M3Net: Efficient Time-Frequency Integration Network with Mirror Attention for Audio Classification on Edge

Xuanming Jiang[1,3], Baoyi An[2,3], Guoshuai Zhao[1,4*], Xueming Qian[1,4]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[2]School of Physical Science and Technology, Lanzhou University, Lanzhou, China
[3]Xi'an Jiyun Technology Co., Ltd., Xi'an, China
[4]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd., Xi'an, China
jiangxm24@stu.xjtu.edu.cn, anby20@lzu.edu.cn, {guoshuai.zhao, qianxm}@mail.xjtu.edu.cn

Abstract

Audio classification plays a crucial role within fields such as human-machine interaction and intelligent robotics. However, high-performance audio classification systems typically demand significant computational and storage resources, posing substantial challenges when deploying to the resource-constrained edge devices with an urgent need for such capabilities. To achieve a new level of balance between model complexity and performance, we introduce a novel multi-view method for the separated time-frequency features extraction and utilization, which exists within the proposed **M**ini **M**irror **M**ulti-View **N**etwork (M3Net) in the form of the Mirror Attention mechanism. M3Net enables reversible spatial transformation of spectral features is capable of efficiently leverages robust local and global features in the time and frequency domains with low requirements for parameters. Experiments based on Mel-Spectrogram without data augmentation and pre-training indicate that M3Net can achieve classification accuracy over 97% on the UrbanSound8K and SpeechCommandsV2 datasets with only 0.03 million parameters. The contribution of each functional segment in M3Net is fully verified and explained in the ablation experiments.

Code — https://github.com/Mental-Scholar/M3Net
Video — https://github.com/Mental-Scholar/AAAI-25-oral

1 Introduction

With the advancements in auditory perception technology, audio classification has become increasingly important in human-machine interaction and intelligent robotics (Silva et al. 2024; Latif et al. 2023; Bingol and Aydogmus 2020). In such applications, audio processing requires high real-time performance for immediate responses. A current method is to move the data processing procedure from the cloud to the edge for reducing latency, which not only places higher demands on model performance, but also imposes greater computation cost on Microcontroller Units (MCUs) with limited resources (Fu et al. 2023; Shuvo et al. 2022).

*Corresponding author.

In recent years, researchers have applied Convolutional Neural Networks (CNNs) to edge devices for addressing multiple audio classification scenarios based on limited resources (Goulão et al. 2024; Lamrini, Chkouri, and Touhafi 2023; Hershey et al. 2017). Obviously, higher model complexity generally results in better performance (Lin and Washington 2024; Abbas et al. 2024). However, State-of-the-art (SOTA) methods currently struggle to achieve a required balance between performance and complexity, such as a high-performance but highly complex Transformer model (Elliott et al. 2021) still have room for improvement when deployed on edge devices. Notably, the widely used CNNs are originally designed for image-based tasks, which may limit the performance of audio classification models based on CNNs (Palanisamy, Singhania, and Yao 2020).

Currently, researchers predominantly focus on improving model performance. For instance, EAT (Chen et al. 2024) has achieved high accuracy on SpeechCommandsV2, but the 88 million parameters limit its feasibility for deployment on edge (Zhang et al. 2024). Many researchers have introduced various "attention" (Li et al. 2023) and compression optimizations (Mishra and Gupta 2023) to deploy models on edge. However, these models' accuracy has been negatively impacted in varying degrees due to the reduced parameters.

In this work, we propose **M**irror **A**ttention (MA) that can be applied to lightweight CNNs and constructed M3Net. Firstly, Mel-Spectrogram features are extracted from raw audio. **D**iagonal **F**eature **P**rocessing (DFP) and **M**irror **F**eature **P**rocessing (MFP) are performed separately after a few convolutions to obtain the multi-view time-frequency feature maps. DFP includes a **C**onvolutional **B**lock (CB) with one-dimensional **E**fficient **C**hannel **A**ttention (ECA) for extracting time-frequency information that is completely opposite to the original features. MFP includes a **M**irror **B**lock (MB) that can perform the time-domain-based **H**orizontal **F**eature **P**rocessing (HFP) and the frequency-domain-based **V**ertical **F**eature **P**rocessing (VFP) while maintaining skip connection with the original time-frequency features.

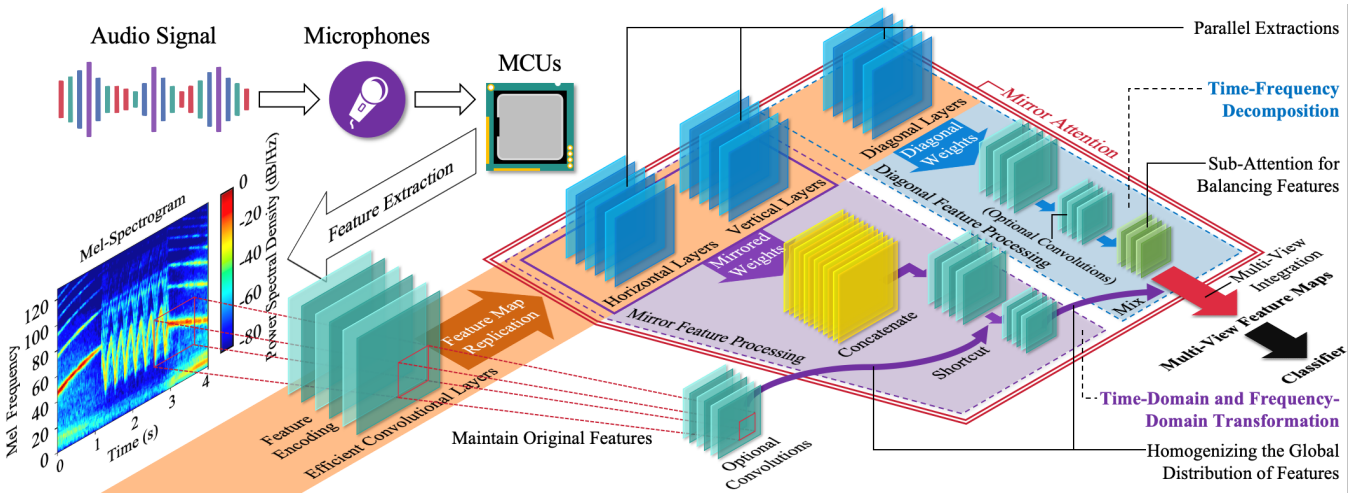

Figure 1: Edge-based audio classification through M3Net and Mel-Spectrogram feature extraction.

Finally, the weighted multi-view feature maps in Figure 1 will be processed through standard operations such as softmax activation to obtain the probability for each audio clip.

M3Net can extract the separated local and global time-frequency features from mixed multi-view feature maps. Compared to other conventional lightweight CNNs, M3Net significantly reduces model complexity and substantially decreases the parameter count by integrating the rare and parallel computational processes in DFP, MFP, and the skip connection from the original features, as shown in Figure 1.

The main contributions are as follows:

- **Time-Frequency Separation Method.** We introduce a method of extracting separated time-domain and frequency-domain features in parallel to enhance the effectiveness of audio content usage. This approach can capture robust features in both time-domain and frequency-domain, which are distinct from the original inputs, during reversible time-frequency transformations.
- **Mirror Attention Mechanism.** We propose a new attention mechanism for the integration of separated time-frequency features through the proposed separation method. It is based on parallel convolutions of mirror and diagonal processes, extracting the robust multi-view features that extend and enhance the original feature set. This mechanism exhibits a low dependency on traditional data augmentation and can be applied to various types of audios.
- **M3Net Architecture.** We introduce M3Net, a lightweight neural network with merely around 0.03 million parameters. M3Net achieves classification accuracy **over 97%** on the UrbanSound8K and SpeechCommandsV2 datasets, with the parameter count being **less than 10%** of those in other accuracy-comparable SOTA methods, and even without any data augmentation or pre-training operations. This achievement in capable of processing a diverse range of audio content indicates that M3Net has the potentiality to facilitate the deployment of high-performance but lightweight models based on edge devices.

2 Methodology

As shown in Figure 1, M3Net with a few convolutions is a lightweight audio processing framework designed to extract separated time-frequency feature maps via a parallel time-frequency transformation structure. It further homogenizes the global distribution of spectral features to facilitate local-global feature weighting using two-stage skip connection.

2.1 Audio Feature Extraction

Feature extraction is the process of transforming raw audio into low-dimensional feature vectors that simplify the audio while retaining essential acoustic information (Salau and Jain 2019; Sharma, Umapathy, and Krishnan 2020).

Recent studies have employed Mel-Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), and Mel-Spectrogram as feature extraction methods (Kranthi Kumar and Alphonse 2022; Gong et al. 2022). Among them, Mel-Spectrogram transforms the time-domain information of audio into a spectrogram, allowing both time and frequency features to be displayed simultaneously. Based on relatively low computational complexity, Mel-Spectrogram extraction has begun to be applied in fields such as audio classification especially on edge devices (Yeow et al. 2024).

The extraction of the Mel-Spectrogram begins with standard preprocessing, which involves framing and windowing. Next, a Fast Fourier Transform (FFT) is applied to derive the spectrogram, followed by the computation of the power spectrum. Finally, the spectrogram is converted into the Mel-Spectrogram using a bank of Mel-filters. All the processes are performed using the Python library Librosa (version: 0.10.2.post1). The sample rate is set to 16 kHz, with 128 Mel-filters and a 512-sample FFT window. The remaining parameters are kept at their default values.

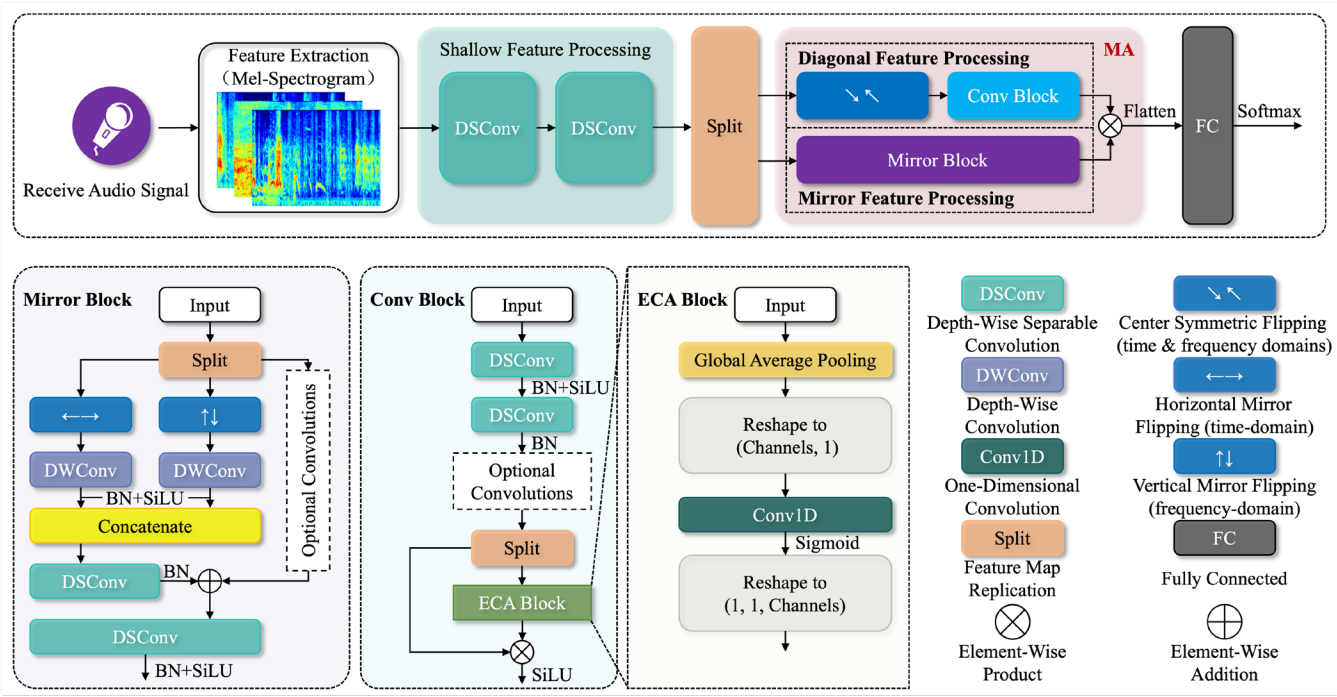

Figure 2: Structure of audio classification system based on M3Net (BN: batch normalization).

2.2 Time-Frequency Integration in M3Net

As demonstrated in Figure 2, the proposed MA aims to perform reversible transformations in the time-domain and frequency-domain of the feature maps extracted from Mel-Spectrogram to learn more robust features in MFP and DFP. Among them, the DFP with weaker feature extraction capability is required to sustain the same feature levels of output as MFP through ECA (Wang et al. 2020).

Time-Domain and Frequency-Domain Transformation. As shown in Figure 2, the MFP branch includes two sub-branches, HMF and VMF, which flip and extract the feature maps based on the time-domain and frequency-domain, respectively. The main transformation process is as follows:

Let the shape of the feature maps input to MA be denoted as $\boldsymbol{X}\in R^{W\times H\times C}$, where C represents the number of channels, W and H denote the width and height of the feature maps, respectively. For any element X in the input time-frequency feature maps $\boldsymbol{X}$, their relationship can be defined as:

$$\boldsymbol{X}=\{X(w,h,c)|w\in[0,W\text{-}1],h\in[0,H\text{-}1],c\in[0,C\text{-}1]\} \quad (1)$$

where the width, height and channel index of X are denoted as w, h, and c, respectively.

The time-domain and frequency-domain transformation results $X_T\in\boldsymbol{X}_T$ and $X_F\in\boldsymbol{X}_F$ can be expressed as:

$$X_{T,c\in[0,C\text{-}1]}(w,h)=M_{HMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{HMF}=\begin{bmatrix}-1&0&W\\0&1&0\\0&0&1\end{bmatrix} \quad (2)$$

$$X_{F,c\in[0,C\text{-}1]}(w,h)=M_{VMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{VMF}=\begin{bmatrix}1&0&0\\0&-1&H\\0&0&1\end{bmatrix} \quad (3)$$

where θ denotes the weight of the selected element.

As presented in Figure 2, the feature maps transformed by M_{HMF} and M_{VMF} in MB are concatenated along the channel dimension. The concatenated results then pass through a depth-wise separable convolution to align the feature dimensions with those from the shortcut connection originating from the original feature maps. Finally, the outcome of the element-wise addition is convolved with a depth-wise separable convolution to obtain the time-frequency double-transformed feature maps $\boldsymbol{X}_{MFP}$, which can be expressed as:

$$\boldsymbol{X}_{MFP}=F_M(M_{HMF}\boldsymbol{X},M_{VMF}\boldsymbol{X}),\quad \boldsymbol{X}\in R^{W\times H\times C} \quad (4)$$

where F_M represents the calculation process in MFP distinct from M_{HMF} and M_{VMF}.

Time-Frequency Decomposition. As presented in Figure 1 and Figure 2, the DFP branch extracts features that are completely opposite to the original features through Center Symmetric Flipping (CSF) within time and frequency domains.

For element $X(w,h,c)$ in the input feature maps $\boldsymbol{X}$, the time-frequency separated result $X_D\in\boldsymbol{X}_D$ can be expressed as:

$$X_{D,c\in[0,C\text{-}1]}(w,h)=M_{CSF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{CSF}=\begin{bmatrix}-1&0&W\\0&-1&H\\0&0&1\end{bmatrix} \quad (5)$$

Dataset: UrbanSound8K	Information entropy: 2.884 ± 0.011 bits/symbol			Zero-crossing rate: 3216.55 ± 2904.85 Hz		
Method	Feature	Accuracy (%)	# Param ($\times 10^6$)	Wilcoxon p-value	Data augmentation	Pre-training
AemNet-DW (Lopez-Meyer et al. 2021)	Log-Mel	82.25	0.9	< 5.0e-2		✓
ULSED (Peng et al. 2022)	Log-Mel	83.5	0.34	< 5.0e-2	✓	
2D CNN (Gupta, Hossain, and Kim 2022)	GFCC	89	1.8	< 5.0e-2	✓	
PhiNets M40 (Paissan et al. 2022)	Mel-Spectrogram	76.3	**0.027**	< 5.0e-2	✓	
SE-TCAM 1D CNN (Xu et al. 2024)	Raw audio	94.04	0.81	< 5.0e-2		
M3Net (Ours)	Mel-Spectrogram	**97.44**	0.029	baseline		

Table 1: Comparison of M3Net and the lightweight SOTA methods on US8K, with the Wilcoxon tests are based on accuracy.

As demonstrated in Figure 2, $M_{CSF}\boldsymbol{X}$ is first individually processed through CB to perform additional depth-wise separable convolutions and ECA. The one-dimensional ECA is designed to obtain the channel weights for balancing time-frequency features' value from DFP and MFP.

The resulting feature maps $\boldsymbol{X}_{DFP}$ can capture the symmetry of global and local features that are completely distinct from $\boldsymbol{X}_{MFP}$, and can be expressed as:

$$\boldsymbol{X}_{DFP}=F_D(M_{CSF}\boldsymbol{X})=F_D(M_{HMF}M_{VMF}\boldsymbol{X}),\quad \boldsymbol{X}\in R^{W\times H\times C} \quad (6)$$

where F_D is the calculation process in DFP beyond M_{CSF}.

Obviously, the transformed $M_{HMF}\boldsymbol{X}$, $M_{VMF}\boldsymbol{X}$, $M_{CSF}\boldsymbol{X}$ and the original $\boldsymbol{X}$ can be related by the learnable transformation weight matrices. After performing operations such as depth-wise separable convolution in F_D, the relationship between the separated time and frequency feature maps in DFP and MFP can be gradually transformed from local correlation to global correlation along the channel dimension, which is a complex but experimentally verifiable process.

Multi-View Integration. The final operation in MA is to multiply the separated feature maps output from DFP and MFP in Figure 2. The output $\boldsymbol{X}_{MA}$ can be expressed as:

$$\boldsymbol{X}_{MA}(w,h,c)=\boldsymbol{X}_{MFP}(w,h,c)\odot\boldsymbol{X}_{DFP}(w,h,c) \quad (7)$$

where $\odot$ denotes the Hadamard product of matrices (Kim et al. 2016). The output of MA exhibits stronger local-global feature correlations in both the time and frequency domains compared to the output from traditional CNNs.

3 Experiments

3.1 Experiment Setup

To demonstrate the audio classification capability of M3Net, we select to conduct experiments on the UrbanSound8K and SpeechCommandsV2 datasets, which are widely used in the field of audio classification. Among them, UrbanSound8K focuses on the classification of environmental noises, whereas SpeechCommandsV2 is specifically designed for classifying vocal commands, which enables them to cover a wide range of practical sound types. Our experiments were conducted based on Python 3.8 (Ubuntu 20.04), TensorFlow 2.9.0, and CUDA 11.2, with the training and testing processes performed on an NVIDIA A800-80GB GPU.

Datasets. UrbanSound8K (US8K) (Salamon, Jacoby, and Bello 2014) is a dataset designed for environmental sound classification tasks. It contains 8,732 audio clips and covers common sounds in urban environments such as sirens, dog barks, footsteps and street music. SpeechCommandsV2 (SCV2) (Warden 2018) serves as a dataset released by Google for research on vocal commands. It contains roughly 105,000 command samples, such as "yes", "no", "up" and "down". These two datasets are randomly partitioned into training, validation and test sets in 6: 2: 2 in our experiments.

Training Details. We converted raw audio with a reset sample rate of 44.1 kHz into Mel-Spectrogram and get an input shape of 256×256×3 for M3Net. After multiple tests, we optimized training processes using a batch size of 64, the Adam optimizer, and cross-entropy loss, with an initial learning rate of 10^{-3}, and a minimum learning rate threshold of 10^{-12}. Each model was trained 14 times independently, with each training session consisting of 100 epochs.

3.2 Analysis of SOTA Comparison

As shown in Table 1, M3Net without data augmentation and pre-training surpasses the relatively higher-accuracy lightweight SOTA methods using 1.5-8.5% of their parameters, and it outperforms SOTA with comparable parameter count by over 20% in accuracy. When the Wilcoxon test's p-value is less than 5.0e-2, the difference between the two sets of data is generally considered significant. Thus, Table 1 indicates that the difference in accuracy between M3Net and the other SOTA methods on US8K is statistically significant.

As indicated in Table 2, although M3Net does not achieve the highest accuracy on SCV2, the other SOTA methods do not considerably outperform M3Net in the Wilcoxon tests. Meanwhile, M3Net still keeps an advantage of at least one order of magnitude in terms of parameter count while maintaining accuracy on par with the other SOTA methods.

Dataset: SpeechCommandsV2	Information entropy: 2.396 ± 0.002 bits/symbol			Zero-crossing rate: 2341.96 ± 1262.64 Hz		
Method	Feature	Accuracy (%)	# Param ($\times 10^6$)	Wilcoxon p-value	Data augmentation	Pre-training
DeLoRes M (Ghosh, Seth, and Umesh 2022)	Log-Mel	89.7	5.3	< 5.0e-2	✓	✓
AdaptFormer (Chen et al. 2022; Selvarajz et al.2023)	Log-Mel	92.3	1.43	< 5.0e-2		✓
DCLS-Delays (3L-2KC) (Hammouamri et al. 2023)	Mel-Spectrogram	95.35	2.5	< 5.0e-2	✓	
SeqBoat (Ren et al. 2023)	Raw audio	97.35	0.293	2.2e-1		
EAT-S-GMME (He et al. 2024)	Raw audio	**97.88**	1.54	1.0e-1	✓	
M3Net (Ours)	Mel-Spectrogram	97.03	**0.031**	baseline		

Table 2: Comparison of M3Net and the lightweight SOTA methods on SCV2, with the Wilcoxon tests are based on accuracy.

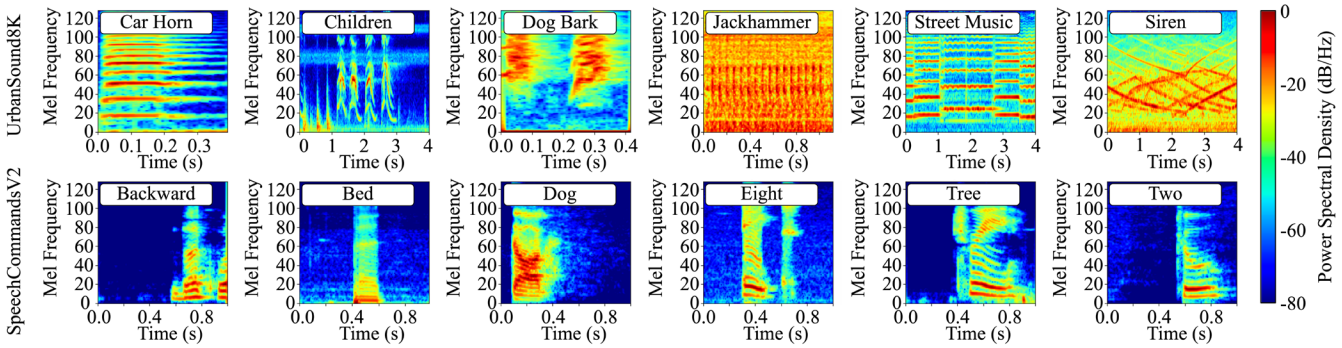

Figure 3: Comparison of Mel-Spectrogram feature distributions for representative audio samples from US8K and SCV2.

To analyze the impact of different types of audio content on M3Net, an examination of the feature images derived from samples in US8K and SCV2 is presented in Figure 3.

As illustrated in Figure 3, there is a significant overall divergence in the visual audio features between US8K and SCV2, with US8K exhibiting generally stronger irregularity in audio content. This is since the time-frequency components of the environmental noise contained in US8K have stronger uncertainty compared to the speech audio of SCV2. It matches the information entropy and zero-crossing rate values stated in Table 1 and Table 2.

The results of the SOTA comparison experiments on different datasets in Table 1 and Table 2 confirm that the multi-view approach within MA can more effectively enrich the relationships between local and global time-frequency features in the more complex US8K compared to the simpler SCV2, thereby enhancing the utilization of the separated features in US8K. Thus, M3Net trained on US8K exhibits superior performance over those trained on SCV2. This phenomenon contradicts the findings of many previous studies.

The adaptability to complex audio is crucial for MA to improve model performance while reducing parameters, as it not only enables the autonomous mining and generation of more critical information with fewer resources through parallel multi-view extractions, but also enhances the robustness to adapt to complex real-world requirements. It makes M3Net considerably advantageous for deployment on the edge compared the other lightweight SOTA methods.

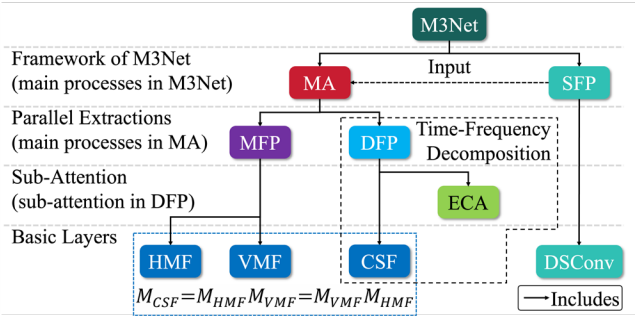

Figure 4: The main processes involved in ablation study.

Therefore, M3Net is anticipated to address the challenges of time-frequency features extraction and utilization in complex audio, as it is more challenging for audio with simple content and uniform patterns to establish relationship between the time and frequency domains, and the contribution of multi-view feature learning to model is relatively low.

3.3 Ablation Study

We conducted ablation experiments to verify the main processes in M3Net as shown in Figure 1 and Figure 2, with the hierarchical relationship of each step presented in Figure 4. In all subsequent experiments, values in "( )" under the Acc. metric indicate the change in accuracy compared to the full M3Net in Table 1 and Table 2, with the baseline for the Wilcoxon tests also being the accuracy of the full M3Net.

Dataset	M3Net w/o MA		M3Net w/o SFP	
	Accuracy (%)	Wilcoxon	Accuracy (%)	Wilcoxon
US8K	75.29 (-22.15)	< 5.0e-2	87.31 (-10.13)	< 5.0e-2
SCV2	68.13 (-28.90)	< 5.0e-2	84.26 (-12.77)	< 5.0e-2

Table 3: Ablation on the framework of M3Net.

Dataset	M3Net w/o MFP		M3Net w/o DFP	
	Accuracy (%)	Wilcoxon	Accuracy (%)	Wilcoxon
US8K	87.26 (-10.18)	< 5.0e-2	93.93 (-3.51)	< 5.0e-2
SCV2	78.94 (-18.09)	< 5.0e-2	93.08 (-3.95)	< 5.0e-2

Table 4: Ablation on the parallel branches of M3Net.

Effect of MA. As shown in Table 3, the accuracy of M3Net without MA is decreased significantly by over 20%. Based on the previous analysis, MA is designed to provide the model with enhanced feature extraction capability, enabling model to discover and learn more critical local and global features within separated time-frequency features. Thus, the model without MA regresses to conventional CNNs, which are incapable of capturing these potential features, resulting in a significant decline in performance, as evidenced by the model's complete underperformance compared to the other SOTA methods and the full M3Net in Table 1 and Table 2.

Effect of Shallow Feature Processing (SFP). To verify the necessity of SFP before MA, we removed one of the depth-wise separable convolutions in SFP in Figure 2. According to Table 3, the model that lost half of SFP experienced a significant drop in accuracy over 10% on both datasets. These results indicate that overly large input size adversely affect the performance of MA, owing to the convolutional operations in MA is insufficient to process such information. We did not remove the entire SFP because the large-sized feature maps would result in extremely high training costs in the absence of preceding convolutions.

Effect of MFP. As depicted in Table 4, the model without MFP suffered a significant decrease in accuracy of over 10%. These results demonstrate that the practical value of the time-frequency decomposed feature maps extracted by DFP can be substantially diminished when the separated time-frequency features directly connected to the original features. For audio with relatively high complexity, DFP can still effectively extract partial multi-view features from inputs, because CSF contains both time and frequency transformation operations. Consequently, the negative impact on M3Net trained on US8K due to the loss of MFP is significantly less than that observed in models trained on SCV2.

Ablation target		1	2	3	4
ECA			✓	✓	✓
CSF				✓	✓
VMF					✓
HMF					
US8K	Accuracy (%)	93.88 (-3.56)	94.56 (-2.88)	95.36 (-2.08)	94.82 (-2.62)
	Wilcoxon	< 5.0e-2	< 5.0e-2	< 5.0e-2	< 5.0e-2
SCV2	Accuracy (%)	94.00 (-3.03)	94.20 (-2.76)	94.89 (-2.14)	91.82 (-5.21)
	Wilcoxon	< 5.0e-2	< 5.0e-2	< 5.0e-2	< 5.0e-2

Table 5: Ablation on the sub-attention and basic layers.

Effect of DFP. As indicated in Table 4, the model without DFP showed a decrease in accuracy of over 3.5%. It is noteworthy that DFP improves accuracy by approximately one-third of MFP's improvement, with DFP and MFP use 5 million and 15 million parameters, respectively. It indicates that although MFP combining the original and transformed features can efficiently improve the model performance, it is worthwhile to construct the full multi-view feature relationship to achieve new breakthroughs in the already high accuracy (over 93%) through more robust local-global weighting.

3.4 Reverse Ablation Study

As shown in Figure 4, HMF, VMF and CSF are basic but critical operations in MFP and DFP for the time-domain and frequency-domain transforming. It is important to verify the relationship between them on different kinds of datasets.

Effect of ECA. As depicted in Table 5, incorporating ECA into the model without any transformation convolutional layers observe an accuracy gain of approximately 0.3-0.7% on both datasets. It confirms that ECA plays a pivotal role in bridging the granularity gap in feature extraction between DFP and MFP. Therefore, ECA within M3Net is to balance the feature levels of the two branches of different scales.

Effect of CSF. As shown in Table 5, the accuracy of the model with CSF and ECA is decreased by roughly 2%. These results are considerably better compared with removing the entire DFP in Table 4, which proves that CSF is a key bridge in establishing separated feature maps from the time-domain and frequency-domain, but traditional convolutional operations cannot effectively provide this function.

Effect of VMF. According to Table 5, the model's accuracy is decreased by 2.6-5.2% without HMF, whereas the accuracy reduction is only 2% when ECA and CSF are restored. It is because DFP remains capable of extracting equivalent time-frequency features simultaneously, but MFP with VMF but without HMF disrupts the balance of the time-frequency features with more complex frequency features.

Effect of HMF. As indicated in Table 5, the model without HMF exhibited a decrease in accuracy by 2.6-5.2%, with a significantly negative impact on SCV2. The reason is that the loss of HMF weaken the capability of robust time-domain features extracting, disrupting the balance of the time-frequency feature levels, and SCV2 with simpler temporal information is more susceptible to greater loss of features compared to the more complex US8K as shown in Figure 3.

4 Related Work

Audio processing often demands high real-time performance. A widely used method involves shifting the collection and inference processes from the cloud to the resource-limited edge devices (Erhan et al. 2021). Therefore, related work is gradually moving towards the direction of improving algorithm performance while reducing its complexity.

The Attention Mechanism Improves Model Efficiency. Recently, attention mechanisms have been proven to have significant potentiality in improving the performance of CNNs (Lau, Po, and Rehman 2024; Peng et al. 2024). When processing complex audio signals, CNNs combined with attention mechanisms can automatically extract critical subsets of features as needed, thereby considerably enhancing models' computational efficiency (Khan et al. 2024).

In the field of audio classification, researchers have recognized that classifying environmental sounds is more complex than classifying structured audio such as speech (Bansal and Garg 2022). Fortunately, using attention mechanisms focus on the time-domain (Wang, Feng, and Anderson 2021) and frequency-domain (Mu et al. 2021) can improve model performance while reducing complexity.

Currently, various attention mechanisms have been used in real-world audio classification. For example, Jung et al. (Jung et al. 2022) introduced an innovative heterogeneous stacking graph attention layer, which model across various time and spectral intervals through heterogeneous attention mechanism and stacked nodes, was successfully applied in an audio spoofing detection system. Similarly, Noumida and Rajan (Noumida and Rajan 2022) proposed a hierarchical attention-based bidirectional gated recurrent unit model to effectively address the challenges of multi-label bird vocalizations classification and have achieved significant results.

Lightweight Models Facilitate Edge Applications. With the rapid development of industries such as autonomous driving and wearable applications, edge devices typically demand high levels of real-time signal processing capability and response speed (Lin et al. 2023). However, due to the performance limitations imposed by unavoidable size constraints, the chips of edge devices are ill-suited for deploying large-scale deep learning models. It creates a critical bottleneck for edge deployment in related applications.

In edge applications, CNNs exhibit remarkable performance in audio processing (Zhang et al. 2023), and a hybrid model combining CNN and LSTM has been proposed recently (Mou and Milanova 2024), which optimizes the trade-off between model size and performance through pruning and compression. Prior to this, knowledge distillation has been explored as a method to improve model efficiency by transferring insights from large-scale models to the more compact edge-based models (Choi and Park 2022). Similarly, Lamrini et al. (Lamrini, Chkouri, and Touhafi 2023) introduced a hybrid model combining CNN and ANN to explore the efficiency limits on edge devices. Recent advancements have proposed a two-stage transfer learning-based network, which effectively reduces model's storage requirements with less accuracy degradation (Zhang et al. 2024). In other aspects, a low complexity denoising method based on sound field imaging was proposed for sound source localization on MCUs (Jiang, Zheng, and Wang 2023), and a simulated artificial chip that integrates DNN was designed to facilitate energy-efficient speech recognition and detection (Ambrogio et al. 2023). Related works have made great contributions to the application of lightweight Audio Detection Technology (ADT) in resource-constrained scenarios.

5 Conclusion

In this paper, we introduce a novel mirror attention-based M3Net that exhibits comparable audio classification accuracy to the other SOTA methods with dozens of times more parameters in both environmental noise and speech commands datasets without data augmentation and pre-training. M3Net achieves this through a special method of efficiently leveraging multi-view time-frequency features derived from the reversible transformation of raw audio. This approach not only enhances model performance by discovering potential time-domain and frequency-domain features which may not be evident in raw audio, but also significantly reduces the parameters through rare lightweight convolutions. It may provide new insights for research and application in ADT, especially in resource-constrained environments. Next, we will focus on deploying M3Net and its improved versions on edge devices for real-world testing.

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China (No. 62372364), Technical Innovation Guidance Plan of Shaanxi Province, China (No. 2024QCY-KXJ-199), and honored to receive initial support from Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment (LZU-JZH2619, LZU-JZH2620). We gratefully honor the memory of **Mrs. Chin** and **Prof. Lee**, whose immortal legacy will inspire generations of scholars.

References

Abbas, S.; Ojo, S.; Al Hejaili, A.; Sampedro, G. A.; Almadhor, A.; Zaidi, M. M.; and Kryvinska, N. 2024. Artificial intelligence framework for heart disease classification from audio signals. *Scientific Reports*, 14 (1): 3123. doi.org/10.1038/s41598-024-53778-7.

Ambrogio, S.; Narayanan, P.; Okazaki, A.; Fasoli, A.; Mackin, C.; Hosokawa, K.; Nomura, A.; Yasuda, T.; Chen, A.; and Friz, A. 2023. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature*, 620 (7975): 768-775. doi.org/10.1038/s41586-023-06337-5.

Bansal, A.; and Garg, N. K. 2022. Environmental Sound Classification: A descriptive review of the literature. *Intelligent systems with applications*, 16: 200115. doi.org/10.1016/j.iswa.2022.200115.

Bingol, M. C.; and Aydogmus, O. 2020. Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot. *Engineering Applications of Artificial Intelligence*, 95: 103903. doi.org/10.1016/j.engappai.2020.103903.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, *35*: 16664-16678.

Chen, W.; Liang, Y.; Ma, Z.; Zheng, Z.; and Chen, X. 2024. EAT: Self-supervised pre-training with efficient audio transformer. arXiv: 2401.03497. doi.org/10.48550/arXiv.2401.03497.

Choi, K.; and Park, H.-M. 2022. Distilling a pretrained language model to a multilingual ASR model. arXiv:2206.12638. doi.org/10.48550/arXiv.2206.12638.

Elliott, D.; Otero, C. E.; Wyatt, S.; and Martino, E. 2021. Tiny transformers for environmental sound classification at the edge. arXiv: 2103.12157. doi.org/10.48550/arXiv.2103.12157.

Erhan, L.; Ndubuaku, M.; Di Mauro, M.; Song, W.; Chen, M.; Fortino, G.; Bagdasar, O.; and Liotta, A. 2021. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, 67: 64-79. doi.org/10.1016/j.inffus.2020.10.001.

Fu, L.; Yan, K.; Zhang, Y.; Chen, R.; Ma, Z.; Xu, F.; and Zhu, T. 2023. EdgeCog: a real-time bearing fault diagnosis system based on lightweight edge computing. *IEEE Transactions on Instrumentation and Measurement*, 72: 1-11. doi.org/10.1109/TIM.2023.3298403.

Ghosh, S.; Seth, A.; and Umesh, S. 2022. Decorrelating Feature Spaces for Learning General-Purpose Audio Representations. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6): 1402-1414. doi.org/10.1109/JSTSP.2022.3202093.

Gong, Y.; Lai, C.-I.; Chung, Y.-A.; and Glass, J. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (10): 10699-10709. doi.org/10.1609/aaai.v36i10.21315.

Goulão, M.; Bandeira, L.; Martins, B.; and L. Oliveira, A. 2024. Training environmental sound classification models for real-world deployment in edge devices. *Discover Applied Sciences*, 6 (4): 166. doi.org/10.1007/s42452-024-05803-7.

Gupta, S. S.; Hossain, S.; and Kim, K.-D. 2022. Recognize the surrounding: Development and evaluation of convolutional deep networks using gammatone spectrograms and raw audio signals. *Expert Systems with Applications*, 200: 116998. doi.org/10.1016/j.eswa.2022.116998.

Hammouamri, I.; Khalfaoui-Hassani, I.; Masquelier, T.; and Masquelier, T. 2023. Learning delays in spiking neural networks using dilated convolutions with learnable spacings. arXiv:2306.17670. doi.org/10.48550/arXiv.2306.17670.

He, B.; Zhang, S.; Wang, X.; Qiu, Z.; Takeuchi, D.; Niizumi, D.; Harada, N.; and Makino, S. 2024. Light Gated Multi Mini-Patch Extractor for Audio Classification. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, 765-769. doi.org/10.1109/ICASSPW62465.2024.10626081.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. 2017. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 131-135. doi.org/10.1109/ICASSP.2017.7952132.

Jiang, X.; Zheng, Y.; and Wang, X. 2023. Adaptive noise reduction method for sound sources based on acoustic imaging. *Physics Experimentation*, 43 (08): 48-55,60. doi.org/10.19655/j.cnki.1005-4642.2023.08.007.

Jung, J.-w.; Heo, H.-S.; Tak, H.; Shim, H.-j.; Chung, J. S.; Lee, B.-J.; Yu, H.-J.; and Evans, N. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6367-6371. doi.org/10.1109/ICASSP43922.2022.9747766.

Khan, M.; Gueaieb, W.; El Saddik, A.; and Kwon, S. 2024. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245: 122946. doi.org/10.1016/j.eswa.2023.122946.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. arXiv: 1610.04325. doi.org/10.48550/arXiv.1610.04325.

Kranthi Kumar, L.; and Alphonse, P. 2022. COVID-19 disease diagnosis with light-weight CNN using modified MFCC and enhanced GFCC from human respiratory sounds. *The European Physical Journal Special Topics*, 231 (18): 3329-3346. doi.org/10.1140/epjs/s11734-022-00432-w.

Lamrini, M.; Chkouri, M. Y.; and Touhafi, A. 2023. Evaluating the performance of pre-trained convolutional neural network for audio classification on embedded systems for anomaly detection in smart cities. *Sensors*, 23 (13): 6227. doi.org/10.3390/s23136227.

Latif, S.; Cuayáhuitl, H.; Pervez, F.; Shamshad, F.; Ali, H. S.; and Cambria, E. 2023. A survey on deep reinforcement learning for audio-based applications. *Artificial Intelligence Review*, 56 (3): 2193-2240. doi.org/10.1007/s10462-022-10224-2.

Lau, K. W.; Po, L.-M.; and Rehman, Y. A. U. 2024. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Systems with Applications*, 236: 121352. doi.org/10.1016/j.eswa.2023.121352.

Li, Y.; Cao, W.; Xie, W.; Li, J.; and Benetos, E. 2023. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. *IEEE Transactions on Multimedia*, 26: 1346-1360. doi.org/10.1109/TMM.2023.3280011.

Lin, J.; Zhu, L.; Chen, W.-M.; Wang, W.-C.; and Han, S. 2023. Tiny machine learning: progress and futures [Feature]. *IEEE Circuits and Systems Magazine*, 23 (3): 8-34. doi.org/10.1109/MCAS.2023.3302182.

Lin, K.; and Washington, P. Y. 2024. Multimodal deep learning for dementia classification using text and audio. *Scientific Reports*, 14 (1): 13887. doi.org/10.1038/s41598-024-64438-1.

Lopez-Meyer, P.; del Hoyo Ontiveros, J. A.; Lu, H.; and Stemmer, G. 2021. Efficient end-to-end audio embeddings generation for audio classification on target applications. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 601-605. doi.org/10.1109/ICASSP39728.2021.9414229.

Mishra, R.; and Gupta, H. P. 2023. Transforming large-size to lightweight deep neural networks for IoT applications. *ACM Computing Surveys*, 55 (11): 1-35. doi.org/10.1145/3570955.

Mou, A.; and Milanova, M. 2024. Performance analysis of deep learning model-compression techniques for audio classification on edge devices. *Sci*, 6 (2): 21. doi.org/10.3390/sci6020021.

Mu, W.; Yin, B.; Huang, X.; Xu, J.; and Du, Z. 2021. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11 (1): 21552. doi.org/10.1038/s41598-021-01045-4.

Noumida, A.; and Rajan, R. 2022. Multi-label bird species classification from audio recordings using attention framework. *Applied Acoustics*, 197: 108901. doi.org/10.1016/j.apacoust.2022.108901.

Paissan, F.; Ancilotto, A.; Brutti, A.; and Farella, E. 2022. Scalable neural architectures for end-to-end environmental sound classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 641-645. doi.org/10.1109/ICASSP43922.2022.9746093.

Palanisamy, K.; Singhania, D.; and Yao, A. 2020. Rethinking CNN models for audio classification. arXiv:2007.11154. doi.org/10.48550/arXiv.2007.11154.

Peng, L.; Yang, J.; Xiao, J.; Yang, M.; Wang, Y.; Qin, H.; Li, X.; and Zhou, J. 2022. ULSED: An ultra-lightweight SED model for IoT devices. *Journal of Parallel and Distributed Computing*, 166: 104-110. doi.org/10.1016/j.jpdc.2022.04.007.

Peng, P.; Chen, Y.; Lin, W.; and Wang, J. Z. 2024. Attention-based CNN–LSTM for high-frequency multiple cryptocurrency trend prediction. *Expert Systems with Applications*, 237: 121520. doi.org/10.1016/j.eswa.2023.121520.

Ren, L.; Liu, Y.; Wang, S.; Xu, Y.; Zhu, C.; and Zhai, C. X. 2023. Sparse modular activation for efficient sequence modeling. *Advances in Neural Information Processing Systems*, 36: 19799-19822. doi.org/10.5555/3666122.3666992.

Salamon, J.; Jacoby, C.; and Bello, J. P. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1041-1044. doi.org/10.1145/2647868.2655045.

Salau, A. O.; and Jain, S. 2019. Feature extraction: a survey of the types, techniques, applications. In *International Conference on Signal Processing and Communication*, 158-164. doi.org/10.1109/ICSC45622.2019.8938371.

Selvaraj, N. M.; Guo, X.; Kong, A.; Shen, B.; and Kot, A. 2023. Adapter Incremental Continual Learning of Efficient Audio Spectrogram Transformers. arxiv:2302.14314. doi.org/10.48550/arXiv.2302.14314.

Sharma, G.; Umapathy, K.; and Krishnan, S. 2020. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158: 107020. doi.org/10.1016/j.apacoust.2019.107020.

Shuvo, M. M. H.; Islam, S. K.; Cheng, J.; and Morshed, B. I. 2022. Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. In *Proceedings of the IEEE*, 111 (1): 42-91. doi.org/10.1109/JPROC.2022.3226481.

Silva, D. A.; Whitehead, S.; Lengerich, C.; and Leather, H. 2023. CoLLAT: on adding fine-grained audio understanding to language models using token-level locked-language tuning. *Advances in Neural Information Processing Systems*, 36: 63197-63209. doi.org/10.5555/3666122.3668881.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534-11542. doi.org/10.1109/CVPR42600.2020.01155.

Wang, Y.; Feng, C.; and Anderson, D. V. 2021. A multi-channel temporal attention convolutional neural network model for environmental sound classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 930-934. doi.org/10.1109/ICASSP39728.2021.9413498.

Warden, P. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv:1804.03209.

Xu, H.; Tian, Y.; Ren, H.; and Liu, X. 2024. A Lightweight Channel and Time Attention Enhanced 1D CNN Model for Environmental Sound Classification. *Expert Systems with Applications*, 249: 123768. doi.org/10.1016/j.eswa.2024.123768.

Yeow, J. W.; Tan, E.-L.; Bai, J.; Peksi, S.; and Gan, W.-S. 2024. Real-Time Sound Event Localization and Detection: Deployment Challenges on Edge Devices. arXiv:2409.11700. doi.org/10.48550/arXiv.2409.11700.

Zhang, W.; Yao, P.; Gao, B.; Liu, Q.; Wu, D.; Zhang, Q.; Li, Y.; Qin, Q.; Li, J.; and Zhu, Z. 2023. Edge learning using a fully integrated neuro-inspired memristor chip. *Science*, 381 (6663): 1205-1211. doi.org/10.1126/science.ade3483.

Zhang, X.; Kou, H.; Xia, C.; Cai, H.; and Liu, B. 2024. Small-footprint automatic speech recognition system using two-stage transfer learning based symmetrized ternary weight network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1-5. doi.org/10.1109/ICASSP48485.2024.10447203.