INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Attentive Convolutional Recurrent Neural Network Using Phoneme-Level Acoustic Representation for Rare Sound Event Detection

Shreya G. Upadhyay[1,2], *Bo-Hao Su*[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare

shreyaupadhyay10@gmail.com, borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

A well-trained Acoustic Sound Event Detection system captures the patterns of the sound to accurately detect events of interest in an auditory scene, which enables applications across domains of multimedia, smart living, and even health monitoring. Due to the scarcity and the weak labelling nature of the sound event data, it is often challenging to train an accurate and robust acoustic event detection model directly, especially for those rare occurrences. In this paper, we proposed an architecture which takes the advantage of integrating ASR network representations as additional input when training a sound event detector. Here we used the convolutional bi-directional recurrent neural network (CBRNN), which includes both spectral and temporal attentions, as the SED classifier and further combined the ASR feature representations when performing the end-to-end CBRNN training. Our experiments on the TUT 2017 rare sound event detection dataset showed that with the inclusion of ASR features, the overall discriminative performance of the end-to-end sound event detection system has improved; the average performance of our proposed framework in terms of f-score and error rates are 97 % and 0.05 % respectively.

Index Terms: sound event detection, convolution recurrent neural network, attention, automatic speech recognition

1. Introduction

Sound event detection (SED) is one of the emerging topics in research that aims at developing algorithms to automatically detect sound events present in an auditory scene. SED has found its use in many applications, including multimedia retrieval [1], context based indexing, automated surveillance systems, and unobtrusive health monitoring [2]. Recent works in SED has focused mainly on identifying improved neural architectures composed with network blocks, such as convolutional neural networks (CNN) [3] and recurrent neural networks (RNN) [4], that models time series sound data and spectral information. Several recent works has shown that using a hybrid composition of these network building blocks, such as long short-term memory(LSTM) with hidden Markov model (HMM) [5], the capsule network model as Bi-LSTM [6], convolution recurrent neural network (CRNN) [7] and convolutional bi-directional recurrent neural network (CBRNN) [8] has all been developed for SED and serve as the recent state-of-the-arts.

Compared to processing other acoustic data, such as music and speech, SED, however, presents its own unique technical challenges due to the limited amount of available information with weakly labelled sound data. Another characteristic of the SED is that the rare events are often important in computational auditory scene analysis, e.g., detecting emergency events with high accuracy. Hence, despite of the effort of developing complex models to improve SED accuracy, the knowledge transfer approach based on generating synthetic sound data has also been considered in [9] to avoid the time-consuming manual labelling process of real world sound data; using a transfer learning network [10] that takes speech data as the source to learn appropriate feature extractor for sound event data has also been shown to demonstrate improved detection accuracy.

While research has started to develop different knowledge transfer approaches in order to mitigate the issue of weak labeling and rare event robustness in SED, most of these works are still limited in using the available SED databases only without considering to leverage other types of high source data, e.g., speech data used for automatic speech recognition (ASR), where not only it is much larger in scale but also has the detailed acoustic properties (phonemic-acoustic characterization) which can be much better studied and well understood. In this paper, we proposed a gradual network architecture that integrates the phoneme-level acoustic representations obtained from a pre-trained ASR for the task of improving SED performance. Our main SED network architecture is a convolutional bidirectional-RNN (CBRNN), i.e., inclusion of both temporal and spectral attention mechanism. The pre-trained ASR acoustic model is based on factorized time-delay neural network (TDNN-F) [11] trained on the Librispeech corpus [12] (resulting in a WER of 3.76%). Our gradual network architecture takes the concatenation of the pre-final layer of ASR features (i.e., from the TDNN-F) and the conventional sound event representations learned from the target SED database together as an input, and the network only updates the SED path while keeps the ASR-based representation frozen.

In this work, we evaluated our framework on the DCASE2017 task 2 dataset [13] which is used often in works for detecting rare sound events, and further compared it with variants of CRNN model proposed recently in different papers as the state-of-the-art. The experiment results surpass the conventional CBRNN model achieving F1 score of 98.0% in baby-cry event, 95.4% in gunshot event and also 97.0% in average performance. This boost in accuracy over the state-of-the-art demonstrated by integrating phoneme-level acoustic representations as predicted by a pre-trained ASR system, it indeed is beneficial in enhancing the capacity in the end-to-end training of SED model. We further analyzed the phoneme distribution of each event to understand how each sound event class is mapped to the respected phonemic class of the pre-trained ASR model.

The rest of the paper is organized as follows: the Section 2 describes the methodology part, which includes detailed description of dataset, feature extraction, the proposed architecture, post-processing steps and the considered metrics. Section 3 illustrates the experimental setup for ASR and SED training with result and analysis done on task 2 of DCASE 2017 dataset. Section 4 draws the conclusion of this work.

http://dx.doi.org/10.21437/Interspeech.2020-2585

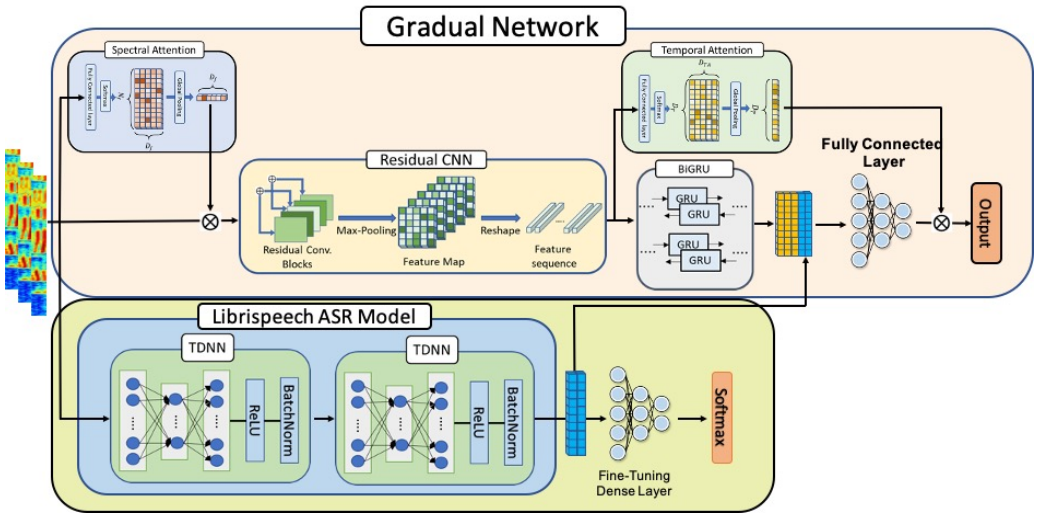

Figure 1: *An illustration of the proposed architecture; the proposed gradual network which concatenates the ASR network representations with the sound feature representations in path of end-to-end SED training.*

2. Methodology

2.1. Dataset

The training of the SED model is done with the task 2 of DCASE 2017 sound data, which consists of monophonic isolated sound events for each target class; it is mixed with the background recording of everyday acoustic scene [13]. In this dataset, there are three target events as: "babycry", "glassbreak", and "gunshot" with background audio contains 15 different scene recordings. They have provided the synthesizer code with synthesized audios which can be used to generate 1500 different mixtures for each class. The event to background ratios are -6, 0, 6dB, the event occurrence probability of 0.5 and all the mixtures are of a 30sec monaural audios. The dataset comprises of two sets, the development set and the evaluation set and the detail information of the task and the dataset is described in [13]. Here we used 90% of training data to train the model with 10% of it is used for validation purpose to prevent over-fitting, and evaluation is done on the given evaluation set. Before further processing, we first converted the sampling rate to 16kHz.

2.2. Feature Extraction

For SED training, the log Mel-filter bank (Fbank) energies with 40 Mel scale filtered acoustic features are used in this work. The sampling rate is set to 16kHz and are extracted within frame size of 20ms with 50% overlapping. The extracted features are normalized using min-max normalization and down-sampled by taking mean of five samples at a time. Here we used the frame concatenation with context window size of 2 on the input features, and after applying context expansion, we obtained a total of 200 concatenated Fbank features for each frame.

2.3. Proposed Model

2.3.1. Sound Event Detection (SED) Model

In this work, we used the CBRNN model as our SED binary classifier for each event. The input of this system is the extracted Fbank features with multi-frame concatenation [15], and the system output is the binary prediction for each frame of size 100ms. Here the CBRNN [8] structure consists of three parts the CNN, RNN and the classification layer.

For the CNN part, the acoustic features are first fed to the consecutive convolution layers and each convolution layer is followed by batch normalization [16] for each feature map, linear activation function and the dropout layer[17]. The max pooling layer is used here to extract the important features on each feature map for both the axes. At the end of the CNN part, the output feature map from CNN is stacked along the frequency axis. The CNN model with max pooling can be described here using function f extracting n features from a data patch w:

$$f(\omega) = \sigma(\sum_{n}(W_1 * \omega + b_1), ..., \sigma(\sum_{n}(W_n * \omega + b_n)) \quad (1)$$

$$y_t = \sigma(W_{hy} h_t + b_y) \quad (2)$$

where b_1,...,b_n are biases and W_1,W_N shows the weight matrices.

In the RNN (GRU) part, the stacked output from the CNN is fed to the recurrent layers followed by activation function. Here, each recurrent layer produces the frame-wise output by taking CNN layer output and the previous frame activation as an input. For each frame the total activation of the GRU layer is the interpolation of previous activation h_{t-1} and the candidate activation h_t. The GRU layer takes a sequence $(x_1, \ldots, x_t)$ as an input and produces a sequence $(h_1, \ldots, h_t)$ of hidden states with a sequence $(y_1, \ldots, y_t)$ as outputs (described in the following equation):

$$h_t = \sigma(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \quad (3)$$

$$y_t = \sigma(W_{hy} h_t + b_y) \quad (4)$$

where σ is a logistic sigmoid function and W_{xh},W_{hh},W_{hy} are weight matrices and b_h,b_y shows the bias.

For classification layer, the output of the bidirectional GRU is fed to the fully connected layer to produce the classification result for each frame, and the sigmoid activation function is used to normalize the probabilities. After the CNN layers, each output frame contains 100ms information, hence, the prediction is for each 100ms data.

Table 1: *System performance of our proposed model as compared with other state-of-the-art models shown in terms of Error Rate and F-score.*

model	CRNN+TA+RA[8]	CRNN+TA[8]	1D-CRNN[14]	CRNN[7]	Finetuning Net	Gradual Net
Babycry	0.18|91.3	0.25|87.4	0.15|92.2	0.18|90.8	0.20|88.6	**0.04**|**98.0**
Glassbreak	**0.04**|**98.2**	0.05|97.4	0.05|97.6	0.10|94.7	0.32|83.5	0.04|97.8
Gunshot	0.17|90.8	0.18|90.6	0.19|89.6	0.23|87.4	0.38|79.0	**0.06**|**95.4**
Average	0.13|93.4	0.16|91.8	0.13|93.4	0.17|91.0	0.30|83.7	**0.05**|**97.0**

2.3.2. Attention Mechanism

In this work we applied two aspects of attention: spectral and temporal attention. The temporal attention is applied here to assign different weights to the positive and negative frames, i.e., to attend to most relevant event occurring frames. Here the CNN output is passed to the fully connected layer with hidden unit N_t followed by activation function and then the global max-pooling is used on the frequency axis to obtain the weights for each frame. After obtaining the weights the element-wise multiplication is done with the output of the fully connected layer. The attention weights is computed as follows:

$$\hat{A}_{n,t} = \max_{n\epsilon\{1,2,..N_t\}} \{\sigma(W_n C_t + b_n)\} \tag{5}$$

where W_n is weights and b_n is bias for hidden units, C_t is output from the CNN, $\hat{A}_{n,t}$ is a temporal attention weights and σ is a activation function.

We also used spectral attentions [8] to assign weights to different spectral characteristic of the frequency component at each frame. The attention weights are calculated using the same structure as the temporal attention, the input of the CNN model is first passed to the fully connected layer with hidden unit N_s followed by activation function and then the global pooling is used on the time axis to obtain attention weights. Then, the element-wise multiplication is done with the input data to give important weights to the spectral characteristics. The weighted features are computed as:

$$\hat{A}_{n,t} = \max_{n\epsilon\{1,2,..N_s\}} \{\sigma(W_n S_t + b_n)\}, \tag{6}$$

$$\hat{S}_t = S_t \otimes \hat{A}_t \tag{7}$$

where $\hat{A}_{n,t}$ is a spectral attention weights, W_n is weights and b_n is bias for hidden units, S_t is a acoustic input feature, σ is a activation function and $\hat{S}_t$ is the weighted feature.

2.3.3. Automatic Speech Recognition (ASR) Model

In our proposed model, we use the factorized time-delay deep neural network (TDNN-F) [11] as our main acoustic modelling structure that is pre-trained on the Librispeech dataset [12]. Unlike the conventional TDNN model, TDNN-F applies the low-rank factorized layers to the TDNN which would dramatically decrease the loading of training parameters with better performances. This pre-train model obtains a word error rate (WER) of 3.76 % on the test-clean and 8.92% on the test-other of Librispeech database. Here, we extracted the event detection audio features from the pre-final layer output of this pre-trained ASR model as the phoneme-level acoustic representations.

2.3.4. Gradual Network

The proposed gradual network combines the ASR representations from ASR pre-train model with sound features representations in the end-to-end training of SED. It contains two parts, the ASR part which is always static and not updated during the training process, and the second part contains the CBRNN architecture for SED which is trained from the scratch. In this structure we concatenated the pre-final layer of the ASR representation in the event detection training path as shown in the gradual network part in Figure 1, and the concatenated vector is fed into the fully connected layer to classify the target events. This is done to improve the performance of the event detection by incorporating ASR features that provides complementary representation to the specific sound event pattern that can be hard to learn from using CBRNN directly in the task 2 of DCASE 2017 sound data.

2.4. Post-processing

In order to reduce the influence of the outliers and to improve the robustness of the binary prediction given for each frame, post-processing steps are necessary [15]. Here we used dynamic thresholding to filter the predicted values to reduce the problems faced by the model because of unbalance positive and negative samples. The dynamic threshold is shown as:

$$T_i = T_{base} + \beta * S_i \tag{8}$$

where T_i is dynamic threshold value, $T_{base}$is the static threshold value, S_i is the average value of the classifier output for each audio and β is ratio value for S_i.

After dynamic thresholding, output is post-processed with a median filter of length 300ms. When performing prediction, several values could appear before the onset and after the offset of the event. Since at most one event would appear in a 30sec log audio, it should show a continuous sequence. With this assumption, we apply a post-processing step to make discontinuous sequence into one continuous sequence and selected the longest continuous sequence of positive predictions to obtain the onset and the offset of the target events.

2.5. Evaluation Metric

The evaluation of the system is done using event-based metrics [18]. Here we considered the event-based F-score and error rate. The f-score is the harmonic mean of the precision(P) and the recall(R) and the error rate is the total number of insertions I, deletion D and substitution S related to the number of references event N. The onset detection is considered accurate only when it is predicted within the range of 500ms of actual onset time. For calculating these performance metrics we have used sed_eval toolbox provided by DCASE organizer. The F-score(F) and the Error-rate(ER) are mathematically defined as

$$ER = \frac{S + D + I}{N} \tag{9}$$

$$F = \frac{2P * R}{P + R} \tag{10}$$

Table 2: *The median time duration is shown with the frequently occurred phonemes for each event.*

Event	Median	Phonemes
babycry	2.33s	SPN_S, EH1_S, M_E, OW1_E, HH_B, N_e, OW1_S, AH1_I L_B, N_B
glassbreak	1.29s	SPN_S, IH1_I, AH0_I, IY1_E, L_B, L_E, M_B, N_E, T_I, AE1_B
gunshot	1.21s	SPN_S, AH1_I, D_E, EH1_I, EH1_S, N_B, W_B, Z_E

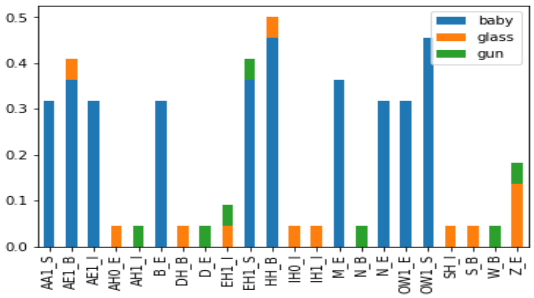

Figure 2: *Phonemes frequency present in each sound event with their probabilities, which shows the important phonemes that contributes more in the prediction of the target events.*

3. Experiment Results and Analysis

The Adam optimizer with stochastic gradient descent algorithm is used in all experiments with a learning rate of 0.0001. The systems are trained by using back-propagation with cross entropy loss function. The network is trained for a max of 100 epochs and a decaying factor 0.01 is set for the learning rate.

For CNN part, four convolution layers with 64 channels are used with different kernel size of (7,7), (5,5), (3,3), (3,3) with the stride of (2,1), (1,1), (1,1), (1,1) for each layer respectively. The three max pooling layers is used after first, third and last convolution layer with kernel size of (4,2), (4,1), (4,1) with the stride of (1,1) for each layer and the dropout is set with the probability of 0.25. We also add the two residual connections [19] to improve the performance of the CNN. The first residual connection is done with the output of first max-pooling with the third convolution layer, and the second residual connection is done with the second max-pooling with the fourth convolution layer. The number of hidden unit used here for GRU is 32 with 2 layers, the hidden layer unit N_t for temporal attention is 200 and for spectral attention the N_s is 32. For spectral attention and temporal attention, we used SoftMax activation function.

Further, we take the pre-final layer features from the ASR model to be combined with the SED features during the training of the CBRNN model for learning. The average performance of the proposed gradual network outperforms other models average performances, which achieves 97.0% f-score and 0.05% error rate where the "babycry" and "gunshot" target events shows an improvement of 6.7% and 4.6% in f-score and 0.14% and 0.11% decrease in error rate respectively as compared to other state-of-the-art models. Here, while the "glassbreak" event detection model does not outperform but it obtained the comparable performance with respect to other state-of-the-art models. The performance of the state-of-the-art models is shown in the Table 1 in terms of f-score(F) and error rate(ER). Furthermore, we have additionally carried out some experiments by creating a fine-tuning network [20] that fine-tune the ASR features. In this fine-tune network, the extracted pre-final output layer of the pre-trained ASR model is fine-tuned to obtain the event prediction for each frame. The fully connected layer is used with softmax activation function to normalize the prediction probabilities to [0,1]. This model by itself already obtains competitive performance without actually training using the SED features (88.6%, 83.5%, 79.0% f-score for "babycry", "glassbreak", and "gunshot" event respectively). By analyzing these results, we can observe that the ASR phoneme-level acoustic features alone may not be enough to completely surpass the discriminatory power SED features, the complementary nature of SER and ASR features are demonstrated in our detection rates.

We further provided an analysis in understanding which phonemes contribute more to the model prediction in predicting these acoustic event classes. We fed our target audio data to ASR model to obtain the most likely phoneme and its alignment duration for analysis. From Figure 2, we can say that the babycry event has high ratio of specific phonemes occurrences than other events. Table 2 shows the median duration of each event happened in the train dataset with the frequently occurred phonemes in that event. It also shows that the duration of the babycry is longer, and it contains more phonemes predicted than other events.

Past research [21] suggests that the babycry comprises of one of the first speech manifestation and represent as the the sound production by larynx and oral cavity movements, this can be seen as precursor to the phonemic production. As the babycry signal production is more likely to similar to human speech production, it may explain the fact that we observe a better detection performance on event like babycry when combining with ASR feature representations as babycry it contains more phonemes-related information as human speech. Events like gunshot and glassbreak are shorter duration events, the mechanism of these sound production are not the same as human speech, which may lead to results in high occurrence in the presence of silence(SIL) and unknown(SNP_S) phonemes.

4. Conclusions

In this work, we proposed a gradual network architecture that takes the advantage of the pre-final layer of ASR representations for SED task. The proposed model is tested on task 2 of DCASE 2017 data. Our experiment results shows that our proposed system which is trained over speech representations can provide useful information in predicting sound events, and ASR-based phoneme-level acoustic representations is indeed beneficial in the detection of different rare sound events, especially for those that would resemble human-like speech. In our future research, we will investigate which patterns of the speech provides the needed contribution to improve the system performance, and what particular sequence of phonemes are contributing in sound event detection for a variety of different sound classes. Additionally, the ASR system can be fine-tuned in parallel with the event detection by updating the layers of the ASR network; moreover, we would extend this framework in handling joint sound event detection and acoustic scene classification task.

5. References

[1] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," *Dept. Electronic Eng., Columbia Univ., New York*, 2001.

[2] J. Schroeder, S. Wabnik, P. W. Van Hengel, and S. Goetze, "Detection and classification of acoustic events for in-home care," in *Ambient assisted living*. Springer, 2011, pp. 181–195.

[3] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[4] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[5] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Bidirectional lstm-hmm hybrid system for polyphonic sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 35–39.

[6] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule based approach for polyphonic sound event detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1853–1857.

[7] E. Cakır and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," *Deep Neural Networks for Sound Event Detection*, vol. 12, 2019.

[8] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, "Learning how to listen: A temporal-frequential attention model for sound event detection," *arXiv preprint arXiv:1810.11939*, 2018.

[9] S. Jung, J. Park, and S. Lee, "Polyphonic sound event detection using convolutional bidirectional lstm and synthetic data-based transfer learning," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 885–889.

[10] H. Lim, M. J. Kim, and H. Kim, "Cross-acoustic transfer learning for sound event classification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2504–2508.

[11] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[14] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 80–84.

[15] J. Wang and S. Li, "Multi-frame concatenation for detection of rare sound events based on deep neural network," in *no. November*, 2017.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing neural speech representations for auditory emotion recognition," *arXiv preprint arXiv:1803.11508*, 2018.

[21] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, "Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies," IEEE, pp. 330–335, 2008.