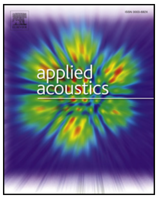
Applied Acoustics 139 (2018) 113–118

Contents lists available at ScienceDirect

Applied Acoustics

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



Global statistical features-based approach for Acoustic Event Detection

S.L. Jayalakshmi[a], S. Chandrakala[b,*], R. Nedunchelian[c]

[a] Department of CSE, Velammal Engineering College, Chennai, Tamil Nadu, India
[b] School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India
[c] Department of CSE, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

ARTICLE INFO

Keywords:
Acoustic Event Detection
Audio surveillance
Support Vector Machine
Hidden Markov model
Mel-frequency cepstral coefficients
Statistical features

ABSTRACT

The analysis of acoustic data typically discusses the problem of segmenting the acoustic events into non-overlapping acoustically compact categories. In Acoustic Event Detection (AED), an acoustic event is categorized into speech and non-speech events. Detection of non-speech sounds such as scream, gun shots, explosions, and glass break events is very helpful in acoustic surveillance, multimedia information retrieval, and acoustic forensic applications. In this paper, we propose global statistical features-based representation for multi-variate varying length acoustic data. A discriminative model-based classifier is then used to classify different acoustic events. The proposed representation is of very less dimension. The proposed approach is evaluated on surveillance-oriented AED datasets such as CICESE (recorded from a smart room scenario), Environmental Sound Classification (ESC), and IEEE AASP/DCASE2013 (Office environment) datasets. The proposed approach gives a better performance when compared with the conventional Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) approaches.

1. Introduction

Audio information retrieval has been a popular topic of research over the last few decades and, being a subfield, Acoustic Event Detection (AED) play an important role. AED typically addresses the difficulty of segregating an isolated audio event into one of the meaningful acoustically compact categories [1,2]. The applications of monitoring for security require automatic audio-based surveillance systems [3–5] in various environments, for instance, recognition of sounds in indoor environments such as an office, auditorium, hospital, smart-homes, and recognition of sounds in outdoor environments such as a beach, road, street, and public places. Some of the sound events include sneezing, laughing, crying of a baby, siren, glass breaking, barking of dogs, or fireworks. The information about non-speech sounds can be aided by the social activities. Examples include the ringing of a phone in a meeting, an explosion in a public place, and rushed footsteps to a hospital. Such information is very helpful in the applications of Health Care Monitoring [6–8], and forensic applications [9]. Hence, it is essential to design an automated sound event detection system.

The analysis of acoustic data has two perspectives: acoustic signal processing and modeling acoustic data. Acoustic signal processing mainly focuses on robust and task-specific features that give better performance. The second perspective involves representation and modeling of acoustic data. This research work mainly focuses on compact representation and discriminative framework for AED using features that are proved to be effective. Various machine learning algorithms were designed to classify sound events in the form of multi-variate varying length acoustic data into some meaningful categories. Most of the approaches focus on finding the best representation of features. AED algorithms are a simple reflection of speech recognition paradigms only. However, on accounting the significance of varying lengths of pseudo-stationary (non-stationary) characteristics of environmental sounds, these algorithms are proved to be ineffective. For example, the speech recognition task often exploits the phonetic structure. But general environmental sounds such as ‘typing’ or ‘gunshot’ do not have any phonetic structure. Another limitation of AED is lack of publicly available datasets. Some of the publicly available datasets are RWCP [10], CICESE [11], ESC-50 [12], and TUT [13].

In this paper, we propose a compact global statistical features-based representation with a discriminative modeling for AED. Thus, the difficult challenge of representing the sound event in the form of multi-variate varying length sequences is converted into fixed dimension representations. The proposed representation reduces the dimensionality of the sound event significantly. The rest of the paper is organized as follows: Section 2 presents a review of methods for isolated AED. The proposed approach is presented in Section 3. Sections 4 presents the studies carried out and the performance analysis of AED.

[*] Corresponding author.
E-mail addresses: sckala@gmail.com (S. Chandrakala), nedun@svce.ac.in (R. Nedunchelian).

https://doi.org/10.1016/j.apacoust.2018.04.026
Received 27 November 2017; Received in revised form 18 February 2018; Accepted 19 April 2018

2. Related work

The detection of audio events includes usually feature extraction and classification. The features can be divided into the following two types: (i) local-domain features and (ii) global features. In the local domain extraction, the input audio signal is segmented into homogeneous frames and then features are extracted from each frame. Some of the local-domain features are called Zero-Crossing Rate (ZCR), Linear Predictor Coefficient (LPC), Linear Predictive Cepstral Coefficient (LPCC), and Log Frequency Cepstral Coefficient (LFCC) [14]. On the other hand, global features are extracted from a whole duration of sound signal that represents an audio event. Some of the global features are pitch, energy, duration and formants [15].

Generative model-based paradigm and discriminative model-based paradigm are the two main paradigms to design pattern classifiers. Generative model-based approaches such as Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) rely on a learned model of the joint probability distribution of the observed data and the corresponding class membership. They use this joint probability model to perform the decision making based on the posterior probabilities of the classes computed using the Bayes rule. These approaches are not suitable for classifying the examples of confusable classes because a model is built for each class using the examples belonging to that class only. The Hidden Markov Model (HMMs) are the most widely used classifiers in Acoustic Event Detection (AED) field [16,17]. Discriminative model-based classifiers [18] such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) focus on modeling the decision boundaries between classes. However, the SVM is a kernel method-based discriminative classifier which requires fixed dimensional feature vectors as an input.

Crocco et al. [3] investigated a number of sound features such as ZCR, spectral flatness, spectral roll-off, spectral centroid, MFCC, LPCC, wavelet coefficients, Gammatone Cepstral Coefficients (GTCC) and Intonation and Teager Energy Operator (TEO) based features. The generative model-based classifiers such as HMM and GMM were used for classifying various acoustic events. A system based on MFCC features and an HMM classifier were used to classify the acoustic events from real-life audio recordings with different background noises [19]. Atrey et al. [4] presented a hierarchical approach using a GMM and four different audio features ZCR, LPC, LPCC, and LFCC. The proposed top-down event detection approach, first classifies a given audio frame into vocal and nonvocal events, and then performs further classification into normal and abnormal events. LFCC performed better in discriminating between the vocal and nonvocal events.

Automatic feature learning is one of the rapid advances in machine learning. The performance of machine learning algorithms usually depends on data representation [1]. Bag-of-words models have been most widely used for acoustic event representation [20]. Plinge et al. [20] demonstrated the Bag of Super Features–Pyramid (BOSF–P) approach using multinomial maximum-likelihood classifier for Acoustic Event Detection. The mel and gammatone frequency cepstral coefficients were used as an input to Bag-of-Features representation. Supervised learning of codebooks with temporal coding was shown to improve the recognition accuracies. Other acoustic atoms-based representations have also been employed, such as exemplar coding [21], sparse sooding [22] and non-negative matrix factorization (NMF) [16]. Recently, Shuyang et al. [23] proposed a novel K-medoid clustering-based active learning method on MFCC features to exclude the annotation effort to train sound event classifiers. In recent years, deep learning-based approach have been explored for automatic feature learning, where the layerwise stacking of extracted features forms a better representation with deep architectures. The resulting deep features can be used as an input to classifiers. However, while it has been shown a development for application of audio event detection to gain the robustness under noisy conditions [24].

Another discriminative classifier SVM provides a better generalization for unseen data. Salamon et al. [25] carried out an experiment on UrbanSound8K dataset with MFCC as features and SVM as classifier. The short temporal scale audio events such as ‘gun shot ’ and ’siren’ were clearly identified compared to other classes such as ‘car horn’ and ‘dog barking’. Rabaoui et al. [5] reviewed the efficiency of various acoustic features as well as the influence of features combinations. The results of optimized one-class support vector machines (1-SVMs) with set of wavelet coefficients, over-performed the conventional HMM-based system.

In this paper, our focus is to propose a compact global statistical features-based representation with a discriminative classifier for AED in order to discriminate between various audio events.

3. Proposed global statistical features-based approach for AED

The choice of global statistical features-based representation influences the outcome of AED system. In general, global statistical features such as mean, variance, skewness, and kurtosis extracted from local features are superior with the following advantages: (i) the number of global statistical features required is less compared with local features; (ii) global statistical features-based representation provides better discrimination among sounds, thus increases the classification accuracy; (iii) compact representation reduces the time required for training and classification; (iv) a range of statistical feature values extracted from examples belonging to one class is expected to be different from that of a range of statistical feature values of examples belonging to other classes.

In this work, we propose compact global statistical features-based representation with an SVM as classifier for AED. MFCC features are predominantly used for acoustic event recognition with good performance results [16,11,25]. MFCCs [3] were calculated from Mel-frequency filter banks because this filter bank imitates the perceptual behavior of human ear by using a nonlinear frequency scale called Mel scale. It also provides a compact representation of the formant structure by discarding the harmonic structure of the sound signal.

The block diagram of MFCC feature extraction process is shown in Fig. 1. The first step is preprocessing of the input sound signal by applying the following operations to it: scaling, frame-blocking, and windowing. Then discrete fourier transform (DFT) is applied to each frame. The output represents the power spectrum of the frame of a sound signal, which is then multiplied by a finite number of triangular Mel-scale filter values. This step filters the time domain values for the corresponding frequency domain input signal and is called as filter-bank energies (FBEs). Then, the logarithm of all FBEs is taken. Finally, the discrete cosine transform (DCT) of the log filter-bank energies are calculated. These DCT coefficients are used as Mel-frequency cepstral coefficients. But only 12 of the 26 DCT coefficients are kept. This is because the higher value of DCT coefficients represents fast changes in the filterbank energies and it turns out that these fast changes actually degrade AED performance, so dropping some of the DCT coefficients will improve the system performance. The first coefficient represents the energy of the signal.

Let $E = \{v_1,v_2,\dots,v_m\dots,v_M\}$ be an acoustic event with M feature vectors, where $v_j = [v_{j1},v_{j2},v_{j3},\dots,v_{jD}]$. D is the number of features in a feature vector. For each acoustic event, we compute the global statistical features across the 13-dimensional MFCC features. Mean, median, minimum, maximum, variance, skewness, and kurtosis are computed as global statistical features. The first two moments of the statistical distribution are known as the mean and the variance. The degree of deviation from the mean is captured by variance.

Skewness and kurtosis are associated with the third and fourth moments of the distributions, respectively. Skewness measures the asymmetry of the distribution and is calculated from normalized values of the third central moment. Skewness for the i-th feature is calculated from the following equation: