Is Your Baby Fine at Home? Baby Cry Sound Detection in Domestic Environments

Tanmay Khandelwal*†, Rohan Kumar Das* and Eng Siong Chng†
*Fortemedia Singapore, Singapore
†Nanyang Technological University, Singapore
E-mail: f20170106p@alumni.bits-pilani.ac.in, rohankd@fortemedia.com, ASESChng@ntu.edu.sg

***Abstract*—Baby cry detection in domestic environments is an essential component in baby monitoring systems, studying sleep cycle patterns in infants, and developing other diagnostic tools. In this work, we explore the state-of-the-art convolutional recurrent neural network (CRNN) and proposed replacement using depth-wise-separable (DWS) convolutions to have a low-complexity model for baby cry sound detection application. The studies are carried out on a dataset curated from AudioSet, which contains baby cry sounds and several other sounds that occur commonly in a domestic household environment. We also perform various data augmentation methods, and a few post-processing techniques to enhance the robustness of our baby cry sound detection system. The studies show that our low-complexity model developed using DWS achieves promising results by only using 3% of the parameters of the standard CRNN system, with the highest F-score of 0.738 on the test set.**

I. Introduction

Crying is the primary way for young infants to communicate and express their needs [1]. This makes infant cry detection a fundamental tool in modern diagnostic applications for research and modern baby monitoring systems designed to alert caregivers. Baby cry detection can be regarded as an application of sound event detection (SED) [2] with the goal to determine the temporal onset and offset of the target sound class in domestic environments. The SED has lately gained popularity and has a wide range of practical uses in smart-homes for security and surveillance [3].

In conventional baby monitoring devices, the monitoring is conducted based on the energy levels given as the input. This can be easily triggered by high-energy sounds, causing a false alert for the parent or caregiver. In the other sort of monitoring device (like walkie-talkies), the parents have to constantly listen to the receiver during their activities. Therefore, the SED-based systems to alert are preferred and can have a more accurate detection in domestic environments.

The research on SED witnessed rapid growth to analyze and recognize target sounds in real-world scenarios in recent years. The approaches to designing a system have shifted from the traditional methods like Gaussian mixture models (GMMs) [4], [5], hidden Markov models (HMMs) [6], and support vector machines (SVMs) [7] to advanced deep learning techniques [8]–[10]. With the recent success of convolutional neural networks (CNNs) in image recognition, they have become the state-of-the-art recipe in the domain of acoustic event detection and classification for feature extraction [11], [12]. This is because of the similarities between image inputs in computer vision and time-frequency representations of the audio signal. However, CNNs cannot store longer temporal context information. To alleviate this limitation, recurrent neural networks (RNNs) with the capacity to learn long temporal context information have been applied to SED. In addition, a hybrid network referred to as a convolutional recurrent neural network (CRNN) [13] was proposed to utilize the advantages of both CNNs and RNNs.

In real-world scenarios, the domestic environment contains various background sounds from vacuum cleaners, human speech, door opening/closing, and many others. The presence of background sounds having similar audio quality and frequency content affects the efficacy of the detection system. The impact is even more when the model is deployed for real-time detection on low computational devices. This shows the importance of having a robust, low-complexity model while designing SED applications. It is also found that the lack of availability of labelled data for training the SED system degrades the system performance. This projects the necessity of carrying out data augmentations and pre/post-processing to enhance the robustness of developed systems.

Literature shows that various studies explored depth-wise-separable (DWS) convolutions to derive low-complexity models with a very small number of parameters. Such approaches have been employed in the MobileNet [14] architecture for image recognition, QuartzNet [15] for automatic speech recognition, and MatchboxNet [16] for speech command recognition, all of which were built exclusively for mobile devices. It is noted that while a VGG16 [17] model takes up around 500 MB of disk space, MobileNet just takes 16-18 MB. Due to the small size, there can be a trade-off of accuracy in a few cases, but that is very minor in comparison to the benefits that we get from the small size of the model.

In this work, inspired by the success of DWS, we proposed to replace the standard 2D convolutions with depth-wise-separable (DWS) convolutions to reduce the number of parameters in a standard CRNN system. In addition, a few popular data augmentation approaches are used to increase the system's robustness. The studies are conducted on a database, which we curated from AudioSet, that contains baby cry sounds and several other sounds that occur commonly in a domestic household environment. The contributions of this work can be summarized as follows:

978-616-590-477-3 ©2022 APSIPA

- Development of a baby cry sound detection dataset in the domestic environment
- Proposal of low-complexity CRNN models for baby cry detection
- Explore various data augmentations for system robustness

The remainder of the paper is structured as follows: Section II introduces the neural network architectures studied to detect baby cry sounds in domestic environments. In Section III, the details of the experimental setup are described. The results and analysis are reported in Section IV. Finally, Section V concludes the work.

II. Convolutional Recurrent Neural Networks

The hybrid system of CNN and RNN is termed the CRNN, which remains state-of-the-art in the polyphonic SED task. We use this CRNN network as a reference system for our studies in this work. We describe the architectures of the baseline model and the proposed low-complexity models in detail in the following subsections.

A. Baseline Model

The 5-Layer CRNN architecture used as the baseline (CNN-5) [18] is depicted in Fig. 1. It consists of four 2D convolutional blocks (Conv-Blocks) with a kernel size of 5×5. In Fig. 1, 5×5 @ 64 depicts a convolutional-block (Conv-Block) employing standard convolution with kernel size of 5×5 and output feature map with size of 64. In the Conv-Block of Fig. 2, batch normalization is applied after each 2D convolution. Thereafter, ReLU function is used as the non-linear activation function. To reduce the feature map size, we used average pooling of 2×2 after each 2D Conv-Block. The baseline (CNN-5) consists of 64, 128, 256, and 512 feature maps in each consecutive block. To learn the temporal context, the Conv-Blocks are followed by a bidirectional gated recurrent unit (Bi-GRU) [19] with 256 hidden units. Finally, the feed-forward layers perform the classification by producing sound event activity probabilities based on the input from the recurrent layer. The total number of parameters (PN) of the baseline is 4.3M. The model learns using gradient-based optimization to minimize the binary cross-entropy loss function (l_{BCE}) given in Eq. (1), where y_k is the label (1 for 'baby-cry' and 0 for 'other') and p_k is the predicted probability of the point being 'baby-cry' for all K points.

$$l_{BCE}(p, y) = \sum_{n=1}^{K} [y_k \ln p_k + (1 - y_k) \ln(1 - p_k)] \quad (1)$$

B. Low-complexity Models

Conventional CNNs have a high computational cost that makes them unsuitable for use in mobile vision and embedded applications. In the CRNN network proposed in this work, depicted in Fig. 1, we replaced the 2D CNN utilized in the baseline with 2D DWS convolutions [20]–[22] which resulted in a low-complexity feature extractor. The DWS convolution, as shown in Fig. 3 factorizes the standard convolution into depth-wise convolution and point-wise convolution. The difference in the complexities of these two convolutional operations is also depicted in Fig. 3. Using those values, we can calculate the Ratio (R) of Complexity (C) of DWS convolution to standard convolution as:

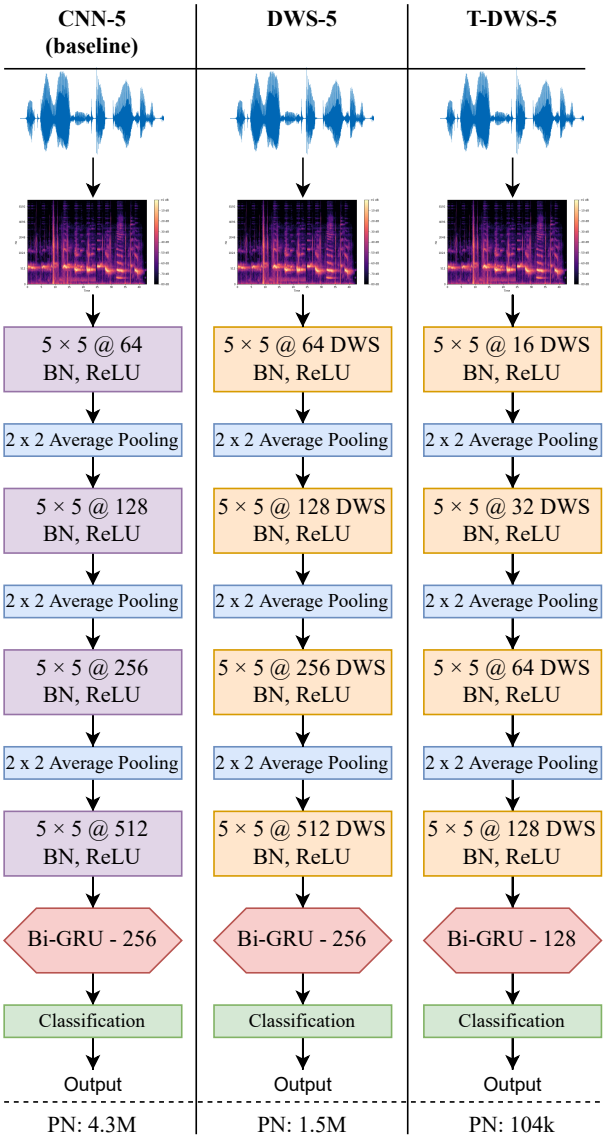

Fig. 1. Architecture of the baseline CNN-5 and the proposed low-complexity models DWS-5 and Tiny (T)-DWS-5.

$$R = \frac{C(\text{DWS Convolution})}{C(\text{Standard Convolution})} = \frac{O^2 \times C_{in} \times (k^2 + C_{out})}{O^2 \times C_{in} \times k^2 \times C_{out}} = \frac{1}{k^2} + \frac{1}{C_{out}} \quad (2)$$

Thus, for the case where $C_{out} = 100$ and $k = 512$, we get Ratio $(R) = 0.01$. This indicates the DWS convolution performs around 100 times fewer multiplications compared to the standard convolution for this case. As a result, we can attain comparable accuracy with less complexity. To investigate

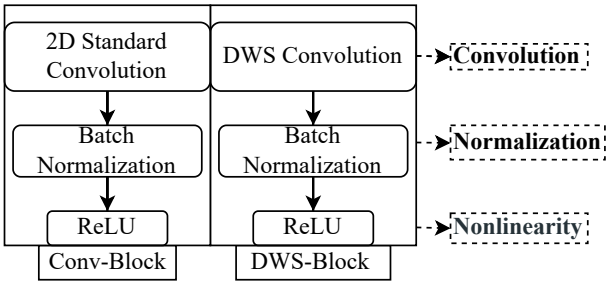

Fig. 2. The structural comparison between Conv-Block and DWS-Block.

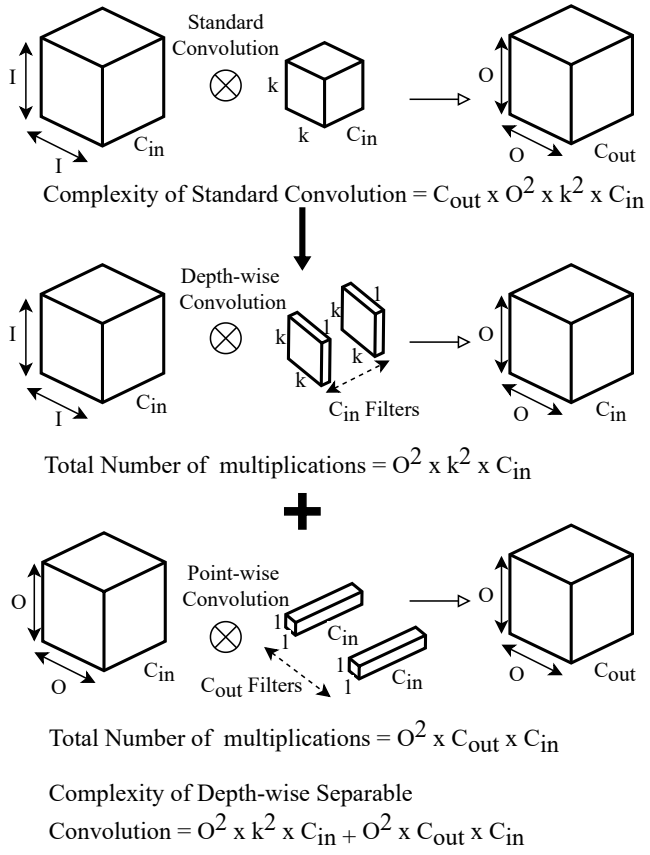

Fig. 3. The comparison between complexities and operations of standard convolution and DWS convolution.

and evaluate the performance of DWS-based convolution with the baseline, we constructed two DWS convolution-based architectures termed DWS-5 and tiny (T)-DWS-5. It is noted that T-DWS-5 has only 6% of the total parameters of DWS-5, as shown in Fig. 1. The 5×5 @ 64 - DWS of Fig. 1 depicts a DWS-block as in Fig. 2 employing DWS convolution with a kernel size of 5×5 and output feature map with a size of 64. Both the architectures used, have four DWS blocks but have a different number of channels in each consecutive block and a different number of hidden cells in the Bi-GRU. The DWS-5 contains four DWS blocks with a kernel size of, 5×5 consisting of the same number of feature maps as in the baseline. As a result, DWS-5 has a total of 1.5M parameters. Similarly, T-DWS-5 has the same number of DWS blocks with the kernel size of 5×5 consisting of 16, 32, 64, and 128 feature maps in each block respectively. The Bi-GRU cell in T-DWS-5 consists of 128 hidden units, resulting in a total of 104k parameters. To summarize, the DWS-5 and T-DWS-5 have 36% and 6% of the parameters of the considered baseline (CNN-5), respectively.

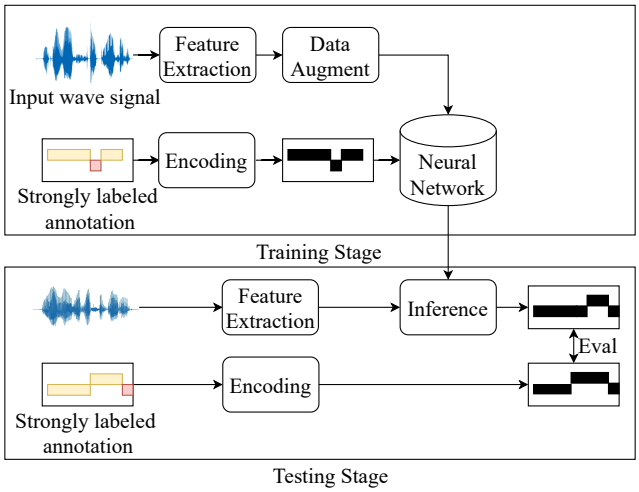

Fig. 4. The detailed framework for baby cry SED system.

III. Experimental Setup

The detailed framework of the developed SED system is illustrated in Fig. 4. It can be viewed under two stages: the training stage and the testing stage. In the training stage, the system uses encoded strongly labelled annotations to train the SED model. The input audio signals are processed to extract mel-spectrograms. We also apply various data augmentation methods to increase the amount of training data and variability. During the testing stage, the trained model is used to infer on the test set and compare it with ground truth annotations to evaluate the performance. This section further contains the database details and other experimental settings in the following subsections.

A. Database

The dataset constructed for this study is a subset of the AudioSet [23]. The AudioSet is a growing ontology comprising 632 audio event classes of human labelled clips with a duration of 10 seconds, derived from YouTube videos. The curated dataset is divided into the development set and evaluation set, incorporating the previously released strongly labelled annotations for the subset of AudioSet. The development dataset is further split into the training (Train) and validation (Val) sets. Since the focus of this work is to detect baby cry sounds in domestic environments, the dataset has been categorized into two primary classes: 'baby-cry' and 'other'. The 'other' class comprises various sound clips from the domestic household environment. The recurrence of the events in the labels for each clip does not overlap. It is also noted that the dataset contains an almost equal distribution of clips between 'baby-cry' and 'other' classes.

After filtering the metadata files for the specific classes from the AudioSet ontology, the dataset was downloaded using

TABLE I
DATA DISTRIBUTION SUMMARY OF THE STRONGLY LABELLED BABY CRY DATASET CURATED FROM AUDIOSET.

Class	Development		Evaluation
	Train	Val	Test
baby-cry	492	39	25
other	480	39	40
Total	972	78	65

YouTube-DL. The audio segments are then trimmed using FFmpeg in accordance with the onset and offset specified in the filtered metadata. The audio clips shorter than 10 seconds are padded with silence to make them 10 seconds long. The dataset described in this paper has been made available on GitHub[1]. Table I summarizes the composition of the database.

B. Pre-Processing and Feature Extraction

The audio clips are re-sampled to 32,000 Hz mono channel audio waveform and packed into hdf5 files, to speed up the training process. Following that, they are segmented using a window size (WS) with consecutive frames of 1024 samples and a hop length of 500 samples. Additionally, we experimented with different window sizes, keeping the hop length constant. Then, the segmented waveforms are transformed into spectrograms by performing Short Time Fourier Transform (STFT). The log-mel spectrogram was obtained by applying 64 mel-filters spanned from 0 to 8kHz in the frequency domain followed by logarithmic operation using librosa. The extracted mel-spectrograms are normalized with global mean and standard deviation over the entire training set. Thus, with these values, each audio signal is represented by 640 frames. With 64 mel-bins and 640 frames, each audio clip is represented with an input dimension of $1 \times 640 \times 64$.

C. Data Augmentation

Data augmentation methods have been proven well to increase the system robustness, which motivated us to apply them in the development of our SED system. We used SpecAugment [24] in all the experiments, including the baselines, which is directly applied to the feature inputs of the neural network. The SpecAugment is widely used in tasks such as speech recognition, where it masks blocks of consecutive frequency channels and time frames. Additionally, we experimented with Time-Shift [25], Pitch-Shift [26], Soft-Mixup [27], and Hard-Mixup [28] as a few other data augmentation methods in this work. The Time-shift shifts the audio clip along the time axis circularly, whereas the Pitch-Shift randomly raises or lowers the pitch of the audio clip. The Hard-Mixup directly adds all the sound clips and labels the mixture with all the classes in the original samples. On the other hand, the Soft Mixup generates pseudo data $(\hat{x}, \hat{y})$ by mixing different data points (x_i, x_j) and their corresponding labels (y_i, y_j) chosen at random as shown below:

$$\hat{x} = \lambda x_i + (1 - \lambda) x_j \tag{3}$$

$$\hat{y} = \lambda y_i + (1 - \lambda) y_j \tag{4}$$

where $\lambda \in [0, 1]$ determines the degree of mixing.

D. Detection Procedure and Post-Processing

We investigated different post-processing steps through smoothing, normalizing, and thresholding on the clip-wise and frame-wise event probabilities. In the first step, we applied a constant threshold of 0.5 to the output (y) of the feed-forward layer to get the binarized results, as depicted below:

$$y = \begin{cases} 1, & \text{if } y \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

In the second step, as different sound events have their own event duration, we experimented with median filtering (MF) by manually searching for the optimal length from 1 to 49 with an increment of 2 for the 'baby-cry' class. After getting the smooth binarized results, we compare the results for the 640 segmented frames with the ground truth labels for performance evaluation. We also experimented with normalizing mel-spectrogram (N-Mel) and reducing noise (NR) in the audio waveform. Further, we experimented with threshold values ranging from 0.1 to 0.9 in an increment of 0.1.

E. SED System Development

The weights of the SED model were initialized using Xavier initialization and all the biases were initialized to zero. The model is implemented in PyTorch and trained using Adam-optimizer (β1=0.9, β2=0.999, ϵ=1e-8, decay=0.0) with a batch size of 32. The learning rate was initialized to 0.001 and then subsequently reduced every 200 iterations by multiplying by 0.9 for a total of 2000 iterations. After each iteration over the training set, we assessed the loss on the validation set. Finally, after completion of training, we evaluated the model on the strongly labelled testing set.

F. Evaluation Metric

We used segment-based evaluation to compare the system output and the reference labels on a fixed temporal duration. This metric is more tolerant compared to the event-based metric to brief pauses and errors in event boundaries. At each segment level, we tallied the detected events in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these values we calculated Precision (P), Recall (R), and F-score (F) for each experiment for comparison, as depicted below:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \tag{6}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP + FN}} \tag{7}$$

$$\mathrm{F} = \frac{2\mathrm{PR}}{\mathrm{P + R}} \tag{8}$$

We placed more emphasis on the F-score as it elegantly summarizes the predictive performance of a model by combining recall and precision, which are otherwise conflicting metrics. Also, we focused on recall so that the system does not miss out on the baby-cry sound.

[1]https://github.com/tanmayy24/Baby_Cry_Detection_Database

TABLE II
PERFORMANCE OF THE BASELINE WITH THE PROPOSED MODELS.

Model	CNN-5 (Baseline)	DWS-5	T-DWS-5
F	0.641	**0.681**	0.666
P	0.829	0.765	0.780
R	0.522	0.613	0.580

TABLE III
EFFECT OF WS IN PERFORMANCE FOR THE PROPOSED MODELS.

WS	512	1024	2048	4096
DWS-5				
F	0.656	0.681	**0.711**	0.620
P	0.836	0.765	0.785	0.748
R	0.539	0.613	0.649	0.528
T-DWS-5				
F	0.559	0.666	**0.674**	0.608
P	0.775	0.780	0.752	0.774
R	0.437	0.580	0.610	0.500

IV. RESULTS AND ANALYSIS

This section presents the results for the studies related to baby cry sound detection in the following subsections.

A. SED System Comparison

In this subsection, we compare the performance of the baseline systems depicted in Fig. 1. As presented in Table II, both of the proposed models DWS-5 and T-DWS-5 perform better on segment-based F-score compared to the CNN-5 baseline system. When compared to CNN-5, the DWS-5 has a 6.24% higher F-score, while T-DWS-5 has a 3.9% higher F-score. Considering the parameters in each, we can observe that DWS-based models outperform the standard CRNN even with fewer parameters. Additionally, the DWS-based models have a higher recall, resulting in more output-sensitive predictions for the baby-cry event.

B. Effect of Window Size

As the average duration of the different sound events varies, we are interested to investigate different window sizes used to segment the audio waveform. Table III reports this analysis, which shows that the maximum F-score and Recall are obtained with a WS of 2048. As a result, with the highest sensitivity for a WS of 2048, the DWS-based models are correctly identifying most of the baby cry events.

C. Effect of Data Augmentation

We carried out various data augmentation methods described in Section III-C on WS of 2048-based proposed models apart from SpecAugment, which is already included in the development of all the systems. Table IV shows the results under this study, where the SpecAugment augmentation method is used in conjunction with various other data augmentation methods during training. We observe from Table IV that the Soft-Mixup provides the best results among the rest of the methods. The Soft-Mixup improved the F-score for both the DWS-based models, while also providing the highest recall among the other data augmentation methods. Next, we are interested in carrying out a few post-processing techniques with this Soft-Mixup with SpecAugment-based best performing system.

TABLE IV
PERFORMANCE COMPARISON WITH VARIOUS DATA AUGMENTATION METHODS FOR THE PROPOSED MODELS WITH WS OF 2048. THE SPECAUGMENT METHOD IS USED IN ALL THE CASES.

Method	No Augmentation	Soft-Mixup	Hard-Mixup	Time-Shift	Pitch-Shift
DWS-5					
F	0.711	**0.764**	0.264	0.563	0.670
P	0.785	0.777	0.928	0.767	0.795
R	0.649	0.750	0.154	0.444	0.578
T-DWS-5					
F	0.674	**0.727**	0.625	0.629	0.706
P	0.752	0.811	0.774	0.712	0.791
R	0.610	0.659	0.523	0.563	0.637

TABLE V
PERFORMANCE COMPARISON OF THE POST-PROCESSING TECHNIQUES NORMALISED-MEL (N-MEL), NOISE-REDUCTION (NR), MEDIAN-FILTERING (MF) AND MANUAL THRESHOLDING (MT) FOR THE PROPOSED MODELS WITH THE BEST DATA ARGUMENTATION (SPECAUGMENT AND SOFT-MIXUP) CASE UNDER WS OF 2048.

Method	None	N-Mel	NR	MF	MT
DWS-5					
F	0.764	0.719	0.744	**0.764**	0.760
P	0.777	0.789	0.775	0.768	0.734
R	0.750	0.660	0.715	0.758	0.788
T-DWS-5					
F	0.727	0.706	0.700	0.728	**0.738**
P	0.811	0.807	0.820	0.801	0.733
R	0.659	0.626	0.610	0.667	0.743

D. Effect of Post-Processing Techniques

In this subsection, we discuss the studies of various post-processing techniques discussed in Section III-D. From Table V, we observe that by applying median filtering to smooth the output event probabilities and a configurable threshold setting for the baby cry class, we obtain the best performance in F-score for DWS-5 and T-DWS-5, respectively. Based on our studies, the optimal value for the median window is 31 and the threshold value for the baby cry class is 0.4. It is also noted that the highest recall of 0.788 for DWS-5 and 0.743 for T-DWS-5 are achieved utilizing manual thresholding.

V. Conclusions

In this work, we focused on developing a low-complexity model for baby cry detection in domestic environments. To construct a low-complexity model, we proposed to replace the standard 2D CNNs with DWS convolutions for the baby cry detection system. We curated a dataset from the recently released strongly labelled subset from AudioSet for our studies. The studies depict the DWS-based models perform effectively with systems having much a lower number of parameters. The studies for data augmentation highlighted SpecAugment with Soft-Mixup as the best-performing model. Post-processing techniques such as median filtering and manual thresholding further improved the F-score for baby cry detection. In summary, this led to a 97% reduction in the number of parameters and a reduction in the average time required per iteration for training processes, with a superior classification performance of 15.13% than the state-of-the-art CRNN.

References

[1] L. L. Lagasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: Acoustic cry analysis and parental perception.," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, pp. 83–93, 2005.

[2] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[4] L. Vuegen, B. V. D. Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. V. hamme, "An MFCC-GMM approach for event detection and classification," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *European Signal Processing Conference (EUSIPCO)*, pp. 1267–1271, 2010.

[6] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17–20, 2008.

[7] L. Lu, F. Ge, Q. Zhao, and Y. Yan, "A SVM-based audio event detection system," *International Conference on Electrical and Control Engineering (ICECE)*, pp. 292–295, 2010.

[8] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, "Sound event detection and separation: A benchmark on DESED synthetic soundscapes," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 840–844, 2021.

[9] N. Tonami, K. Imoto, Y. Okamoto, T. Fukumori, and Y. Yamashita, "Sound event detection based on curriculum learning considering learning difficulty of events," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 875–879, 2021.

[10] J. Wu, Z. Pan, M. Zhang, R. K. Das, Y. Chua, and H. Li, "Robust sound recognition: A neuromorphic approach," *Interspeech*, pp. 3667–3668, 2019.

[11] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," *International Society of Music Information Retrieval (ISMIR)*, pp. 805–811, 2016.

[12] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge baseline with convolutional neural networks," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2018.

[13] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1291–1303, 2017.

[14] 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," *ArXiv*, vol. abs/1704.04861, 2017.

[15] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6124–6128, 2020.

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

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019.

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS Workshop on Deep Learning*, 2014.

[20] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *ArXiv*, vol. abs/1403.1687, 2014.

[21] J. Guo, Y. Li, W. Lin, Y. Chen, and J. Li, "Network decoupling: From regular to depthwise separable convolutions," *British Machine Vision Conference (BMVC)*, 2018.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2016.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[24] 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," *Interspeech*, pp. 2613–2617, 2019.

[25] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4 technical report," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019.

[26] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation," *International Society for Music Information Retrieval (ISMIR)*, pp. 248–254, 2015.

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *ArXiv*, vol. abs/1710.09412, 2017.

[28] N. Shao, E. Loweimi, and X. Li, "RCT: Random consistency training for semi-supervised sound event detection," *ArXiv*, vol. abs/2110.11144, 2021.