

PAPER • OPEN ACCESS

Environmental acoustic transformation and feature extraction for machine hearing

To cite this article: R Catanghal Jr *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **482** 012007

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the **collection** - download the first chapter of every title for free.

Environmental acoustic transformation and feature extraction for machine hearing

R Catanghal Jr¹, T Palaoag², and C Dayagdag³

¹University of Antique, Sibalom, Antique, Philippines

²University of the Cordilleras, Baguio, Philippines

³Romblon State University, Odiongan, Romblon, Philippines

catanghal.ric@gmail.com, tpalaoag@gmail.com, carlwindayagdag@gmail.com

Abstract. This paper explores the transformation of environmental sound waveform and feature set into a parametric type representation to be used in analysis, recognition, and identification for auditory analysis of machine hearing systems. Generally, the focus of the research and study in sound recognition is concentrated on the music and speech domains, on the other hand, there are limited in non-speech environmental recognition. We analyzed and evaluated the different current feature algorithms and methods explored for the acoustic recognition of environmental sounds, the Mel Filterbank Energies (FBEs) and Gammatone spectral coefficients (GSTC) and for classifying the sound signal the Convolutional Neural Network (CNN) was used. The result shows that GSTC performs well as a feature compared to FBEs, but FBEs tend to perform better when combined with other feature. This shows that a combination of features set is promising in obtaining a higher accuracy compared to a single feature in environmental sound classification, that is helpful in the development of the machine hearing systems.

1. Introduction

A significant number of information is carried by sounds in connection with our daily surroundings and physical happenings that take place in it. In terms of efficiency, humans are very effective and accurate in identifying the typical general feature of the soundscape surrounding him, whether it is an office environment, a quiet park, bustling city lanes, beach fronts and deciphering individual sources of sound in the scenes like coughing, rooster crowing, crying baby, glass breaking, running cars, dog barks, or thunderstorm. The automatic recognition of sounds in the environment for example in the multimedia applications is an expanding and increasing problem of research. The comprehension of this sounds in the environment are an important factor in the perception of the content in multimedia. This environmental sounds are very much divergent and varied sound events occurrence from day to day and cannot just be confined between the music or speech. For this reason, the technological development of the classification of sounds in the environment is best for describing large numbers of applications in multimedia, such as classification of scenes in the audio, hearing aid systems, intelligent house monitoring, generation and highlighting of video content, surveillance system through audio, and a lot more [1].

Sound is frequently a valuable supplement to modalities, for example, a video conveying information not generally present, for instance, data from discourse and birdsong and it can likewise be more easier to gather or collect, such as on a cell phone or portable recording device. Collected



information from a semantic acoustic examination can be valuable for additional transformation, for example, robotic navigational route, user alarms, or prediction and analysis in pattern occurrence of events. Apart from being a listening machine, similar advancements have applications in cataloging/searching sound files, whose digital compilations have developed massively in decades. Often sound or acoustic collection contains a rich assorted and variety of music, bio-acoustics, speeches, city and rural soundscapes, environmental sounds, ethnographic archives to name only some, in spite of the accessibility at present, it is behind compare to text [2].

Researchers have attempted to use information extracted from various sounds to enrich various applications, such as monitoring systems for older adults or infants, automatic surveillance systems, automatic lifelogging systems, and advanced multimedia retrieval systems. The techniques used in these applications include acoustic scene analysis that analyzes scenes regarding the places, situations, and user activities they depict and acoustic event analysis that analyzes various sounds, such as footsteps, the rumble of traffic, and scream [3].

This paper explores and presents the current feature extraction algorithm and methods for environmental sound recognition that would be utilized in machine hearing systems, further objective of this research paper is the comparative feature extraction and methods. A summary of the architecture of the machine hearing system is discussed. A thorough analysis of the most current developed feature was discussed, namely the gammatone filterbank and the Mel filterbank energies (FBEs). The most relevant result of its evaluation shall be shown and suggested implementation for the environmental sound recognition.

2. The Machine Hearing

2.1. The Machine Hearing Design

Even what particular type of problem is applied in the acoustic field, the fundamental structure of the system can be characterized using a general or universal pattern design as shown on Figure 1.



Figure 1. Architecture of Machine Hearing

The process of windowing is the first stage; this is accomplished by capturing a continuous stream of audio using a device usually a microphone and partitioning this into a block of the shorter signal. To accomplish the process of windowing, a hypothetically endless sample of streams from the signal input will be slid by a window function. Moreover, in order to facilitate the successive signal analysis, contingent on the window function duration it is presumed that a commonly non-static sound signal within the individual frame is quasi-stationary.

The problem of buffering can be overcome with the use of the window function. This window function works by fetching the audio chunks at a certain intervals of time, even not knowing with regards to their completeness. To further address the issue, the edge of the buffer is being polished using the window functions. Through this, the entire important periods in the audio signal is required to be completed. In this study the Hamming window was used and is characterized as the equation 1 below.

$$\text{HammingWindow}(s) = 0.54 - 0.46 \cos\left(\frac{2\pi s}{\text{BufferSize}}\right) \quad (1)$$

2.2. Feature Set Extraction

Features extraction, for the most part, is carried out in three main stages and is the series of feature vector calculation that accomplishes for a given signal a condensed representation. The first stage is

called acoustic front-end or analysis of the sound, in here where the raw feature are produced characterizing the container of the power spectrum of abrupt intervals of sounds, and in this stage where the sound signal spectra temporal analysis is carried. In the next stage is where the compilation of a continued feature vector which comprises of the two features: dynamic and static. Then lastly the transformation stage, where the continued feature vectors is converted further to a much compressed and concentrated vectors as a recognizer input.

2.3. The Mel Filterbank Energies (FBEs)

Research experiment noted that human ear works as a filterbank, that is a bank of subband filters. In this human perception experiment where the analysis of the Mel frequency was based. This Mel frequency focuses only on some particular or specific components of frequency. It is in the frequency-axis where the filters are space unevenly or overlapped, further in research of processing of sounds; it exhibited that the signal in the range of 10 – 30 ms duration is considered to be in static and in this instance a window that is smaller is considered. A proposed phase encoded filterbank energies (PEFBEs) were even proposed as an enhancement technique [4] as shown in Figure 2.

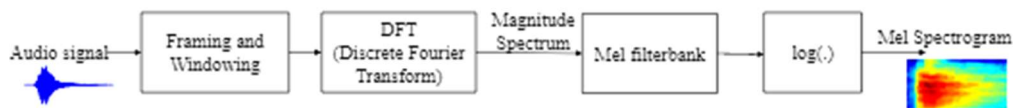


Figure 2. Mel spectrogram diagram of sound signal [4].

2.4. The Gammatone Filterbank and Teager Energy Operator

The Gammatone filter impulse response is a multiplication of the Gamma distribution function and a sinusoidal tone centered at a particular frequency [6]. Given by the equation 2:

$$g(f, t) = t^{a-1} e^{-2\pi b t} \cos(2\pi f t), t > 0, \quad (2)$$

where f is defined as the center frequency, b is rectangular bandwidth, and a is the filter order. The concept of the Gammatone filter was motivated by the biological studies it is where it stems. In the modeling of the filter response of the auditory of humans, this is where the Gammatone function is utilized [7]. There is a great similarity on the how the reox function (human auditory filter response in the cochlea) to the Gammatone filter for its magnitude response. In the human hearing system the Gammatone filter bandwidth resembles the basilar membrane placement of filters. The scale to where it is measured is called equivalent rectangular bandwidth (ERB) [5].

The temporal modulations, among the many perceptual and acoustic features of the sound, is one of the essential parametric representations of the sound signal and it describes the signal variations of sounds regarding the frequency modulation (FM) and amplitude modulation (AM). Using the cochlea as the filter bank model, where the feedback of the AM-FM can be drawn, and it was noted that AM-FM is an integral combined feature of the sound signal and occurs at the same time. The researcher [5] believed that adopting an operator like Teager Energy Operator (TEO) which will trace the energy coming from the two elements which is the AM and FM. This Teager Energy Operator represents the necessary energy for the generation of a signal or the system energy that produces the signal. The Teager Energy Operator not like the standard or typical energy, it is roughly equivalent to the result of squaring the product of frequency and amplitude. The signal was separated for refinement through the narrowband filter bank before applying the teager energy operator, the application filtering was done because teager energy operator is not applicable for direct application in the sound signal. The reason for this process is that Teager energy operator, for the most part, functions on the bandpass filtered

signal or on the monocomponent. The directly short-term spectral feature used in the CNN is called as TEO-based Gammatone spectral coefficients (TEO-GTSC) [5] as shown in Figure 3.

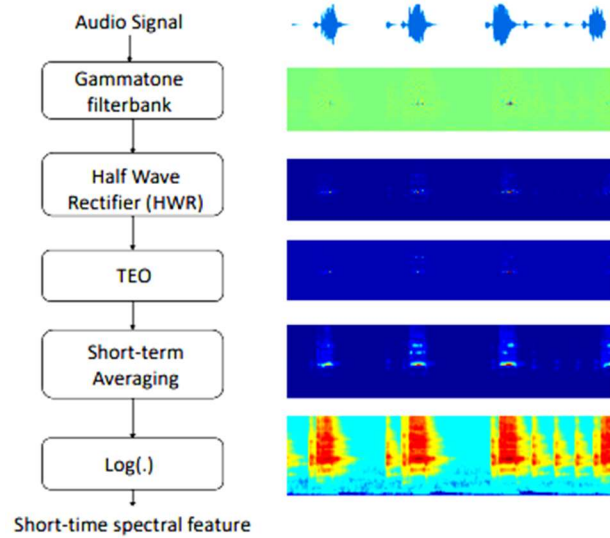


Figure 3. Block diagram of TEO-GTSC [5]

2.5. The Convolutional Neural Networks

The bottom line of the convolutional neural networks is that they are an elementary extended version of the model multilayer perceptron. Nevertheless, there are important and substantial pragmatic implication and results due to their design or architectural differences. The usual and conventional neural network in a deep architecture is composed of a few distinct layers stacked collectively. That is a layer for an input, an array of pooling and convolutional layers (there are several ways it can be mixed), completely linked layers that are hidden which are limited in numbers, and a layer for output (loss) as shown in Figure 4. In contrast with the multilayer perceptron, the actual distinction lies in addition of the pooling and convolution procedures [8].

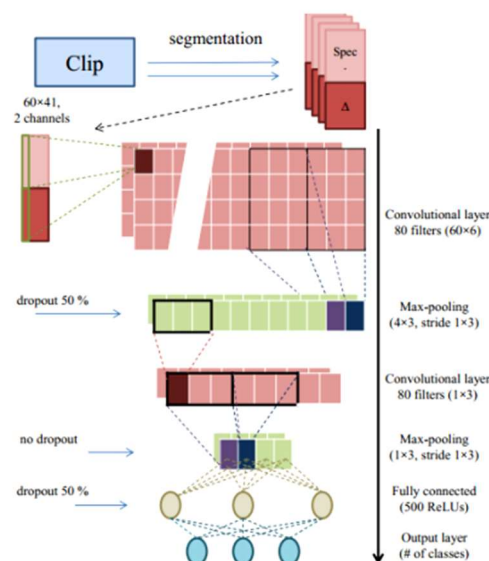


Figure 4. CNN architecture for ESC. After [5], [8]

3. Results and Discussions

A meta-analysis of the different experiments implemented [9][4][5][8] in the literature using the environmental sound classification dataset [1], with 2000 labeled environmentally recordings that are sampled at 44.1kHz and comprises of 50 equally balanced classes; each clip is about 5 seconds. The 50 classes was categorize into five (5) main divisions: natural soundscape, the animal sounds, different water sounds, urban/exterior sounds, domestic/interior noises and the non-speech by humans. An array of diversified audio sources from the dataset was used as a training, several are very recognizable such as the brushing of teeth, while for example airplane and helicopter noise in which the distinction or differentiation is somewhat slight, there are very familiar sounds like that of a barking dog, laughter or cat meowing are among the few examples.

An acoustic transformation was made before feature extraction was executed. To have an analogous comparison with the baseline, the entire sound files were resampled/down sampled to 22.05kHz [8]. The feature extraction of the sound files where accomplish by dividing the frames into windows by applying the 25 ms Hamming method and using a fifty percent (50%) window overlap. In removing the regions of silence, an algorithm called simple energy thresholding was used. The algorithm works by removing the frame if there exist silence greater than three frames, otherwise, frames are preserved.

Table 1. Classification accuracy(%) of different feature with CNN classifier

Feature Sets	Accuracy (%)
FBE \oplus ConvRBM-BANK [9]	86.50
FBE \oplus PEFBEs [4]	84.15
GTSC \oplus ConvRBM-BANK [9]	83.00
GTSC \oplus TEO-GTSC [5]	81.95
GTSC [5]	79.10
FBEs [5]	67.80
Baseline [8]	64.50

The performances of various feature sets were evaluated through 5-fold cross-validation. Table 1 shows that using Mel Filterbank Energies alone yield the lowest accuracy which is 67.80%, although higher compared to the reported baseline [8] of 64.50%. When compared as feature set without combination with other systems Gammatone spectral coefficients yields higher accuracy 79.10% compared to Mel Filterbank Energies which is 67.80%. It is noteworthy that even though Mel Filterbank Energies yields lower accuracy taken as the only feature when combined with other features it yields the highest accuracy which is 86.50%, higher compared to Gammatone spectral coefficients even when combined with other feature sets.

4. Conclusion and Recommendation

This paper aim is to explore the transformation of environmental sound waveform and feature set into a parametric type representation to be used in analysis, recognition, and identification for auditory analysis of machine hearing systems. We analyzed and evaluated the different most current developed feature algorithms and methods explored for the acoustic recognition of environmental sounds, the Mel Filterbank Energies (FBEs) and Gammatone spectral coefficients (GSTC) using Convolutional Neural Network (CNN) as a pattern classifier.

It seems that a combination of features set results to better performance as compared to single features. The Mel Filterbank Energies (FBEs) combined with Convolutional Restricted Boltzmann Machine filterbank (ConvRBM-BANK) shows a promising result as the highest accuracy in

classifying the environmental sound. Studies show that a combination of features improves the performance in the classification of the environmental sounds.

Several probable questions that needs to be answered for future investigation and analysis is whether these features can be successfully applied in other machine learning classifiers and how will they perform in terms of accuracy. Given the complexity of the process and model, how will it perform regarding processing time when applied with the machine hearing systems.

References

- [1] Piczak K J 2015 ESC: Dataset for environmental sound classification *Proc. of the 23rd ACM Int. Conf. on Multimedia (ACM)* pp 1015-8
- [2] Stowell A D, Giannoulis D, Benetos E, Lagrange M and Plumbley M D 2015 Detection and classification of acoustic scenes and events *IEEE Trans. on Multimedia* vol **17** no **10** pp 1733-46
- [3] Imoto K 2018 Introduction to acoustic event and scene analysis *Acoustical Sci. and Technol.* vol **39** no **3** pp 182-8
- [4] Tak RN, Agrawal DM and Patil HA 2017 Novel phase encoded mel filterbank energies for environmental sound classification *Pattern Recognition and Machine Intelligence PReMI 2017* ed Shankar B et al (Cham: Springer)
- [5] Agrawal D M, Sailor H B, Soni M H and Patil H A 2017 Novel teo based gammatone features for environmental sound classification *European Signal Processing Conf. 2017* pp 1809-13
- [6] Valero X and Alias F 2012 Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification *IEEE Trans. on Multimedia* vol **14** no **6** pp 1684-9
- [7] Carney L H and Yin T 1988 Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model *J. of Neurophysiology* vol **60** no **5** pp 1653-77
- [8] Piczak K J 2015 Environmental sound classification with convolutional neural networks *25th Int. Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 2015* pp 1-6
- [9] Sailor H B, Agrawal D M and Patil H A 2017 Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification *Proc. Interspeech 2017* pp 3107-11