Applied Acoustics 130 (2018) 270–282

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Audio sounds classification using scattering features and support vectors machines for medical surveillance

Sameh Souli [a,b,*], Zied Lachiri [a]

[a] LR-SITI Laboratory, National Engineering School of Tunis, Tunis, Tunisia
[b] Arab University of Science, Tunisia

ARTICLE INFO

Article history:
Received 5 May 2017
Received in revised form 14 July 2017
Accepted 1 August 2017

Keywords:
Classification
Audio sounds
Scattering transform
MFCCs
SVM multiclass

ABSTRACT

This paper proposes a new approach to recognize environmental sounds for audio surveillance and security applications. The sounds are extremely versatile, including sounds generated in domestic, business, and outdoor environments. Since this variability is hard to model, investigations concentrate mostly on specific classes of sounds. Among those, the system that is able to recognize indoor environmental sounds may be of great importance for surveillance and security applications. These functionalities can also be used in portable teleassistive devices to inform disabled and elderly persons affected in their hearing capabilities about specific environmental sounds (door bells, alarm signals, etc.). We propose to apply an environmental sounds classification method, based on scattering transform and the principal component analysis (PCA). Our method integrates ability of PCA to de-correlate the coefficients by extracting a linear relationship with what of scatter transform analysis to derive feature vectors used for environmental sounds classification. The performance evaluation shows the superiority of this novel sound recognition method. The support vector machines method based on Gaussian kernel is used to classify the datasets due to its capability to deal with high-dimensional data. Our SVM−based multiclass classification approach seems well suited for real-world recognition tasks. Experimental results have revealed the good performance of the proposed system and the classification accuracy is up to 92.22%.

1. Introduction

Several speech, music and environmental sounds classifiers use mel-frequency cepstral coefficients (MFCCs) which are an efficient audio features delivering spectral energy measurements over short time windows of length 23 ms. Consequently, they lose information on signal structures such as transients or time-varying structures. These signal structures of spectral information are non-stationary over this time interval [1]. As a solution to minimize this loss, in the majority of applications most signals used are locally stationary. The descriptors of audio classification on larger time scales follow the aggregation of cosine transforms of mel-frequency spectral coefficients (MFSCs) in time, with multiple ad-hoc approaches like mel-frequency cepstral coefficients (MFCCs) segments [2]. The study in [1] shows that the spectral information lost by MFSC coefficients can be retrieved like spectral co-occurrence coefficients. Scattering transform computes multi-scale co-occurrence coefficients with a cascade of wavelet filter banks and modulus rectifiers. Moreover, a scattering transform illustrates a locally translation invariant representation. Thus, this transformation expands MFCCs descriptors using calculation of modulation spectrum coefficients of multiple orders, by means of cascades of wavelet convolutions and modulus operators [3].

A scattering transformation represents efficient characterizations for environmental sounds classification. The use of second-order cooccurrence coefficients enhance results obtained by MFCCs features.

The information lost by spectral energy can be retrieved by a scattering wavelets [4]. It shows that mel-frequency spectrograms and MFCCs are limited to such short time intervals. The scattering representation recovered the information lost by mel-frequency spectrogram and MFCCs with multiple layers of wavelet coefficients. Moreover, a scattering transform has strong resemblances with physiological patterns of the cochlea and of the auditory pathway [5,6], as well as utilized for audio processing [7]. The scattering transform calculates co-occurrence coefficients by cascading wavelet filter banks and rectifiers computed with modulus

* Corresponding author at: Arab University of Science, BP 37, 1002, Le Belvédère, Tunisia.

E-mail addresses: soulisameh@yahoo.fr (S. Souli), zied.lachiri@enit.rnu.tn (Z. Lachiri).

https://doi.org/10.1016/j.apacoust.2017.08.002
0003-682X/

operators. Besides, a scattering transform has high similarities with auditory physiological models depend on cascades of constant-Q filter banks and rectifiers [8,9]. It is proven that second-order co-occurrence coefficients bear a significant part of the signal information. Moreover, scattering transforms have demonstrated effective for image [10,11], and also for audio classification [1]. Moreover, scattering transforms have an excellent frequency resolution that is suitable to audio frequency structures. This paper is organized as follows: Section 2, we present a review of previous works of audio sounds recognition. Section 3 presents MFCCs, Mel-frequency spectrum and scattering representation. In Section 4, we present two classifiers used in our study namely multiclass SVM and HMM. Section 5 depicts our proposed approach. Section 6 provides an application to environmental sounds classification, indicating that the combination of scattering co-occurence coefficients with MFCCs enhance classification performance. Classification results are attained with a Gaussian kernel SVM devoted to scattering feature vectors. Finally, we conclude presenting some future works in Section 7.

2. Review of previous works

In this part, we will present previous works which contains the tasks about acoustic scene classification of sound events. We discuss recent evaluation research on general unstructured audio-based scene recognition.

Audio-based scene classification looks for classifying the acoustic environment of an audio stream by choosing a semantic label for it [12].

In the literature, computational auditory scene analysis (CASA) [13] is interested in many applications such as speech or music recognition [14,15].

Its objective is the separation of various types of sounds like speech, music or environmental sounds to be classified using pattern recognition approaches. Research on environmental sound recognition has evolved like areas of music [16,17], or speech classification [18,19].

In [20], CASA study shows that speech and speaker classification can attain a high performance using nonstationary (time-frequency) approaches.

According to [21], stationary and non-stationary time-frequency-based feature extraction was applied for environmental sound classification. Besides, the audio surveillance has been used for detecting and classifying various types of acoustic events such as humans coughing in an office environment [22], impulsive sounds such as gunshot detection [23], glass breaking, explosions, or door alarms [24].

To classify time-based audio sound, the important question is how to analyze temporally-structured data to yield a single label representing the audio sound. It exist two main areas in the literature.

The first area proposes to represent the scene as a single object using spectral characteristics. The most popular spectral used features is the Mel-frequency Cepstral Coefficients (MFCCs) which have been used to make quite well [12]. The classifier GMM was used for the discrimination between sound classes. In [25] Peltonen's used as feature the averaging of band-energy ratio and a K-nearest neighborhood (kNN) as classifier. Also Peltonen choose MFCCs characteristics and a Gaussian Mixture Model (GMM) classifier. He remarked the shortcomings of MFCCs for courant life-sounds, thus he proposed to use the band-energy ratio in order to perform sounds occurring in different frequency ranges.

Many efforts were provided to classify audio sounds. Thus, these studies did not carefully present the time-directional change of the frequency [26]. In [27] Goldhor proposed to use MFCC feature for environmental sound recognition such as a "barking dog sound". Study of Kraft et al. [28] uses kitchen sounds for a humanoid robot. Eronen et al. [29] are interested in classifying sounds of trains and streets. They considered an optimal recording times for recognition. Smith et al. [30] enhanced their extraction method by combining differential MFCC value of velocity and acceleration with various spectral and temporal features.

Dargie [31] investigated an audio sound classification for laptop computers. The classification performance based on MFCC was high, but the specific sound result was poor.

El-Maleh [14] tried various pattern classification models like the quadratic Gaussian classifier, the nearest-neighbor (NN) classifier, and least-square linear classifier, nearest-neighbor classifier, to classify environmental noises. This classification allows control noise pollution. The used features derived from linear prediction coefficients. The best result is obtained with QGC classifier, of the order of 13.6% as error rate.

The study in [32] emphasized the differences between urban environments, and polyphonic music. The objective is to model the distribution of MFCCs using 50-component GMMs.

In [24], the aim is to classify environmental sounds for surveillance applications. It used Gaussian mixture models (GMM) and hidden Markov models (HMM) as classifiers. It was shown that HMM obtained better recognition rate than GMM. The studies in [33,34] used several impulsive sound techniques for extraction features likes traditional time-frequency transforms, speech processing acoustic features, and psychoacoustical features. For classification, it performed with various classifiers. In [35], the authors concentrate on detecting a set of events. A Gaussian mixture model (GMM) is used as classifier.

The audio classification systems are robust when they are resistant to noise. Hence, various works have been studied to reduce the effect of noise on the classification performance such as multiband and multiresolution approaches [36]. In [24], Dufaux implemented detection and classification system of environmental sounds for a real-world application. The evaluation is performed using two classifiers GMM and HMMs. The approach used multimodels, so adapted to noisy environment.

The second area uses a set of higher level features which are generally extracted by a dictionary of audio atoms called also acoustic atoms. We can mention as example the non-negative matrix factorization (NMF), the authors in [37] proposed to use NMF in order to obtained bases. However, these bases are converted into MFCCs.

Benetos [38] proposed to use shift-invariant probabilistic latent component analysis (SIPLCA) and hidden Markov models (HMMs) to classify audio sounds. This approach enhances performance compared with spectral features. Furthermore, the authors in [35] proposed also an approach based on convolutive non-negative matrix factorization (NMF) in order to classify acoustic events. The goal of NMF is to capture parts-based decompositions of data. Using NMF is more efficient than the MFCC-based system, and a combination of the two systems assures even better than using individually features. Chu et al. [40] used as feature the matching pursuit (MP) to get a robust time-frequency descriptor but the combination of (MP) with MFCCs accomplish better audio sound recognition. Also, Ebenezer et al. [41] are interested to use MP for signal classification. Umapathy et al. [42] considered also as features an adaptive time-frequency transform, that is based on MP with Gaussian functions. They used the parameters of their sound decomposition.

Our aim in this paper is to study environmental sounds classification system using Scatter wavelet as features extracted by a dictionary of audio atoms in order to discriminate between various sound classes.