IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 November 2023, accepted 27 November 2023, date of publication 30 November 2023,
date of current version 8 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337992

RESEARCH ARTICLE

Infant Cry Detection With Lightweight Wavelet Scattering Networks

HAITAO CAO[1], HAISHAN CHEN[1], AND JUNYING YUAN[2]
[1]School of Information Engineering, Guangzhou Panyu Polytechnic, Guangzhou 511483, China
[2]School of Electrical and Computer Engineering, Guangzhou Nanfang College, Guangzhou 510970, China

Corresponding author: Junying Yuan (yuanjy6@mail2.sysu.edu.cn)

This work was supported in part by the Collaborative Innovation Center for Intelligent Educational Technology of Guangzhou under Grant 2023B04J0002, in part by the Guangdong Scientific Research Promotion Project for Key Construction Disciplines under Grant 2021ZDJS132, and in part by the Guangdong Engineering Technology Center of Regular Universities under Grant 2021GCZX001.

ABSTRACT Devices equipped with algorithms for detecting infant cries allow parents to respond immediately to their crying babies. However, improving detection performance often requires complex algorithms that consume more computing resources, leading to increased power consumption and prices. In this study, we propose a novel approach that leverages first-order Wavelet Scattering Coefficients (WSC) as translation-invariant and deformation-stable representations of infant crying sounds. Based on this approach, we introduce an end-to-end Deep Neural Network (DNN) architecture designed to detect crying using merely 17K parameters and 22.7M MACs. The accuracy results, with a 96.98% accuracy rate on open-source datasets, demonstrate the effectiveness and robustness of our model for detecting infant cries in real-world environments.

INDEX TERMS Infant cry detection, wavelet scattering, lightweight neural networks.

I. INTRODUCTION

Crying is the primary method through which infants communicate their needs and feelings to their parents. However, many parents face challenges in responding promptly to their infant's cries due to their busy work [1], [2], [3]. In such situations, auxiliary devices offer a valuable tool for parents to detect cries effectively. Studies have shown that using an auxiliary device greatly reassures parents who may feel anxious about meeting their baby's needs [4], [5].

Most affordable cry detection devices rely solely on the amplitude or energy of the audio signal, resulting in inferior performance in noisy real-life environments. Recently, the advancements in Machine Learning (ML) have led to the development of sophisticated systems utilizing complex algorithms like Deep Neural Networks (DNN), which deliver exceptional detection results [6], [7], [8], [9]. Nevertheless, these advanced systems come at the cost of increased price and power consumption, as a large number of model parameters require more computing resources.

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

In real-world scenarios, devices employed for infant cry detection constantly monitor the audio streams in their surroundings, resulting in continuous power consumption. As a result, infant cry detection algorithms should exhibit high accuracy scores while maintaining low power consumption. In the ML field, detecting cries can be categorized as an audio classification task [3]. The prevalent approach involves extracting acoustic features from the cry sounds and feeding these features into robust classifiers, such as DNN models, to yield a discriminate result [10].

Given the stability of the audio signal within a short duration [11], the Fourier Transform (FT) is commonly used to extract acoustic features in Mel-scale, such as Mel-scale spectrogram and Mel-Frequency Cepstral Coefficients (MFCC). However, these so-called Short-time Fourier Transform (STFT) based features prevent extracting large-scale features, which limits classification performance [12], [13].

To overcome this limitation, researchers have discovered that the wavelet scattering method offers a breakthrough by extending the frame beyond its short length without sacrificing substantial information. A wavelet scattering network can

be viewed as a collection of specialized convolutional layers with fixed parameters, thereby facilitating the extraction of crucial acoustic features [14], [15], [16]. Numerous studies have demonstrated the advantageous properties of wavelet scattering for audio classification tasks [17], [18], [19].

On the other hand, despite recent advancements, there are challenges associated with applying DNN models in real-world scenarios. Firstly, achieving good performance typically requires a large number of training data. However, collecting infant cry samples can be difficult due to various reasons [3], [20], [21]. Secondly, the limited computing resources available on our devices necessitate the use of lightweight DNN models. Hence, complex structures that involve numerous parameters and operators, such as VGG [22] and ResNet [23], are not suitable for our specific case. Consequently, it is crucial to propose efficient DNN models that are suitable for mobile applications [24], [25], [26] and embedded systems [27], [28], [29], [30], [31].

In this paper, we present a lightweight and efficient model for infant cry detection utilizing wavelet scattering. Our contributions are outlined as follows:

1) We propose the utilization of first-order Wavelet Scattering Coefficients (WSC) as acoustic features for infant cry detection, which has been demonstrated to be more efficient than STFT-based features.
2) We combine the feature extractor and Neural Network Blocks (NNB) to construct a lightweight end-to-end wavelet scattering network. This design ensures ease of implementation and facilitates further engineering improvements.
3) The proposed model is trained with the infant cry samples collected from open-source datasets. These datasets are readily accessible to allow for result replication and comparison.

The subsequent sections of this paper are structured as follows. Section II provides a comprehensive overview of the related work in the field. In Section III, we outline the three desirable properties for acoustic features in audio classification tasks, and we also compare the effectiveness of WSC with STFT-based features. The proposed model is introduced in Section IV. To validate its performance, we conduct experiments in Section V, followed by discussions in Section VI. Finally, we present our conclusions in Section VII.

II. RELATED WORK

Extensive efforts have been dedicated to the detection of infant cries. Recently, advanced detection algorithms have emphasized two main steps. Firstly, extracting acoustic features aims to provide more discriminate representations of raw cry sounds. Subsequently, powerful DNN models are trained using these acoustic features to capture the crucial characteristics of infant cries further. This two-step approach has been a focal point in recent advancements in infant cry detection algorithms.

Although temporal patterns have shown promising performance [32], STFT-based features are widely used to detect infant cries. Many pieces of research have demonstrated the effectiveness of feeding Mel-scale spectrograms into Convolutional Neural Networks (CNNs). For example, [7], [33] proposed non-symmetric kernels with height (frequency content) larger than width (time content) in CNNs. This method achieved high frequency resolution compared to low temporal resolution for better performance. Significant advancements in ML approaches and DNN models for infant cry detection have been made in [9], where CNNs performed best. Furthermore, algorithms have been designed to address real-world scenarios by considering background noise. Reference [34] focused on noise reduction before cry detection, and [8] evaluated model performance using real-world infant cry datasets.

In terms of feature extraction, wavelet scattering has been utilized for time series classification [35], [36], and it has shown better performance than STFT-based features [13], [37], [38]. Researchers such as [39] and [40] have employed WSC in the time domain as features to identify ECG signals, achieving high accuracy scores. Furthermore, the application of joint time-frequency scattering transform has shown benefits in audio classification tasks [18], [41]. In recent years, more and more applications of WSC have been found in various tasks including image classification [42], indoor localization [43], speech emotion recognition [44], spoken language identification [45], and speaker identification [46]. However, to the best of our knowledge, there is no existing work on applying wavelet scattering for infant cry detection. Therefore, exploring the performance of wavelet scattering on infant cry sounds is both interesting and promising.

III. ACOUSTIC FEATURES IN AUDIO CLASSIFICATION TASKS

As previously mentioned, infant cry detection can be considered an audio classification task. Analyzing the properties of acoustic features is crucial for understanding the characteristics of audio signals and improving classification performance. In this section, we will briefly discuss three desirable properties of acoustic features in audio classification. Subsequently, we will compare the mathematical properties of STFT-based features and WSC.

A. TRANSLATION-INVARIANCE

The translation in time is generally desired to be uninformative for the classification of audio signals. For an audio signal $x(t)$, its representation Φx is considered to be invariant to global translations $x_\tau(t) = x(t - \tau)$ by $\tau \in \mathbb{R}$ if

$$\|\Phi x_\tau - \Phi x\| \leq C \tag{1}$$

for some small constant $C > 0$. Note that we may consider only local translation-invariance by restricting the distance $|\tau|$ for which Eqn. (1) holds.

B. STABILITY TO NOISE

The property of stability to additive noise provides robustness to the algorithm in noisy environments. It is defined in terms of Lipschitz continuity which is a strong form of uniform continuity for functions. To be stable to additive noise, it is necessary to ensure that for a signal $x_\epsilon(t) = x(t) + \epsilon(t)$, there must exist a bounded $C > 0$ s.t.

$$\|\Phi x_\epsilon - \Phi x\| \leq C\|x_\epsilon - x\|. \tag{2}$$

C. STABILITY TO DEFORMATIONS

The representation Φx is considered stable to deformation if it remains relatively unchanged by small deformations in the signal $x(t)$. This property helps to minimize intra-class variations while allowing significant variations for inter-class. Formally, for a signal $x_\tau(t) = x(t - \tau(t))$, where $\tau(t)$ is a non-constant displacement field (i.e., not just a translation) that deforms $x(t)$, we require a $C > 0$ s.t.

$$\|\Phi x_\tau - \Phi x\| \leq C\|x\| \sup_x |\nabla \tau(t)|. \tag{3}$$

The term $|\nabla \tau(t)|$ measures the deformation amplitude, so its supremum limits the global deformation amplitude.

D. STFT-BASED FEATURES

The acoustic features discussed in this section are compared in Fig. 1. From the left panel, we observe that the STFT-based features are the results obtained by applying the FT to the framed and windowed audio signal, then averaged by Mel-scale filters. Specifically, to mitigate the loss of high-frequency information, a common technique called pre-emphasis is employed. This involves applying a high-pass filter to the framed audio signal $x_0(t)$, resulting in $x(t)$. The signal $x(t)$ is then localized by a window function ϕ of duration T such that $\int \phi(t)dt = 1$. These batches of time series, denoted as $x(t)\phi(t)$, are transformed into the frequency domain using the FT, producing a 2D-like spectrogram $\hat{x}(t, \omega)$ that is measured in terms of time t and frequency ω. Finally, the energy of the spectrogram is averaged using Mel-scale filters $\hat{\psi}_\lambda$, where λ represents the center frequency of each filter:

$$\mathrm{M}x(t, \lambda) = \frac{1}{2\pi} \int |\hat{x}(t, \omega)|^2 |\hat{\psi}_\lambda(\omega)|^2 d\omega, \tag{4}$$

where $\mathrm{M}x(t, \lambda)$ is the Mel-scale spectrogram. In practice, the logarithmic representation is often beneficial for sound classification, as it better separates different types of signals with similar frequency content [7].

The Mel-scale spectrogram aims to mimic the non-linear human ear perception of sound by being more discriminate at lower frequencies and less discriminate at higher frequencies. The MFCC is obtained from the Discrete Cosine Transform (DCT) on the Mel-scale spectrogram to decorate the data [7].

The Mel-scale filter, denoted as $\hat{\psi}_\lambda$, is commonly designed as a bandpass filter with a constant-Q frequency bandwidth at high frequencies. The filter's frequency support is centered at λ with a bandwidth of the order of λ/Q. While at lower frequencies, the bandwidth of $\hat{\psi}_\lambda$ remains constant and equal to $2\pi/T$.

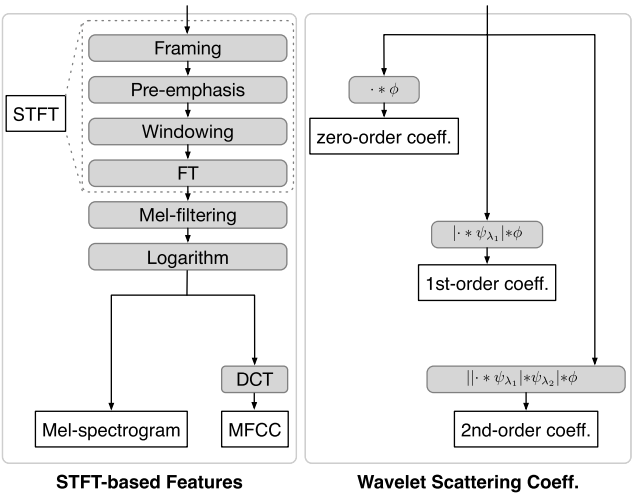

FIGURE 1. The flowchart illustrates the process of generating acoustic features from an audio signal. The left panel depicts STFT-based features, including the Mel-scale spectrogram and MFCC. The right panel shows the zero-order, first-order, and second-order WSC.

It has been established that the Fourier modulus representation $\Phi(x) = |\hat{x}|$ is not stable to deformation, as small deformations can severely distort high frequencies [13]. In Eqn. (4), the Mel-scale averaging addresses this deformation instability issue but comes at the cost of information loss. The research conducted by [13] demonstrated that frequency averaging approximates the time averaging of a filtering output, which can be expressed as:

$$\mathrm{M}x(t, \lambda) \approx |x * \psi_\lambda|^2 * |\phi|^2(t), \tag{5}$$

when $\lambda \gg Q/T$. The operation "$*$" represents convolution.

In Eqn. (5), the window function ϕ operates as a low-pass filter, ensuring local invariance for time-shifts smaller than T. However, this process leads to the removal of fine-scale information like vibratos and attacks. To reduce information loss, the Mel-scale spectrogram is often computed over small time windows of duration $T = 20 \sim 30$ ms. As a result, the extraction of large-scale features becomes challenging, thus limiting the performance of audio classification. To address this limitation, wavelet scattering has been proposed to capture the amplitude modulations of $|x * \psi_\lambda(t)|$ at scales smaller than T, which play a crucial role in auditory perception. This approach enables the possible increase of T without significant information loss.

E. WAVELET SCATTERING

The large coefficients in Eqn. (5) can be considerably amplified by the square operator. If we remove the square and calculate $|x * \psi_\lambda| * \phi$ instead, the high frequencies removed by the low-pass filter ϕ can be recovered by a new set of wavelet modulus coefficients. This gives rise to the concept of wavelet scattering. In theory, we define the zero-order scattering coefficients as a local translation-invariant descriptor of the

signal $x(t)$:

$$S_0 x(t) = x * \phi(t). \tag{6}$$

S_0 is a time-averaging that removes all high frequencies. Audio signals have little energy at low frequencies, so $S_0 x(t) \approx 0$ and it is often neglected [13].

The high frequencies that are not captured in the zero-order scattering coefficients can be recovered using a wavelet modulus transform, which forms the basis for defining the first-order scattering coefficients:

$$S_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t). \tag{7}$$

The first-order scattering coefficients are computed using wavelets ψ_{λ_1} with an octave frequency Q_1. In the case of audio signals, it is common to set $Q_1 = 8$, which corresponds to wavelets with the same frequency resolution as Mel-scale filters [13]. Expanding on this idea, we can further express the second-order scattering coefficients as follows:

$$S_2 x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t). \tag{8}$$

The wavelets ψ_{λ_2} in this context have an octave resolution of Q_2. Studies have shown that setting $Q_2 = 1$ results in wavelets with narrower time support, making them adapt to characterize transients and attacks [47].

The iterative process described above allows for calculating scattering coefficients at any order m. For $m \geq 1$, the iterated wavelet modulus convolutions can be expressed as:

$$U_m x(t, \lambda_1, \ldots, \lambda_m) = ||\ldots|x * \psi_{\lambda_1}| * \ldots| * \psi_{\lambda_m}(t)|, \tag{9}$$

where mth-order wavelets ψ_{λ_m} have an octave resolution Q_m. Averaging $U_m x$ with ϕ gives the scattering coefficients of order m:

$$S_m x(t, \lambda_1, \ldots, \lambda_m) = U_m x(t, \lambda_1, \ldots, \lambda_m) * \phi(t), \tag{10}$$

Therefore, a scattering decomposition of maximum order l is defined by initializing $U_0 x = x$, where x is the input signal. The final scattering vector is obtained by aggregating all the scattering coefficients of different orders.

$$S_x = \{S_m x\}_{0 \leq m \leq l}. \tag{11}$$

Interestingly, the cascaded convolutions and non-linearities in Eqn. (9) can be interpreted as a CNN, where $U_m x$ represents the output of the mth internal layer and the modulus operation acts as the activation function. However, when observing the right panel of Fig. 1, we can find that each such layer has an output $S_m = U_m * \phi$, and not just the last layer as in standard CNNs. This characteristic allows us to choose an arbitrary order of WSC, reducing computational complexity. Moreover, it is worth noting that all the filters ψ_λ are predefined wavelets and not learned from training data, which avoids the black box nature present in DNN models.

Fig. 2 depicts the STFT-based features and WSC of three samples sourced from an open-source dataset. Upon inspection, it is evident that the Mel-scale spectrogram provides the most information, whereas the MFCC appears to have the least informative content. The WSC yields moderate information, which will be demonstrated to be sufficient for discriminating between infant cries and other sounds.

IV. THE PROPOSED MODEL

This paper presents a novel lightweight model designed specifically for infant cry detection, as depicted in Fig. 3. The proposed model adopts an end-to-end architecture that comprises three types of blocks:

1) **Feature Extractor**: It follows the flowchart of wavelet scattering, as depicted in the right panel of Fig. 1, and serves as the basis for our model. However, for the sake of computational efficiency, we have integrated these operations within our model itself. As a result, we only utilize the first-order wavelet coefficients as the acoustic features, thereby significantly reducing the overall computational complexity.
2) **CNN Block**: The main component of this block is the concatenation of a depthwise 2D convolution and a pointwise 2D convolution. This design has been shown to significantly reduce computational complexity compared to traditional 2D convolutions during the training process [49]. The output of the pointwise 2D convolution is then normalized using batch normalization [50] and activated using the ReLU function [51].
3) **Residual Block**: This block draws inspiration from Matchboxnet [27], but it is designed to be more straightforward with fewer parameters. It consists of a learning pipeline on the left to strengthen pattern recognition and a residual pipeline on the right [23] to alleviate gradient vanishing in deep structures. The outputs of each pipeline are separately normalized using batch normalization, and then their sum is activated using the ReLU function. The aforementioned CNN and residual blocks are referred to as NNB in this paper.

V. EXPERIMENTS

In this section, we conducted experiments to assess the performance of our proposed model and compared the findings with previous research. We implemented the algorithms using the TensorFlow deep learning framework [52]. To accelerate the training process, we utilized the NVIDIA GeForce RTX 3060. The comprehensive block diagram depicting the experiments is presented in Fig. 4. Details of the experiments are to be described next.

A. DATA COLLECTION AND SAMPLE SELECTION

We collected a comprehensive set of infant cry samples from various open-source datasets, as outlined in Tab. 1. For the background noise, we utilized a part of the DESED dataset [53], [54], which contains recordings from domestic environments where infant crying events commonly occur.

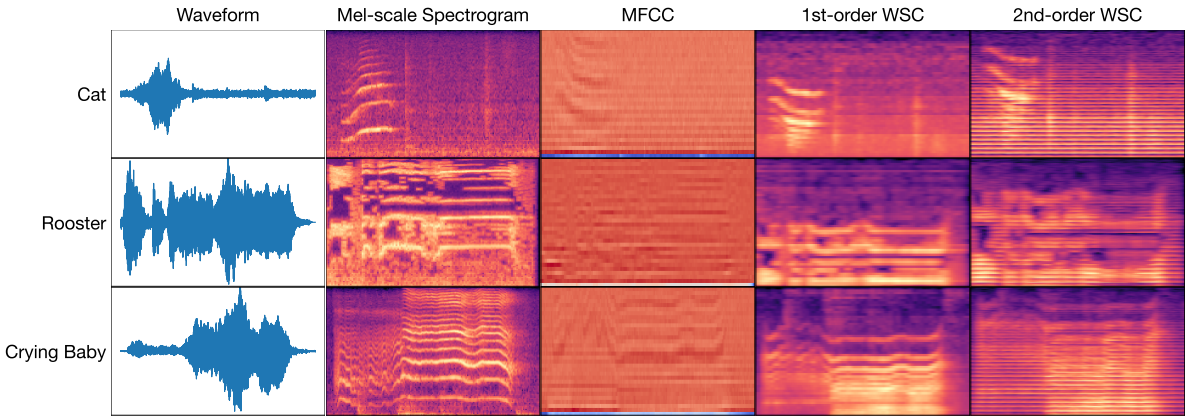

FIGURE 2. The first column displays the waveform, followed by the Mel-scale spectrogram in the second column, MFCC in the third column, first-order WSC in the fourth column, and second-order WSC in the fifth column. Each row corresponds to a specific sound class: the first represents a cat, the second represents a rooster, and the third showcases a crying baby. These sounds are associated with the respective files "1-47819-B-5.wav," "1-34119-B-1.wav," and "1-22694-A-20.wav" in ESC-50 dataset [48].

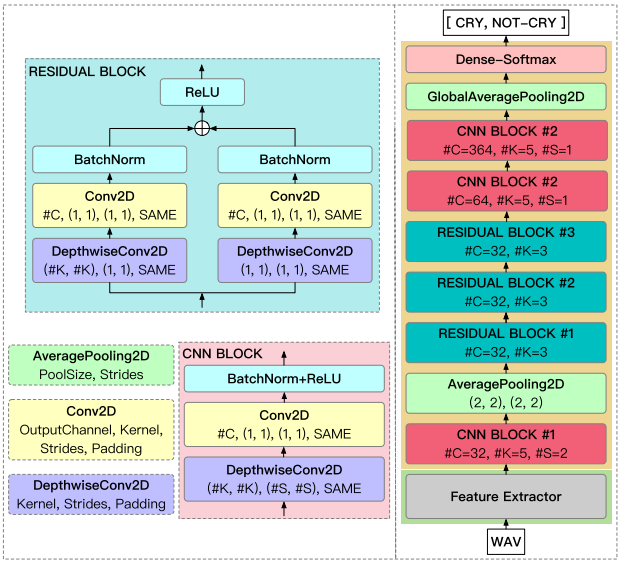

FIGURE 3. The proposed model. The left panel shows the blocks or layers of the model, and the right panel is the proposed structure. The parameters of each layer or block are also presented.

All the samples were classified into two categories: Cry and Not-Cry. Furthermore, we selected samples within a specific range of duration and volume for data augmentation.

In total, we gathered 9260 samples, which were then divided into three parts following an approximate ratio of 7:2:1: the training set (6346 samples), validation set (1956 samples), and test set (958 samples). It's important to note that we associated each sample with a specific part based on the hash of its file name, following the approach described by [55]. This method ensures the samples remain consistent across sets, enabling fair comparisons with other works. Among the 9260 samples, 4500 originate from the DESED dataset, while the remaining 5937 DESED samples were employed as background noise for data augmentation.

B. TRAINING PROCESS

Throughout the training process, the model's parameters were optimized using the training dataset. At every 100 training steps, the model's performance was assessed on the validation dataset. Ultimately, the model that attained the highest accuracy score on the validation dataset was selected for evaluation on the test dataset to determine its test accuracy score.

However, before directly feeding the training samples into the model, we applied three data augmentation techniques [58], [59], [60]. These policies transformed the original training samples into variations under different conditions, thereby improving the model's robustness and generalizability. These data augmentation techniques are considered "online" since they are performed during training.

To ensure consistency in the input audio sample duration, we fixed it at 4 seconds. This involved randomly extracting 4 seconds of data from longer-duration samples, while shorter-duration samples were padded with zeros before being passed to the model. For more information on the hyper-parameters specific to the data augmentation techniques and training process, please refer to Tab. 2.

C. THE EFFECTS OF THE INVARIANCE SCALES

As discussed in section III-E, we employed 8 wavelets per octave in the first-order wavelet scattering for the feature extractor of our proposed model. However, determining the appropriate invariance scale T is not universally standardized, and it typically varies across different applications [44]. The invariance scale plays a crucial role in audio classification tasks as it establishes the time scale of the scaling (lowpass) filter in wavelet scattering. Hence, we conducted several experiments using different values of T to make comparisons, and the corresponding results are presented in Fig. 5.

The results indicate that as the invariance scale T increased, the number of first-order WSC decreased. A smaller number of coefficients reduces the computational burden during

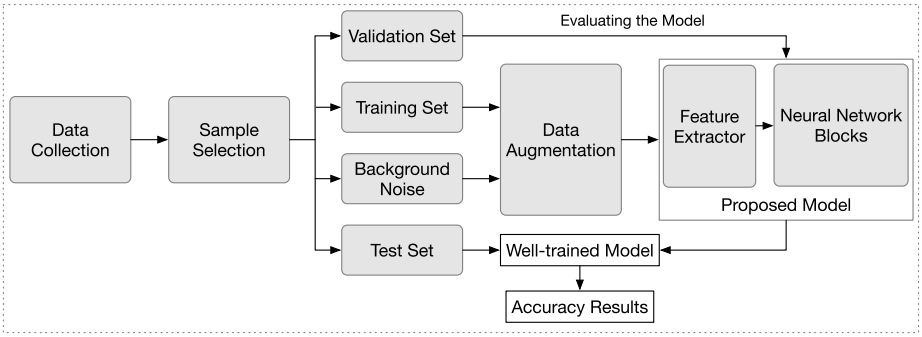

FIGURE 4. The block diagram of the experiments.

TABLE 1. The datasets used in the experiment.

Dataset	Class	Label[1]	Augmentation[2]	#Official	Constraints		#Used
					Duration (s)	Volume (dB)	
Baby cry detection project[3]	Crying baby	C	Y	108	$4 \sim 6$	$-30 \sim -10$	120
	Silence	NC	Y	108	$4 \sim 6$	$-50 \sim -30$	
	Noise	NC	N	108	$4 \sim 6$	$-50 \sim -30$	
	Baby laugh	NC	N	108	$4 \sim 6$	$-50 \sim -30$	
Donate a cry corpus[4]	hungry	C	Y	382	$6 \sim 8$	$-30 \sim -10$	273
	tired	C	Y	24			
	burping	C	Y	8			
	belly pain	C	Y	16			
	discomfort	C	Y	27			
iFLYTEK Challenge[5]	awake	C	Y	160	$5 \sim 20$	$-30 \sim -10$	852
	diaper	C	Y	134			
	hug	C	Y	160			
	hungry	C	Y	160			
	sleepy	C	Y	144			
	uncomfortable	C	Y	160			
	test set	C	Y	228			
ESC-50 [48]	crying baby	C	Y	40	$4 \sim 6$	$-30 \sim -10$	579
	other	NC	N	1960	$4 \sim 6$	$-50 \sim -30$	
VOICe[6] [56]	clean baby cry	C	Y	812	$2 \sim 5$	$-30 \sim -10$	1884
	snr-3 indoor	C	N	905	$2 \sim 5$	$-40 \sim -20$	
	snr-3 outdoor	C	N	908	$2 \sim 5$	$-40 \sim -20$	
	snr-3 vehicle	C	N	864	$2 \sim 5$	$-40 \sim -20$	
	snr-9 indoor	C	N	888	$2 \sim 5$	$-40 \sim -20$	
	snr-9 outdoor	C	N	853	$2 \sim 5$	$-40 \sim -20$	
	snr-9 vehicle	C	N	814	$2 \sim 5$	$-40 \sim -20$	
TUT Rare Sound Events 2017[7] [57]	ebr0	C	N	230	$2 \sim 5$	$-50 \sim -20$	1052
	ebr6	C	N	212	$2 \sim 5$	$-50 \sim -20$	
	ebr-6	C	N	219	$2 \sim 5$	$-50 \sim -20$	
	clean baby cry	C	Y	52	$5 \sim 50$	$-30 \sim -10$	
	background	NC	N	1508	$29 \sim 31$	$-50 \sim -30$	
DESED[8] [53], [54]	train: unlabel	NC	N	14412	$9 \sim 11$	$-50 \sim -20$	10437
	train: weak	NC	N	1578			
	validation	NC	N	1168			
						Total	15197
						#DESED for Classification	4500
						#DESED for Data Augmentation	5937

[1] C: Cry; NC: Not-Cry
[2] If adding DESED samples for data augmentation: Yes/No.
[3] https://github.com/giulbia/baby_cry_detection
[4] https://github.com/gveres/donateacry-corpus
[5] http://challenge.xfyun.cn/topic/info?type=baby-crying
[6] The cry clips were extracted from the mixture files.
[7] The cry clips were extracted from ebr0, ebr6, ebr-6 mixture files.
[8] There are 2346 missing files resulting from the unavailable URLs or internet problems.

the training process. While there was no linear correlation between the accuracy score and T, it was observed that smaller invariance scales (less than 128ms) tended to yield better performance, with accuracy scores surpassing 95%.

TABLE 2. The hyper-parameters of the training process and data augmentation.

Process	Hyper-parameters	Values
Training	Batch size	16
	Training steps	1000,1000,2000,2000
	Learning rate for steps	0.001,0.0008,0.0005,0.0001
	Weights initialization	Xavier
	Bias initialization	Zeros
	Loss function	Cross Entropy
	Optimizer	Adam
Augmentation in time	Timeshift	$-100 \sim 100$ms
	Background noise volume	$0.4 \sim 0.8$
	Crying volume	$0.7 \sim 1.3$
Augmentation in time-frequency	#Time masks	2
	Time mask size	$0 \sim 10$
	#Frequency masks	2
	Frequency mask size	$0 \sim 5$
Augmentation with Cutout [60]	#Masks	3
	Time mask size	$0 \sim 10$
	Frequency mask size	$0 \sim 5$

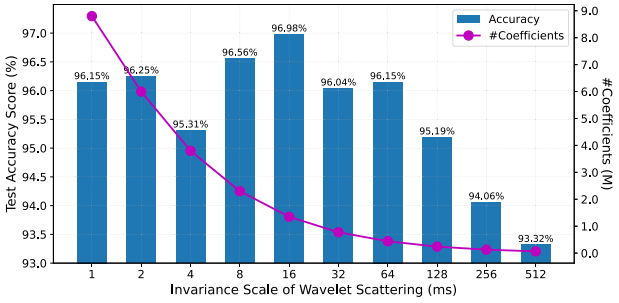

FIGURE 5. The test accuracy results and the number of the first-order WSC for the proposed models using different invariance scales.

TABLE 3. The hyper-parameters for STFT-based features in the ablation experiments.

Mel-scale Spectrogram		MFCC	
#Mel-filters	128	#Coefficients	32
Window type	hann		
Window length	32ms		
Overlap length	16ms		
FFT length	32ms		
Pre-emphasis	0.97		
Mel-filter lower frequency	20Hz		
Mel-filter upper frequency	8000Hz		

Based on these findings, we chose an invariance scale of $T = 16$ms for the subsequent experiments.

D. THE ABLATION EXPERIMENTS

To highlight the benefits of wavelet scattering, we conducted ablation experiments where the WSC of the feature extractor was substituted with STFT-based features. The hyper-parameters employed to extract the STFT-based features are presented in Tab. 3. The outcomes of these ablation experiments are outlined in Tab. 4 (under "Proposed-Mel" and "Proposed-MFCC" rows) and visualized in Fig. 6.

E. THE COMPARISON OF THE PREVIOUS WORK

Because previous work often evaluated performance on different datasets, we attempted to replicate their methods and assess their performance on our datasets (refer to Tab. 1). During this process, we aimed to closely align the model architectures and hyper-parameters for extracting acoustic features with the corresponding literature.[1] However, other strategies remained consistent with our paper, including data augmentation policies (see Tab. 2), hyper-parameters of the training process (see Tab. 2), and the evaluation method (refer to section V-B). Eventually, we compared these results with our proposed model, taking into account the model size and Multiply-Accumulate Operations (MACs), as indicated in Tab. 4 and Fig. 6.

VI. DISCUSSIONS

One of the advantages of WSC is that there is no requirement for the duration T of the analysis window. Fig. 5 illustrates that WSC can capture features within a duration of 128ms, as indicated by a test accuracy score higher than 95%. While the STFT method restricts the analysis window duration to a maximum of 32ms, as shown in Tab. 4. The results of the ablation experiments further confirm the effectiveness of WSC in capturing crucial acoustic features.

Another advantage of WSC is its flexibility in choosing the order of coefficients for different purposes. In this paper, our goal is to develop an efficient algorithm with low computational complexity suitable for resource-limited devices. As a result, we opted to utilize the first-order WSC as the acoustic feature, eliminating the need for additional wavelet scattering transformations. The remarkable results in Tab. 4 demonstrate that the first-order WSC is sufficient for extracting crucial features from infant cry sounds.

[1] Due to the potential unavailability of complete details in the literature, the reported results from previous work might not be entirely accurate, including aspects like the number of model parameters and MACs.

TABLE 4. The experimental results compared to those of previous work. The number of model parameters and MACs were calculated using NeSsi[1]. It is worth noting that the feature extraction operations of the previous work were also incorporated within the model, similar to our approach. Furthermore, the remaining hyper-parameters of the previous work for the Mel-scale spectrogram were aligned with the configuration presented in Tab. 3.

Reference	Feature Type	Hyper-parameters			#Parametes (M)	#MACs (M)	Test Accuracy
		Window (ms)	Overlap (ms)	#Filters			
[33]	Mel[2]	32	8	40	0.56	1024.39	96.35%
[32]	Mel	32	16	40	20.34	100.24	95.71%
[34]	Mel	20	10	40	32.47	160.03	96.25%
[6]	Mel+Delta	30	15	60	0.43	1124.33	96.35%
Proposed-Mel	Mel	32	16	128	0.017	42.07	95.62%
Proposed-MFCC	MFCC	32	16	32	0.017	17.75	95.31%
Proposed-WSC	WSC[3] (1st-order)	Invariance scale T: 16ms #Wavelets per octave: 8			0.017	22.74	96.98%

[1] https://github.com/AlbertoAncilotto/NeSsi#nessi
[2] Mel: Mel-scale spectrogram
[3] WSC: Wavelet Scattering Coefficients

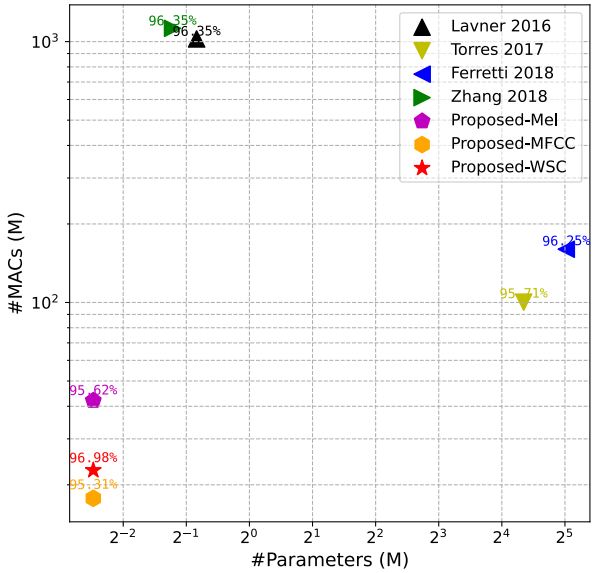

FIGURE 6. The comparison of the proposed model with the previous work regarding model size and MACs.

The structure of the proposed model is characterized by its simplicity and efficiency. Two key considerations guide the design of this structure. First, the incorporation of convolution leverages its powerful learning capability. Second, the utilization of residual connections makes deep learning possible. By combining these two components, the model achieves a parameter count of approximately 17K and a MAC count of around 23M. This aspect proves highly advantageous in reducing computational complexity, making the model well-suited for real-world applications.

Furthermore, based on Fig. 6, we can observe a lack of balance between the number of model parameters and the number of MACs in previous works. In other words, it has been challenging to simultaneously achieve both a reduced parameter count and a low MAC count along the traditional approach. However, this paper demonstrates the feasibility of an efficient model that achieves high-accuracy results while maintaining low computational complexity. This finding holds great significance for similar research.

During our experiments, we observed a phenomenon where certain noise was mistakenly identified as crying sound, especially in extremely noisy environments. Therefore, our current research is primarily focused on developing noise reduction techniques to address this issue. We intend to present the outcomes of our research in future work.

VII. CONCLUSION

This paper highlights the benefits of employing first-order WSC for extracting crucial features from infant cries with low computational complexity. The impressive accuracy result of 96.98% demonstrates the effectiveness of integrating wavelet scattering with a specialized NNB. This combination results in a lightweight end-to-end architecture comprising merely 17K parameters and 22.7M MACs. Such a design is highly advantageous for future engineering implementations.

REFERENCES

[1] V. Hiremath and P. Venkataratnam, "Automatic cradle system with measurement of baby's vital biological parameters," in *Proc. Int. Conf. Smart Technol. Smart Nation (SmartTechCon)*, Aug. 2017, pp. 480–485.
[2] W. A. Jabbar, H. K. Shang, S. N. I. S. Hamid, A. A. Almohammedi, R. M. Ramli, and M. A. H. Ali, "IoT-BBMS: Internet of Things-based baby monitoring system for smart cradle," *IEEE Access*, vol. 7, pp. 93791–93805, 2019.
[3] C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan, "A review of infant cry analysis and classification," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, pp. 1–17, Dec. 2021.
[4] A. F. Symon, N. Hassan, H. Rashid, I. U. Ahmed, and S. M. T. Reza, "Design and development of a smart baby monitoring system based on raspberry Pi and Pi camera," in *Proc. 4th Int. Conf. Adv. Electr. Eng. (ICAEE)*, Sep. 2017, pp. 117–122.
[5] M. P. Joshi and D. C. Mehetre, "IoT based smart cradle system with an Android app for baby monitoring," in *Proc. Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Aug. 2017, pp. 1–4.
[6] X. Zhang, Y. Zou, and Y. Liu, "Aicds: An infant crying detection system based on lightweight convolutional neural network," in *Proc. 7th Int. Conf. Artif. Intell. Mobile Services (AIMS)*, Seattle, WA, USA. Cham, Switzerland: Springer, Jun. 2018, pp. 185–196.

[7] R. Cohen, D. Ruinskiy, J. Zickfeld, H. IJzerman, and Y. Lavner, "Baby cry detection: Deep learning and classical approaches," in *Development and Analysis of Deep Learning Architectures*. Cham, Switzerland: Springer, 2020, pp. 171–196.

[8] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro, "Infant crying detection in real-world environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 131–135.

[9] R. Jahangir, "CNN-SCNet: A CNN net-based deep learning framework for infant cry detection in household setting," *Eng. Rep.*, Oct. 2023, Art. no. e12786.

[10] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020.

[11] L. R. Rabiner, *Digital Processing of Speech Signals*. Pearson Education India, 1978.

[12] J. Andén and S. Mallat, "Scattering representation of modulated sounds," in *Proc. 15th Int. Conf. Digit. Audio Effects (DAFx)*, vol. 9, 2012, pp. 17–21.

[13] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[14] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[15] E. Oyallon, S. Mallat, and L. Sifre, "Generic deep networks with wavelet scattering," 2013, *arXiv:1312.5940*.

[16] S. Mallat, "Understanding deep convolutional networks," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, Art. no. 20150203.

[17] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proc. Int. Soc. Music Inf. Retr. Conf.* Miami, FL, USA, 2011, pp. 657–662. [Online]. Available: https://api.semanticscholar.org/CorpusID:8275457

[18] J. Andén, V. Lostanlen, and S. Mallat, "Joint time-frequency scattering for audio classification," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[19] M. K. Reddy, Y. M. Keerthana, and P. Alku, "End-to-end pathological speech detection using wavelet scattering network," *IEEE Signal Process. Lett.*, vol. 29, pp. 1863–1867, 2022.

[20] S. Sharma, S. Asthana, and V. K. Mittal, "A database of infant cry sounds to study the likely cause of cry," in *Proc. 12th Int. Conf. Natural Lang. Process.*, 2015, pp. 112–117.

[21] M. S. Rusu, S. S. Diaconescu, G. Sardescu, and E. Bratila, "Database and system design for data collection of crying related to infant's needs and diseases," in *Proc. Int. Conf. Speech Technol. Human-Comput. Dialogue (SpeD)*, Oct. 2015, pp. 1–6.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[26] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[27] S. Majumdar and B. Ginsburg, "MatchboxNet: 1D time-channel separable convolutional neural network architecture for speech commands recognition," in *Proc. Interspeech*, Oct. 2020, pp. 3356–3360.

[28] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proc. Interspeech*, 2021.

[29] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, "Temporal convolution for real-time keyword spotting on mobile devices," 2019, *arXiv:1904.03814*.

[30] Z. Qiu Lin, A. G. Chung, and A. Wong, "EdgeSpeechNets: Highly efficient deep neural networks for speech recognition on the edge," 2018, *arXiv:1810.08559*.

[31] J. Lin, W.-M. Chen, Y. Lin, J. Cohn, C. Gan, and S. Han, "MCUNet: Tiny deep learning on IoT devices," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, Jul. 2020, pp. 11711–11722.

[32] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in *Proc. Int. Conf. Eng. Appl. Neural Netw. (EANN)*, Athens, Greece, Aug. 2017, pp. 168–179.

[33] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," in *Proc. IEEE Int. Conf. Sci. Electr. Eng. (ICSEE)*, Nov. 2016, pp. 1–5.

[34] D. Ferretti, M. Severini, E. Principi, A. Cenci, and S. Squartini, "Infant cry detection in adverse acoustic environments by using deep neural networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 992–996.

[35] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8667–8671.

[36] V. Bruni, M. L. Cardinali, and D. Vitulano, "An MDL-based wavelet scattering features selection for signal classification," *Axioms*, vol. 11, no. 8, p. 376, Jul. 2022.

[37] J. Bruna and S. Mallat, "Classification with scattering operators," in *Proc. CVPR*, Jun. 2011, pp. 1561–1566.

[38] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21413

[39] Z. Liu, G. Yao, Q. Zhang, J. Zhang, and X. Zeng, "Wavelet scattering transform for ECG beat classification," *Comput. Math. Methods Med.*, vol. 2020, Oct. 2020, Art. no. 3215681.

[40] A. Sepùlveda, F. Castillo, C. Palma, and M. Rodriguez-Fernandez, "Emotion recognition from ECG signals using wavelet scattering and machine learning," *Appl. Sci.*, vol. 11, no. 11, p. 4945, May 2021.

[41] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3704–3718, Jul. 2019.

[42] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5619–5628.

[43] B. Soro and C. Lee, "A wavelet scattering feature extraction approach for deep neural network based indoor fingerprinting localization," *Sensors*, vol. 19, no. 8, p. 1790, Apr. 2019.

[44] P. Singh, G. Saha, and M. Sahidullah, "Deep scattering network for speech emotion recognition," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 131–135.

[45] S. Dey, P. Singh, and G. Saha, "Wavelet scattering transform for improving generalization in low-resourced spoken language identification," 2023, *arXiv:2310.00602*.

[46] W. Ghezaiel, L. Brun, and O. LéZoray, "Wavelet scattering transform and CNN for closed set speaker identification," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[47] F. Cotter, "Uses of complex wavelets in deep convolutional neural networks," Ph.D. dissertation, Cambridge, MA, USA, Univ. Cambridge, Aug. 2019.

[48] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[49] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[51] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[52] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Design Implement.*, vol. 16, Nov. 2016, pp. 265–283.

[53] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 1–6.

[54] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 86–90.

[55] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.
[56] S. Gharib, K. Drossos, E. Fagerlund, and T. Virtanen, "VOICe: A sound event detection dataset for generalizable domain adaptation," 2019, *arXiv:1911.07098*.
[57] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2017, pp. 85–92.
[58] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming keyword spotting on mobile devices," 2020, *arXiv:2005.06720*.
[59] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617.
[60] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

HAITAO CAO received the B.Sc. and M.Sc. degrees from Guangzhou University, China, in 2013 and 2016, respectively, and the Ph.D. degree from the University of Padova, Italy, in 2020. He has three years of working experience with Midea, specializing in the research and development of AI speech processing. He is currently with Guangzhou Panyu Polytechnic, China. His research interests include wavelet theory and machine learning.

HAISHAN CHEN received the Ph.D. degree from Sun Yat-sen University, in 2017. He had seven years of enterprise work experience with Nortel and Ericsson and four years of work experience in universities. He is currently an Associate Professor with Guangzhou Panyu Polytechnic. His research interests include multimedia processing and deep learning, wherein he has published more than 15 articles.

JUNYING YUAN received the Ph.D. degree from Sun Yat-sen University, in 2023. She is currently the Associate Dean of the School of Electrical and Computer Engineering, Guangzhou Nanfang College, where she is also a Professor. Her research interests include multimedia processing and machine learning.