DETECTING AND CLASSIFYING SEPARATED SOUND EVENTS USING WAVELET_BASED SCALOGRAMS AND DEEP LEARNING

Technical Report

Abigail Copiaco

University of Wollongong
Engineering and Information Sciences Dept.,
Northfields Wollongong, NSW 2522, Australia
abigailc@uow.edu.au

Christian Ritz

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
Northfields Wollongong, NSW 2522, Australia
critz@uow.edu.au

Stefano Fasciani

University of Oslo
Musicology Dept., 1072 Blindern
0316 Oslo, Norway
stefano.fasciani@imv.uio.no

Nidhal Abdulaziz

University of Wollongong in Dubai
Blocks 5,14, &15, Dubai Knowledge Park,
Dubai, UAE
nidhalabdulaziz@uowdubai.ac.ae

ABSTRACT

This report describes our proposed system submitted to the DCASE 2020 Task 4 challenge that includes sound separation and detection. In this work, we examine the use of a Deep Neural Network (DNN) based sound separation system as a pre-processing technique to the sound event classification technique used. For the detection of sound events, a combination of signal energy and spectral centroid features with 0.05-s of time windowing was utilized. Along with this, spectro-temporal features extracted from a Fast Fourier Transform (FFT) based wavelet coefficients of the audio files were used for classification. These coefficients are mapped into images called scalograms, which are fed into the layers of AlexNet, a pre-trained Deep Convolutional Neural Network (DCNN), for transfer learning. Through the validation set, this method gathered an average F1-score of 70% amongst the 10 classes of the DESED database for weak labelling However, this technique is not deemed to be suitable for classification with strong time stamps labelling, gathering an F1-score of a mere 8.73%.

***Index Terms*—** DCASE 2020, Scalogram, Deep Learning, Neural Network, Sound Event Detection, Classification

1. INTRODUCTION

In line with the problem of sound event separation, detection, and classification associated with the DCASE 2020 Task 4 challenge, we propose a novel technique that utilizes neural networks along with features that are spectral and spectro-temporal in nature. This paper details our proposed methodology, its strengths and justifications, as well as the results that we gathered using this system.

2. PROPOSED METHODOLOGY

In this section, we discuss the steps conducted in order to produce the labels and timestamps among the audio files provided, considering the possibility of overlapping sound events within the entire audio duration. In order to promote uniformity within the sampled signals, all signals are initially resampled to 16 kHz prior to feature extraction. The reason behind this is that the sound separation system was trained using a 16 kHz sampling rate. Hence, the signals are adjusted as such in order to promote overall consistency.

2.1. Sound Separation System

The Sound Separation system is implemented in this work as a pre-processing method prior to the extraction of features for sound event detection. Our proposed method is based on Short Time Fourier Transform (STFT) coefficients as features to the neural networks. This method is based on the MATLAB implementation of the Cocktail party problem [1], but has been improvised through the following ways:

- Taking into consideration that the overlapping sound events can belong to multiple varying classes, this performs a multi-level sound separation, as opposed to bi-level sound separation. This is done by training multiple neural networks that each generates a unique soft mask, as opposed to simply inverting the mask generated by a single network.

- The original implementation utilizes a biased sigmoid activation function. However, in this design, we used a hyperbolic tangent (tanh) activation function, due to its benefits in terms of sound separation algorithms [2].The tanh function, defined by (1), generally provides stronger gradients when compared to the sigmoid function, due to its wider range that also takes into consideration the negative components [2].

$$\tanh(x)=2\cdot\sigma(2x)-1 \qquad (1)$$

- Aside from using a different activation function, the CNN network used for this design also contains more layers than the MATLAB model. This network consists of 16 layers of fully connected, tanh activation functions, batch normalization, and drop out layers, while the cocktail party CNN consists of only 12 layers.

The overall process for this algorithm is summarized in Fig. 1. As observed, the normalized STFT of mixed signals, along with the soft masks of each source that exist within those signals, are used in order to train the neural network. The normalization method used for this case is through the mean and standard deviation measures. In order to train the neural network model produced for sound separation, the Free Universal Sound Separation (FUSS) dataset is used [3]. This composes of signals sampled at 16 kHz, with mixtures of up to 4 different sound sources.

Once the networks were trained, these are used in order to generate soft masks for other mixed signals that need to be separated. The soft masks are convolved with the STFT of the mixed signal, prior to performing an inverse STFT in order to generate the separated signals. An example of such separated signals can be observed in Fig. 2, where it is compared against the actual individual signals present within the mixture.

After pre-processing, the sound events are detected and labelled according to their specific classes. The methodologies used in this work for these steps are summarized in Fig. 3.

2.2. Sound Event Detection

In order to detect sound events within a span of an audio signal, two feature sequences are utilized, namely: the Signal energy, and the Spectral centroid. These features were selected, as the former is advantageous for identifying silent periods throughout a signal, and is also useful in differentiating audio classes [4]. The signal energy is identified by (2),

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} |x_i(n)|^2 \qquad (2)$$

where $x_i(n)$, n = *1,….,N* are the audio samples of the *i-th* frame, of the length *N*.

Further, spectral centroids correspond to the central gravity of the audio spectrum. Since sound events are often identified by sudden bursts of sounds, such as footsteps, clapping, or alarms, spectral centroids are beneficial features for identifying such signals. Spectral centroids are defined by (3),

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N} X_i(k)} \qquad (3)$$

where $X_i(k)$, k = *1,….,N,* are the Discrete Fourier Transform (DFT) coefficients of the *i-th* short-term frame, of the frame length *N.*

A thresholding technique is then applied in order to identify the sound segments of the sound events detected from the signal. Thresholds are computed through the first and second local maxima of the histograms, as per (4).

$$T = \frac{WM_1 + M_2}{W+1} \qquad (4)$$

where W is the time windowing parameter selected by the user. In the case of this experiment, this was chosen to be 0.05-s.

Once the thresholds are computed for both signal energy and spectral centroid, segments of sound events are identified according to whether the values of features derived for each frame exceeds the pre-computed thresholds. This methodology for detecting the presence of sound is adapted from [4], with our contribution being the estimation of time stamps indicating the start and end of the identified sound events.

In the case of this work, we apply a time windowing of 0.05-s with median filtering throughout the audio signal. The series of segments of sound events within the signal, along with the time onset and offset wherein they occur, are then sent to the sound event classification block in order to identify the classes in which they belong in.

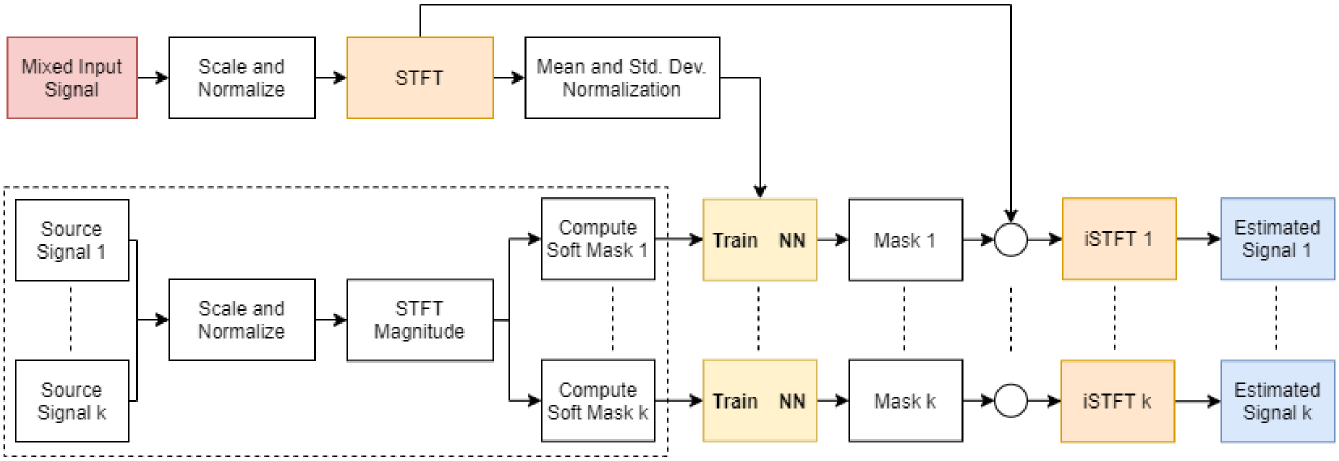

Figure 1. Sound Source Separation System

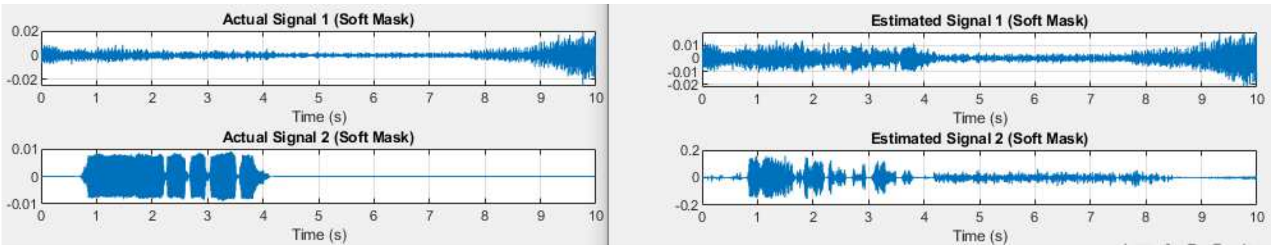

Figure 2: (L to R), Individual Signals present within a mixture versus the Estimated Signals

2.3. Sound Event Classification

For the classification and labeling portion of the system, spectro-temporal features from a Fast Fourier Transform (FFT) based Wavelet coefficients were utilized. Wavelets are time-localized, and are advantageous for audio classification due to its ability to separate mixed audio sources, which allows detailed examination of individual audio sources [5]. These coefficients are calculated per segment, identified by the sound event detection system.

In this work, calculation is achieved with the aid of the MATLAB Audio System and Data Communications toolboxes. the designed system uses a Morlet (Gabor) mother wavelet with a value of 6, a spacing of 0.4875, and computes 31 coefficient scales per channel. The minimum and maximum scales are automatically defined through the energy spread of the wavelet on a spectro-temporal basis. The wavelet coefficients are then separated through low-dimensional models that resonated from harmonic template models [6].

Morlet wavelets are utilized due to its computational efficiency, requiring less calculations and resources compared to other types of wavelets, which is made possible through the implementation of the FFT. Furthermore, Morlet wavelets are characterized by a Gaussian shape, which eliminates any sharp edges that may be misconceived as oscillations [7]. Lastly, the convolutional results of the Morlet wavelet retains the temporal properties of the original signal [7].

Once computed, the coefficients were then mapped into a time-frequency plot called the scalogram image. The scalogram results from the absolute value of the CWT coefficients plotted against the time and frequency. Its spectro-temporal nature considers both the time and frequency components of the signal, which is beneficial for mapping the properties of the constant movement of signals while maintaining the loss of information at a minimum.

The scalogram plots, however, are first resized into 227x227 RGB plots through a bi-cubic interpolation algorithm coupled with antialiasing techniques, in order to accommodate the requirement of the AlexNet Deep Convolutional Neural Network (DCNN) classifier, which will be used to train and classify the images via transfer learning.

2.4. Neural Network Training

AlexNet is a type of a pre-trained deep convolutional neural network consisting of 8 layers in total, 5 of which are convolutional layers and 3 are fully-connected layers [8]. It has achieved a Top-5 error rate of 15.3%, and has around 60 million parameters. As inputs, it requires RGB images of the size 227x227.

This network was selected for the purpose of classification and sound event labeling for this experiment, due to several advantages associated with its default activation function. The Rectified Linear Unit (ReLU) function represents positive values as they are, while negative values are represented by a 0. This mechanism allows for a faster duration in training [8]. Furthermore, the dropout layer application resolves some issues faced by other common activation functions, such as over-fitting, and the vanishing gradient problem [8].

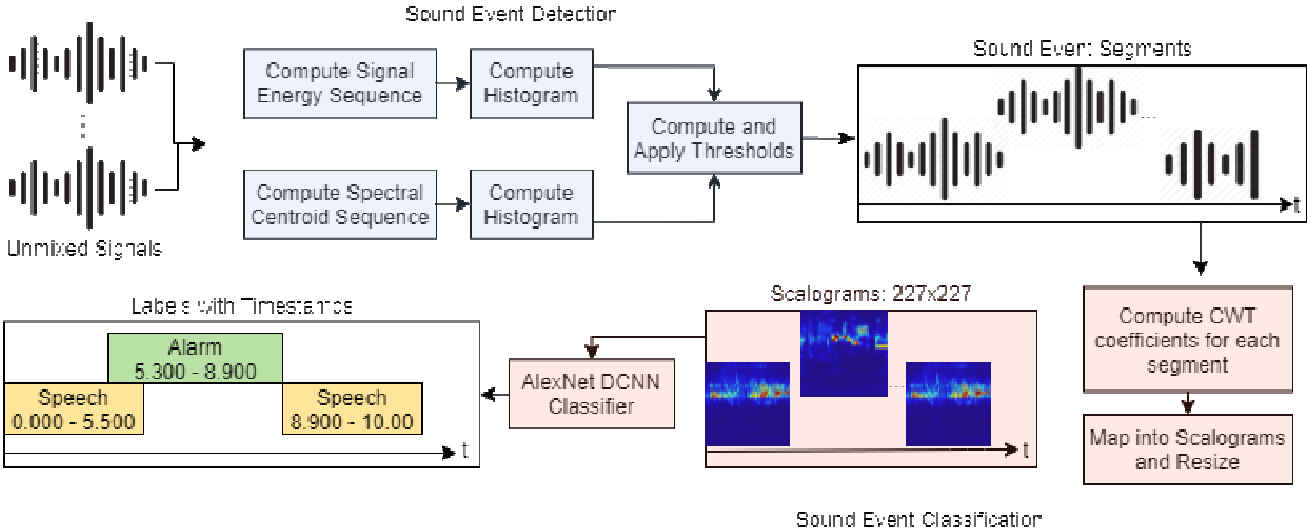

Figure 3: Overall Proposed Methodology Block Diagram

In our previous work, we also examined the performance of AlexNet against other pre-trained models for domestic acoustic scene classification purposes through the SINS database [9,10], and had proven AlexNet to have the optimum performance for similar problems when it comes to pre-trained neural networks.

3. RESULTS

This experiment was conducted using the Domestic Environment Sound Event Detection (DESED) database [11], provided by the DCASE 2020 challenge [12]. Contents of the dataset were a combination of real and synthetic recordings that were either single- or dual-channel, sampled at 44.1 kHz but with different durations. This database allows classification into ten categories: alarm bell ringing, blender, cat, dishes, dog, frying, electric shaver or toothbrush, speech, running water, and vacuum cleaner.

Applying our proposed methodology, which incorporates a sound separation technique along with the sound event detection and labeling, produced an average F1-score of around 70% for weak labelling, for a uniform sampling rate of 16 kHz. This provides a slight improvement when compared to the network trained without a sound separation pre-processing technique, which uses the same sampling rate. A comparison of these results are shown in Table 1.

Table 1: Proposed System Results on Weak Labelling

Class	F-score (with SS)	F-score (w/out SS)
Alarm_bell	78.43%	77.24%
Blender	61.22%	65.32%
Cat	79.30%	75.76%
Dishes	83.53%	78.29%
Dog	75.95%	44.22%
Frying	38.52%	31.17%
Speech	98.22%	97.92%
Vacuum	89.10%	72.53%
Electric_shaver	42.00%	46.88%
Running_water	54.95%	39.77%
Average	**70.12%**	**62.91%**

However, after taking into consideration the time stamps through the sed_eval toolbox [13], the results gathered are shown in Table 2.

Table 2: Proposed System Results, signals resampled at 16 kHz

Class	Precision	Recall	F-score
Alarm_bell_ringing	9.50%	17.60%	12.30%
Blender	5.00%	5.20%	5.10%
Cat	8.40%	19.10%	11.70%
Dishes	3.00%	3.00%	3.00%
Dog	3.80%	9.80%	5.50%
Frying	0.00%	0.00%	0.00%
Speech	22.30%	10.60%	14.40%
Vacuum	3.70%	20.70%	6.20%
Electric_shaver	11.10%	1.50%	2.70%
Running_water	2.50%	8.90%	3.90%
Overall metrics (micro-average)	**7.48%**	**10.48%**	**8.73%**
Error Rate	**2.04**	**Substitution Rate**	**0.15**
Deletion Rate	**0.74**	**Insertion Rate**	**1.14**

As observed, the system was not able to show satisfactory results for strongly labelled data, when onset and offset time stamps are taken into consideration. This can be due to the following reasons:

- The sound separation algorithm was not trained with a large database due to current memory limitations in the computing resources available to the team. Results for the sound separation algorithm are expected to improve when trained with a larger data.
- The utilization of signal energy and spectral centroids as features for sound event detection are particularly useful for identifying voiced sections within an audio duration. However, since the dataset contains foreground and background sounds, this may negatively affect the accuracy of the timestamps. Further adaptation of the timestamp and sound event detection algorithm may be needed so that it is more robust to background noise and overlapping sounds.
- Once the sound events are detected, the segments returned are not all of equal length. Hence, for the much shorter segments, the number of features extracted may not be enough for the classifier to provide an accurate prediction of the sound class.

4. CONCLUSION

In conclusion, although our proposed methodology displays promising results for SED with weak labelling, this fails to do so when the timestamps are taken into consideration. In the future work, the performance of our system can be further improved through providing more training data to the sound separation system, in order to remove noise and prevent any loss of information that could occur within the process. Furthermore, the implementation for other techniques for generating audio timestamps in SED could also be looked into.

5. REFERENCES

[1] https://www.mathworks.com/help/deeplearning/ug/cocktail-party-source-separation-using-deep-learning-networks.html.

[2] A, Olgac, Karlik, Bekir. "Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks" *International Journal of Artificial Intelligence And Expert Systems*, vol 1, pp 111-122, 2011.

[3] E. Fonseca, et al., *FSD50k: an open dataset of human-labeled sound events.* In arXiv. 2020.

[4] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in Matlab", 2010.

[5] S. Mallat and S. Shamma, "Audio Source Separation with Time-Frequency Velocities," ScatBSS - Supported by ERC Invariant Class 320959.

[6] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Ziegr and M. Omolgo, "Acoustic Event Detection and Classification," in Springer London, London, 2009.

[7] M. Cohen, "A better way to define and describe Morlet wavelets for time-frequency analysis," Biorxiv, Donders Institute for Neuroscience, n.d.

[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS Proceedings, 2012.

[9] A. Copiaco, C. Ritz, S. Fasciani and N. Abdulaziz, "Scalogram Neural Network Activations with Machine Learning for Domestic Multi-channel Audio Classification," 2019 IEEE ISSPIT, Ajman, United Arab Emirates, 2019, pp. 1-6

[10] A. Copiaco, C. Ritz, N. Abdulaziz and S. Fasciani, "Identifying Optimal Features for Multi-channel Acoustic Scene Classification," 2019 ICSPIS, Dubai, United Arab Emirates, 2019, pp. 1-4.

[11] N. Turpault, R. Serizel, AP. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis”, In Workshop on Detection and Classification of Acoustic Scenes and Events. New York City, United States, October 2019.

[12] http://dcase.community/workshop2020/.

[13] C. Bilen, et al. “A framework for the robust evaluation of sound event detection”. arXiv preprint arXiv:1910.08440, 2019.