mathematics

MDPI

Article

Voiceprint Recognition under Cross-Scenario Conditions Using Perceptual Wavelet Packet Entropy-Guided Efficient-Channel-Attention–Res2Net–Time-Delay-Neural-Network Model

Shuqi Wang [1], Huajun Zhang [1,*], Xuetao Zhang [1], Yixin Su [1] and Zhenghua Wang [2]

[1] School of Automation, Wuhan University of Technology, Wuhan 430070, China; shuqiw@whut.edu.cn (S.W.); zxtao@whut.edu.cn (X.Z.); suyixin@whut.edu.cn (Y.S.)
[2] Wuhan DaSoundGen Technologies Co., Ltd., Wuhan 430070, China; wang.zhenghua@dasoundgen.com
* Correspondence: zhanghj@whut.edu.cn

Abstract: (1) Background: Voiceprint recognition technology uses individual vocal characteristics for identity authentication and faces many challenges in cross-scenario applications. The sound environment, device characteristics, and recording conditions in different scenarios cause changes in sound features, which, in turn, affect the accuracy of voiceprint recognition. (2) Methods: Based on the latest trends in deep learning, this paper uses the perceptual wavelet packet entropy (PWPE) method to extract the basic voiceprint features of the speaker before using the efficient channel attention (ECA) block and the Res2Net block to extract deep features. The PWPE block removes the effect of environmental noise on voiceprint features, so the perceptual wavelet packet entropy-guided ECA–Res2Net–Time-Delay-Neural-Network (PWPE-ECA-Res2Net-TDNN) model shows an excellent robustness. The ECA-Res2Net-TDNN block uses temporal statistical pooling with a multi-head attention mechanism to weight frame-level audio features, resulting in a weighted average of the final representation of the speech-level feature vectors. The sub-center ArcFace loss function is used to enhance intra-class compactness and inter-class differences, avoiding classification via output value alone like the softmax loss function. Based on the aforementioned elements, the PWPE-ECA-Res2Net-TDNN model for speaker recognition is designed to extract speaker feature embeddings more efficiently in cross-scenario applications. (3) Conclusions: The experimental results demonstrate that, compared to the ECAPA-TDNN model using MFCC features, the PWPE-based ECAPA-TDNN model performs better in terms of cross-scene recognition accuracy, exhibiting a stronger robustness and better noise resistance. Furthermore, the model maintains a relatively short recognition time even under the highest recognition rate conditions. Finally, a set of ablation experiments targeting each module of the proposed model is conducted. The results indicate that each module contributes to an improvement in the recognition performance.

Keywords: voiceprint recognition; perceptual wavelet packet entropy; efficient channel attention; Res2Net; TDNN

MSC: 92-08

Citation: Wang, S.; Zhang, H.; Zhang, X.; Su, Y.; Wang, Z. Voiceprint Recognition under Cross-Scenario Conditions Using Perceptual Wavelet Packet Entropy-Guided Efficient-Channel-Attention–Res2Net–Time-Delay-Neural-Network Model. *Mathematics* **2023**, *11*, 4205. https://doi.org/10.3390/math11194205

Academic Editor: Jonathan Blackledge

Received: 8 September 2023
Revised: 2 October 2023
Accepted: 5 October 2023
Published: 9 October 2023

1. Introduction

A voiceprint refers to the spectrum of sound waves that convey speech information and is a biometric feature characterized by uniqueness and stability. The technology used to identify speaker identities with voiceprints is known as voice recognition technology. This technology is currently one of the popular methods for identity authentication. It has been applied in the judiciary, finance, and security fields, among others [1,2]. However, it still has a long way to go before achieving widespread implementation, particularly considering

that engineers frequently encounter significant discrepancies in the performance of voice recognition systems in real-world scenarios compared to their performance on standard test sets. For instance, the prevalent mainstream x-vector model generally exhibits an equal error rate of approximately 2% on the SITW test set, but the rates of false acceptance and false rejection significantly increase in real-life applications [3–5]. The cause of this phenomenon is cross-environmental recognition. If the domain from which the training data for voice recognition are collected is the source domain, and the domain where the test data are collected is the target domain, then, if there are notable disparities between the source and target domains regarding the feature space, the class space, or marginal distribution, then the efficiency of voice recognition models trained on the source domain will significantly degrade when applied to the target domain [6]. This makes it challenging for voice recognition technology to achieve large-scale applications. Cross-environment recognition is likely to be the most prevalent form of recognition in practical systems, as users tend to register their voiceprints in quiet environments where they can produce relatively stable and clear pronunciations, while, during verification, they may encounter various complex environments where environmental noises interfere with the recognition effectiveness. Most studies adopt scenarios that are overly simplistic, such as HI-MIA, which categorizes scenarios based on the distance between the recording device and the speaker [7–9]. These excessively idealistic research findings fail to reflect the true level of complexity involved in cross-environment recognition [10].

To investigate the problem of cross-environment recognition in practical settings, Tsinghua University's Center for Speech and Language Technology has recently launched CNCeleb, a database of Chinese celebrity voiceprints. This database comprises voice clips that have been publicly accessible on the internet from 3000 Chinese celebrities. An essential aspect of this database is its varied range of settings, covering singing, interviews, speeches, vlogs, and eleven other different scenarios [11,12]. It includes abundant cross-environment data, as speakers are exposed to varying levels of interference noise and may use different recording devices in each scene, as depicted in Figure 1.

Figure 1. Speaker and noise sources in vlog, singing, interview, and speech scenes.

In the singing scene, the sound of various musical instruments being played may interfere with speaker recognition. In the speech scene, applause from the audience may impact the recognition of the speaker. In vlog and interview scenes, car horns and background noise can potentially affect speaker recognition, and so on.

So far, cross-environment recognition remains one of the most important aspects in speaker recognition research. The improvement of cross-environment recognition can be approached from two perspectives: conducting research on speech feature extraction algorithms and investigating speaker modeling [13].

1. Speech feature extraction algorithms are one of the key technologies in voiceprint recognition. They are used to extract speech features from speech signals. Currently, the most widely used feature extraction methods are MFCC [14] and Fbank [15]. However, these traditional feature extraction methods have limited expressiveness for non-stationary signals and are sensitive to noise [16]. Some researchers have proposed using wavelet transform (WT) for extracting features from speech signals. Algorithms based on wavelet analysis, such as cepstral features, eliminate the need for Mel filtering and compress the spectral information of speech using the perceptual characteristics of wavelet transform. This simplifies the feature extraction process [17,18]. Additionally, utilizing Shannon entropy, which has a good stability, discriminability, and resistance to interference, researchers have further proposed the use of Shannon entropy extraction algorithms based on wavelet packet transform (WPT). This type of feature is composed of the entropy of the sub-band power spectrum of the signal's wavelet, which is capable of representing abrupt changes in the signal and has a low dimensionality, making it suitable for representing non-stationary signals [19,20]. To enhance the analysis capability of WPT for speech signals and reduce its computational complexity, researchers have also introduced the perceptual wavelet packet entropy (PWPE) feature extraction algorithm. It accurately analyses speech information, suppresses acoustic noise, reduces the number of parameters, and shortens the feature extraction time [21–23].
2. The objective of the research on speaker modeling is to develop speaker models capable of extracting speaker identity information from speech features. Popular speaker recognition models, such as ECAPA-TDNN utilizing time-delay neural networks and r-vector models employing deep residual networks, have demonstrated outstanding performances in text-independent speaker recognition tasks [24,25]. ECAPA-TDNN, proposed by Desplanques et al. from the University of Mons in Belgium in 2020, introduced the squeeze-excitation (SE) module and the channel attention mechanism for the first time [26–28]. This approach won first place in the international speaker recognition competition. However, ECAPA-TDNN is susceptible to noise interference, leaving room for improvement. Firstly, integrating noise reduction techniques can be beneficial in pre-processing speech features during feature extraction. Additionally, feature enhancement techniques can be employed to enhance the characteristics of the speech. Moreover, enhancing the network depth and attention pooling in the ECAPA-TDNN model can further improve its robustness. These enhancements strive to mitigate the impact of noise on the model's performance.

In light of the aforementioned analysis, this paper proposes two improvements to the existing voiceprint recognition models to make them more robust. Firstly, it employs perceptual wavelet entropy (PWE) for feature extraction and applies threshold denoising to the extracted features [29]. Voiceprint feature extraction methods such as MFCC and Fbank have a limited expressive ability for non-stationary signals and are sensitive to noise. However, the speech features processed via PWE can compensate for the shortcomings of MFCC and Fbank and improve the recognition performance robustly [30]. Secondly, this paper makes improvements to the classical speaker recognition model. It combines the non-dimensional-reducing efficient channel attention (ECA) block with Res2Net to assign weights to feature channels and introduces a multi-head attention mechanism. This allows the model to learn different behaviors without focusing attention excessively on its own position. The paper also introduces the sub-center ArcFace loss function module to mitigate the effects of noise in the data and enhance the robustness of the model. Based on these designs, a perceptual wavelet packet entropy-guided ECA–Res2Net–Time-Delay-Neural-Network (PWPE-ECA-Res2Net-TDNN) speaker recognition model is developed, and improves the model's recognition performance and robustness. Extensive experiments have been conducted to validate these improvements. The major contribution of this work is summarized as follows:

1. The ECAPA-TDNN model is improved, utilizing denoised PWPE features as input. This effectively reduces the noise interference in the model's input features and enhances its ability to resist noise.
2. The ECAPA-TDNN model structure is improved through increasing the network depth and learning channel weights, incorporating attention-based statistical pooling, and optimizing the loss function. These modifications result in a model with a better robustness. This paper conducts speaker recognition experiments in different scenarios using the proposed method and model. The results of the experiments demonstrate that the improved model outperforms the baseline ECAPA-TDNN model based on MFCC features in terms of robustness in cross-scene recognition.
3. In order to investigate the recognition performance of the current mainstream models in different scenarios, a set of experiments is designed for both single- and multi-scene conditions. Based on the experimental results, improvements are made to the ECAPA-TDNN model, which shows a better cross-scene recognition performance. The ECAPA-TDNN model is used as the baseline for comparison with the designed model.

The organization of this paper is as follows. In Section 2, the overall architecture of the proposed model is depicted. Furthermore, a detailed description of the implementation process for each component is provided. In Section 3, the dataset is introduced, and comparative experiments on different feature extraction methods and models are conducted. In Section 4, the conclusion summarizes the findings, with an overview of the advantages and limitations of the study, and gives recommendations for further research.

2. Methods

In this section, an overview of the model's overall architecture is provided. It is compared with the ECAPA-TDNN model that uses MFCC as the input feature in Section 3.1. Then, in Sections 3.2 and 3.3, the extraction process for the PWPE and the feature enhancement procedure is explained. Finally, in Section 3.4, the structure of the proposed ECA-Res2Net-TDNN model is described.

2.1. Model Architecture

Inspired by Section 1, the PWPE-ECA-Res2Net-TDNN model is proposed, and its architecture and comparison with the MFCC-ECAPA-TDNN structure are illustrated in Figure 2.

The pre-processed speech signal undergoes feature extraction; for this, traditional methods such as MFCC extract spectral features by applying discrete Fourier transform (DFT) and Mel filter banks. However, these processes result in the loss of some information from the original audio signal. In particular, MFCC has weak capturing capabilities for detailed information in the high-frequency range and is sensitive to noise. When noise is present in the audio, the reliability of the MFCC coefficients decreases, which can have a negative impact on subsequent speech processing tasks.

On the other hand, PWPE is a feature extraction method specifically designed for speech. It prunes the WPT features based on the auditory range of human hearing, applies threshold denoising, and calculates the non-normalized Shannon entropy (NSE) coefficients to obtain PWPE feature vectors. These PWPE features are then fed into a TDNN layer for feature enhancement. Furthermore, the model incorporates a Res2Net structure that includes lightweight attention mechanisms such as the SE block and the ECA block, which is a non-reducing dimension channel attention mechanism. The SE net, which compresses channels, can result in data information loss. To address this issue, the non-reducing dimension ECA is utilized to avoid such problems. The outputs of the four ECA-Res2Net modules are concatenated and fed into the TDNN network for multi-feature fusion. In ECAPA-TDNN, attention pooling is used to weight the input features. Unlike single-attention pooling, this paper uses multi-attention pooling to simultaneously capture the relationships between different feature subspaces and fuse this information during the

pooling process. This increases the expressiveness of the model and extracts richer feature representations. The model also includes a sub-center ArcFace loss function layer, which is commonly used in large-scale datasets with noise. It requires intra-class cohesion and inter-class separability, while not being overly influenced by the noise in the dataset.

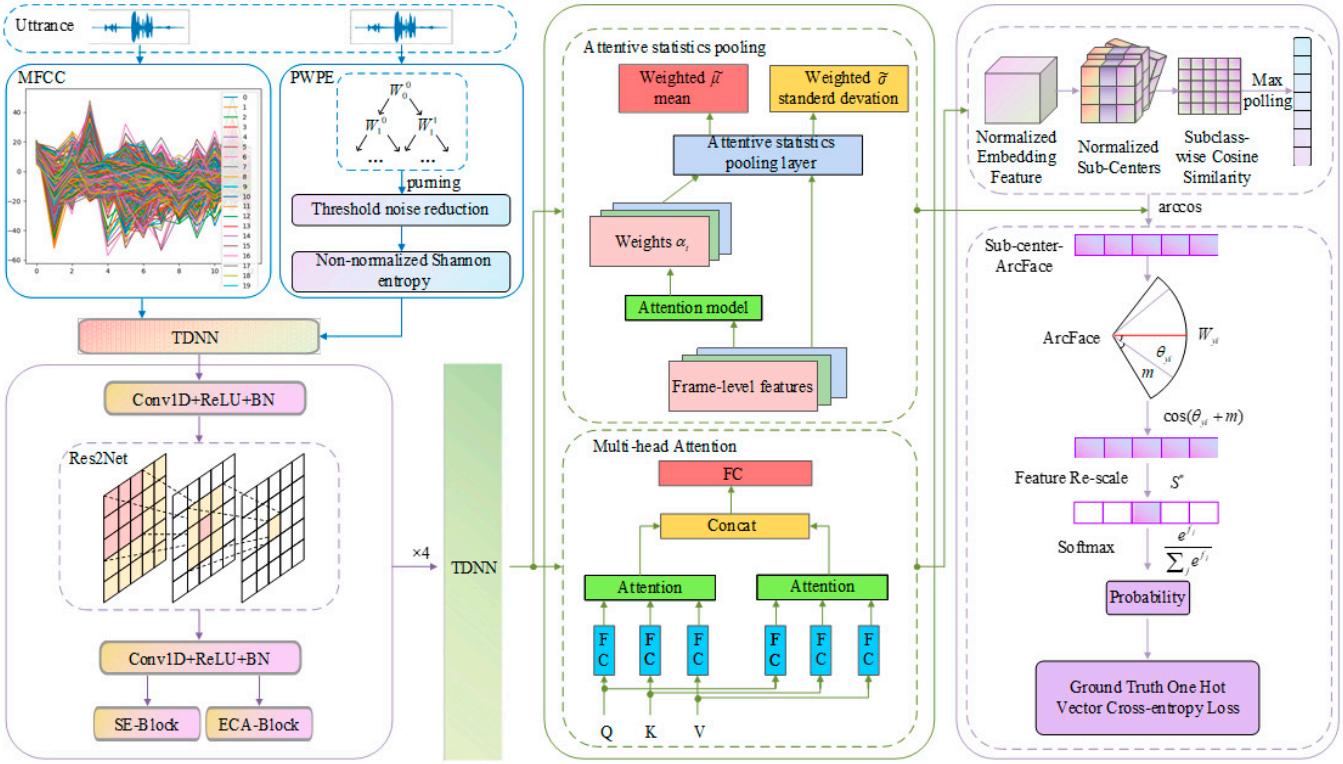


Figure 2. The structure of the PWPE-ECA-Res2Net-TDNN model and its comparison with the ECAPA-TDNN model.

2.2. The Feature Extraction Method for PWPE

Before the extraction of acoustic features from speech signals, a series of preprocessing operations are usually applied to the input one-dimensional speech signal, known as speech signal preprocessing. The purpose is to remove interfering information and obtain relatively clean and pure speech signals [31]. After preprocessing, the speech signals are subjected to acoustic feature extraction, which is also a crucial step in speaker recognition systems. The extracted feature parameters should describe speaker characteristics as much as possible and offer strong discriminative and high stability properties. In the case of speech recorded in complex environments, it is desirable to extract feature vectors with a good noise robustness and discriminability.

Wavelet transform uses a wavelet basis, which has a finite length and decaying function. Its advantage is that it can analyze any part of the signal by adjusting it through scaling and translation, thereby obtaining frequency and time information [32]. In contrast, the Fourier transform uses trigonometric basis functions and does not provide time information during signal analysis. The scaling of wavelet basis corresponds to the signal frequency, while the translation corresponds to the signal time. Due to this characteristic of the wavelet transform, it has been used as an alternative to the Fourier transform for signal processing. Wavelet denoising techniques can effectively suppress global noise, while the localized nature of wavelets limits the impact of local noise. However, wavelet-based cepstral feature extraction algorithms have high-dimensional features and are not suitable for combining with speaker models to form speaker recognition systems [33–35]. To reduce the influence of noise during feature extraction and computational complexity, researchers have proposed the perceptual wavelet entropy short-time spectral feature based

on wavelet transform, specifically designed for analyzing speech signals. This feature is composed of the entropy of the power spectrum of wavelet sub-bands of the signal and is good at representing sudden changes in the signal. It also has a low dimensionality, making it suitable for representing non-stationary signals. Commonly used entropy feature extraction algorithms include Shannon entropy, energy entropy, and threshold entropy based on wavelet packet transform. Among them, the Shannon-entropy-based feature extraction algorithm is widely used due to its good stability, discriminability, and resistance to interference [36,37].

The PWPE method utilizes perceptual wavelet packet transform (PWPT) to analyze speech signals while employing threshold denoising techniques and the auditory properties of PWPT to suppress both global and local noise. The essence of PWPT lies in pruning the decomposition process of wavelet packet transform using an auditory model. The decomposition process is as follows:

$$\Psi_{m,n}(t) = 2^{-m/2}\psi\left(2^{-m}t - n\right), m, n \in Z \tag{1}$$

where $\psi(t)$ represents a square-integrable function. To reduce the computational complexity, only the sound features within the human auditory range are retained. A mathematical model of the cochlear auditory filter bank is used to prune the WPT. The model can be expressed as follows:

$$f_c = A(10^{\alpha x} - k) \tag{2}$$

where f_c represents the center frequency of the auditory filter, and A, α, and k are constant and related to the specific biological species. For humans, the value of k is set to 0.88. The values of A, α are determined based on the auditory range of the specific biological species.

$$A = \frac{f_{\min}}{1-k} \tag{3}$$

$$\alpha = \log_{10}\left(\frac{f_{\max}}{A} + 1\right) \tag{4}$$

where $[f_{\min}, f_{\max}]$ represents the auditory range of the biological species. For humans, $f_{\min}$ = 20 Hz and $f_{\max}$ = 20 kHz, and the frequency range is generally from 30 Hz to 4 kHz. Using the auditory model, 16 auditory filters are constructed within this range. To construct PWPT, a seven-layer WPT is created, and its decomposition tree is shown in Figure 3a.

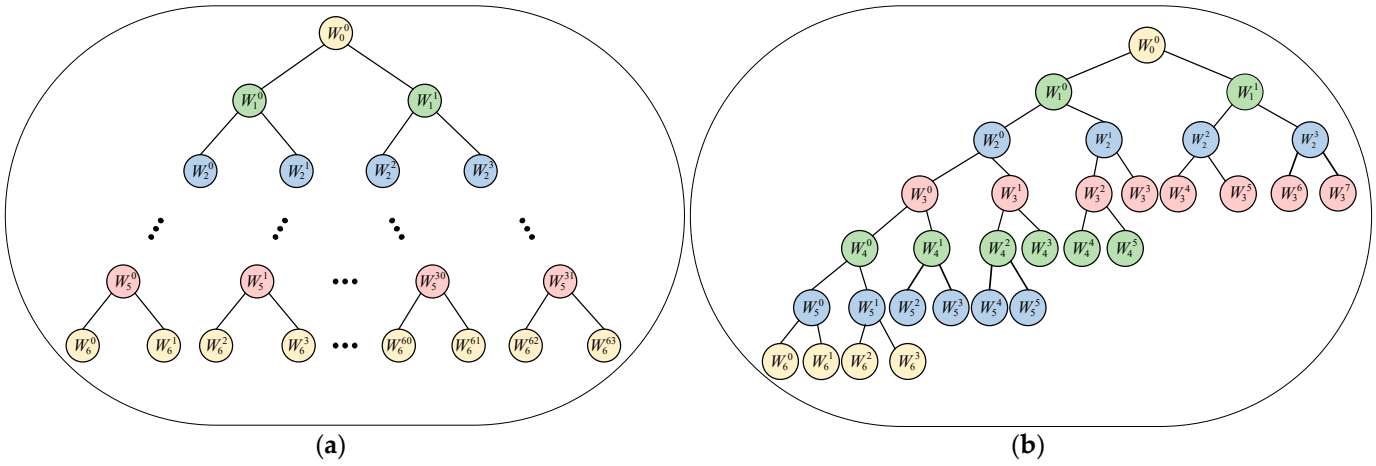


Figure 3. (**a**) WPT decomposition tree, (**b**) PWPT decomposition tree.

W represents the decomposed signal, and each node's left and right branches represent the low-pass and high-pass filtering processes, respectively. Their definitions are as follows:

$$\begin{cases} W_j^{2p}[n] = \sum_{n=-\infty}^{n=+\infty} W_j^p[n]h[n-2p] \\ W_j^{2p+1}[n] = \sum_{n=-\infty}^{n=+\infty} W_j^p[n]g \mid [n-2p] \quad j = 0,1,2,\ldots,J, p = 0,1,2,\ldots,2^j-1 \\ W_0^0[n] = s[n] \end{cases} \tag{5}$$

where J represents the total number of decomposition levels in the wavelet transform. In Figure 3a, J = 6, h represents the high-pass filter, and g represents the low-pass filter. Then, a "pruning" operation is performed via selecting nodes based on whether their center frequency is close to the center frequency of the auditory filter, and whether their energy is greater than 20% of the total energy of the node. For a node W_j^p with center frequency f_c and energy E, the calculation is given by the following equation:

$$f_c = \left[\frac{2000}{2^j}\right] + \left[\frac{4000}{2^j}\right] p \tag{6}$$

$$E = \sum_{n=1}^{L_j} \left| W_j^p[n] \right|^2 \tag{7}$$

where L_i represents the length of signal W_j^p. After the "pruning" operation, WPT is transformed into PWPT, and its decomposition tree is shown in Figure 3b, where the leaf nodes represent the 16 sub-signals obtained from the PWPT decomposition, denoted as $W_1 \sim W_{16}$. These sub-signals have center frequencies that approximate the 16 critical frequencies obtained from the auditory model.

2.3. Speech Enhancement Based on Wavelet Thresholding

To enhance the resistance of entropy features to environmental noise, denoising is applied to each subframe. This paper chooses the wavelet thresholding method, where the low-frequency component mainly contains the transformed coefficients of the original speech signal, while the high-frequency component contains the transformed coefficients of the noise. This is based on the correlation characteristics between noise and the useful signal, which are approximately separated after scale transformation. The next step is to process the wavelet coefficients of the two components by comparing them with a pre-determined threshold. It can be considered that the wavelet coefficients smaller than the threshold in either component represent the noise to be removed. A significant amount of noise can be removed by performing a multi-scale wavelet transform. Further decomposition of the selected signal from the first decomposition stage with a second or multi-scale decomposition results in a cleaner useful signal. The key to this process is the selection of an appropriate threshold, as the choice of threshold directly affects the final enhancement performance. Different threshold functions can be selected based on different environments to effectively remove noise signals while preserving useful signals [38,39]. Commonly used threshold functions include:

1. The hard thresholding function; the denoising process can be represented as follows:

$$D_j[i] = \begin{cases} W_j[i], & \left|W_j[i]\right| < \lambda \\ 0, & \left|W_j[i]\right| > \lambda \end{cases} \quad j = 1,2,3,\ldots,16 \tag{8}$$

where $W_j[i]$ represents the coefficients of the subframe W_j, $D_j[i]$ represents the coefficients after denoising, and λ represents the denoising threshold. It can be observed that the function is discontinuous in the interval $(-\lambda, +\lambda)$. The discontinuity of the hard thresholding function in this interval can lead to oscillations in the resulting signal after inverse wavelet transform.

2. The soft thresholding function; the denoising process can be represented as follows:

$$D_j[i] = \begin{cases} \operatorname{sgn}(|w_j[i]| - \lambda) & |w_j[i]| >= \lambda \\ 0 & |w_j[i]| < \lambda \end{cases} \tag{9}$$

where sgn() represents the sign function. The soft thresholding function is continuous within the specified interval. However, during the wavelet reconstruction process, the soft thresholding function may result in an incomplete representation of the useful signal, leading to significant differences between the reconstructed signal and the original function and, thus, causing signal distortion.

3. The compromise thresholding function, represented in the denoising process, is expressed as follows:

$$D_j[i] = \begin{cases} \operatorname{sgn}(|w_j[i]| - a\lambda) & |w_j[i]| >= \lambda \\ 0 & |w_j[i]| < \lambda \end{cases} \tag{10}$$

In Equation (10), $a \in [0, 1]$. The threshold selection methods mainly include unbiased likelihood estimation, fixed threshold estimation, and heuristic threshold estimation. These methods aim to find a compromise between two thresholding functions to improve the performance of wavelet denoising [40,41]. However, there is still significant room for improvement in the denoising performance. In this case, an improved threshold function is chosen, with the following mathematical form:

$$D_j[i] = \begin{cases} \mu w_j[i] + (1-\mu)\operatorname{sgn}(w_j[i]) \left[|w_j[i]| - \lambda - \frac{a\lambda \log \frac{\lambda}{w_j[i]}}{1+e^{(w_j[i]^2-\lambda^2)}} \right] & |w_j[i]| >= \lambda \\ 0 & |w_j[i]| < \lambda \end{cases} \tag{11}$$

where $D_j[i]$ represents wavelet coefficients, and a is the adjustment factor. $\mu = 1 - e^{-a(|w_j[i]|-\lambda)^2}$. The commonly used general threshold is represented as:

$$\lambda = \frac{M(W_j)}{C}\sqrt{2\ln(L(W_j))} \tag{12}$$

where $L(W_j)$ represents the length of W_j, $M(W_j)$ represents the absolute median deviation of W_j, and $C = 0.675$ is the noise coefficient. In the general threshold, the threshold is positively correlated with the number of signal sampling points. The larger the number of sampling points, the larger the threshold. However, a high threshold can cause signal distortion, indicating a significant flaw in the general threshold. In order to tackle this limitation, some researchers have optimized it [42,43], as represented by:

$$\lambda_j = \frac{M(W_j)}{C}\frac{\sqrt{2\ln(L(W_j))}}{\ln(j+1)} \tag{13}$$

where λ_j is the threshold for the j-th layer of wavelet coefficients. This threshold function optimizes the selection of thresholds. When there is a large number of sampling points, increasing the decomposition level helps to mitigate the rate at which the threshold rises, ensuring that λ_j does not become excessively large. We compute the NSE coefficient for D_j (j = 1, 2, 3, ……, 16) after denoising:

$$H(D_j) = -\sum_{i=1}^{I} |D_j[i]|^2 \log |D_j[i]|^2 \tag{14}$$

where I represents the length of D_j. The feature vector of PWPE is computed as follows:

$$v_{\text{pwpe}} = [H(D_1), H(D_2), \ldots, H(D_{16})] \tag{15}$$

2.4. Speaker Recognition Model ECA-Res2Net-TDNN

The ECAPA-TDNN model captures global properties by introducing the SE module and the channel attention mechanism. Through multi-layer feature aggregation and summation, as well as channel- and context-dependent statistics pooling, the output features of all SE-Res2Blocks are concatenated and used to generate attention-based statistical pooling features. ECAPA-TDNN has achieved state-of-the-art results on several benchmark speech recognition datasets and is considered one of the most effective approaches to speech recognition currently available [44,45]. However, ECAPA-TDNN is noise-sensitive, and its performance can be affected by acoustic noise, such as environmental noise.

This paper improves the ECAPA-TDNN model to enhance its robustness. Firstly, instead of using MFCC features as the input, it replaces them with the proposed perceptual wavelet entropy features that have undergone speech enhancement. Meanwhile, ECAPA-TDNN utilizes SE-Net to weight the feature channels, which involves compression that leads to the loss of some features. As the input features have already been pruned, the parameter size is reduced by 75%. In this case, feature compression is avoided, and the ECA network is utilized to learn channel weights, thus mitigating the impact of dimensionality reduction on channel attention learning. The ECA network consists of non-reduced GPA-aggregated convolutional features, where the kernel size 'k' can be determined adaptively and implemented using fast one-dimensional convolution [46]. Finally, the channel attention mechanism is acquired through the utilization of the sigmoid function. The structure is illustrated in Figure 4a.

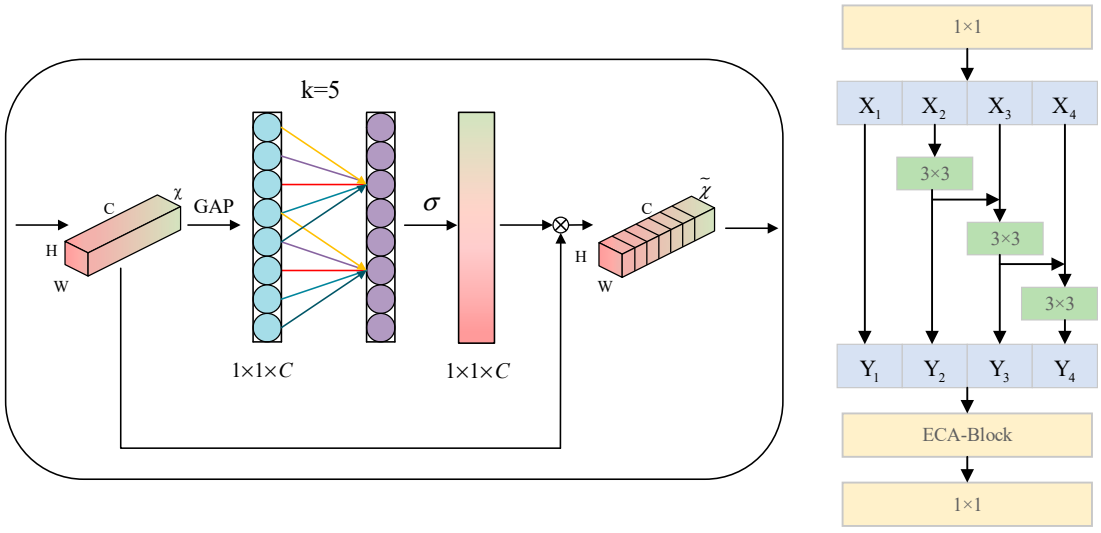


Figure 4. (**a**) The structure of the ECA network, (**b**) the structure of the ECA-Res2Net network.

When the ECA network is connected to Res2Net, the structure of ECA-Res2Net is shown in Figure 4b. The first layer performs convolutional operations with a kernel size of 1, which transforms the feature channels at each pixel. Then, the first layer's feature maps are evenly divided into n identical feature maps in terms of the feature channels. In the second layer, the first feature map remains unchanged, while each subsequent feature map undergoes convolutional operations with a kernel size of 3. Each output is then connected through residual connections. Finally, the last layer uses a convolutional operation with a kernel size of 1 to restore the feature channels for each pixel.

The integration of the ECA-Block layer improves the performance of the system while significantly reducing the number of model parameters. In the ECAPA-TDNN architecture, the SE-Res2Net is replaced with ECA-Res2Net, and the network depth is increased. In the ECAPA-TDNN network structure with ECA-Res2Block, the fourth expansion factor dilation = 5 is added to improve the performance of the model in complex scenarios.

Instead of using attentive stat pooling in ECAPA-TDNN, this paper modifies it to a structure with a multi-head attention pooling and batch normalization (BN) structure. Multi-query multi-head attention pooling is a time-series pooling with multiple attention heads, dividing the model into multiple subspaces, each of which can obtain a better expressive power. The multi-head attention mechanism can also enhance the robustness of the model, so that even if some features are not well represented in some subspaces, they can still be well represented in other subspaces.

The structure of the multi-head attention is shown in Figure 5. First, it calculates attention separately in four different attention heads. X is multiplied with the weight matrices. Then, the attention is calculated using the resulting Q/K/V matrices. Finally, the resulting Z matrices are calculated and multiplied with weight matrix W^0 to produce the output of the layer.

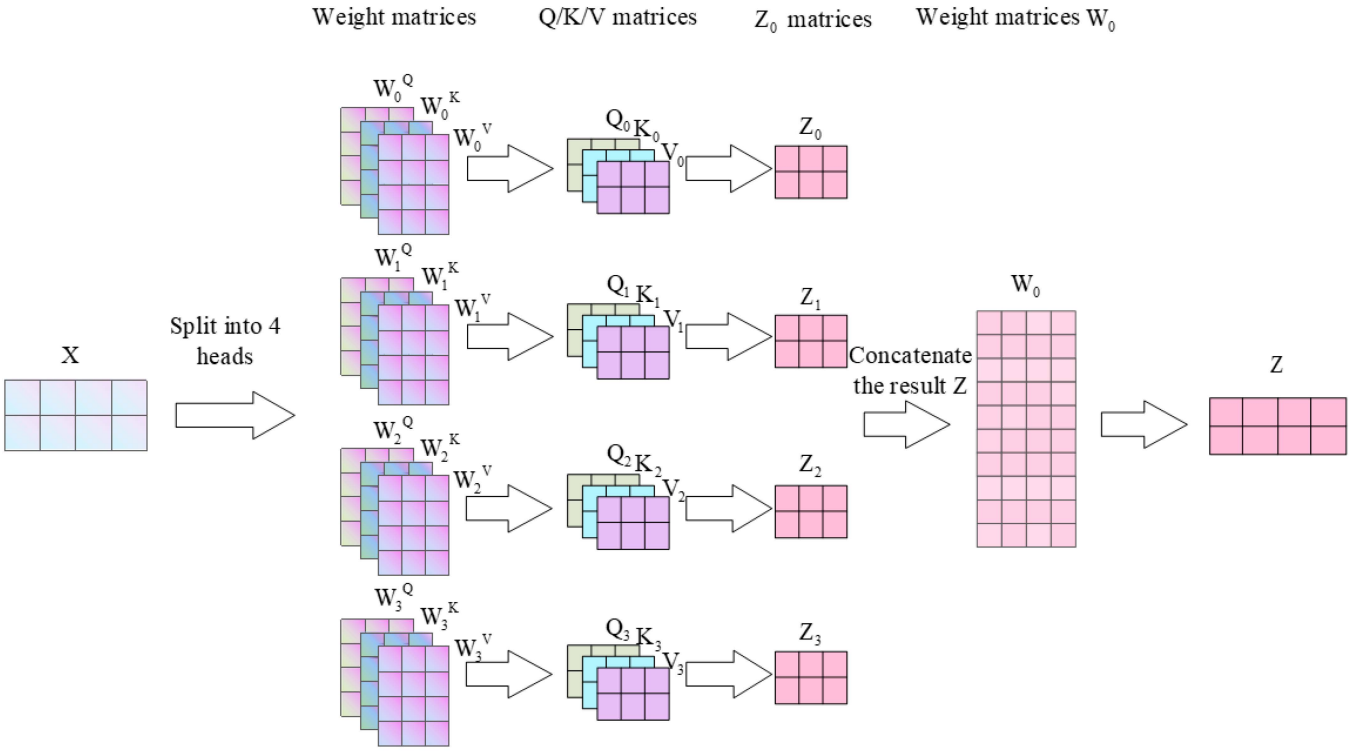


Figure 5. The structure of the multi-head attention.

In ECAPA-TDNN, the AAM-softmax loss function is used. This paper modifies it to the sub-center ArcFace loss function layer, which is not affected by noise in the data. The weight matrix W consists of rows representing the learned center points of each speaker. Out of the K center points, one is designated as the dominant center point, while the K-1 center points represent non-dominant noise samples. The cosine distance is calculated between each center point and the embedding code to derive a similarity matrix. During the training, a penalty strategy is applied to the angle between the embedding code and the speaker center points, promoting inter-class separability and intra-class compactness. The similarity matrix is pooled along each row, and then the softmax loss is computed. The formula of the loss computing is as follows:

$$L = -\log\left(\frac{e^{s\cos\theta_j}}{\sum_{i=0}^{N} e^{s\cos\theta_i}}\right) \tag{16}$$

The proposed structure of the ECA-Res2Net TDNN model is shown in Figure 6. The first TDNN layer consists of one-dimensional convolution, the ReLU activation function, and batch normalization, with the structural parameters $k = 5$ and $d = 1$. The second, third, fourth, and fifth layers adopt the Res2Net structure with an ECA lightweight attention mechanism. The outputs of the four modules are concatenated along the feature dimension. The sixth layer is a structure with multi-head attention statistical pooling. The seventh layer is a fully connected layer with batch normalization, used for the linear transformation of the final features. The last layer is the sub-center ArcFace loss function layer.

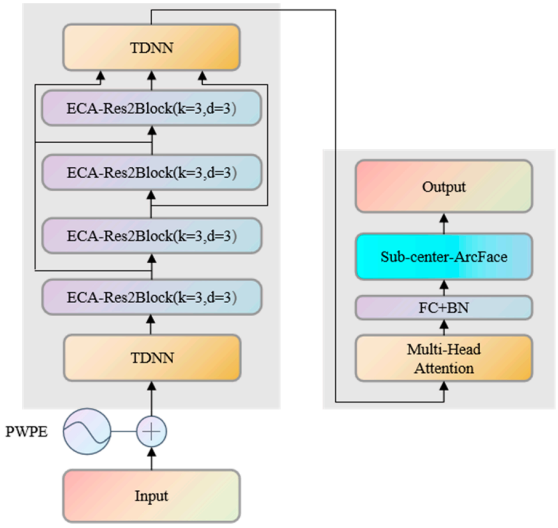


Figure 6. The proposed ECA-Res2Net-TDNN model structure.

3. Results and Discussion

This section first introduces the construction and partitioning of the dataset, analyzing single-scene and cross-scene voiceprint recognition for each scenario separately. Based on the recognition results and the noise decibels of each scene, they are categorized into three noise levels: low, medium, and high. The cross-scene recognition performance of the different models is compared using the data from each noise level. Finally, the recognition time of each model is compared, and ablation experiments are conducted on the individual modules of the proposed model.

3.1. Dataset Construction and Analysis

In order to study cross-scene recognition problems, this paper used a Chinese celebrity voiceprint database called CNCeleb, published by the Center for Speech and Language Technology at Tsinghua University. This database collects voice clips of 3000 Chinese celebrities published on the Internet. One notable characteristic of this database is its diverse range of scenes, including singing, interviews, speeches, entertainment, and 11 other scenes. It provides a rich dataset for cross-scene analysis because celebrities are more likely to appear in multiple scenes. For example, a singer may also participate in interviews, and a comedian may also act in movies. To evaluate the performance of state-of-the-art voiceprint recognition models on the CNCeleb database, this paper conducted separate tests for each individual scene with EER, a widely used performance metric in speaker recognition tasks. It represents the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR) in a recognition system. The calculation method for the TPR and FPR is as follows:

$$TPR = \frac{TP}{TP + FN} \tag{17}$$

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

where *TP* stands for true positive, which refers to the instances where the model correctly identifies positive samples as positive. *FN* stands for false negative, which indicates cases where the model incorrectly classifies positive samples as negative.

The results are shown in Table 1. It can be observed that the recognition rates vary significantly across different scenes. Clear and simple scenes, such as speeches, interviews, and live broadcasts, perform well (EER $\approx$ 5%), while the recognition rates for movies and songs are much lower (EER $\approx$ 10–20%). However, when the models are trained and tested on a mixture of overall scenes, the results are not satisfactory. Despite achieving a good performance in individual scenes, the performance in mixed scenes is subpar. This performance discrepancy is attributed to the challenge of cross-scene recognition.

Table 1. The data partition of the CNCeleb dataset and the single/mixture scene recognition performance (EER%) of the ECAPA-TDNN and x-vector.

Genres	Speakers	Utterances	ECAPA-TDNN	x-Vector
Advertisement	75	781	8.16	9.37
Drama	377	4521	10.22	11.70
Entertainment	1020	18,931	6.75	7.31
Interview	1253	41,586	6.12	6.98
Live broadcast	496	154,249	4.87	5.42
Movie	165	1495	10.35	11.47
Play	170	5476	10.61	11.56
Recitation	259	58,839	15.66	16.55
Singing	683	32,279	19.12	20.86
Speech	331	39,792	3.10	3.21
Vlog	524	120,812	4.63	5.31
Overall	3000	485,361	26.78	27.43

From Table 1, it can be observed that the singing scene has the highest EER, while the speech scene has the best recognition rate. This could be related to the different noise levels in each scene. Then, the noise decibel range is calculated for each scene, and the results are shown in Figure 7.

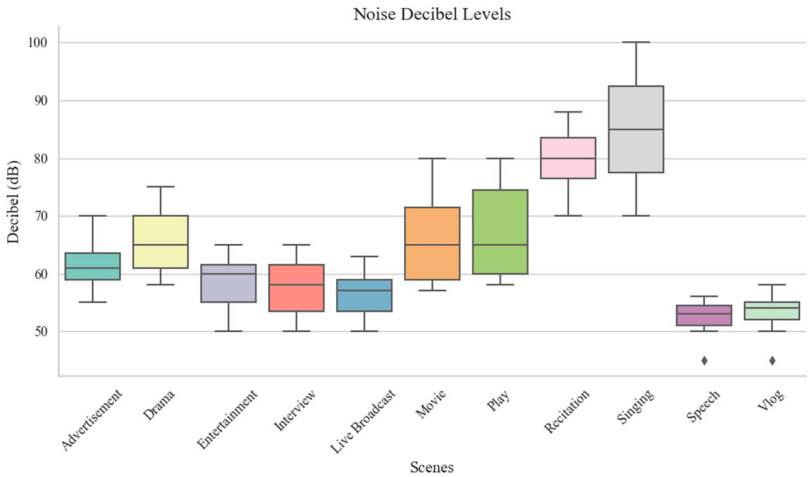


Figure 7. The distribution of noise levels in different scenes.

It can be observed that the singing scene has the highest noise levels, while the speech scene has the lowest noise levels. This is in line with expectations, as higher noise levels tend to result in a lower recognition accuracy. The points below the box plot represent outliers. In addition to the single-scene speaker recognition discussed above, cross-scene recognition holds greater significance. The CNCeleb database provides abundant cross-scene speech data. In the CNCeleb database, there are 558 celebrities who appear in two different scenes,

405 celebrities who appear in three different scenes, 361 celebrities who feature in four different scenes, and 79 celebrities who appear in five or more scenes.

Figure 8 presents the cross-scene recognition performance of the ECAPA-TDNN system on the CNCeleb test set. Each column represents a registration scene, each row represents a verification scene, and the numbers in each cell represent the corresponding EER results for registration and verification scenes. It can be observed that the cross-scene recognition performance is significantly lower than the single-scene recognition performance in almost all recognition tasks. For example, when registered in the speech scene and verified in the same speech scene, the EER is 3.1%. However, when verified in the singing scene, the EER increases to 20.24% and, when verified in the advertisement scene, the EER reaches 23.12%. The singing testing scene, which has the highest noise levels, exhibits the highest EER.

Enroll \ Test	Ad.	Drama	Enter.	Interview	L. B.	Movie	Play	Recitation	Singing	Speech	Vlog
Advertisement	8.16	18.61	14.4	9.73	5.42	28.67	16.53	11.35	23.34	23.12	7.98
Drama	20.33	10.23	19.78	20.13	12.57	23.77	7.62	29.85	30.42	28.25	15.87
Entertainment	12.37	18.21	6.31	8.37	11.35	16.77	12.34	8.29	25.12	10.36	9.76
Interview	18.11	19.72	10.45	5.71	10.34	16.96	11.68	6.25	27.36	9.42	10.23
Live Broadcast	6.35	15.61	12.11	10.98	4.97	16.48	10.12	21.23	20.65	15.33	8.67
Movie	27.59	23.79	13.92	16.34	10.38	10.02	10.87	12.46	27.32	12.72	23.18
Play	8.69	14.38	12.86	18.16	13.59	22.56	10.49	4.49	26.19	10.44	16.46
Recitation	10.23	18.86	21.89	7.23	8.46	18.26	3.25	14.68	32.72	4.36	31.25
Singing	23.45	26.22	30.12	27.55	19.86	27.87	16.83	34.69	19.72	20.24	22.64
Speech	28.12	24.01	10.79	9.56	11.22	18.97	9.34	2.34	23.89	3.1	35.12
Vlog	4.98	18.73	25.65	9.55	9.46	14.13	8.99	19.1	24.13	35.63	5.06

Figure 8. The cross-scene recognition performance (EER%) of the ECAPA-TDNN model on the CNCeleb dataset.

3.2. Dataset Partitioning and Recognition Performance of Different Scenes

The division of noise decibel levels is typically determined based on specific applications, domains, and regulations. Therefore, there is no standardized international classification for noise decibel levels. In general, the noise levels in various environments can be described as follows:

1. Quiet environments, such as normal indoor conversations and silent libraries, are typically around 40–60 decibels. At this point, speech can be heard clearly.
2. Normal conversations, TV volume, office noise, city traffic noise, etc., are generally around 60–70 decibels, and speech can be heard normally at this level.
3. High-volume music, vehicle noise, construction site noise, etc., are typically around 70–100 decibels, making it difficult to hear speech clearly at this level.

Based on the analysis above, the 11 scenes were divided into three levels: low noise, medium noise, and high noise, as shown in Table 2.

Scenes with low noise have clearer pronunciation and a better recognition performance, while scenes with high noise have more background noise, making recognition relatively difficult. All the scenes were tested with the proposed PWPE-ECA-Res2Net-TDNN model, and the results are shown in Figure 9. In the figure, the cases where "Speech" is used as the registration scene, and "Advertisement" is used as the verification scene, as well as the cases where "Advertisement" is used as the registration scene, and "Speech" is used as the verification scene, have the highest EER. At the same time, speech recognition in a single scene has the lowest EER, which contradicts previous finding that higher noise levels lead to a lower accuracy. It can be observed that the EERs of the test scenes with medium and high noise levels are relatively consistent, indicating that the proposed model has a certain level of robustness for robust recognition in noisy cross-scene scenarios. Another set of experiments was designed to validate the advantages of the proposed model at a

high level of noise. The 11 scenes were merged into low, medium, and high noise levels, and the EER and accuracy of the PWPE-ECA-Res2Net-TDNN model and the traditional MFCC-ECAPA-TDNN model were tested through registering and verifying scenes with the three noise levels. Accuracy is a commonly used performance metric in classification tasks. The formula to calculate accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where *TP* represents true positive and denotes the number of correctly predicted positive samples. *TN* represents true negative and denotes the number of correctly predicted negative samples. *FP* represents false positive and denotes the number of negative samples incorrectly predicted as positive. *FN* represents false negative and denotes the number of positive samples incorrectly predicted as negative.

Table 2. Noise level classification of the different scenes.

Scenes	Noise Level	Decibel Interval
Speech Vlog Live broadcast Interview	Low	40–60 db
Entertainment Advertisement Drama Movie Play	Medium	60–70 db
Recitation Singing	High	>70 db

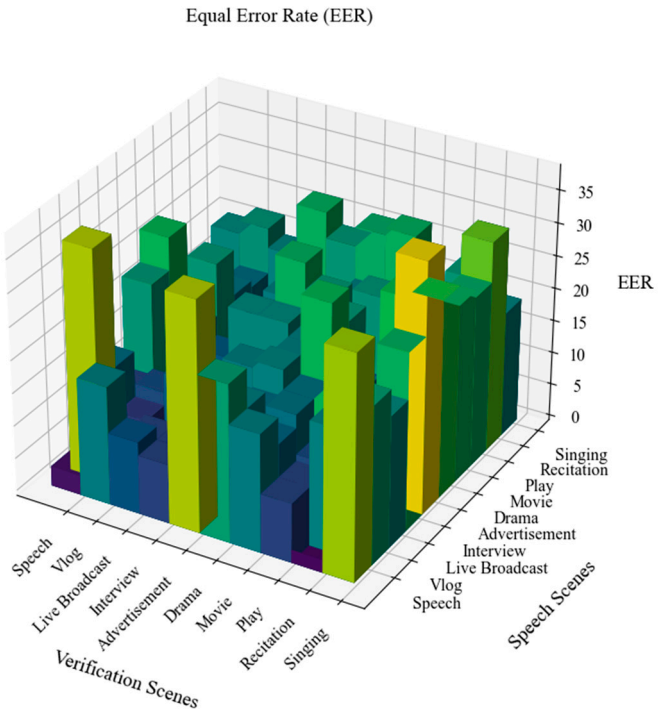


Figure 9. The recognition performance (EER%) of different scenes using the PWPE-ECA-Res2Net-TDNN model.

The results are shown in Figures 10 and 11. It can be seen that both models achieve good recognition results in the low-noise scenes. However, in the medium- and high-noise test scenes, the traditional MFCC-based ECAPA-TDNN model with feature extraction experiences a significant drop in recognition performance, with an EER reduction of approximately 10%. On the other hand, the PWPE-ECA-Res2Net-TDNN model, which uses PWPE as the feature extraction, maintains an EER of less than 8%, indicating that the proposed model is more suitable for cross-scene recognition in medium- and high-level noise scenarios than the traditional model.

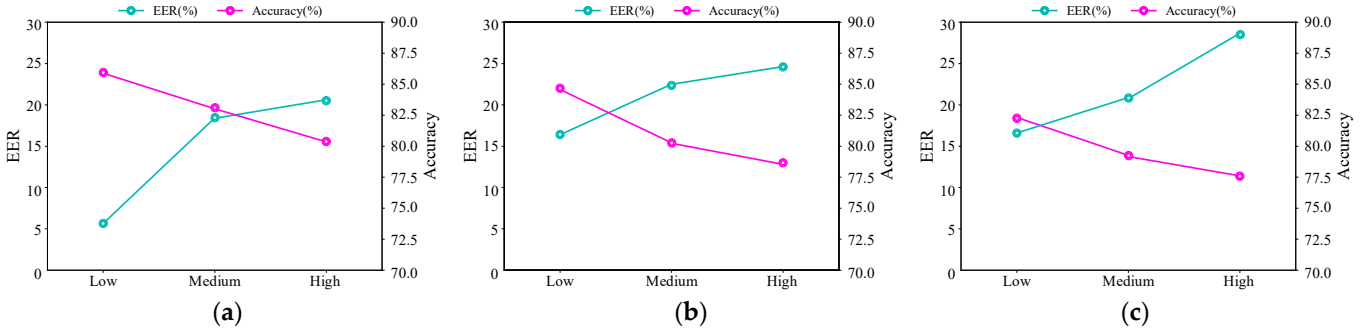


Figure 10. The recognition EER (%) and accuracy (%) of the PWPE-ECA-Res2Net-TDNN model in low-, medium-, and high-level noise scenes for three different scenarios. (**a**) Low-level noise, (**b**) medium-level noise, and (**c**) high-level noise.

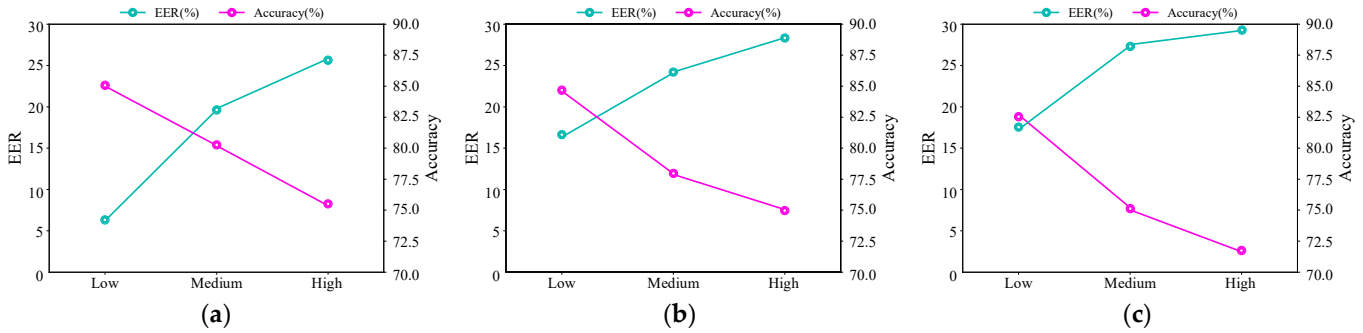


Figure 11. The recognition EER (%) and accuracy (%) of the MFCC-ECAPA-TDNN model in low-, medium-, and high-level noise scenes for three different scenarios. (**a**) Low-level noise, (**b**) medium-level noise, (**c**) high-level noise.

3.3. Ablation Experiments on Different Feature Extraction Methods and Models

Based on the above analysis, comparative experiments on several representative scenarios were conducted to validate the effectiveness of the proposed modules in cross-scene recognition. The speech and advertisement scenes were selected as the registration speech, and then the advertisement, speech, and singing scenes were chosen as validation scenes. The recognition performance (EER%) of the ECAPA-TDNN model using MFCC as the input features was compared with the ECAPA-TDNN model using PWPE as the input features, both with speech enhancement. And they were also compared with the proposed ECA-Res2Net-TDNN model using PWPE as the enhanced input features. The results are shown in Table 3.

Based on the above results, it can be observed that the ECAPA-TDNN model with PWPE extracted features has an EER reduction of approximately 1% compared to the ECAPA-TDNN model with MFCC extracted features. This indicates that PWPE is more suitable for feature extraction in noisy environments. Furthermore, the proposed ECA-Res2Net-TDNN model outperforms the ECAPA-TDNN model with the same PWPE feature extraction. It is also evident that the proposed model exhibits significant improvement in cross-scene scenarios, with an increase of approximately 2% in EER. However, the improve-

ment in recognition performance is not substantial in single-scene scenarios, reaffirming that the model is particularly suitable for cross-scene recognition.

Table 3. Comparison of recognition performance (EER%) in typical cross-scene scenarios for ECAPA-TDNN with MFCC and PWPE input features, and ECA-Res2Net-TDNN with PWPE input features.

Enroll	Speech	Speech	Speech	Advertisement	Advertisement
Test	**Advertisement**	**Singing**	**Speech**	**Advertisement**	**Singing**
MFCC/ECAPA-TDNN	28.36	23.43	3.10	8.16	23.66
PWPE/ECAPA-TDNN	26.78	22.69	3.02	8.06	21.27
PWPE/ECA-Res2Net-TDNN	25.50	21.52	2.99	7.98	20.45

3.4. Model Complexity and Recognition Time of Different Models

The ECA-Res2Net-TDNN model proposed in this paper consists of a total of 214 layers in the network, comprising 9024 neurons. The TDNN layer consists of four layers in total: an input layer, two hidden layers, and an output layer, with a total of 64 neurons. There are four ECA-Res2Net modules, each containing 50 layers in each block, with 2048 neurons per block. The second TDNN layer has three network layers, totaling 128 neurons. The multi-head attention module comprises four attention heads, with each head including 64 neurons, totaling 256 neurons. The fully connected and batch normalization layers, together, comprise 256 neurons. The sub-center ArcFace has a total of 128 neurons.

While the complexity of this model has increased compared to the ECAPA-TDNN model with 163 layers, this increase in complexity allows us to better capture the rich features in the voiceprint data and, thus, further improve the recognition accuracy. Although this may result in some additional computational costs and resource consumption, from a performance perspective, it is justifiable, as it leads to significant performance improvements. In practical applications, we believe that the improvement in accuracy outweighs the increase in complexity, especially for critical tasks, such as voiceprint recognition. Next, we will compare how the increase in model complexity affects the recognition time.

Figure 12a compares the recognition accuracy (DCF) and recognition time of the three speaker recognition models on test utterances of different durations. The recognition time refers to the time taken by the speaker recognition models to identify the speaker from 3, 5, and 7 s of test speech, respectively. DCF represents the detection cost function. The formula to calculate the detection cost function (DCF) is as follows:

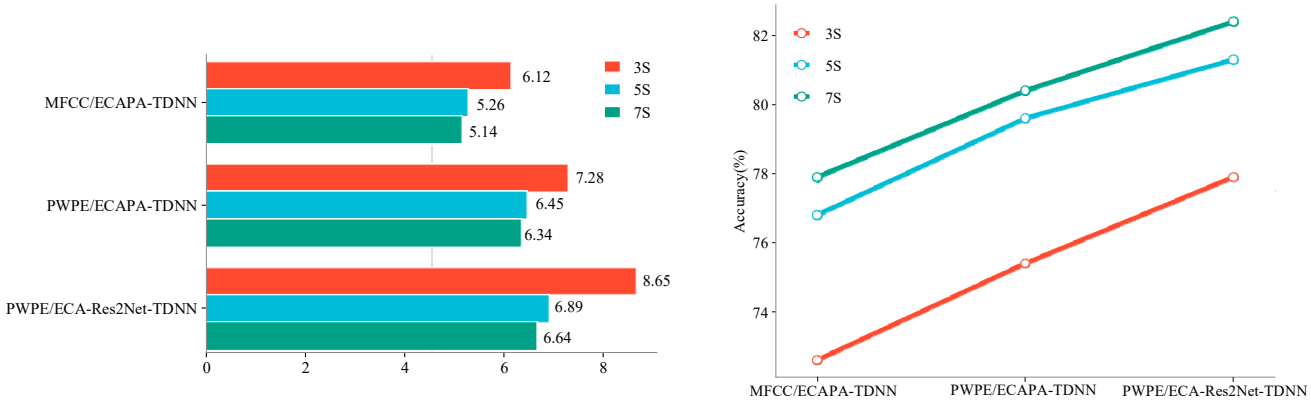
$$DCF = C_{\text{miss}} \cdot P_{\text{miss}} \cdot P_{\text{target}} + C_{\text{false alarm}} \cdot P_{\text{false alarm}} \cdot \left(1 - P_{\text{target}}\right) \tag{20}$$

where C_{miss} is the cost of missing detection, and P_{miss} is the probability of missing detection, which is the probability of classifying actual positive samples as negative. P_{target} is the target prior probability, representing the actual occurrence probability of positive samples. $C_{\text{false alarm}}$ is the cost of a false alarm, and $P_{\text{false alarm}}$ is the probability of a false alarm, which is the probability of classifying actual negative samples as positive.

Figure 12a shows that the MFCC/ECAPA-TDNN model has the shortest overall recognition time, approximately 5.26 s. The PWPE/ECAPA-TDNN model has a medium recognition time of around 6.45 s. The recognition time of the proposed PWPE/ECA-Res2Net-TDNN model is 6.89 s, which is only about 1.5 s longer than the traditional model, indicating that the recognition time of the PWPE/ECA-Res2Net-TDNN model is still relatively short, even under the highest recognition rate conditions.

In addition, Figure 12b shows that when a 3 s speech is used in the experiment, all speaker verification models achieve the lowest accuracy. This is because a 3 s speech is too short to contain sufficient speaker information, which affects the performance of the speaker verification models. When the test speech duration increased from 3 s to 5 s, the accuracy of the speaker verification models improved by around 4%. However, when the test speech duration increased from 5 s to 7 s, the accuracy of the speaker verification

models only improved by around 1%. Considering the trade-off between model performance and computational cost, it is more appropriate to use a 5 s test speech as input for speaker identification.



(**a**) (**b**)

Figure 12. (**a**) The recognition time of speaker verification models on test speech of different durations. (**b**) The accuracy of speaker verification models on test speech of different durations.

3.5. Ablation Experiments on Different Modules of PWPE/ECA-Res2Net-TDNN

Ablation experiments on the components were further conducted, which are introduced in Section 4. Speech was used as the training data, and singing as the test data, with PWPE as the input feature. Table 4 provides an overview of the results of these experiments.

Table 4. Comparative analysis of ablation experiments on various modules.

	CAM	Attention	Loss	EER (%)
A	ECA	Multi-head	Sub-center ArcFace	21.52
A1	SE	Multi-head	Sub-center ArcFace	21.90
A2	ECA	Attentive statistics	Sub-center ArcFace	21.63
A3	ECA	Multi-head	AAM softmax	21.72

To evaluate the channel attention mechanism (CAM) module, Experiment A1 was conducted, where the proposed ECA module was not utilized. The results showed an increase of approximately 0.38% in EER, demonstrating the benefits of learning channel weights with an uncompressed ECA network for improving system performance. To investigate the proposed attention module, Experiment A2 was performed, and incorporated a multi-head attention mechanism. It resulted in a reduction of approximately 0.11% in EER, indicating the effectiveness of incorporating the multi-head attention mechanism. To study the effect of the loss function on the model, Experiment A3 was conducted, and the AAM softmax loss function was used. The EER increased by 0.2%, indicating that the utilization of the sub-center ArcFace loss function is effective in improving the recognition performance.

4. Conclusions

This paper proposes a cross-scenario speaker recognition model based on PWPE and ECA-Res2Net-TDNN. Through the improvement of the feature extraction method of ECAPA-TDNN, the perceptual wavelet entropy feature is combined with the proposed speaker recognition model, and wavelet threshold denoising is applied to reduce the influence of input noise. The experimental results show that, compared to the ECAPA-TDNN model using MFCC features, the PWPE-based ECAPA-TDNN model performs better in terms of recognition accuracy and has better noise resistance. Furthermore, this paper also made improvements to the classical speaker recognition model. It combined the non-dimensional-reducing ECA block with Res2Net to assign weights to feature channels and introduced a multi-head attention mechanism. This allowed the model to learn

different behaviors without focusing attention excessively on its own position. The paper also introduced the sub-center ArcFace loss function module to mitigate the effects of noise in the data and enhance the robustness of the model.

A set of experiments is designed, where the 11 scenes in the dataset are divided into three levels of high, medium, and low noise. Firstly, the MFCC-ECAPA-TDNN and PWPE-ECA-Res2Net-TDNN models are tested for their speaker recognition performance in these three levels of environmental noise. The experimental results show that the ECAPA-TDNN model with PWPE extracted features shows an EER reduction of approximately 1% compared to the ECAPA-TDNN model with MFCC extracted features. Furthermore, the proposed ECA-Res2Net-TDNN model outperforms the ECAPA-TDNN model based on the same PWPE feature extraction, with an increase of approximately 2% in EER. This results validate the theory that the PWPE-ECA-Res2Net-TDNN model significantly improves the robustness of the model in moderate-to-high-noise environments. Secondly, a comparison of the recognition time across different models was conducted. The result shows the model maintains a relatively short recognition time even under the highest recognition rate conditions. Finally, a set of ablation experiments targeting each module in the proposed model is conducted. The results indicate that each module contributes to an improvement in the recognition performance, demonstrating the effectiveness of the proposed modules in enhancing the system.

The limitation of this work is that, although the proposed network architecture shows improvements in cross-scene recognition, it does not have any advantage in terms of computational time compared to other speaker recognition models. There is a significant bottleneck in terms of the computational performance, and further consideration is required of how this can be addressed through alternative approaches with a lower computational cost.

Author Contributions: Conceptualization, S.W. and H.Z.; methodology, S.W. and Z.W.; software, S.W.; validation, S.W., H.Z. and X.Z.; formal analysis, S.W.; investigation, S.W. and Z.W.; resources, S.W.; data curation, Z.W. and Y.S.; writing—original draft preparation, S.W. and H.Z.; writing—review and editing, X.Z. and Z.W.; visualization, S.W. and X.Z.; supervision, Y.S.; project administration, X.Z. and Y.S.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are available in a publicly accessible repository. The data used in this study are available at http://www.openslr.org/82/, accessed on 1 February 2023.

Acknowledgments: The authors thank He Zhao and Ziyang Chen for their careful review and advice.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gui, S.; Zhou, C.; Wang, H.; Gao, T. Application of Voiceprint Recognition Technology Based on Channel Confrontation Training in the Field of Information Security. *Electronics* **2023**, *12*, 3309. [CrossRef]
2. Li, S.-A.; Liu, Y.-Y.; Chen, Y.-C.; Feng, H.-M.; Shen, P.-K.; Wu, Y.-C. Voice Interaction Recognition Design in Real-Life Scenario Mobile Robot Applications. *Appl. Sci.* **2023**, *13*, 3359. [CrossRef]
3. Cheng, S.; Shen, Y.; Wang, D. Target Speaker Extraction by Fusing Voiceprint Features. *Appl. Sci.* **2022**, *12*, 8152. [CrossRef]
4. Ye, F.; Yang, J. A Deep Neural Network Model for Speaker Identification. *Appl. Sci.* **2021**, *11*, 3603. [CrossRef]
5. Yao, W.; Xu, Y.; Qian, Y.; Sheng, G.; Jiang, X. A Classification System for Insulation Defect Identification of Gas-Insulated Switchgear (GIS), Based on Voiceprint Recognition Technology. *Appl. Sci.* **2020**, *10*, 3995. [CrossRef]
6. Shi, Y.; Zhou, J.; Long, Y.; Li, Y.; Mao, H. Addressing Text-Dependent Speaker Verification Using Singing Speech. *Appl. Sci.* **2019**, *9*, 2636. [CrossRef]
7. Uyulan, Ç.; Mayor, D.; Steffert, T.; Watson, T.; Banks, D. Classification of the Central Effects of Transcutaneous Electroacupuncture Stimulation (TEAS) at Different Frequencies: A Deep Learning Approach Using Wavelet Packet Decomposition with an Entropy Estimator. *Appl. Sci.* **2023**, *13*, 2703. [CrossRef]
8. Sun, T.; Wang, X.; Zhang, K.; Jiang, D.; Lin, D.; Jv, X.; Ding, B.; Zhu, W. Medical Image Authentication Method Based on the Wavelet Packet and Energy Entropy. *Entropy* **2022**, *24*, 798. [CrossRef]

9. Zhang, Y.; Xie, X.; Li, H.; Zhou, B. An Unsupervised Tunnel Damage Identification Method Based on Convolutional Variational Auto-Encoder and Wavelet Packet Analysis. *Sensors* **2022**, *22*, 2412. [CrossRef]
10. Lei, L.; She, K. Identity Vector Extraction by Perceptual Wavelet Packet Entropy and Convolutional Neural Network for Voice Authentication. *Entropy* **2018**, *20*, 600. [CrossRef]
11. Lei, L.; She, K. Speaker Recognition Using Wavelet Cepstral Coefficient, I-Vector, and Cosine Distance Scoring and Its Application for Forensics. *J. Electr. Comput. Eng.* **2016**, *2016*, 462–472. [CrossRef]
12. Daqrouq, K.; Sweidan, H.; Balamesh, A.; Ajour, M.N. Off-Line Handwritten Signature Recognition by Wavelet Entropy and Neural Network. *Entropy* **2017**, *6*, 252. [CrossRef]
13. Dawalatabad, N.; Ravanelli, M.; Grondin, F.; Thienpondt, J.; Desplanques, B.; Na, H. ECAPA-TDNN Embeddings for Speaker Diarization. *arXiv* **2021**, arXiv:2104.01466.
14. Jung, S.-Y.; Liao, C.-H.; Wu, Y.-S.; Yuan, S.-M.; Sun, C.-T. Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features. *Diagnostics* **2021**, *11*, 732. [CrossRef]
15. Joy, N.M.; Oglic, D.; Cvetkovic, Z.; Bell, P.; Renals, S. Deep Scattering Power Spectrum Features for Robust Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 1673–1677.
16. Gao, Z.; Song, Y.; McLoughlin, I.; Li, P.; Jiang, Y.; Dai, L.-R. Improving Aggregation and Loss Function for Better Embedding Learning in End-to-End Speaker Verification System. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 361–365.
17. Bousquet, P.-M.; Rouvier, M.; Bonastre, J.-F. Reliability criterion based on learning-phase entropy for speaker recognition with neural network. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 281–285.
18. Sang, M.; Hansen, J.H.L. Multi-Frequency Information Enhanced Channel Attention Module for Speaker Representation Learning. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 321–325.
19. Stafylakis, T.; Mosner, L.; Plchot, O.; Rohdin, J.; Silnova, A.; Burget, L.; Černocký, J. Training speaker embedding extractors using multi-speaker audio with unknown speaker boundaries. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 605–609.
20. Luu, C.; Renals, S.; Bell, P. Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 610–614.
21. Zhu, H.; Lee, K.A.; Li, H. Serialized Multi-Layer Multi-Head Attention for Neural Speaker Embedding. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 106–110.
22. Li, G.; Liang, S.; Nie, S.; Liu, W.; Yang, Z.; Xiao, L. Deep Neural Network-Based Generalized Sidelobe Canceller for Robust Multi-Channel Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 51–55.
23. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798. [CrossRef]
24. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
25. Liu, Y.; He, L.; Liu, W.; Liu, J. Exploring a unified attention based pooling framework for speaker verification. In Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan, 26–29 November 2018; pp. 200–204.
26. Cai, D.; Wang, W.; Li, M. An iterative framework for selfsupervised deep speaker representation learning. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6728–6732.
27. Yang, J.; Jiang, J. Dilated-CBAM: An Efficient Attention Network with Dilated Convolution. In Proceedings of the IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 15–17 October 2021; pp. 11–15.
28. Liu, R.; Cai, W.; Li, G.; Ning, X.; Jiang, Y. Hybrid Dilated Convolution Guided Feature Filtering and Enhancement Strategy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5508105. [CrossRef]
29. Yang, L.; Chen, W.; Wang, H.; Chen, Y. Deep learning seismic random Noise attenuation via improved residual convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *59*, 7968–7981. [CrossRef]
30. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3830–3834.
31. Haitao, C.; Yu, L.; Yun, Y. Research on voiceprint Recognition system based on ECAPA-TDNN-GRU architecture. In Proceedings of the International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 24–26 February 2023; pp. 1508–1513.
32. Li, J.; Xu, Q.; Kadoch, M. A Study Of Voiceprint Recognition Technology Based on Deep Learning. In Proceedings of the International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, 30 May–3 June 2022; pp. 24–27.
33. Dong, X.; Song, J. Application of Voiceprint Recognition Based on Improved ECAPA-TDNN. In Proceedings of the International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 9–11 December 2022; pp. 1196–1199.
34. Bayerl, S.P.; Wagner, D.; Baumann, I.; Bocklet, T.; Riedhammer, K. Detecting Vocal Fatigue with Neural Embeddings. *J. Voice* **2023**, *1*, 3428–3439. [CrossRef]

35. Zhu, H.; Lee, K.A.; Li, H. Discriminative speaker embedding with serialized multi-layer multi-head attention. *Speech Commun.* **2022**, *144*, 89–100. [CrossRef]
36. Strake, M.; Defraene, B.; Fluyt, K.; Tirry, W.; Fingscheidt, T. INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 2467–2471.
37. Li, Y.; Zhang, X.; Zhang, X.; Li, H.; Zhang, W. Unconstrained vocal pattern recognition algorithm based on attention mechanism. *Digit. Signal Process.* **2023**, *136*, 103973. [CrossRef]
38. Lin, S.; Zhang, M.; Cheng, X.; Zhou, K.; Zhao, S.; Wang, H. Hyperspectral anomaly detection via sparse representation and collaborative representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 946–961. [CrossRef]
39. Lin, S.; Zhang, M.; Cheng, X.; Wang, L.; Xu, M.; Wang, H. Hyperspectral Anomaly Detection via Dual Dictionaries Construction Guided by Two-Stage Complementary Decision. *Remote Sens.* **2022**, *14*, 1784. [CrossRef]
40. Zi, Y.; Xiong, S. Joint filter combination-based central difference feature extraction and attention-enhanced Dense-Res2Block network for short-utterance speaker recognition. *Expert Syst. Appl.* **2023**, *233*, 1–12. [CrossRef]
41. Hanifa, R.M.; Isa, K.; Mohamad, S. A review on speaker recognition: Technology and challenges. *Comput. Electr. Eng.* **2021**, *90*, 1–14.
42. Lin, S.; Zhang, M.; Cheng, X.; Zhou, K.; Zhao, S.; Wang, H. Dual Collaborative Constraints Regularized Low-Rank and Sparse Representation via Robust Dictionaries Construction for Hyperspectral Anomaly Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 2009–2024. [CrossRef]
43. Lin, S.; Zhang, M.; Cheng, X.; Zhao, S.; Shi, L.; Wang, H. Hyperspectral Anomaly Detection Using Spatial–Spectral-Based Union Dictionary and Improved Saliency Weight. *Remote Sens.* **2023**, *15*, 3609. [CrossRef]
44. Tsao, Y.; Lin, T.-H.; Chen, F.; Chang, Y.-F.; Cheng, C.-H.; Tsai, K.-H. Robust S1 and S2 heart sound recognition based on spectral restoration and multi-style training. *Biomed. Signal Process. Control* **2019**, *49*, 173–180. [CrossRef]
45. Lee, J.; Nam, J. Multi-Level and Multi-Scale Feature Aggregation Using Pretrained Convolutional Neural Networks for Music Auto-Tagging. *IEEE Signal Process. Lett.* **2017**, *24*, 1208–1212. [CrossRef]
46. Le, X.; Lei, T.; Chen, K.; Lu, J. Inference Skipping for More Efficient Real-Time Speech Enhancement With Parallel RNNs. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2411–2421. [CrossRef]