electronics

MDPI

Article

Enhanced-Deep-Residual-Shrinkage-Network-Based Voiceprint Recognition in the Electric Industry

Qingrui Zhang [1], Hongting Zhai [1], Yuanyuan Ma [2], Lili Sun [1], Yantong Zhang [1], Weihong Quan [1], Qi Zhai [1], Bangwei He [2] and Zhiquan Bai [2,*]

[1] Information and Telecommunications Branch, State Grid Shandong Electric Power Company, Jinan 250001, China; zhangqingrui1993@126.com (Q.Z.); zhaihongtingbupt@163.com (H.Z.); sddlsll@163.com (L.S.); 13791051765@139.com (Y.Z.); 2120180468@mail.nankai.edu.cn (W.Q.); zhaiqi10690@163.com (Q.Z.)

[2] School of Information Science and Engineering, Shandong University, Qingdao 266237, China; myy98@mail.sdu.edu.cn (Y.M.); hbw017@mail.sdu.edu.cn (B.H.)

* Correspondence: zqbai@sdu.edu.cn; Tel.: +86-13355319215

Abstract: Voiceprint recognition can extract voice features and identity the speaker through the voice information, which has great application prospects in personnel identity verification and voice dispatching in the electric industry. The traditional voiceprint recognition algorithms work well in a quiet environment. However, noise interference inevitably exists in the electric industry, degrading the accuracy of traditional voiceprint recognition algorithms. In this paper, we propose an enhanced deep residual shrinkage network (EDRSN)-based voiceprint recognition by combining the traditional voiceprint recognition algorithms with deep learning (DL) in the context of the noisy electric industry environment, where a dual-path convolution recurrent network (DPCRN) is employed to reduce the noise, and its structure is also improved based on the deep residual shrinkage network (DRSN). Moreover, we further use a convolutional block attention mechanism (CBAM) module and a hybrid dilated convolution (HDC) in the proposed EDRSN. Simulation results show that the proposed network can enhance the speaker's vocal features and further distinguish and eliminate the noise features, thus reducing the noise influence and achieving better recognition performance in a noisy electric environment.

Keywords: voiceprint recognition; deep learning; deep residual shrinkage network; convolutional block attention mechanism; hybrid dilated convolution

Citation: Zhang, Q.; Zhai, H.; Ma, Y.; Sun, L.; Zhang, Y.; Quan, W.; Zhai, Q.; He, B.; Bai, Z. Enhanced-Deep-Residual-Shrinkage-Network-Based Voiceprint Recognition in the Electric Industry. *Electronics* **2023**, *12*, 3017. https://doi.org/10.3390/electronics12143017

Academic Editors: Yu-Chen Hu, Praveen Kumar Donta, Piyush Kumar Pareek and Chinmaya Kumar Dehury

Received: 6 June 2023
Revised: 4 July 2023
Accepted: 5 July 2023
Published: 10 July 2023

1. Introduction

With the rapid development of the social economy and urban construction, the high efficient operation and management of the electric industry becomes more and more important, and the limitations of the traditional electric industry are gradually appearing. Nowadays, intelligent electricity is widely and rapidly developing. It is a new generation of electric systems based on traditional electricity and it integrates new materials, new equipment, and new technologies, such as information and control technology and artificial intelligence (AI), which has the characteristics of having a high informatization and automation and which ensures an efficient and reliable operation of electric systems.

As the scale of the electric industry gradually expands, its applications tend to be extensive, and electricity dispatching plays an important role among them. At present, the dispatching commands are mainly given by human voice, which makes the dispatching system have certain security risks due to the lack of personnel verification and the inaccurate voice recognition. Voiceprint recognition [1], as a biometric verification method, can improve the voice recognition security even in the voice-dispatching scenario. In electric communication systems, voice dispatching and identity authentication can be implemented by voiceprint recognition technology, so that the dispatching commands can be given without additional identity authentication steps, which ensures the reliability and security

of electricity dispatching. However, due to the presence of large mechanical noise and an electromagnetic noise environment in the electric industry [2], the voice signals are easily disturbed, which significantly reduces the accuracy of traditional voiceprint recognition algorithms. As a main branch of AI, deep learning (DL) can be applied to the electricity dispatching system because of its inherent characteristics, especially for noisy scenarios. Combined with voiceprint recognition technology, we can use DL to extract the noise features and reduce the voice noise and then feed the noise-reduced voice signals into the designed voiceprint recognition network to realize highly reliable dispatcher authentication and improve the security of electricity dispatching. With the development of DL techniques, more and more network models have been proposed for speaker recognition. However, the following problems still exist in the field of DL:

(1) The continuous deepening of the model leads to the degradation of the network performance.
(2) The error gradient used to update the network weights keeps increasing or decreasing, which results in a gradient explosion and gradient disappearance.
(3) A low accuracy for short-time speech recognition and a low recognition accuracy in noisy scenarios.
(4) A poor robustness.

To solve the above problems in the noisy electric industry, this paper takes full advantage of DL and combines it with traditional voice preprocessing methods to improve the recognition accuracy. Our contributions are summarized below.

(1) We divide the voiceprint recognition in a noisy environment into two main parts: the first part performs the noise reduction on the voice dataset, and the second part aims to realize identity matching by voiceprint recognition. In the noise reduction part, we model the voice signal features using a dual-path convolutional recurrent network (DPCRN) [3] in a noisy electric environment and set the model learning target as the complex ratio mask (CRM). First, we use the spectrogram of the noisy voice signals as the inputs to the encoder. Subsequently, an RNN is used in the frequency domain to capture the long-term speech's harmonic correlations [4]. Finally, the real and imaginary parts of the CRM are output at the decoder. The learning objective is optimized by a signal approximation (SA), and the multiplication of the estimated CRM with the noisy signal spectrogram is performed to achieve the noise reduction process.
(2) For the voiceprint recognition, an enhanced deep residual shrinkage network (EDRSN) is proposed in a noisy electric environment. The proposed EDRSN scheme reconstructs the network structure based on a deep residual shrinkage network (DRSN) [5] and combines the convolutional block attention mechanism (CBAM) [6] and the hybrid dilated convolution (HDC) [7]. Meanwhile, we combine the EDRSN with traditional voice processing methods to accomplish the voiceprint recognition in noisy environments. After the noise reduction by DPCRN, we enhance the voice segments in the voice signals by pre-emphasizing and eliminating the silent segments by an endpoint detection to facilitate the extraction of voiceprint features. Finally, the soft thresholding mechanism of EDRSN is utilized to further distinguish and eliminate the noisy features, and the CBAM and HDC are taken to extract the effective vocal features and improve the recognition accuracy.
(3) Simulation results are provided to verify the accuracy of the proposed scheme in voice recognition. Based on the comparison of the time–frequency spectrograms before and after noise reduction, it is shown that the DPCRN is able to reduce the background noise in the voice signals. Furthermore, numerical results show that the proposed EDRSN model has a better accuracy in voiceprint recognition than the other neural network models while ensuring a lower complexity.

The remaining sections of this paper are organized as follows. In Section 2, we present some related work in the field of voiceprint recognition. In Section 3, we introduce the

noise reduction scheme for voice signals, including the DPCRN model and the principle of noise reduction. In Section 4, the preprocessing and feature extraction process of the voiceprint recognition are given, followed by the identity recognition method. In Section 5, the voiceprint recognition network EDRSN is proposed, and its basic composition and structural parameters are presented. The simulation results and conclusions of this paper are presented in Sections 6 and 7, respectively.

2. Related Work

A method for speech feature extraction, linear prediction coding (LPC), was presented in [8], where the linear sampling prediction was obtained a linear fitting and local minimization algorithm. In ref. [9], Bing-Hwang Juang first used Gaussian mixture models (GMM) to represent the relationship between hidden Markov models (HMM) states and acoustic inputs. Ref. [10] provided a joint factor analysis (JFA) method, which further modeled the space where the mean supervector obtained by the GMM-based method was located to compensate for the effects of channel variation. However, this model still had a very high complexity. A Gaussian mixture model–universal background model (GMM-UBM) was proposed in [11] for speaker verification, and the proposed system included preprocessing, feature extraction, modeling and classification stages. Pitch frequencies and Mel frequency cepstrum coefficients (MFCC) were used as feature vectors. It utilized a large quantity of data for the computation and had a high computational complexity, but its recognition accuracy was not sufficient.

At present, DL is widely used in the field of voiceprint recognition and has achieved good results due to its operation mechanism. D-vector was proposed by Google in 2014 to convert the training process into a classification problem [12], which took the hidden layer output of the neural network instead of the I-Vector method and proved the feasibility of using a DL method in voiceprint recognition. A recurrent neural network (RNN) [13] takes the sequence data as its input and recurs in the direction of sequence evolution, where all the nodes (recurrent units) are connected in a chainlike manner. RNNs are able to handle speech sequences with variable length well and are widely used in speech recognition. Nowadays, RNNs are also applied in the field of voiceprint recognition. D. Snyder et al. used the time-delay neural network to extract the frame-level features and the aggregated statistical pooling layers to extract the utterance-level features in [14], where the probabilistic linear discriminant analysis (PLDA) was used for back-end scoring, and offline data were added for data enhancement. The overall effectiveness of the model surpassed that of the I-Vector scheme. The work in [15] replaced the traditional Gaussian mixture model (GMM) with a deep neural network (DNN) and proposed a new GMM-derived (GMMD) algorithm to train the DNN acoustic model. This work took each voice frame of the speaker as the input and took a hidden layer to extract the speaker's voice features for regularization with a d-vector vocal recognition model. The analysis and comparison of the related works is shown in Table 1.

Table 1. Analysis of the related works.

Algorithms	Technologies	Datasets	Results
GMM-UBM	GMM	TIMIT corpora	Accuracy: 80.83%
JFA	GMM	NIST	EER [1]: 5.2%
RNN	DL	Self-collected datasets	FAR [2]: outperformed GMM by 26%
d-vector DNN	DL	Self-collected datasets	EER: outperformed i-vector by 14%
End-to-end DNN	DL	US English speech	EER: outperformed i-vector by 29%
GMMD DNN	DL	CHiME	WER [3]: outperformed baseline by 16%

[1] Equal error rate (EER). [2] False accept rate (FAR). [3] Word error rate (WER).

3. Noise Reduction

The noise interference in the electric industry may affect the voiceprint information and degrade the recognition accuracy. To reduce the impact of noise on the performance of voiceprint recognition and overcome the problems of incomplete or nonideal noise

reduction of traditional noise elimination methods, we employed a DPCRN to separate the clean speech from the background interference and increase the speech intelligibility and perceptual quality. The DPCRN model consists of an encoder, a two-path RNN module, and a decoder, as shown in Figure 1. The core of DPCRN is an RNN structure, which has two types of RNNs, intrablock and interblock RNNs. The intrablock RNN is used to simulate the spectrum of a single time period, and the interblock RNN is used to simulate the change in the spectrum over time.

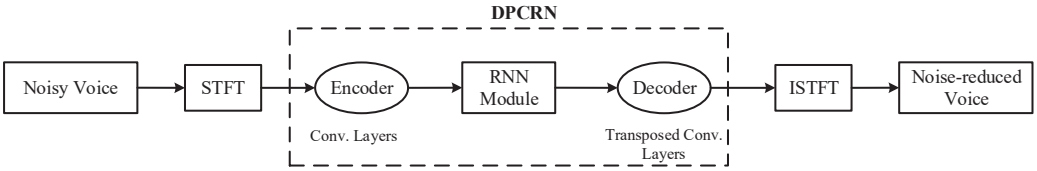


Figure 1. Noise reduction process of DPCRN.

3.1. Noise Reduction Theory

For a voice containing noise, its time domain expression can be expressed as

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ and $n(t)$ represent the pure voice and the noise in the time domain, respectively.

Then, the short-time Fourier transform (STFT) is used to convert the time-domain noisy voice signal $x(t)$ into the time–frequency-domain signal $X(t, f)$ as

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-j2\pi f \tau} d\tau, \tag{2}$$

where $h(t)$ is the STFT analysis window function. The time-frequency domain expression of Equation (1) becomes

$$X(t, f) = S(t, f) + N(t, f), \tag{3}$$

where $S(t, f)$ and $N(t, f)$ are the STFTs of the pure voice $s(t)$ and the noise $n(t)$ at time t and frequency f, respectively. In order to recover the pure voice signal, we can make the estimation by a CRM as

$$M(t, f) = M_r(t, f) + jM_j(t, f), \tag{4}$$

where $M(t, f)$ is the complex signal and consists of a real part $M_r(t, f)$ and an imaginary part $M_j(t, f)$.

Then, we can obtain the recovered voice signal $\hat{S}(t, f)$ by multiplying $M(t, f)$ with the noisy voice signal $X(t, f)$ as

$$\hat{S}(t, f) = X(t, f) \odot M(t, f), \tag{5}$$

where $\odot$ denotes the element-by-element multiplication of two vectors. Finally, the noise-reduced voice signal in the time domain obtained through the inverse short-time Fourier transform (ISTFT) can be expressed as

$$\hat{s}(t) = \int_{-\infty}^{\infty} \hat{S}(t, f) e^{j2\pi t f} df. \tag{6}$$

3.2. Noise Reduction Process

Unlike the traditional noise reduction modeling method in the time domain, the DPCRN utilizes the harmonic structure of the voice signals and models it based on frequency characteristics, thus providing better speech noise reduction performance. The RNNs in DPCRN can overcome the disadvantages of a partial absence of information in

convolutional neural networks (CNNs) [16] and can capture the harmonic correlation of voice signals over a long time.

We feed the noisy voice signals into the DPCRN and send its real and imaginary parts to the encoder as two data streams. The encoder employs a two-dimensional convolutional layer to extract the voice features from the noise spectrogram and compress the feature resolution. The decoder is symmetric with respect to the encoder and uses the transposed convolutional layers to restore the low-resolution features to their original size. The DPCRN outputs a CRM in the last transposed convolutional layer by learning the voice features and using the signal approximation algorithm. Finally, the noise is removed by CRM and the time-domain noise-reduced voice signals can be obtained by an ISTFT.

4. Voiceprint Recognition

After the processing of DPCRN, we can obtain the noise-reduced voice signals and start the voiceprint recognition process. The purpose of voiceprint recognition is the identification of the speaker based on the uniqueness of the voiceprint features due to the specific vocal fold construction. The process of voiceprint recognition consists of four parts: preprocessing, audio feature extraction, voiceprint feature extraction, and identity evaluation. Among them, preprocessing is a necessary prerequisite for audio feature extraction, and these two parts are usually referred to as preprocessing in voiceprint recognition. Therefore, we introduce the process of voiceprint recognition from three aspects: preprocessing, voiceprint feature extraction, and identity recognition evaluation, as shown in Figure 2.

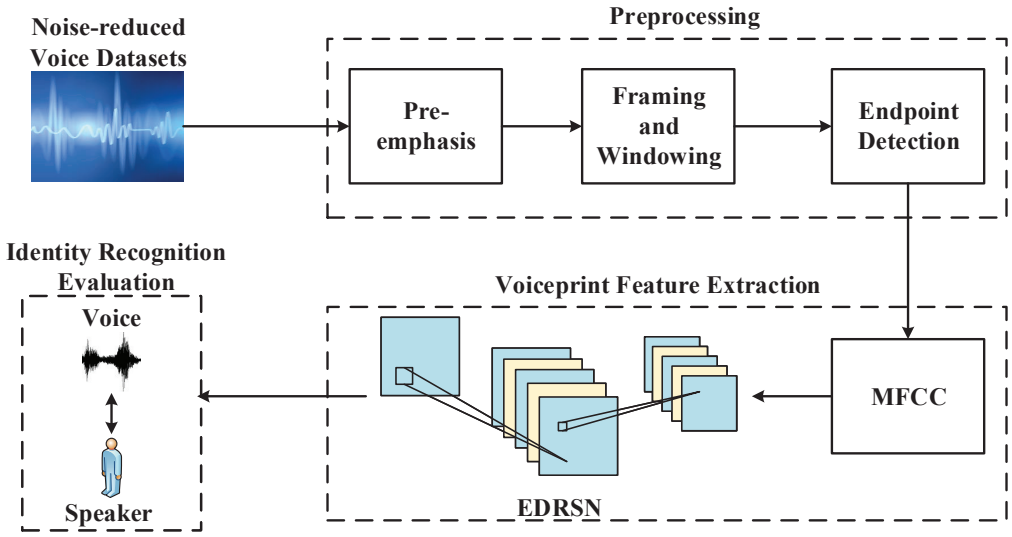


Figure 2. Diagram of the voiceprint recognition system.

4.1. Preprocessing

The preprocessing of the voice signals has three main steps, including a pre-emphasis, framing and windowing, and endpoint detection [17].

When a speaker makes a sound, the voice is radiated by the mouth and lips. Meanwhile, the air is used as a medium in which the voice signals consume energy as they travel. The higher the frequency of the voice signals, the greater the energy loss. Therefore, a pre-emphasis is needed for the processing of voice signals, which can alleviate the effect of voice radiation to a certain extent and compensate for the loss of high-frequency voice signals. In voiceprint recognition, a high-pass filter is generally used to achieve this purpose, which can be written as

$$H(Z) = 1 - \mu Z^{-1}, \tag{7}$$

with $\mu \in [0.9, 1]$. If the input signal is $x[n]$, the output $y[n]$ through the filter can be expressed as

$$y[n] = x[n] - \mu x[n-1]. \tag{8}$$

Framing is the process of splitting the voice signals in very small intervals and treating these intervals as a smooth signal. This is because in the feature extraction process of voiceprint recognition, smooth voice signals are needed for the Fourier transform, while pre-emphasized voice signals are fluctuating. When we split the voice signals into voice frames with time intervals of about 20 ms, each small segment can be considered to be smooth. The frame shift is the difference between the starting positions of two adjacent voice frames, and the ratio of the frame shift to the frame length is generally less than one-half. The voice frame's splitting diagram is shown in Figure 3.

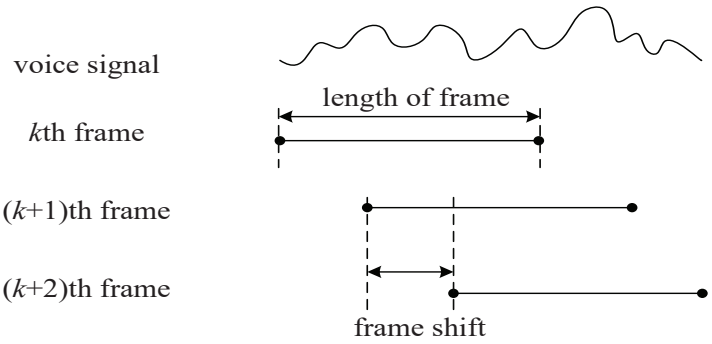


Figure 3. Voice framing model.

After framing, there exist discontinuities in the starting and ending positions of the signal, and if the feature extraction is performed directly on the signal by a Fourier transform, a Gibbs phenomenon will occur and cause spectral loss. To deal with the above problems, we can perform a windowing operation to multiply the original voice signals with the window function. In the process of voiceprint recognition, we generally take the Hemming window [18] as

$$\omega(n) = [0.54 - 0.46\cos(\frac{2\pi n}{N-1})]R_M(n), \tag{9}$$

where M is the length of the Hemming window function and $R_M(n)$ is the rectangular window that can be denoted as

$$R_M(n) = \begin{cases} 1, & 0 \le n \le M-1 \\ 0, & otherwise. \end{cases}. \tag{10}$$

Finally, we need to perform the endpoint detection on the voice signals, which aims to distinguish the silent part from the nonsilent part, so as to filter the invalid information in the voice signals and keep the valid one. The endpoint detection marks the start and end points of the voice segments, removes the silent and noisy parts, and gets the valid voice information.

In this paper, a double-threshold method for endpoint detection was used, which contained a short-time energy detection and short-time overzero detection to further distinguish the voice segment from the noise segment and to distinguish the voice segment from the silent segment, respectively. For the short-time energy of a voice signal y at a moment n, we can consider it as the sum of the squares of the samples of the frame, which can be expressed as

$$E_n = \sum_{m=n-(N-1)}^{n} [y(m)\omega(n-m)]^2 \tag{11}$$

We define the short-time overzero as the number of voice signals passing through the zero value per second. When the window function starts from zero, the short-time per-zero rate Z_n can be calculated as

$$Z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[y(m)] - \mathrm{sgn}[y(m-1)]|\omega(n-m) \tag{12}$$

where sgn() is the step function, which can be expressed as

$$\mathrm{sgn}(n) = \begin{cases} 1, & n \geq 0 \\ -1, & n < 0 \end{cases} \tag{13}$$

Based on the characteristic parameters, such as short-time energy and short-time overzero rate, we can set the high and low thresholds to detect the signal changes and complete the endpoint detection process.

4.2. Voiceprint Feature Extraction

When the preprocessing process is completed, the voice signals become voice frames with a fixed time interval, which subsequently need to be subjected to a voiceprint feature extraction, and their features are used as the input to the neural network. In the feature extraction, acoustic features are commonly obtained by typical methods, such as MFCC, linear predictive cepstral coefficients (LPCC), and a spectrogram. In this paper, we adopted the speech spectrogram method in the voiceprint recognition process and applied the discrete Fourier transform (DFT) [19] for the voice frames as

$$x[k] = \sum_{n-0}^{N-1} x[n]e^{-j\frac{2\pi}{N}nk}, 0 \leq k \leq N-1, \tag{14}$$

where $x[k]$ is the frequency-domain signal and N is the length of the discrete signal $x[n]$.

The STFT is generally used in the calculation of the time-frequency spectrum [20]. The voice signals are split into many time segments with small time intervals, and the STFT is used in each time segment as

$$\sum_{n=-\infty}^{+\infty} x(n)\omega(n - mL_R)e^{-j\omega n}, \tag{15}$$

where $x(n)$ is a continuous voice signal, m is the minimum length interval of the window, and L_R is the window movement distance.

In Section 4, we describe the neural network designed for the extraction of the voiceprint features in detail.

4.3. Identity Recognition Evaluation

We input two voice signals and obtained their acoustic features by the speech spectrogram method. With these feature data, the diagonal cosine value of the two voice signals can be calculated as [21]

$$\cos\theta = \frac{\sum_{i=1}^{n} (s_{1,i} \times s_{2,i})}{\sqrt{\sum_{i=1}^{n} s_{1,i}^2} \sqrt{\sum_{i=1}^{n} s_{2,i}^2}}, \tag{16}$$

where n represents the number of input voice signal pairs, $s_{1,i}$ is the first signal of the ith pair of the voice signals, and $s_{2,i}$ denotes the second signal of the ith pair of the voice signals.

The obtained cosine value $\cos\theta$ can be used to evaluate their similarity. We set the threshold of the similarity to be 0.7, and if $\cos\theta > 0.7$, the two voice signals can be considered as emitted by the same speaker, otherwise they are from two different speakers.

5. Enhanced Deep Residual Shrinkage Network

In this section, we redesign the network structure based on DRSN and introduce the CBAM module and HDC to build the EDRSN. By employing the EDRSN for voiceprint recognition, we can further enhance the extraction capability of voiceprint features.

5.1. Deep Residual Shrinkage Network

To solve the problems in Section 1, a constant mapping was considered in CNNs [22]. In 2017, Li et al. used residual networks for acoustic recognition, proposed a deep residual convolutional neural network (Res-CNN) to extract voiceprint features, and then trained the model using a triplet loss method based on the cosine similarity. Compared to the DNN-based i-vector recognition method, its recognition accuracy was improved by 60% on a text-independent dataset [23].

The DRSN is an optimization of the residual network (ResNet) [24], which introduces a soft thresholding mechanism, while retaining the residual module and the constant path in ResNet. As a classical denoising method, the input signal is firstly decomposed by the convolutional layers, then the signal is filtered by automatically generated thresholds, and finally, the filtered signal can be reconstructed. Zhao et al. introduced the soft thresholding method in the voiceprint fault diagnosis [25] and proved that the DRSN model with soft thresholding improved the accuracy performance by 2% over the ResNet model, due to the fact that the soft thresholding could ignore the noise-related features contained in different channels of the feature map. The basic structure of the residual shrinkage module is shown in Figure 4.

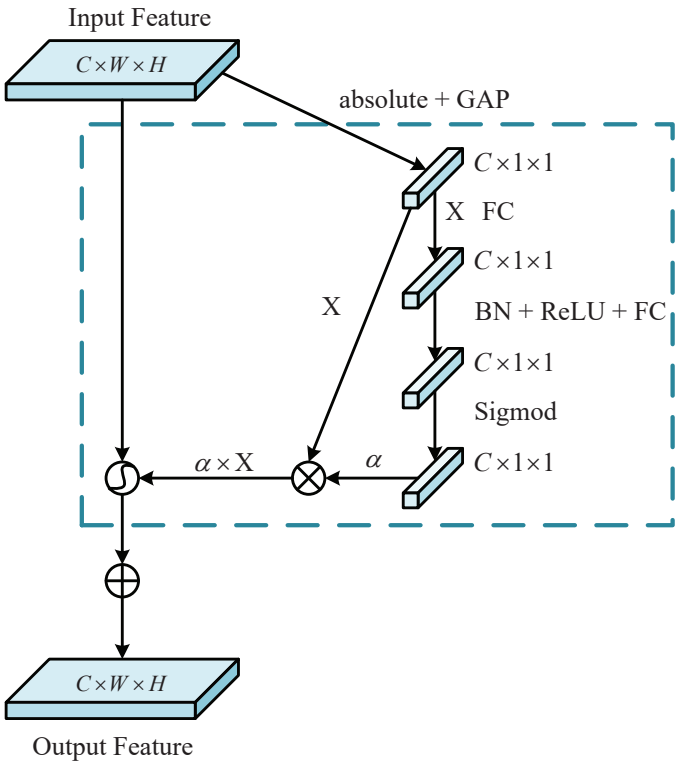


Figure 4. Diagram of the deep residual shrinkage module.

5.2. Convolutional Block Attention Mechanism Module

The core of the CBAM module is to focus on the key features and ignore the irrelevant features. As an attention module applied to feedforward neural networks, the CBAM module serially incorporates a channel attention module (CAM) and a spatial attention module (SAM) to perform the feature extraction in two independent dimensions for the adaptive feature optimization. The CBAM module is a lightweight attention module that can be easily integrated into the DRSN architecture, while reducing the use of large

numbers of parameters and the complexity of the model. It was discussed in [26] how the CBAM was able to improve the feature extraction without increasing the complexity, outperforming the traditional SE attention module with a 22.66% error in ImageNet-1K classification experiments [27]. The structure of the CBAM module is shown in Figure 5.

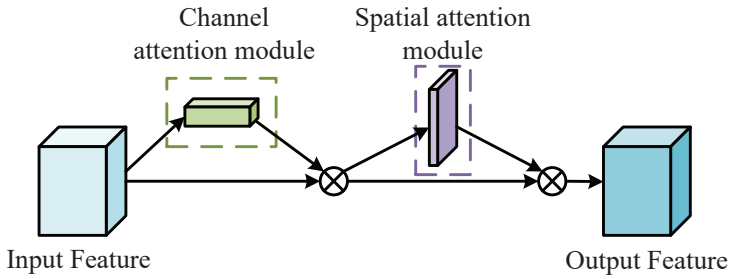


Figure 5. Structure of CBAM Module.

In the CAM, the input feature map $\mathbf{X}$ is first subjected to a global maximum pooling and a global average pooling operation. Then, the results are fed into a two-layer multilayer perceptron (MLP), and the output results are summed based on the elementwise summation operation. Finally, the obtained results are fed into the activation function to obtain the feature map of the channel attention $\mathbf{M}_c(\mathbf{X})$ as

$$\mathbf{M}_c(\mathbf{X}) = \sigma\left(\mathbf{W}_2\left(\mathbf{W}_1\left(\mathbf{X}^c_{avg}\right)\right) + \mathbf{W}_2(\mathbf{W}_1(\mathbf{X}^c_{max}))\right), \tag{17}$$

where σ is the sigmoid activation function, $\mathbf{W}_1$ is the weight matrix of the first layer of the MLP, and $\mathbf{W}_2$ denotes the weight matrix of the second layer of the MLP. $\mathbf{X}^c_{max}$ and $\mathbf{X}^c_{avg}$ are the max pooling and global average pooling operations on $\mathbf{X}$ in the channel dimension, respectively.

For the SAM, the feature map output $\mathbf{Y}$ from CAM is used as its input. First, the channel-based global max pooling and the global average pooling operations are performed on the input $\mathbf{Y}$. Then, the two feature maps are concatenated according to the multilayer fusion operation, and the result is further fed into the convolutional layer for dimensionality reduction. Finally, the reduced-dimensional result is passed through the activation function to obtain the feature map of the spatial attention $\mathbf{M}_s(\mathbf{Y})$ as

$$\mathbf{M}_s(\mathbf{Y}) = \sigma\left(conv^{7\times7}\left(\left[\mathbf{Y}^s_{avg}; \mathbf{Y}^s_{max}\right]\right)\right), \tag{18}$$

where $conv^{7\times7}$ represents a convolution operation with a kernel size of 7×7, $\mathbf{Y}^s_{max}$ and $\mathbf{Y}^s_{avg}$ denote the max pooling and global average pooling operations on $\mathbf{Y}$ in the spatial dimension, respectively.

5.3. Hybrid Dilated Convolution

In the design of the EDRSN, the use of multiple convolutional layers subsequently led to a loss of local information and a resolution degradation of the input information. To cope with these issues, we introduced a null convolution in the convolutional layer by changing the size of the perceptual field. In the convolutional layer, we can change the perceptual field by setting the size of the dilation rate. Taken Figure 6 as an example, for a 3×3 convolutional kernel with a dilation rate of 1, the perceptual field size is the same as the original convolutional kernel. Therefore, for successive convolutional layers in a neural network, increasing the dilation rate can increase the receptive field size. However, the continuous increase in the dilation rate may lead to data loss due to the kernel discontinuity. In this case, we can apply an HDC so that the dilation rate of the superimposed convolution cannot get a factor greater than 1.

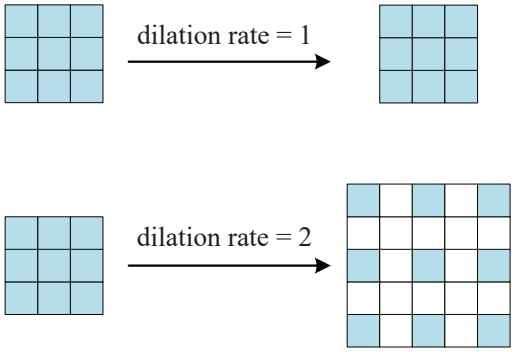


Figure 6. Introduction diagram of the dilated convolution.

Figure 7 describes the architecture of the proposed EDRSN in detail. We improved the network structure based on a DRSN and introduced a CBAM and an HDC to form the EDRSN. The EDRSN consisted of one input layer, three enhanced residual shrinkage block units (ERSBU), one average layer, one affine layer, and one output layer. As shown in Table 2, each ERSBU contained three convolutional layers, two CBAM modules, one residual path, and one soft-threshold module. We put the CBAM module between every two convolutional layers as a way to enhance the extraction of acoustic features and set the dilation rates of the three consecutive convolutional layers to be 1, 2, and 5, respectively, to expand the perceptual field. Finally, the noise effect on the voiceprint was further reduced through the soft thresholding. We set three consecutive ERSBUs, added an average layer as well as a fully connected layer at the end of the network, and output the recognition results through the output layer.

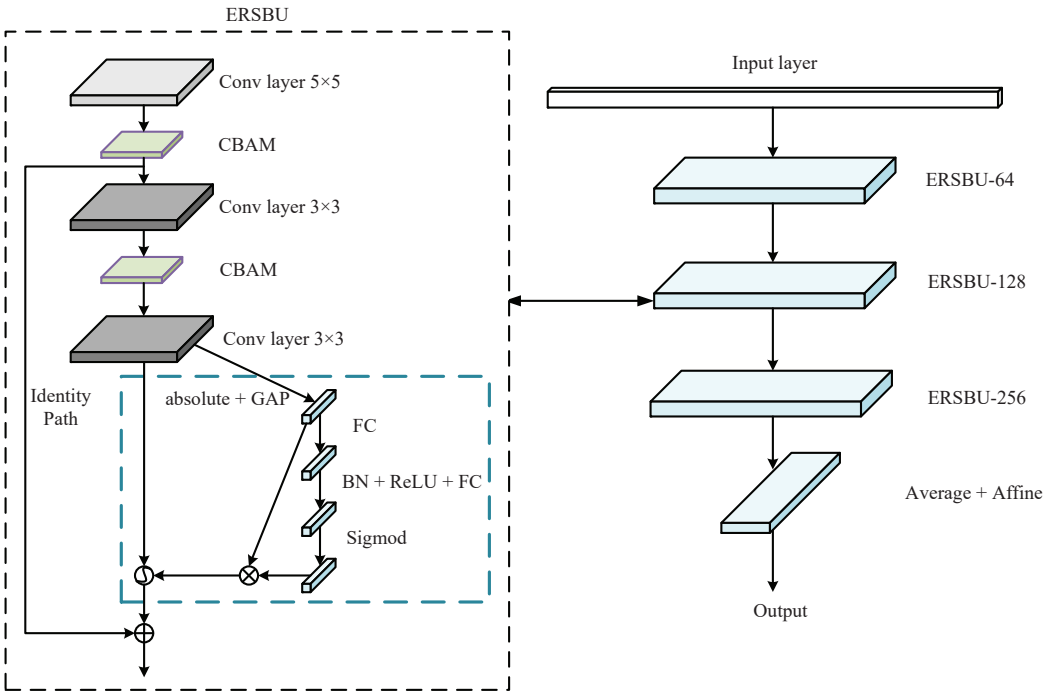


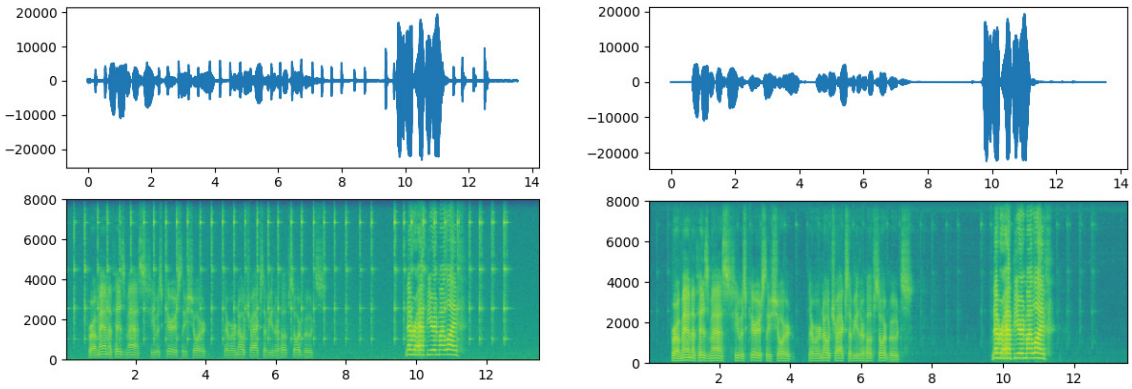
Figure 7. Structure of the EDRSN.

Table 2. Architecture of the EDRSN.

Layer	Structure	Stride	Dim
Input	-	-	-
ERSBU-64	5×5, 64	2×2	2048
	$[3 \times 3, 64] \times 2$	1×1	2048
ERSBU-128	5×5, 128	2×2	2048
	$[3 \times 3, 128] \times 2$	1×1	2048
ERSBU-256	5×5, 256	2×2	2048
	$[3 \times 3, 256] \times 2$	1×1	2048
Average	-	-	2048
Affine	2048×256	-	256
K.l2normalize	-	-	256
Output	-	-	-

6. Simulation Results

In this section, we provide the simulation results of the proposed voiceprint recognition system in a noisy electric environment. In the experiments, we collected 5×10^4 voice data from 1000 speakers. First, we demonstrate the noise cancellation effect of the DPCRN. Taking a segment of one voice signal as an example, we show the spectrum diagrams of the voice signal in the time-frequency domain before and after noise reduction. Comparing Figure 8a and Figure 8b, we can see that for the regular mechanical noise in the electric industry, the curve of the voice signal becomes smoother after DPCRN processing, and the noise effect is significantly reduced.



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

Figure 8. Comparison of voiceprint images before and after noise reduction. (**a**) Time-domain waveform and spectrogram of a noisy voice. (**b**) Time-domain waveform and spectrogram of a noise-reduced voice.

After the noise reduction completed, we performed identity matching based on the voiceprint information. First, we created a data list labeled by the speaker index. Then, we processed the voice signals by removing the voice segments with a mute length greater than 1.5 s through the endpoint detection, making voice enhancement for the rest of the data, and converting the voice signals to the amplitude spectrum by an STFT. Finally, the voice training set and the validation set were input into the EDRSN for network training, and the model was saved at the end of each training round until the end of training. We set the batch size to 20 and the number of epochs to 50, and the detailed parameters of the EDRSN are shown in Table 3.

Table 3. Simulation Parameters.

Parameter	Value
Input Shape	(160, 64, 1)
Size of batch	20
No. of epochs	50
Learning rate	0.001
Optimization function	Adam
Loss function	Triplet

Moreover, we used the triplet loss [28] as the loss function for training different speakers' voiceprints. The triplet loss has three inputs, a segment of a particular speaker's voice, a segment of that speaker's voice, and a segment of another speaker's voice. We represented these three inputs as an anchor, a positive example, and a negative example. The triplet loss calculates the intersample similarity by continuously optimizing the distance between the anchor and the positive example so that it is smaller than the distance between the anchor and the negative example. In the voiceprint recognition, we used the cosine similarity, as in Equation (13), between the examples to represent the intersample distance. The triplet loss is shown as $L = \max\left(\cos_{ap} - \cos_{an} + \alpha, 0\right)$, where cos_{ap} is the cosine similarity of the anchor and the positive example, and cos_{an} is the cosine similarity of the anchor and the negative example. To prevent the model from training the distance from the anchor point to the positive and negative examples to be very similar, we set a minimum bound α between the similarities. Thus, we could correctly distinguish the positive and negative examples of voice signals and prevent the case where $cos_{ap} = cos_{an}$.

For N triplets, the loss function can be expressed as

$$Loss = \sum_{i=1}^{N} \max\left(\cos_{ap,i} - \cos_{an,i} + \alpha, 0\right), \tag{19}$$

where $cos_{ap,i}$ is the distance between the anchor and the positive example in the ith triplet, and $cos_{an,i}$ is the distance between the anchor and the negative example in the ith triplet.

The number of model parameters of the CNN, Res-CNN [29], and the proposed EDRSN are shown in Table 4. It can be seen that the proposed EDRSN used the least number of parameters. The CBAM module in the EDRSN was a lightweight attention module and only slightly increased the number of model parameters.

In addition, we give the loss, accuracy, and F-measure performance comparison of the above models in Table 5. The experiment results showed that the proposed EDRSN scheme enhanced the extraction of vocal features thanks to the use of the CBAM and HDC. Meanwhile, the utilization of the soft thresholding reduced the noise interference on the voice signals. As seen from Table 5, the EDRSN scheme achieved the minimum training loss during the model training process. Its training accuracy exceeded 96%, which was much better than the Res-CNN and CNN schemes and reflected its obvious advantages in voiceprint recognition. Meanwhile, the size of the F-measure indicated the degree of model fit, and we found that the EDRSN model with a CBAM also achieved the best model fit.

Table 4. Comparison of model parameters.

Models	Parameters	Trainable Parameters	Nontrainable Parameters
CNN	4.936 M	4.926 M	0.010 M
Res-CNN	4.936 M	4.926 M	0.010 M
EDRSN	3.804 M	3.800 M	0.004 M
EDRSN + CBAM	3.849 M	3.845 M	0.004 M

Table 5. Comparison of model performance.

Models	Loss	Accuracy	F-Measure
CNN	0.7629	86.54%	0.7124
Res-CNN	0.3507	92.38%	0.7799
EDRSN	0.2146	94.83%	0.8235
EDRSN + CBAM	0.1785	96.02%	0.8462

The loss curves of the EDRSN and Res-CNN models are shown in Figure 9. The results are from at most 10,000 iterations. From the above two figures, it can be seen that the loss value of the EDRSN with 4000 iterations was similar to that of the Res-CNN with 8000 iterations, which indicated that the EDRSN model converged faster than the Res-CNN model. Moreover, the loss value of the Res-CNN model became 1.21 after 10,000 iterations, while it was 0.94 for the EDRSN model, which proved that the proposed EDRSN model was more accurate.

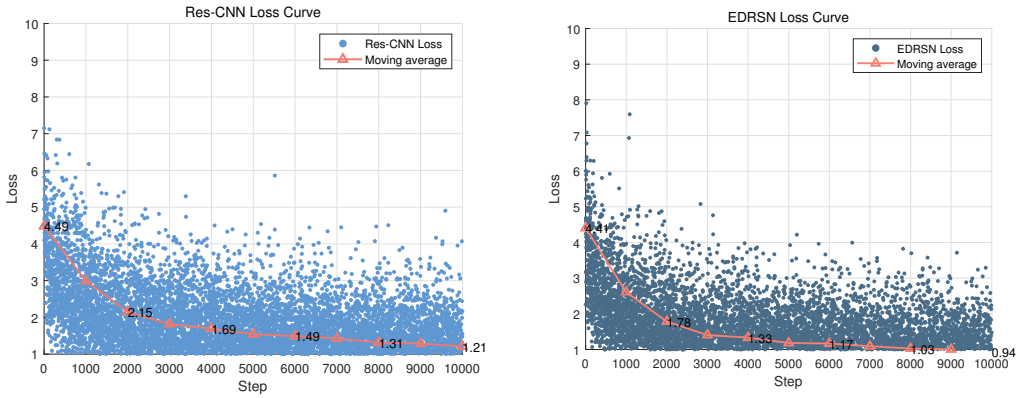


Figure 9. The loss curves of the Res-CNN and EDRSN models.

In Figure 10, we investigated the accuracy of three different models in voiceprint recognition with the collected voice dataset. It is clearly seen that the EDRSN exhibits a better accuracy than the Res-CNN. Meanwhile, the network converged faster and improved the recognition accuracy to some extent due to the introduction of the CBAM in the EDRSN, which also proves the superiority of the EDRSN and CBAM in voiceprint recognition applications.

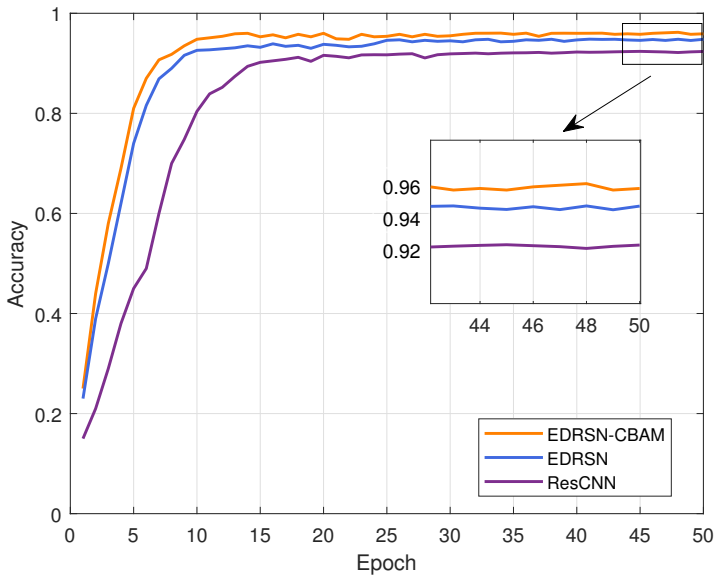


Figure 10. The accuracy curves of EDRSN-CBAM, EDRSN and Res-CNN models.

7. Conclusions

Voiceprint recognition is gradually being applied to daily life due to its unique advantages. Nowadays, researchers use neural networks such as CNNs, DNNs and ResNet for voiceprint recognition. However, due to the special nature of the electric industry and the complex noise in its working space, research on voiceprint recognition for the electric industry has not yet been carried out. Therefore, it is meaningful and necessary to study a high-accuracy recognition scheme for the noisy electric environment.

In this paper, we presented the process of voiceprint recognition. Firstly, considering the noisy environment in the electric industry, we utilized the DPCRN to model the harmonic structure of the voice signals for the problem of noise-induced recognition-accuracy degradation. Secondly, we use traditional pre-emphasis, framing and windowing, and endpoint detection steps to preprocess the voice signals. Finally, we improved the network structure based on a DRSN and proposed an EDRSN-based voiceprint recognition scheme. By further combining CBAM and HDC, our proposed EDRSN scheme achieved better performance in terms of noise reduction and feature extraction. Simulation results showed that our proposed EDRSN scheme could reduce the number of model parameters and achieve a recognition accuracy of 96.02% , which was much higher than other schemes.

The training process of the proposed EDRSN and other networks is time-consuming. It is necessary to study a low-complexity and lightened model in the future, while ensuring a similar recognition accuracy. Moreover, to further improve the feature extraction capability of the model, we plan to optimize the network structure by introducing other new attention mechanisms and loss functions for better performance.

Author Contributions: Conceptualization, Q.Z. (Qingrui Zhang) and W.Q.; methodology, Y.M.; validation, B.H. and Q.Z. (Qi Zhai); formal analysis, Y.M. and B.H.; investigation, L.S.; resources, Q.Z. (Qingrui Zhang) and Y.Z.; data collation, H.Z.; writing, Y.M. and B.H.; writing—review and editing, Z.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a science and technology project of State Grid Corporation of China (Research on Dispatching Fusion Communication Oriented to Power Communication Network and Its Cooperative Control with Power Network Operation, 520627220008).

Data Availability Statement: The data are not publicly available due to privacy restrictions of State Grid Shandong Electric Power Company.

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

References

1. Huang, Z.; Zhang, X.; Wang, L.; Li, Z. Study and implementation of voiceprint identity authentication for Android mobile terminal. In Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 14–16 October 2017; pp. 1–5. [CrossRef]
2. Jin, X.; Zhang, Y.; Wang, X. Strategy and coordinated development of strong and smart grid. In Proceedings of the IEEE PES Innovative Smart Grid Technologies, Tianjin, China, 21–24 May 2012; pp. 1–4. [CrossRef]
3. Le, X.; Chen, H.; Chen, K.; Lu, J. DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. *arXiv* **2021**, arXiv:2107.05429.
4. 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]
5. Lin, N.; Chen, G.; Zhou, Q.; Liu, C. Dilated Residual Shrinkage Network for SAR Image Despeckling. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; pp. 503–507. [CrossRef]
6. Yang, J.; Jiang, J. Dilated-CBAM: An Efficient Attention Network with Dilated Convolution. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 15–17 October 2021; pp. 11–15. [CrossRef]
7. 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]
8. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
9. Juang, B.; Levinson, S.; Sondhi, M. Maximum likelihood estimation for multivariate mixture observations of markov chains (Corresp). *IEEE Trans. Inf. Theory* **1986**, *32*, 307–309. [CrossRef]

10. Kenny, P.; Boulianne, G.; Ouellet, P.; Dumouchel, P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1435–1447. [CrossRef]
11. Jagtap, S.S.; Bhalke, D.G. Speaker verification using gaussian mixture model. In Proceedings of the 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 8–10 January 2015; pp. 1–5. [CrossRef]
12. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056. [CrossRef]
13. Hughes, T.; Mierle, K. Recurrent neural networks for voice activity detection. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7378–7382. [CrossRef]
14. Snyder, D.; Ghahremani, P.; Povey, D.; Garcia-Romero, D.; Carmiel, Y.; Khudanpur, S. Deep neural network-based speaker embeddings for end-to-end speaker verification. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 165–170. [CrossRef]
15. Nathwani, K.; Vincent, E.; Illina, I. DNN Uncertainty Propagation Using GMM-Derived Uncertainty Features for Noise Robust ASR. *IEEE Signal Process. Lett.* **2018**, *25*, 338–342. [CrossRef]
16. Yuan, W.; Dong, B.; Wang, S.; Unoki, M.; Wang, W. Evolving multi-resolution pooling CNN for monaural singing voice separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 807–822. [CrossRef]
17. Berdibaeva, G.K.; Bodin, O.N.; Kozlov, V.V.; Nefed'ev, D.I.; Ozhikenov, K.A.; Pizhonkov, Y.A. Pre-processing voice signals for voice recognition systems. In Proceedings of the 2017 18th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM), Erlagol, Russia, 29 June–3 July 2017; pp. 242–245. [CrossRef]
18. Üstün, B.; Avci, K. A new hybrid window based on cosh and hamming windows for nonrecursive digital filter design. In Proceedings of the 2015 23nd Signal Processing and Communications Applications Conference (SIU), Malatya, Turkey, 16–19 May 2015; pp. 2282–2285. [CrossRef]
19. Serbes, A. Fast and efficient sinusoidal frequency estimation by using the DFT coefficients. *IEEE Trans. Commun.* **2019**, *67*, 2333–2342. [CrossRef]
20. Wang, X.; Ying, T.; Tian, W. Spectrum representation based on STFT. In Proceedings of the 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Chengdu, China, 17–19 October 2020; pp. 435–438. [CrossRef]
21. Wu, B.; Wu, H. Scalable similarity-consistent deep metric learning for face recognition. *IEEE Access* **2019**, *7*, 104759–104768. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
23. Li, C.; Ma, X.; Jiang, B.; Li, X.; Zhang, X.; Liu, X.; Cao, Y.; Kannan, A.; Zhu, Z. Deep speaker: An end-to-end neural speaker embedding system. *arXiv* **2017**, arXiv:1705.02304.
24. Zhang, K.; Sun, M.; Han, T.X.; Yuan, X.; Guo, L.; Liu, T. Residual networks of residual networks: Multilevel residual networks. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 1303–1314. [CrossRef]
25. Zhao, M.; Zhong, S.; Fu, X.; Tang, B.; Pecht, M. Deep Residual Shrinkage Networks for Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2020**, *16*, 4681–4690. [CrossRef]
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
28. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [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]