electronics

MDPI

Article

An Investigation of ECAPA-TDNN Audio Type Recognition Method Based on Mel Acoustic Spectrograms

Jian Wang [1,2,3], Zhongzheng Wang [4], Xingcheng Han [1,3] and Yan Han [1,2,3,*]

[1] Shanxi Key Laboratory of Signal Capturing & Processing, North University of China, Taiyuan 030051, China
[2] The State Key Laboratory for Electronic Testing Technology, North University of China, Taiyuan 030051, China
[3] School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
[4] Xuzhou Communication Workshop of Shanghai Communication Section, China Railway Shanghai Bureau Group Co., Ltd., Xuzhou 221000, China
* Correspondence: hanyan@nuc.edu.cn

Abstract: Audio signals play a crucial role in our perception of our surroundings. People rely on sound to assess motion, distance, direction, and environmental conditions, aiding in danger avoidance and decision making. However, in real-world environments, during the acquisition and transmission of audio signals, we often encounter various types of noises that interfere with the intended signals. As a result, the essential features of audio signals become significantly obscured. Under the interference of strong noise, identifying noise segments or sound segments, and distinguishing audio types becomes pivotal for detecting specific events and sound patterns or isolating abnormal sounds. This study analyzes the characteristics of Mel's acoustic spectrogram, explores the application of the deep learning ECAPA-TDNN method for audio type recognition, and substantiates its effectiveness through experiments. Ultimately, the experimental results demonstrate that the deep learning ECAPA-TDNN method for audio type recognition, utilizing Mel's acoustic spectrogram as features, achieves a notably high recognition accuracy.

Keywords: Mel's acoustic spectrogram; deep learning; ECAPA-TDNN; audio type recognition

Citation: Wang, J.; Wang, Z.; Han, X.; Han, Y. An Investigation of ECAPA-TDNN Audio Type Recognition Method Based on Mel Acoustic Spectrograms. *Electronics* **2023**, *12*, 4421. https://doi.org/10.3390/electronics12214421

Academic Editors: Gerardo Di Martino and Manohar Das

Received: 6 September 2023
Revised: 28 September 2023
Accepted: 23 October 2023
Published: 27 October 2023

1. Introduction

Sound serves as a crucial pathway for conveying information, allowing humans to comprehend the conditions and changes in their surroundings through auditory cues. All audible sounds that humans can perceive are collectively referred to as audio. In practical environments, during the acquisition and transmission of audio signals, we often encounter a multitude of diverse noise sources. Strong background noise gets intertwined with the intended audio signals, resulting in a significant masking of the inherent features of the target audio signals. Amidst strong noise interference, the recognition of noise segments enables the implementation of audio enhancement procedures, such as noise reduction and echo elimination, thus enhancing the quality and audibility of the audio. When analyzing environmental sounds, identifying relevant sound segments contributes to detecting specific events; sound patterns; or isolating abnormal sounds like sirens, passing vehicles, and vocal conversations. Concurrently, the recognition of noise segments allows for a precise analysis of the attributes of environmental noise, thereby enhancing safety measures.

Audio type recognition poses a significant challenge within the realm of pattern recognition. The early 1990s saw the initiation of research into methodologies for audio type recognition. Notably, in 1994, B. Feiten and S. Gunzel employed a technique based on Self-Organizing Neural Nets to automatically identify auditory features with similar acoustic qualities [1]. As computational power has advanced and the volume of audio feature data has expanded significantly, recognition models rooted in machine learning and deep learning have

become indispensable for audio signal recognition [2]. These models encompass a range of approaches, including convolutional neural networks (CNNs) [3–6], recurrent neural networks (RNNs) [7], convolutional recurrent neural networks (CRNNs), randomized learning [8], deep convolutional neural networks (DCNNs) [9], support vector machines (SVMs) [4], Gaussian mixture models (GMMs), deep attention networks [10], transfer learning [9], and ensemble learning [11], among others. These methods can be applied independently or in combination to enhance the performance of audio category recognition.

Enhancing the accuracy of audio type recognition hinges on two pivotal considerations: firstly, selecting the optimal feature or feature combination that captures the fundamental characteristics of the audio signal; and secondly, choosing the appropriate method or model for recognizing audio signal types [12,13]. The Mel acoustic spectrogram aligns with the perceptual attributes of the human ear, enabling a more effective capture of crucial audio signal information. It finds applicability across various audio processing tasks, contributing to heightened efficiency and performance in feature representation. Our study proposes the adoption of the Mel acoustic spectrogram to characterize audio signals.

Neural networks operate akin to the human brain, yet their initial performance lagged behind that of traditional machine learning models during the same era. Given that a majority of neural networks encountered challenges in effectively handling the dynamic attributes inherent in audio signals and that recognizing phonemes necessitated the incorporation of contextual information, the Time Delay Neural Network (TDNN) was introduced as a solution by Hinton et al. in 1989 [14]. The TDNN boasts two remarkable attributes: its capacity to dynamically adapt to temporal changes in features and its minimal parameter count [15,16]. There is only one hidden layer node linked to each input in traditional deep neural networks. In contrast, the features of the hidden layer are collaboratively influenced by both the present-moment inputs and future-moment inputs in the modified TDNN. This approach effectively leverages the temporal context information within audio signals by processing multiple consecutive frames of audio input. The Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) is a neural network model designed for speech recognition tasks, introduced in the year 2020 [17].This model amalgamates the conventional TDNN architecture with attention mechanisms, emphasizing channel attention, propagation, and aggregation in TDNN-based speaker verification [18–21]. Furthermore, it enhances feature extraction and representation capabilities by incorporating extended context aggregation and introducing expansion layers. To discern the type of both target audio and noise, our research employs an audio type recognition method founded on the deep learning model ECAPA-TDNN. The experimental data were sourced from the THCHS-30 dataset of Tsinghua University [22], comprising speech signals serving as event signals. Additionally, the dataset includes noise signals encompassing 12 distinct types from the NoiseX-92 dataset [23].

2. Mel Sound Spectrogram

The time-domain waveform represents the most straightforward and readily obtainable feature for recognizing audio signal types. However, due to its susceptibility to influences and inherent instability, time-domain information tends to yield suboptimal results as a recognition feature. Conversely, the frequency-domain information of audio signals offers greater accuracy in capturing audio characteristics and is less prone to interference. Currently, the conversion of an audio signal's time domain information into frequency-domain information can be achieved through methods like Fourier transform or wavelet transform. However, these approaches often lead to the loss of certain signal features. The time–frequency characterization of a signal encompasses both time-domain and frequency-domain information, endowing it with heightened identification capabilities. An acoustic spectrogram that is built on spectral analysis with the incorporation of the time dimension offers a more intuitive depiction of signal changes. Essentially, it embodies a time–frequency characterization of the audio signal [24].

The spectrum portrays signal distribution across various frequencies. However, the human auditory system discriminates between frequencies with varying sensitivity. Research reveals that the frequency resolution of the human ear is not linear but logarithmic. This means that two pairs of frequencies situated at equal distances in the frequency domain might not be perceived equally by the human ear [25]. This issue finds an effective resolution through the introduction of Mel frequency. Mel frequency characterizes the human ear's sensitivity to audio signal frequencies [26]. The logarithmic relationship between linear frequency and Mel frequency is defined by Equation (1) [27].

$$F_{mel} = 2595\lg(1 + f/700) \quad (1)$$

where F_{mel} is the perceived frequency in Mel, and f the actual frequency in Hz.

Figure 1 illustrates a schematic representation of the relationship between Mel frequency and actual frequency. Notably, as the frequency decreases, the Mel frequency exhibits a more rapid alteration concerning linear frequency, resulting in a steeper curve slope. Conversely, at higher frequencies, the Mel frequency experiences a gentler ascent, leading to a smaller curve slope. This phenomenon underscores the concept that high-frequency sounds are less distinguishable to the human ear, while low-frequency sounds are more easily discerned. This variation in perceptual sensitivity by the human ear is distinctly portrayed.

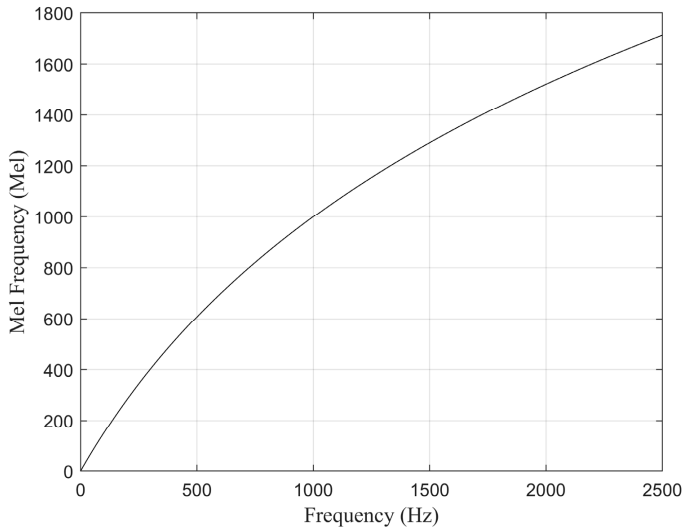


Figure 1. Schematic diagram of the relationship between Mel frequency and linear frequency.

Building upon the investigation into Mel frequency, the Mel filter is introduced to simulate the phenomenon where higher frequencies are perceived less distinctly by the human ear, exhibiting a more gradual auditory response. This involves constructing numerous triangular filters, with a greater emphasis on low-frequency filters and fewer high-frequency filters, forming a filter bank aligned with their frequency distribution. The frequency response characteristic curve of the Mel filter bank is depicted in Figure 2.

By subjecting the spectrogram to Mel-scaled filtering through a bank of Mel filters, the transformation to Mel spectrogram for the audio signal is achieved. Similar to the spectrogram, the Mel spectrogram is also a representation in the time–frequency domain. Figure 3 displays the time-domain waveform, spectrogram, and Mel spectrogram corresponding to a segment of noise signal.

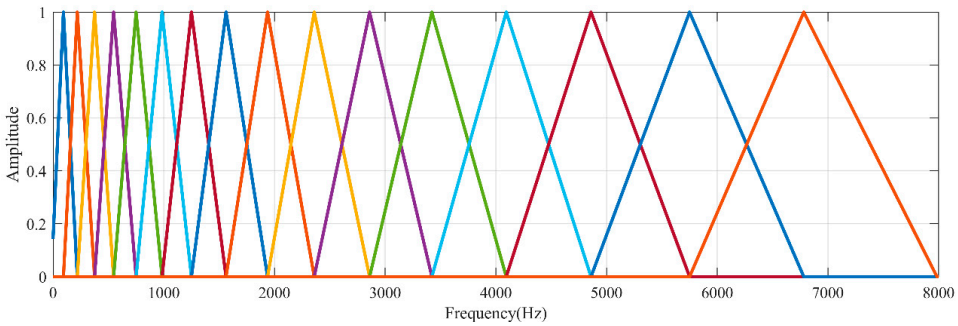
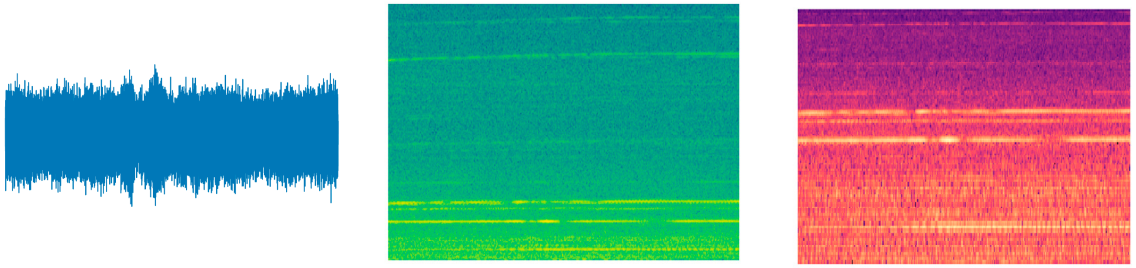


Figure 2. Mel filter bank frequency response characteristic curve, where lines of different colors represent filters of different frequencies.



(**a**) Time-domain waveform. (**b**) Spectrogram. (**c**) Mel spectrogram.

Figure 3. Three representation methods of noise signa.

The Mel acoustic spectrogram employs the Mel scale on the frequency axis, with the Mel filter bank designed to align with the human ear's sound perception characteristics. This design facilitates the mapping of denser frequency regions to sparser ones, effectively reducing unnecessary redundancy within the spectrogram. By remapping frequency axis information, the spectrogram's dimensionality is reduced. This not only lessens the computational complexity of the features but also accelerates the training and inference processes of the model. Ultimately, the Mel sound spectral feature parameters are chosen for audio event detection and noise type identification, effectively accommodating various signal types.

Figure 4 presents the Mel sound spectral characterizations for the 12 types of noise sourced from the NoiseX-92 dataset. The first and second plots depict Mel acoustic spectrograms of white noise and pink noise, respectively. In these plots, time is represented on the horizontal axis, frequency on the vertical axis, and the plot color corresponds to signal amplitude. A comparison across the plots highlights significant disparities in the Mel acoustic spectrograms across different scenarios. For instance, there is minimal high-frequency information in the Volvo vehicle noise scenario, while the f16 fighter noise scenario prominently features a higher high-frequency component. Additionally, the Mel acoustic spectrograms of the f16 fighter noise scenario exhibit distinct horizontal stripes, whereas the factory1 and factory2 scenarios display predominant vertical stripes. The irregular "speckling" observed in the factory2 scene arises from the recording's context in an automobile production plant, where abrupt acoustic events like knocks are common.

Figure 5 shows the Mel acoustic spectrograms of three distinct audio event signals: speech, alarm, and explosion. Evidently, the variations in characteristics among different audio events are pronounced, enabling clear differentiation between several audio events.

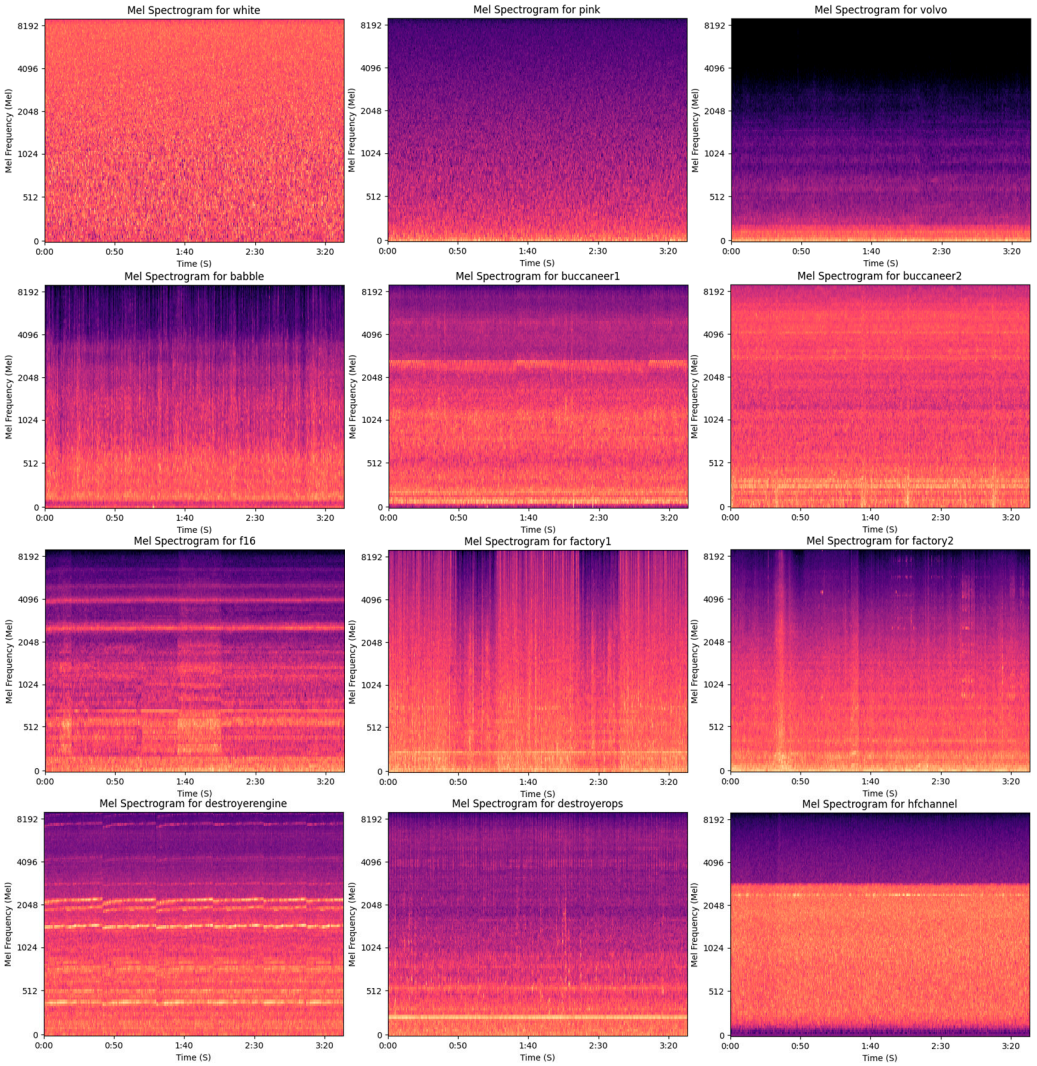
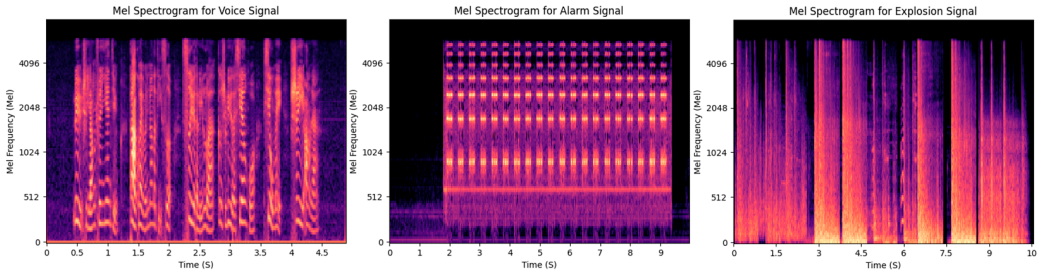


Figure 4. Mel spectrogram of 12 types of noise.



(**a**) Voice signal. (**b**) Alarm signal. (**c**) Explosion signal.

Figure 5. Mel spectrograms of three typical audio events.

With deep learning models demonstrating their prowess across diverse domains, many challenges that elude effective resolution via traditional machine learning find improved outcomes through deep neural networks. This effect is particularly prominent in image recognition [28]. The Mel spectrogram encapsulates fundamental audio features, leveraging the neural networks' proficiency in image processing; inputting Mel spectrograms into deep neural network models enables the recognition of audio types.

3. ECAPA-TDNN Deep Learning Model

The ECAPA-TDNN model was introduced by Desplanques et al. at the University of Goethe, Belgium, in 2020. Drawing on the latest advancements in computer vision-related fields, the ECAPA-TDNN model brings forth several enhancements to the TDNN model. This model places heightened emphasis on inter-channel attention and multilayer feature aggregation [17,18].

Figure 6 illustrates the structure of the ECAPA-TDNN model, comprising key components such as TDNN+ReLU+BN, SE-Res2Block, Attentive Statistics Pooling (ASP), and Multilayer Feature Aggregation (MFA). The model employs the AAM-Softmax loss function for optimization.

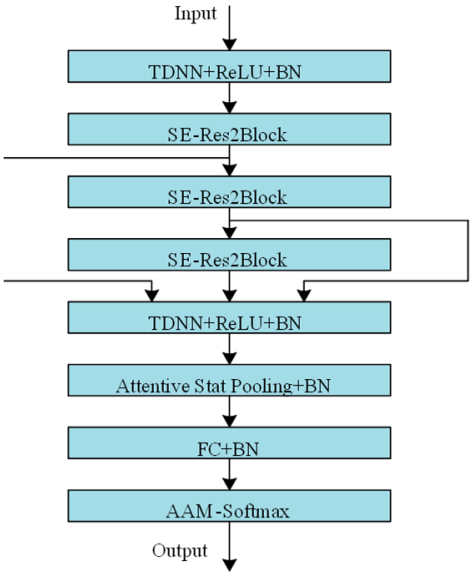


Figure 6. ECAPA-TDNN model network structure.

3.1. SE-Res2Block

The ECAPA-TDNN model includes a section composed of multiple SE-Res2Block modules linked sequentially. The core constituents of this module encompass TDNN, SE-Net, and Res2Net components. Figure 7 presents an illustrative representation of the network structure of the SE-Res2Block modules.

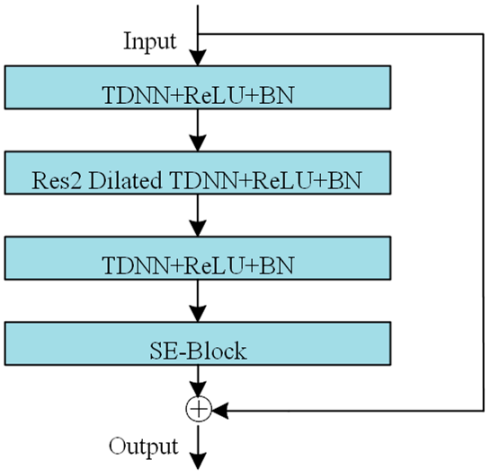


Figure 7. Schematic diagram of SE-Res2Block structure.

In recent years, to enhance the effectiveness of deep learning models, researchers have introduced concepts like inception structures and attention mechanisms. These innovations focus on optimizing the spatial dimensions of input feature maps. By aggregating features from various receptive fields and adeptly capturing both global and local connections, these enhancements enhance the overall performance of deep learning models. The distinctive aspect of SE-Net lies in its approach of modeling the channel dimensions of input feature maps. This enables a recalibration of the feature maps, thereby boosting the model's performance [29]. Figure 8 provides an illustrative depiction of the SE-Net structure.

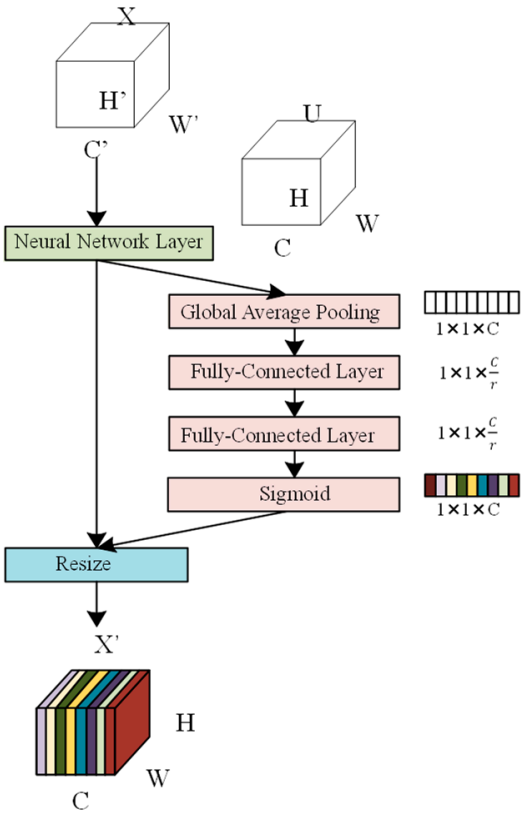


Figure 8. Schematic diagram of SE-Net structure.

The input X of the SE-Net possesses a dimension of $H' \times W' \times C'$, which can be processed and translated into a feature map U with the dimensions $H \times W \times C$. Subsequently, the feature map U undergoes additional compression through global average pooling, resulting in a $1 \times 1 \times C$ channel vector that encapsulates global information for each channel. The subsequent step involves excitation, entailing two fully connected layer operations performed on the channel vectors. The initial fully connected layer executes dimensionality reduction, which curtails parameters to lower computational complexity. The subsequent fully connected layer conducts dimensionality enhancement, aiming to restore the dimensions of channel numbers and weight vectors. Using two fully connected layers often reveals channel correlations more effectively than a single layer, offering augmented nonlinear capabilities, while trimming parameters to bolster computational efficiency. Employing the Sigmoid activation function, the output from the fully connected layers is employed to compute weights for each channel feature. Consequently, the original features can be adjusted based on these weights, generating a feature map, X', that more accurately captures type-specific characteristics.

Diverging from preceding network architectures that rely on features of varying resolutions to enhance multiscale capabilities, Res2Net captures features across distinct receptive fields and scales, thus acquiring a comprehensive blend of global and local

information [30,31]. Res2Net is an advancement built upon the foundation of the Bottleneck structure by segmenting the 3 × 3 convolutional layers within each residual block into multiple sub-branches. This is then followed by the fusion of features through residual connections. The Bottleneck structure is depicted in Figure 9a, comprising sequentially connected 1 × 1, 3 × 3, and 1 × 1 convolutions that are integrated through residual connections. In Figure 9b, the Res2Net structure for s = 4 is illustrated. The distinct differentiation between Res2Net and Bottleneck lies in the approach of Res2Net, which splits the feature map subsequent to the 1 × 1 convolution into s sub-branches. Notably, x_1 serves directly as the output of y_1 without any modifications. Meanwhile, x_2 undergoes a 3 × 3 convolution; a portion becomes y_2's output, while the remaining part connects to x_3. In a cascading manner, x_3 and x_4 execute the same sequence of operations.

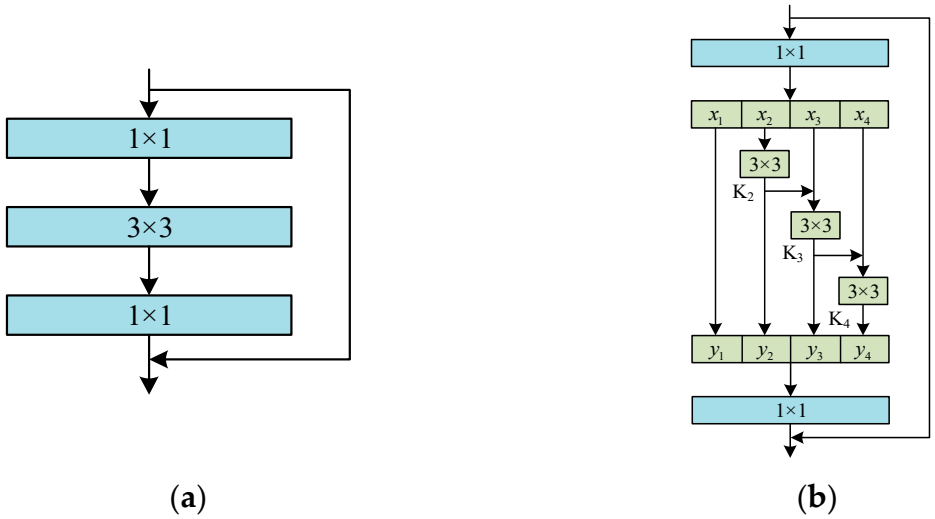


Figure 9. Schematic diagram of two network structures. (**a**) Schematic diagram of Bottleneck structure. (**b**) Schematic diagram of Res2Net structure with s = 4.

3.2. Attention Statistical Pooling

To address the limitation of the average pooling layer, which is susceptible to information loss, attention statistic pooling was introduced. This approach simultaneously considers disparities in information across both time and channel dimensions. Thus, the network can allocate attention to crucial details across various time intervals and feature maps. The implementation of attention statistic pooling is realized through Equation (2).

$$e_{tc} = v_c^T f(Wh_t + b) + k_c \tag{2}$$

In Equation (2), h_t represents the activation value of the preceding layer's network at time step t. Following the weight matrix, $W \in W^{R \times C}$, and bias transformation, $b \in R^{R \times 1}$, the dimensionality of h_t is reduced from C channels to R channels. This streamlines the parameter configuration and mitigates the potential for overfitting.

Moreover, v_c^T signifies that the R dimensional vector derived from the activation function, $f(\cdot)$, undergoes linear transformation and projection, resulting in a C dimensional spatial representation. To calculate the attention weight of time step t on channel c, the Softmax transformation is applied to e_{tc}. Equation (3) represents the formula for this computation.

$$\alpha_{tc} = \frac{e^{e_{tc}}}{\sum_{\tau} e^{e_{\tau c}}} \tag{3}$$

Equations (4) and (5) yield the weighted mean vector, $\mu_c^{\sim}$, and the weighted standard deviation vector, $\sigma_c^{\sim 2}$, on channel c, respectively. These vectors are then concatenated to yield the ultimate output of the attention statistic pooling layer.

$$\mu_c^{\sim} = \sum_t^T \alpha_{tc} h_{tc} \tag{4}$$

$$\sigma_c^{\sim 2} = \sum_t^T \alpha_{tc} h_{tc}^2 - \mu_c^{\sim 2} \tag{5}$$

3.3. Multilayer Feature Aggregation

Compared to earlier TDNN-based systems, ECAPA-TDNN stands out because of its pioneering approach to multilayer feature aggregation. It not only incorporates features solely from the last frame layer but also integrates those from the other two layers. This is achieved by amalgamating the features produced by the first, second, and final SE-Res2Block modules in the channel dimension, facilitated by residual connections. Subsequently, the deep features are further extracted through a fully connected layer, and these features are then utilized in the computation of attention statistics pooling. The specific progression is elucidated in Figure 6.

In the realm of deep learning, a variety of feature types and sources exist that necessitate integration to enhance the model's effectiveness and generalization capability. Among the commonly employed methods for integration are merge operations and element-level summation. Figure 6 illustrates a multilayer feature aggregation process employing the merge operation to integrate features across different levels.

3.4. AAM-Softmax Loss Function

Refining the choice of a more effective loss function remains a persistent challenge in the realm of deep learning. Particularly for classification tasks, the selected loss function must strike a delicate balance between maximizing the distance between different classes while minimizing the distance within the same class. While neural network classification models often adopt the Softmax loss function, this approach overlooks the absence of information regarding the angular relationships between classes within the feature space, consequently yielding suboptimal results. To address this problem, researchers have introduced a fixed angular interval as a penalty term into the Softmax loss function, proposing the Additive Angular Margin Loss Softmax (AAM-Softmax). This approach effectively narrows intra-class gaps and enlarges inter-class distances [32,33]. The precise formulation of AAM-Softmax is presented in Equation (6).

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cdot(\cos(\theta_{yi+m}))}}{e^{s\cdot(\cos(\theta_{yi+m}))} + \sum\limits_{j=1, j\neq yi}^{n} e^{s\cdot\cos\theta_j}} \tag{6}$$

In Equation (6), N represents the total number of samples, n denotes the number of classes, y_i is the i sample, θ_{yi} stands for the angle between the samples y_i and the corresponding weight vectors of j class, s represents the scaling factor, and m refers to the edge angle. The edge angle serves the purpose of fostering more closely knit samples within the same class while simultaneously enhancing the disparities between different classes. This serves to enhance the efficacy of classification or recognition.

ECAPA-TDNN entails a slightly elevated computational load in comparison to alternative neural network models. The training and inference processes of ECAPA-TDNN demand a greater allocation of computational resources and time as opposed to conventional DNN and LSTM (Long Short-Term Memory Network) models. Nevertheless, the computational overhead of ECAPA-TDNN remains relatively modest when contrasted with certain more recent speech recognition models, such as Transformers.

4. Experimental Paradigm

The experimental simulations were conducted on the Shenzhou laptop platform, featuring the following specific configuration: CPU—Intel Core i5-8400; graphics card—NVIDIA GTX 1060 6 GB; operating system—Windows 10; simulation software—PyCharm Community Edition; development language—Python 3.9; and deep learning framework—PyTorch 1.10.

In the experimental simulation, the noise signals were chosen from the 12 available types in the NoiseX-92 dataset. Simultaneously, due to the scarcity of audio event-related datasets and the representative nature of speech signals in the audio domain, audio event signals were sourced from the speech signals within Tsinghua University's THCHS-30 dataset. These signals were employed for speaker recognition testing. The data were standardized into 16 kHz mono audio files, with sample data randomly allocated for training and testing sets. The frame length was set at 20 ms (320 samples), while the frame shift was established at 10 ms (160 samples) during frame segmentation.

For the speaker recognition test in the experiment, speech data from nine individuals in the THCHS-30 dataset were randomly chosen and designated as participants A to I. These individuals' speech recordings were utilized for both training and testing. The frame segmentation was performed with a frame length of 20 ms (320 samples) and a frame shift of 10 ms (160 samples).

The effectiveness of noise signal type recognition is assessed using the accuracy metric R. Prior to computing the accuracy rate, the recognized noise signals need to be manually labeled with their corresponding frame noise signal types. The accuracy rate, R, for each noise type recognition can be computed using Equation (7).

$$R = \frac{N_{1,1}}{N_{\text{frame}}} \times 100\% \tag{7}$$

where $N_{1,1}$ represents the count of frames in which the manually labeled noise type matches the noise type obtained through the deep learning model's recognition, while N_{frame} denotes the total number of frames containing noise signals.

5. Results

5.1. Noise Type Recognition Test

In deep learning, model training involves multiple stages, including hyperparameter optimization, tuning model training efficiency, feature selection, and more. Training process curves provide an intuitive way to understand the model's performance at different iterations. The accuracy and loss function transformation curves during the training process of the ECAPA-TDNN algorithm are shown in Figure 10.

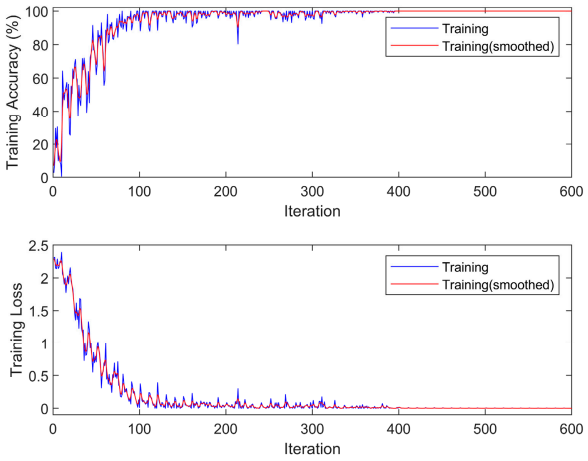


Figure 10. Training curve.

Initially, a noise type test is executed for the 12 categories of noise signals within the NoiseX-92 dataset, leveraging the trained deep learning model to assess its performance. The accuracy of recognizing the 12 noise types from NoiseX-92 is presented in Table 1, compared to advanced classification and recognition methods such as Convolutional Neural Network, Random Forest, and Naive Bayes. While these methods may outperform ECAPA-TDNN in specific application scenarios, nevertheless, ECAPA-TDNN consistently demonstrates significantly superior classification results in the majority of real-world scenarios.

When analyzing the recognition accuracy of ECAPA-TDNN in Table 1, it is evident that the average recognition accuracy achieved by the audio type recognition model proposed in the paper stands at an impressive 98.15% across the 12 noise categories. Among these, the recognition accuracy for the factory1 noise type appears relatively low at 88.49%, while the remaining 11 noise types exhibit recognition accuracy exceeding 95%. In fact, most of the noise types boast recognition accuracy surpassing 99%. Notably, the recognition accuracy for three noise types—f16, hfchannel, and Volvo—attains a perfect 100%. When drawing from the experimental outcomes, it is apparent that the proposed audio type recognition method, based on ECAPA-TDNN, effectively accomplishes precise noise type identification.

Table 1. Accuracy of 12 types of noise recognition in NoiseX-92 (%).

Type of noise	white	pink	Volvo	babble	buccaneer1	buccaneer2
Naive Bayes	74.77	74.02	67.54	67.56	80.30	82.46
Random Forest	93.03	97.69	89.79	98.27	**100.00**	**99.94**
CNN	97.36	97.43	91.71	95.62	94.37	87.75
ECAPA-TDNN	**99.85**	**99.24**	**100.00**	**99.08**	99.55	99.26
Type of noise	factory1	factory2	f16	destroyer engine	destroyerops	hfchannel
Naive Bayes	71.47	93.54	71.09	88.72	78.48	73.19
Random Forest	**90.08**	**95.33**	86.26	97.66	84.41	79.61
CNN	83.36	92.05	74.91	**99.39**	83.12	77.36
ECAPA-TDNN	88.49	95.04	**100.00**	98.78	**98.50**	**100.00**

Precision and recall are interrelated metrics in a way that pursuing one often comes at the expense of the other. Ideally, we would like to maximize both precision and recall, but in practice, they tend to "constrain" each other. In other words, if we aim for high precision, the recall tends to decrease, and vice versa. Precision and recall are generally in conflict, and this conflict is resolved through the introduction of the F1 Score, which serves as a comprehensive measure for balancing the trade-off between precision and recall and evaluating a classifier more holistically. The F1 Score is calculated as the harmonic mean of precision and recall. A higher F1 Score indicates a higher-quality model, as it takes into consideration both false positives and false negatives, offering a well-rounded assessment of the classifier's performance. The F1 Score for ECAPA-TDNN in recognizing 12 types of noise from the NoiseX-92 dataset is presented in Table 2 as a percentage.

Table 2. F1 Score of ECAPA-TDNN of 12 types of noise recognition in NoiseX-92 (%).

Type of noise	white	pink	Volvo	Babble	buccaneer1	buccaneer2
F1 Score	98.88	98.87	99.35	98.86	98.85	98.82
Type of noise	factory1	factory2	f16	destroyer engine	destroyerops	hfchannel
F1 Score	93.13	95.43	98.94	98.51	98.52	99.31

The ESC-50 dataset is a meticulously annotated collection of 2000 environmental sound recordings, specifically curated for the evaluation of algorithms in the field of environmental sound classification [34]. Within this dataset, there are ten distinct semantic classes, which include helicopter, chainsaw, siren, car horn, engine, train, church bells, airplane, fireworks, and handsaw. These classes are categorized under the broader umbrella of exterior/urban noise. The primary purpose of utilizing this dataset is to assess and verify the repeatability

and effectiveness of the ECAPA-TDNN algorithm in the recognition and classification of these sound categories. The accuracy and F1 Score results can be found in Table 3.

Table 3. Accuracy and F1 Score of ECAPA-TDNN of 10 types of noise recognition in ESC-50 (%).

Type of noise	Helicopter	Chainsaw	Siren	Car horn	Engine
Accuracy	93.21	92.99	97.55	93.98	94.87
F1 Score	71	71.96	89.69	70.86	76.76
Type of noise	Train	Church bells	Airplane	Fireworks	Handsaw
Accuracy	91.92	98.39	94.31	98.64	97.07
F1 Score	68.75	93.9	72.73	94.79	87.88

5.2. Speaker Recognition Test

Speaker recognition experiments are executed in interference-free conditions, utilizing the deep learning model trained earlier to evaluate its competence in recognizing the intended audio type. The speaker recognition accuracy is depicted in Table 4 for when the identified signal comprises purely speech,

Table 4. Speaker recognition accuracy without noise interference (%).

Speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Accuracy	95.62	94.49	95.83	96.43	92.86	97.02	95.23	94.48	96.54

When analyzing the recognition accuracy outcomes displayed in Table 4, it becomes evident that, in comparison to the more elevated accuracy achieved in noise type recognition, the speaker recognition accuracy experiences a marginal reduction. Nonetheless, in the absence of noise interference, the average accuracy of speaker recognition still reaches a commendable 95.39%. Speaker E's recognition accuracy is relatively lower at 92.86%, while the remaining speakers exhibit recognition accuracy hovering around 95%. Remarkably, speakers D, F, and I attain a recognition accuracy surpassing 96%.

In practical application scenarios, the identification of target audio segment types often collaborates with audio endpoint detection methods. In the presence of noisy signals, the audio endpoint detection method is utilized to pinpoint the commencement and conclusion of the target audio segment, facilitating the extraction of relatively unpolluted audio signals. Subsequent identification endeavors are then conducted on these purified audio segments. However, owing to the attributes of noise-infused frequency signals and the performance limitations inherent in endpoint detection methods, there is a possibility that noise components persist within the target audio segments, even after undergoing noise reduction via multi-window spectral subtraction. In such operational contexts, it becomes imperative for the type of recognition model to exhibit a certain degree of resilience against interference. As the benefits of the maritime economy increase, verbal interaction and communication within the ship environment have become more frequent. Nevertheless, the pronounced noise emanating from various ship equipment during operation significantly disrupts the smooth flow of interactions and communication among crew members. In many instances, this formidable noise directly obscures spoken words. To assess this, nine sets of speaker speech signals were amalgamated with destroyer engine noise from the NoiseX-92 dataset, at signal-to-noise ratios (SNRs) of 10 dB and 5 dB. This process engendered noise-laden speech signals for testing purposes. The resultant recognition accuracy is enumerated in Table 5.

Table 5. Speaker recognition accuracy under noise interference (%).

SNR (dB)	A	B	C	D	E	F	G	H	I
5	81.45	81.37	82.16	83.23	79.68	83.97	81.82	80.65	81.15
10	94.36	94.21	94.64	95.66	91.12	96.04	93.61	92.64	93.33

When analyzing the speaker recognition test outcomes under noise interference conditions in Table 5, we observe that a signal-to-noise ratio of 10 dB has relatively minor effects on the model's recognition performance when contrasted with a noiseless environment. The average recognition accuracy across the nine sets of speakers stands at 93.95%. With exceptions being speakers E and H, which demonstrate slightly lower recognition accuracy, the remaining groups exceed 93%. It is noteworthy that the accuracy of groups D and F remains above 95%. In contrast, when considering a signal-to-noise ratio of 5 dB, the recognition accuracy notably decreases compared to the noise-free environment. The average accuracy in this context drops to 81.72%. Within this realm, group E's recognition rate decreases to 79.68%, falling below the 80% threshold. However, the recognition accuracy for the remaining groups still manages to stay above 80%.

Figure 11 illustrates the recognition outcomes of the ECAPA-TDNN-based classification model across nine sets of speech signals, both in the absence and presence of noise. In comparison to the noise-free scenario, the accuracy of speaker recognition diminishes by 1.44% and 13.67% at signal-to-noise ratios of 10 dB and 5 dB, respectively. These results from the speaker recognition test underline the ECAPA-TDNN model's capacity to exhibit certain resistance to interference, enabling more precise type recognition within audio signals containing residual noise.

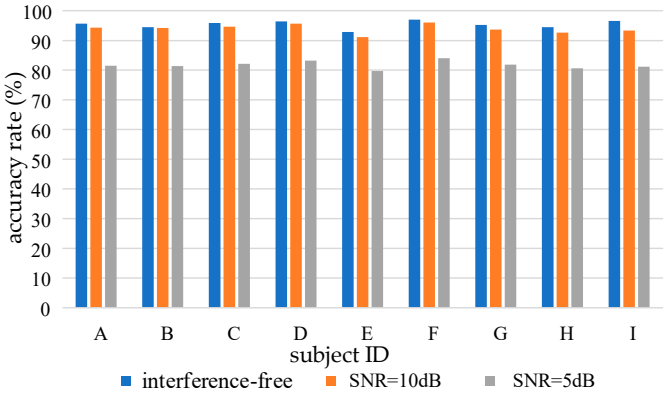


Figure 11. Statistical chart of speaker-recognition test results.

6. Conclusions

Through a comparative analysis of spectrogram and Mel's acoustic spectrogram features in noise and audio signals, the distinctiveness of various noise signal and audio event signal features was established. Based on these findings, Meier's acoustic spectrogram features were employed for recognizing noise and target audio types. This paper introduces a deep learning-based method for audio type recognition. The experimental results demonstrate that this method achieved an impressive 98.15% accuracy in recognizing 12 types of noise signals. In noise-free conditions, the recognition rate for nine groups of speaker voices reached 95.39%. Even when residual noise interference was present, the method maintained an average speaker recognition accuracy of over 80%, highlighting its ability to deliver high recognition accuracy in noisy environments.

Author Contributions: Conceptualization, J.W., Z.W., X.H. and Y.H.; methodology, J.W., Z.W. and Y.H.; software, J.W. and Z.W.; investigation, Z.W., J.W. and Y.H.; writing, J.W. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the National Natural Science Foundation Youth Science Foundation Project, under Grant 62203405; in part by the Higher Education Science and Technology Innovation Project of Shanxi Province, under Grant 2020L0301; in part by the Fundamental Research Program of Shanxi Province, under Grant 20210302124545; and in part by the Youth Science and Technology Research Fund Project of Shanxi Province, under Grant 201901D211250.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

References

1. Feiten, B.; Gunzel, S. Automatic Indexing of a Sound Database Using Self-Organizing Neural Nets. *Comput. Music. J.* **1994**, *18*, 53. [CrossRef]
2. Presannakumar, K.; Mohamed, A. Deep learning based source identification of environmental audio signals using optimized convolutional neural networks. *Appl. Soft Comput.* **2023**, *143*, 110423. [CrossRef]
3. Cheng, K.W.; Chow, H.M.; Li, S.Y.; Tsang, T.W.; Ng, H.L.B.; Hui, C.H.; Lee, Y.H.; Cheng, K.W.; Lee, C.K.; Tsang, S.W. Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks. *Appl. Acoust.* **2023**, *205*, 109254. [CrossRef]
4. Cinyol, F.; Baysal, U.; Köksal, D.; Babaoğlu, E.; Ulaşlı, S.S. Incorporating support vector machine to the classification of respiratory sounds by Convolutional Neural Network. *Biomed. Signal Process. Control.* **2023**, *79*, 104093. [CrossRef]
5. Özseven, T. Investigation of the effectiveness of time-frequency domain images and acoustic features in urban sound classification. *Appl. Acoust.* **2023**, *211*, 109564. [CrossRef]
6. İnik, Ö. CNN hyper-parameter optimization for environmental sound classification. *Appl. Acoust.* **2023**, *202*, 109168. [CrossRef]
7. Yassin, A.I.; Shariff, K.K.M.; Kechik, M.A.; Ali, A.M.; Amin, M.S.M. Acoustic Vehicle Classification Using Mel-Frequency Features with Long Short-Term Memory Neural Networks. *TEM J.* **2023**, *12*, 1490–1496. [CrossRef]
8. Kang, Y.; Lee, J. Randomized learning-based classification of sound quality using spectrogram image and time-series data: A practical perspective. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105867. [CrossRef]
9. Harimi, A.; Ameri, M.A.; Sarkar, S.; Totaro, M.W. Heart sounds classification: Application of a new CyTex inspired method and deep convolutional neural network with transfer learning. *Smart Heal.* **2023**, *29*, 100416. [CrossRef]
10. Dong, S.; Xia, Z.; Pan, X.; Yu, T. Environmental sound classification based on improved compact bilinear attention network. *Digit. Signal Process.* **2023**, *141*, 104170. [CrossRef]
11. Bansal, A.; Garg, N.K. Environmental Sound Classification using Hybrid Ensemble Model. *Procedia Comput. Sci.* **2023**, *218*, 418–428. [CrossRef]
12. Zhang, T.; Liu, Y.; Ren, X. Voice Activity Detection Based on Long-Term Power Spectrum Variability. *J. Front. Comput. Sci. Technol.* **2019**, *13*, 1534–1542.
13. Zhang, T.; Ren, X.; Liu, Y.; Geng, Y. Acoustic Features Extraction of Speech Enhancement Based on Auto-Encoder Feature. *J. Front. Comput. Sci. Technol.* **2019**, *13*, 1341–1350.
14. Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K.J. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 328–339. [CrossRef]
15. Chen, X.H.; Bao, C.C. Phoneme-Unit-Specific Time-Delay Neural Network for Speaker Verification. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1243–1255. [CrossRef]
16. Lang, K.J.; Waibel, A.H.; Hinton, G.E. *A Time Delay Neural Network Architecture for Speech Recognition*; Elisvier: Amsterdam, The Netherlands, 1989.
17. Dawalatabad, N.; Ravanelli, M.; Grondin, F.; Thienpondt, J.; Desplanques, B.; Na, H. ECAPA-TDNN Embeddings for Speaker Diarization. *arXiv* **2021**, arXiv:2104.01466.
18. Sigona, F.; Grimaldi, M. Validation of an ECAPA-TDNN system for Forensic Automatic Speaker Recognition under case work conditions. *arXiv* **2023**, arXiv:2104.01466.
19. Singh, V.P.; Sahidullah, M.; Kinnunen, T. Speaker Verification Across Ages: Investigating Deep Speaker Embedding Sensitivity to Age Mismatch in Enrollment and Test Speech. *arXiv* **2023**, arXiv:2306.07501.
20. Zhao, Z.; Li, Z.; Wang, W.; Zhang, P. PCF: ECAPA-TDNN with Progressive Channel Fusion for Speaker Verification. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
21. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *arXiv* **2020**, arXiv:2005.07143.
22. Wang, D.; Zhang, X. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882.
23. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
24. Johnson, A. An integrated approach for teaching speech spectrogram analysis to engineering students. *J. Acoust. Soc. Am.* **2022**, *152*, 1962–1969. [CrossRef] [PubMed]

25. Chen, Z.; Wu, W.; Xia, S. Voice activity detection algorithm based on Mel cepstrum distance order statistics filter. *J. Univ. Chin. Acad. Sci.* **2014**, *31*, 524–529.
26. Zhang, T.; Feng, G.; Liang, J.; An, T. Acoustic scene classification based on Mel spectrogram decomposition and model merging. *Appl. Acoust.* **2021**, *182*, 108258. [CrossRef]
27. Ancilin, J.; Milton, A. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl. Acoust.* **2021**, *179*, 108046. [CrossRef]
28. Li, G.; Qiao, Y.; Wu, W.; Zheng, Y.; Hong, Y.; Zhou, X. Review of deep learning and its application in computer vision. *Appl. Res. Comput.* **2019**, *12*, 3521–3529.
29. Wei, S.; Qu, Q.; Wu, Y.; Wang, M.; Shi, J. PRI Modulation Recognition Based on Squeeze-and-Excitation Networks. *IEEE Commun. Lett.* **2020**, *24*, 1047–1051. [CrossRef]
30. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
31. Aysa, Z.; Ablimit, M.; Hamdulla, A. Multi-Scale Feature Learning for Language Identification of Overlapped Speech. *Appl. Sci.* **2023**, *13*, 4235. [CrossRef]
32. Deng, J.K.; Guo, J.; Yang, J.; Xue, N.N.; Kotsia, I.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5962–5979. [CrossRef]
33. Zhong, Q.H.; Dai, R.N.; Zhang, H.; Zhu, Y.S.; Zhou, G.F. Text-independent speaker recognition based on adaptive course learning loss and deep residual network. *Eurasip J. Adv. Signal Process.* **2021**, *2021*, 45. [CrossRef]
34. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015. [CrossRef]