Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Classification of audio scenes with novel features in a fused system framework

Shefali Waldekar, Goutam Saha

Dept of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Block-based MFCC
CQCC
Environmental sounds
Fusion
Machine listening
SCFC

ABSTRACT

The rapidly increasing requirements from context-aware gadgets, like smartphones and intelligent wearable devices, along with applications such as audio archiving, have given a fillip to the research in the field of Acoustic Scene Classification (ASC). The Detection and Classification of Acoustic Scenes and Events (DCASE) challenges have seen systems addressing the problem of ASC from different directions. Some of them could achieve better results than the Mel Frequency Cepstral Coefficients – Gaussian Mixture Model (MFCC-GMM) baseline system. However, a collective decision from all participating systems was found to surpass the accuracy obtained by each system. The simultaneous use of various approaches can exploit the discriminating information in a better way for audio collected from different environments covering audible-frequency range in varying degrees. In this work, we show that the frame-level statistics of some well-known spectral features when fed to Support Vector Machine (SVM) classifier individually, are able to outperform the baseline system of DCASE challenges. Furthermore, we analyzed different methods of combining these features, and also of combining information from two channels when the data is in binaural format. The proposed approach resulted in around 17% and 9% relative improvement in accuracy with respect to the baseline system on the development and evaluation dataset, respectively, from DCASE 2016 ASC task.

1. Introduction

The research in audio processing is mostly concentrated around speech and music signals. However, speech or music recorded in natural environments consist of other additional sounds, too. In the state-of-the-art information retrieval systems for real-time speech/music signals, the background information is either considered useless and so it gets discarded in the pre-processing stage [1], [2], or at the most, it is marked as *environmental sounds* without any further analysis [3]. Nonetheless, for real-life systems, such signals can provide useful information about the acoustic scene from where the audio is captured. *Acoustic scene classification* (ASC) [4] is a closed-set classification task, where semantic labels are assigned to audio streams according to the environments they belong to. These environments could be indoor (home, office, library etc.), outdoor (busy-street, forest, beach etc.), or a moving vehicle (car, bus, train etc.). ASC is getting attention these days due to an increased need of context or situational awareness. It is related to *machine listening* process, which also includes similar research fields like computational auditory scene analysis (CASA) [5], soundscape cognition [6], and audio event detection (AED) [7]. As compared to the well-established field of automatic speaker recognition [8], ASC could be considered analogous to speaker identification [9], while AED resembles speaker diarization [10,11].

Nowadays, a large section of the world population uses mobile phones, in which smartphone users are rapidly growing. Intelligent wearable devices are also gaining popularity at a fast pace. Information extracted from sound, by such portable devices, can help them make sense of the surroundings. As compared to video, audio has the advantages of robustness towards changing ambient conditions and ease of recording, storing and analysis. Moreover, unlike video, audio recordings are hardly affected by the device's position. Thus, ASC is a useful technology for such context-aware devices that continuously monitor the environment around them and accordingly perform specific tasks without the need of human intervention. Other important applications where ASC can be directly useful are hearing-aids, robotic navigation systems, and audio archive management systems.

Among the first recorded works in ASC, speech and audio features such as RASTA analysis, power spectral density (PSD) and frequency bands from a filter bank, were used, along with recurrent neural networks and k-nearest neighbor as classifiers [12]. In [13], Hidden Markov models (HMM) were used for modeling

E-mail addresses: shefaliw@ece.iitkgp.ernet.in (S. Waldekar), gsaha@ece.iitkgp.ernet.in (G. Saha).

https://doi.org/10.1016/j.dsp.2017.12.012
1051-2004/

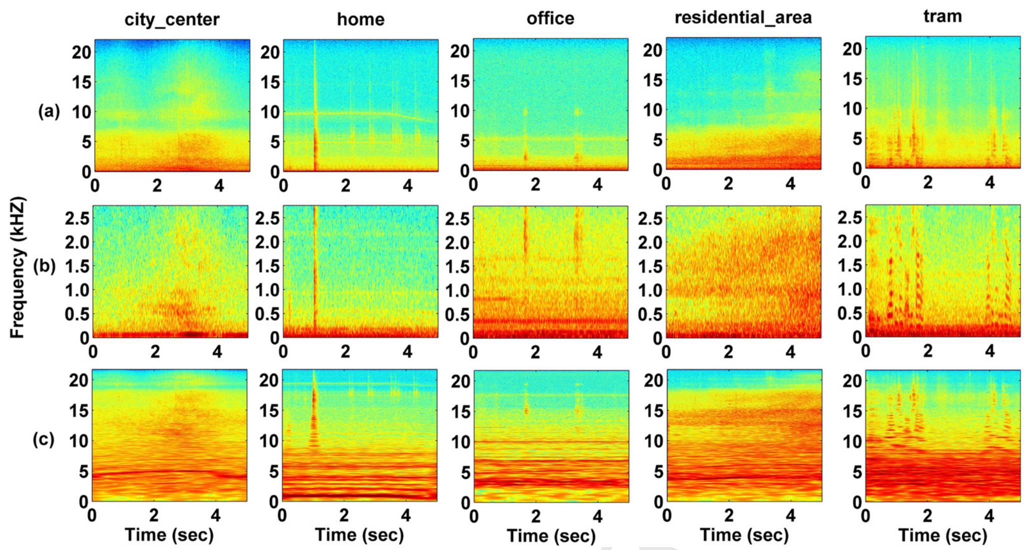

Fig. 1. Spectrograms for randomly selected 5 s data from five classes of DCASE 2016: (a) STFT with maximum frequency $F_s/2$, (b) STFT with maximum frequency $F_s/16$, (c) CQT with maximum frequency $F_s/2$. F_s is the sampling frequency.

and classification, while the features employed were Mel-scaled filter-bank coefficients (MFCs) and pitch estimates. The analysis of temporally structured data, such as audio streams, requires a method that can efficiently produce a single label that represents the time-based media object appropriately. In the so-called "bag-of-frames" (BoF) approach, a scene is represented by a long-term statistical distribution of some set of short-term spectral features. Most commonly used features are Mel frequency cepstral coefficients (MFCCs), while Gaussian mixture models (GMMs) are used for comparison of distributions [14]. This is a relatively simple and the most widely used approach for audio classification, be it speech, music or other sounds. It was claimed in [14] that this system is sufficient for recognizing urban soundscapes, with a 96% accuracy of classification obtained on a dataset covering four classes. However, in a recent study [15], it was shown that the BoF system is no better than the much simpler one-point average approach when evaluated on other three currently available audio scenes datasets with less within-class variability. The relatively high performance obtained in [14] was attributed to an exceedingly appreciative dataset that consisted of only 16 recordings spanning four outdoor scenes, with abnormally low within-class variability. In other words, the system was over-fit and therefore could not perform satisfactorily on other larger datasets. A combination of time, frequency and wavelet domain features with GMM-HMM as the universal background model for acoustic surveillance of urban environments was used in [16].

Another strategy to represent a time-based media object is to use higher level features, captured using a vocabulary or dictionary of "acoustic atoms", as intermediate representations to model the scene before classification. The atoms are usually learned from the data in an unsupervised manner. Better discrimination and classification can be obtained with the help of sparsity or other constraints. For example, non-negative matrix factorization was applied to train-station scene classification in [17] in order to extract bases which were later converted to MFCCs. Time-frequency (TF) features obtained from matching pursuit algorithm (where the dictionary atoms are generated from Gabor functions), complemented MFCCs for environment sound classification in [18]. The experimental results showed promising performance, which was comparable to human classification results, on a database collected from 14 different environments. In [19], the acoustic scene signals were transformed into TF representations and estimated features were based on the histogram of gradients (HoG). These features carried information about the shape and evolution of the TF structures, and they resulted in improved performance on multiple databases when fed to a linear support vector machine (SVM) classifier.

The challenge on *detection and classification of acoustic scenes and events* (DCASE) was organized in 2013 and 2016 with the goal of stimulating research in the field of machine listening with respect to general environmental sounds [20,21]. The baseline system for ASC provided with DCASE challenges is the BoF system. The best performing algorithm for DCASE 2013, which employed recurrence quantification analysis (RQA) of MFCC features in addition to MFCC features [22], showed a mean accuracy of 76%. This was comparable to the median accuracy of human listeners and far better than the 55% mean accuracy of the baseline. In the 2016 challenge, the reported baseline performance was 72.5% for the development set and 77.2% for the evaluation set. Of the 48 submissions in this challenge, the top-ranking system used late-fusion of binaural i-vector and deep convolutional neural network (DCNN) architecture trained on spectrograms of audio excerpts in an end-to-end fashion, and it reportedly achieved 89.9% and 89.7% accuracy on the development and the evaluation datasets, respectively [23].

1.1. Motivation and contributions

It was shown in the analysis of the results of DCASE 2013 [4] that a majority vote of the classification decisions from all eleven participating systems had higher accuracy than the best performing participant in the challenge. It was also shown that discriminative methods (e.g. SVM) of classification performed better than the generative methods (e.g. GMM). The results from previous research in this field and the analysis in [4] show the need for greater clarity on a suitable feature-classifier combination. Also, the variety of environmental sounds and consequently the acoustic scenes that they can form is large. Therefore, a particular kind of feature may not be sufficient to effectively and also discriminatively represent them. This can also be observed from Fig. 1, which shows different spectrograms of randomly selected five seconds from samples of five classes (two indoor, two outdoor, and one moving vehicle) from the development data of DCASE 2016. It can be seen in the first panel (Fig. 1(a)), which shows full STFT spectrograms, that although the energy concentration is more in lower frequencies (like any natural signal), all classes differ from each other in their spectral characteristics. A closer look is provided in Fig. 1(b), where STFTs were evaluated for one-eighth of the sampling frequency. It

Please cite this article in press as: S. Waldekar, G. Saha, Classification of audio scenes with novel features in a fused system framework, Digit. Signal Process. (2018), https://doi.org/10.1016/j.dsp.2017.12.012