrated enhanced effectiveness compared to P_TCEF. This improved performance is because M_TCEF and G_TCEF can capture the varied frequency components and intensity levels of emotional speech. Extending our analysis to the RAVDESS-NR dataset, simulating emotional speech with added noise and reverberation, G_TCEF outperforms M_TCEF and P_TCEF. M_TCEF still surpasses P_TCEF in the RAVDESS-NR dataset.

This study indicates that G_TCEF outperforms other feature extraction techniques across diverse recording conditions, making it the most appropriate choice for speaker identification in various acoustic environments.

VI. EVALUATION REPRODUCIBILITY

The Python code for the reproducibility of the evaluations presented in this paper is available on GitHub at <https://github.com/YassinTERRAF/TCEF>. This repository includes the necessary code to implement our TCEF approach and to conduct its evaluations. The datasets GRID-NR and RAVDESS-NR, which were created for use in this research, are available on Kaggle at <https://www.kaggle.com/datasets/yassinterraf/grid-nr> and <https://www.kaggle.com/datasets/yassinterraf/ravdess-nr> respectively. The dependencies for this project include Python,¹ along with essential libraries

such as pyroomacoustics,² librosa,³ scipy,⁴ gammatone,⁵ and tensorflow.⁶

VII. CONCLUSION

In this research, we introduced the TCEF approach, which uses a context window to average out features over adjacent frames. This technique is designed to mitigate short-term variations caused by noise, reverberation, and fluctuations in emotional speech and neutral speech recording to enhance speaker identification in diverse acoustic environments. Our comprehensive evaluations that use the GRID dataset for neutral speech, GRID-NR for neutral speech with noise and reverberation, RAVDESS for emotional speech, and RAVDESS-NR for emotional speech with noise and reverberation, have consistently shown that TCEF, particularly when integrated with derivatives, outperforms conventional feature extraction methods in frame-level and sequence-level analyses. These results validate the effectiveness of TCEF in extracting robust features for speaker identification. Further analysis of popular feature extraction techniques, including MFCC, GTCC, and PNCC, shows that TCEF with derivatives provides a better practical balance between high performance and computational efficiency, making it an advantageous choice for speaker identification under various acoustic recording conditions. However, the scope of the study is bounded by the specific nature of the datasets used. While comprehensive and covering a range of scenarios, these datasets mainly simulate controlled acoustic environments and may not cover the full spectrum of challenges present in real-world settings. In particular, elements such as real-life noise variations and overlapping speech scenarios are not represented. Therefore, future research is encouraged to extend the application of TCEF to environments that include these real-world acoustic challenges. Investigating how TCEF performs in situations with overlapping speech and

²<https://github.com/LCAV/pyroomacoustics>

³<https://librosa.org/>

⁴<https://scipy.org/>

⁵<https://github.com/detly/gammatone>

⁶<https://tensorflow.org/>

TABLE 6. Speaker identification performance on the GRID-NR dataset for context window sizes 1 to 10: comparative analysis of frame-level analysis with 1D CNN and sequence-level analysis with LSTM using TCEF and conventional features under neutral conditions with noise and reverberation - Accuracy (%).

Table with 28 columns: Feature Extraction, Technique, CW1, CW2, CW3, CW4, CW5, CW6, CW7, CW8, CW9, CW10, and 20 sub-columns for ID CNN and LSTM for each CW. Rows include MFCC, GTCC, and PNCC techniques with various feature combinations.

TABLE 7. Speaker identification performance on the RAUDES2 dataset for context window sizes 1 to 10: comparative analysis of frame-level analysis with 1D CNN and sequence-level analysis with LSTM using TCEF and conventional features under emotional conditions - Accuracy (%).

Table with 28 columns: Feature Extraction, Technique, CW1, CW2, CW3, CW4, CW5, CW6, CW7, CW8, CW9, CW10, and 20 sub-columns for ID CNN and LSTM for each CW. Rows include MFCC, GTCC, and PNCC techniques with various feature combinations.

TABLE 8. Speaker identification performance on the RAUDES2-NR dataset for context window sizes 1 to 10: comparative analysis of frame-level analysis with 1D CNN and sequence-level analysis with LSTM using TCEF and conventional features under emotional conditions with noise and reverberation - Accuracy (%).

Table with 28 columns: Feature Extraction, Technique, CW1, CW2, CW3, CW4, CW5, CW6, CW7, CW8, CW9, CW10, and 20 sub-columns for ID CNN and LSTM for each CW. Rows include MFCC, GTCC, and PNCC techniques with various feature combinations.

more naturalistic noise and reverberation will provide deeper insight into its practicality for speaker identification tasks in real-world scenarios.

APPENDIX

The detailed performance results with the considered datasets comparing TCEF with conventional features for both frame-level and sequence-level across different context window sizes, ranging from 1 to 10, are presented in this appendix.

A. GRID DATASET RESULTS

Table 5 presents the performance results with the GRID dataset.

B. GRID-NR DATASET RESULTS

Table 6 presents the performance results with the GRID-NR dataset.

C. RAUDES2 DATASET RESULTS

Table 7 presents the performance results with the RAUDES2 dataset.

D. RAVDESS-NR DATASET RESULTS

Table 8 presents the performance results with the RAVDESS-NR dataset.

REFERENCES

- [1] Z. Kh. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.
- [2] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating speaker height and subglottal resonances using MFCCs and GMMs," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 159–162, Feb. 2014.
- [3] R. Chakroun and M. Frikha, "A deep learning approach for text-independent speaker recognition with short utterances," *Multimedia Tools Appl.*, vol. 82, no. 21, pp. 33111–33133, Sep. 2023.
- [4] F. Z. Chelali and A. Djeradi, "Text dependant speaker recognition using MFCC, LPC and DWT," *Int. J. Speech Technol.*, vol. 20, no. 3, pp. 725–740, Sep. 2017.
- [5] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2020.
- [6] T. G. Clarkson, C. C. Christodoulou, Y. Guan, D. Gorse, D. A. Romano-Critchley, and J. G. Taylor, "Speaker identification for security systems using reinforcement-trained pRAM neural network architectures," *IEEE Trans. Syst., Man Cybern., C*, vol. 31, no. 1, pp. 65–76, Mar. 2001.
- [7] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "The grid audio-visual speech corpus," Tech. Rep., Jan. 2020.
- [8] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 342–350, Jun. 1981.
- [9] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier," *IEEE Access*, vol. 8, pp. 96994–97006, 2020.
- [10] S. Hamsa, I. Shahin, Y. Iraqi, E. Damiani, A. B. Nassif, and N. Werghi, "Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG," *Exp. Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 119871. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S095741742300372X>
- [11] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134018–134028, 2022.
- [12] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.
- [13] W. Hong and P. Jin'gui, "Modified MFCCs for robust speaker recognition," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 1, Oct. 2010, pp. 276–279.
- [14] A. K. Hunt and T. B. Schalk, "Simultaneous voice recognition and verification to allow access to telephone network services," U.S. Patent 5499288, Mar. 12, 1996.
- [15] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Exp. Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114591.
- [16] N. P. Jawarkar, R. S. Holambe, and T. K. Basu, "Text-independent speaker identification in emotional environments: A classifier fusion approach," in *Frontiers in Computer Education*. Berlin, Germany: Springer, 2012, pp. 569–576.
- [17] K. P. Bharath, "ELM speaker identification for limited dataset using multi-taper based MFCC and PNCC features with fusion score," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 28859–28883, Oct. 2020.
- [18] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.
- [19] A. Lauraitis, R. Maskeliunas, R. Damaševicius, and T. Krilavicius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96162–96172, 2020.
- [20] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall Signal Processing Series). 1993.
- [21] S. Lee, I. Seo, J. Seok, Y. Kim, and D. S. Han, "Active sonar target classification with power-normalized cepstral coefficients and convolutional neural network," *Appl. Sci.*, vol. 10, no. 23, p. 8450, Nov. 2020.
- [22] X. Liu, M. Sahidullah, and T. Kinnunen, "Optimized power normalized cepstral coefficients towards robust deep speaker verification," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 185–190.
- [23] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS)," Tech. Rep., Apr. 2018, doi: [10.5281/zenodo.1188976](https://doi.org/10.5281/zenodo.1188976).
- [24] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. ISMIR*, 2000, vol. 270, no. 1, p. 11.
- [25] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2016, pp. 1–6.
- [26] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement," *Multimedia Tools Appl.*, vol. 79, nos. 33–34, pp. 24013–24028, Sep. 2020.
- [27] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, and C. G. Dorny, "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic Sci. Int.*, vol. 263, pp. 92–100, Jun. 2016.
- [28] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.
- [29] E. W. Noreen, *Computer-Intensive Methods for Testing Hypotheses*. New York, NY, USA: Wiley, 1989.
- [30] S. O. Sadjadi and J. H. L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Commun.*, vol. 72, pp. 138–148, Sep. 2015.
- [31] M. E. Safi and E. I. Abbas, "Isolated word recognition based on PNCC with different classifiers in a noisy environment," *Appl. Acoust.*, vol. 195, Jun. 2022, Art. no. 108848.
- [32] D. Salvati, C. Drioli, and G. L. Foresti, "A late fusion deep neural network for robust speaker identification using raw waveforms and gammatone cepstral coefficients," *Exp. Syst. Appl.*, vol. 222, Jul. 2023, Art. no. 119750.
- [33] P. Sangwan, D. Deshwal, D. Kumar, and S. Bhardwaj, "Isolated word language identification system with hybrid features from a deep belief network," *Int. J. Commun. Syst.*, vol. 36, no. 12, p. e4418, Aug. 2023.
- [34] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.
- [35] I. Shahin, A. B. Nassif, N. Nemmour, A. Elnagar, A. Alhudhaif, and K. Polat, "Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments," *Neural Comput. Appl.*, vol. 33, no. 23, pp. 16033–16055, Dec. 2021.
- [36] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7554–7558.
- [37] X. Shi, H. Yang, and P. Zhou, "Robust speaker recognition based on improved GFCC," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2016, pp. 1927–1931.
- [38] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.
- [39] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1293–1302, 2020.
- [40] Y. Tang, C. Liu, Y. Leng, W. Zhao, J. Sun, C. Sun, R. Wang, Q. Yuan, D. Li, and H. Xu, "Attention based gender and nationality information exploration for speaker identification," *Digit. Signal Process.*, vol. 123, Apr. 2022, Art. no. 103449.
- [41] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.
- [42] K. Velayuthapandian and S. P. Subramoniam, "A focus module-based lightweight end-to-end CNN framework for voiceprint recognition," *Signal, Image Video Process.*, vol. 17, no. 6, pp. 2817–2825, Sep. 2023.

- [43] R. Wang, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, and Y. Zhang, "Multi-view self-attention based transformer for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6732–6736.
- [44] X. Wang, F. Xue, W. Wang, and A. Liu, "A network model of speaker identification with new feature extraction methods and asymmetric BLSTM," *Neurocomputing*, vol. 403, pp. 167–181, Aug. 2020.
- [45] Z. Wang and J. H. L. Hansen, "Multi-source domain adaptation for text-independent forensic speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 60–75, 2022.
- [46] M. Yousefi and J. H. L. Hansen, "Block-based high performance CNN architectures for frame-level overlapping speech detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 28–40, 2021.
- [47] Y. Zouhir, M. Zarka, and K. Ouni, "Power normalized gammachirp cepstral (PNGC) coefficients-based approach for robust speaker recognition," *Appl. Acoust.*, vol. 205, Mar. 2023, Art. no. 109272.



YASSIN TERRAF (Student Member, IEEE) received the engineering degree from the National School of Applied Sciences, Berrechid, Morocco, in 2021. He is currently pursuing the Ph.D. degree in signal processing with a focus on speaker recognition with the College of Computing, University Mohammed VI Polytechnic, Morocco. His Ph.D. degree was a big step toward his goal of becoming an expert in speaker recognition and audio analysis. He is a Data Scientist with

Henceforth, Rabat, Morocco, in addition to his academic pursuits. In this position, he uses his knowledge of data analysis and signal processing to address real-world problems in the field. His research efforts are intended to make a significant contribution to the knowledge and progress in this important field of signal processing.



YOUSSEF IRAQI (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Montreal, Canada, in 2000 and 2003, respectively. He is currently an Associate Professor with the College of Computing, University Mohammed VI Polytechnic, Morocco. Before that, he was with the Department of Electrical Engineering and Computer Science, Khalifa University (KU), United Arab Emirates, for 12 years. Before joining KU, he was the Chair

of the Department of Computer Science, Dhofar University, Oman, for four years. From 2004 to 2005, he was a Research Assistant Professor with the David R. Cheriton School of Computer Science, University of Waterloo, Canada. He has published over 130 research papers in international journals and refereed conference proceedings. His research interests include resource management in wireless networks, blockchain, trust and reputation management, cloud computing, and stylometry. In 2008, he received the IEEE Communications Society Fred W. Ellersick Paper Award in the field of communications systems. He is on many technical program committees of international conferences and is often approached for his expertise by international journals in his field.

• • •