SOUND EVENT DETECTION AND CLASSIFICATION USING CWT 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
Department of Musicology,
0371 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 for the DCASE 2020 Task 4 challenge. In this work, we examine the combination of signal energy and spectral centroid features with 0.05 s of time windowing for the detection of sound events. Along with this, spectro-temporal features extracted from 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 74% 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 11.21%.

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

1. INTRODUCTION

Common methodologies utilized for sound event classification often involve spectral and cepstral features, such as the Log-mel energies and Mel Frequency Cepstral Coefficients (MFCC) [1]. In this paper, we propose the integration of signal energy and spectral centroid features with thresholding for sound event detection, to the use of spectro-temporal features in the form of Scalograms for classification [2].

2. PROPOSED METHODOLOGY

In this section, we discuss the process applied in order to produce the labels and timestamps among the audio files provided. The general algorithm is also summarized in Fig. 1. Note that in order to promote uniformity within the sampled signals, all signals are initially resampled to 16 kHz sampling rate for Model 1, and 44.1 kHz for Model 2, prior to feature extraction.

2.1. Sound Event Detection

In order to detect the presence of sound events within a span of an audio signal, two feature sequences are utilized as shown in Fig. 1, 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 [3]. The signal energy is identified by (1),

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

where $x_i(n)$, $n = 1,\ldots,N$ are the audio samples of the *i-th* frame, of the audio 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 (2),

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N} X_i(k)} \quad (2)$$

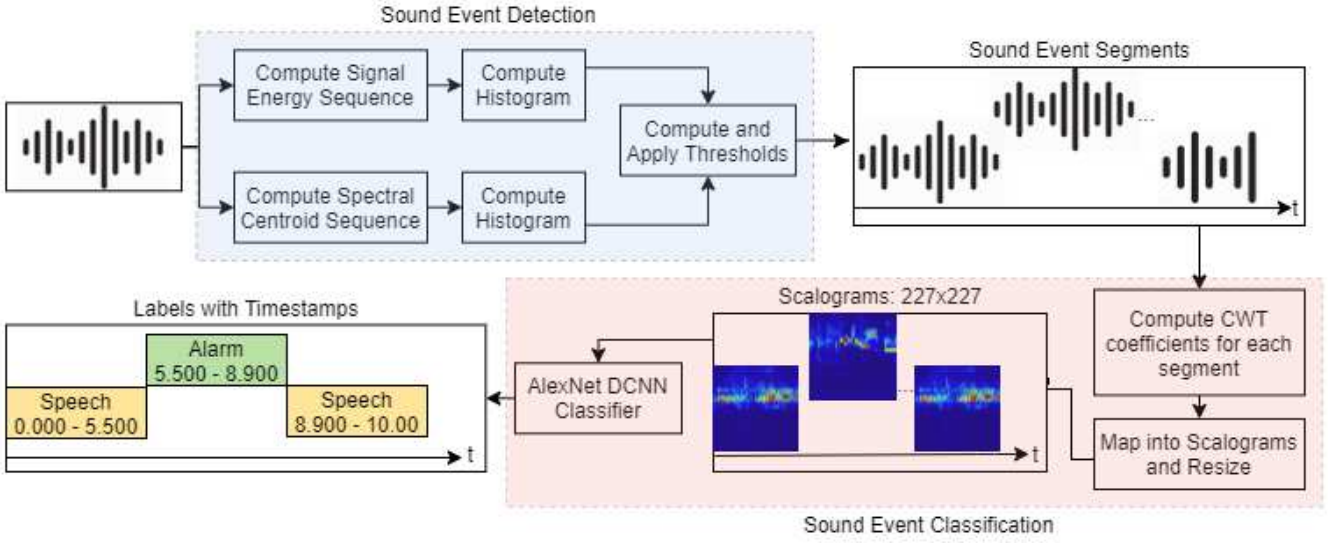

Figure 1: Overall Proposed Methodology Block Diagram

where $X_i(k)$, $k = 1,\ldots,N$, 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 (3).

$$T = \frac{WM_1 + M_2}{W + 1} \tag{3}$$

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 events (Sound Event Segments of Fig. 1) is adapted from [3], 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 onset and offset time for each segment are then sent to the sound event classification block of Fig. 1.

2.2. Sound Event Classification

For the classification 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 their ability to separate mixed audio sources, which allows detailed examination of individual audio sources [4]. These coefficients are calculated from each 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. These parameters were selected based on trial and error, through observing visual differences between each sound classes via scalograms. The minimum and maximum scales are automatically defined through the energy spread of the wavelet on a spectro-temporal basis. All of the wavelet coefficients generated are then separated through low-dimensional models that resonated from harmonic template models [5].

Morlet wavelets are utilized due to their 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 [6]. Lastly, the convolutional results of the Morlet wavelet keeps the temporal properties of the original signal [6], which is necessary in order to keep the consistency with the time stamps.

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 temporal properties of sound signals whilst also capturing discriminative spectral information about the signal that can be used for classification.

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.

2.3. 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 [7]. 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 this experiment due to several advantages associated with its default activation function, the Rectified Linear Unit (ReLU). This function represents positive values as they are, while negative values are represented by a 0. This mechanism allows for a faster duration in training [7]. Furthermore, the dropout layer application resolves some issues faced by other common activation functions, such as overfitting, and the vanishing gradient problem [7].

In our previous work, we also examined the performance of AlexNet against other pre-trained neural networks for domestic acoustic scene classification purposes through the SINS database [2,8], and found that AlexNet had the optimum performance for similar problems when using pre-trained networks.

3. RESULTS

This experiment was conducted using the Domestic Environment Sound Event Detection (DESED) database [9], provided by the DCASE 2020 challenge [10]. The dataset contains 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.

The system was trained using the strongly labelled synthesized soundscapes, weakly labelled recorded soundscapes, and unlabeled recorded soundscapes, whose labels were generated for using a mean-teacher algorithm, with the two previously mentioned sets as the teacher. For validation, the strongly labelled validation set was used, which comprises of 1168 clips (4062 sound events) [9].

Applying our proposed methodology produced an average F1-score of 74% for weak labelling, without generating the timestamps, for a uniform sampling rate of 44.1 kHz. For a uniform sampling rate of 16 kHz, this provides an average F1-score of 63% for weak labelling, considering all 10 classes. More detailed results on the F1-score for each class shown in Table 1.

Table 1: Proposed System Results on Weak Labelling

Class	F-score (16 kHz)	F-score (44.1 kHz)
Alarm_bell	77.24%	80.19%
Blender	65.32%	64.00%
Cat	75.76%	71.36%
Dishes	78.29%	95.11%
Dog	44.22%	77.37%
Frying	31.17%	45.39%
Speech	97.92%	99.77%
Vacuum	72.53%	91.55%
Electric_shaver	46.88%	46.60%
Running_water	39.77%	67.56%
Average	**62.91%**	**73.89%**

However, following the consideration of the onset and offset time stamps, as per the sed_eval toolbox provided [11], the results of Tables 2 and 3 were established using the validation set provided, in terms of precision, recall, and F-score. Table 2 entails the results for signals uniformly resampled at 16 kHz (Model 1), while Table 3 shows that of signals consistently resampled at 44.1 kHz (Model 2).

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

Class	Precision	Recall	F-score
Alarm_bell_ringing	7.20%	22.40%	10.90%
Blender	3.20%	5.20%	3.90%
Cat	12.10%	18.80%	14.70%
Dishes	2.70%	8.10%	4.00%
Dog	7.20%	3.00%	4.20%
Frying	1.30%	7.40%	2.30%
Speech	23.20%	7.10%	10.90%
Vacuum	6.30%	19.60%	9.50%
Electric_shaver	0.90%	3.10%	1.40%
Running_water	7.50%	13.50%	9.70%
Overall metrics (micro-average)	**6.91%**	**9.68%**	**8.06%**
Error Rate	**2.05**	**Substitution Rate**	**0.16**
Deletion Rate	**0.74**	**Insertion Rate**	**1.14**

Table 3: Proposed System Results, signals resampled at 44.1 kHz

Class	Precision	Recall	F-score
Alarm_bell_ringing	26.10%	14.00%	18.30%
Blender	5.90%	9.40%	8.70%
Cat	13.80%	20.20%	16.40%
Dishes	10.90%	2.60%	4.30%
Dog	5.90%	8.20%	6.90%
Frying	0.00%	0.00%	0.00%
Speech	23.90%	9.70%	13.80%
Vacuum	7.50%	28.30%	11.80%
Electric_shaver	0.00%	0.00%	0.00%
Running_water	6.00%	10.50%	7.70%
Overall metrics (micro-average)	**12.86%**	**9.94%**	**11.21%**
Error Rate	**1.48**	**Substitution Rate**	**0.09**
Deletion Rate	**0.81**	**Insertion Rate**	**0.58**

As observed, although the method works well for weak labelling, these do not show promising results when time stamps are considered. This may be due for the following reasons:

- 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 works well for SED with weak labelling, the same is not observed when the timestamps are taken into consideration. In the future work, the performance of our system can be further improved through the application of a sound separation system, provided that some of the audio files contain overlapping sound events. Further, other techniques for generating more accurate audio timestamps for SED could be observed.

5. REFERENCES

[1] R. Serizel, et. al, "Acoustic Features for Environmental Sound Analysis," Computational Analysis of Sound Scenes and Events, pp. 71-101, 2017.

[2] 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

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

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

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

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

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

[8] 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.

[9] 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.

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

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