WAVELET BASED MEL-SCALED FEATURES FOR DCASE 2019 TASK 1A AND TASK 1B

Technical Report

Shefali Waldekar, Goutam Saha

Electronics and Electrical Communication Engineering Dept.,
Indian Institute of Technology Kharagpur, India,
{shefaliw, gsaha}@ece.iitkgp.ernet.in

ABSTRACT

This report describes a submission for IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 for Task 1 (acoustic scene classification (ASC)), sub-task A (basic ASC) and sub-task B (ASC with mismatched recording devices). The system exploits time-frequency representation of audio to obtain the scene labels. It follows a simple pattern classification framework employing wavelet transform based mel-scaled features along with support vector machine as classifier. The proposed system relatively outperforms the deep-learning based baseline system by almost 8% for sub-task A and 26% for sub-task B on the development dataset provided for the respective sub-tasks.

Index Terms— Haar function, spectral features, SVM, wavelet transform.

1. INTRODUCTION

Acoustic scene classification (ASC) [1] is a supervised classification task, where semantic labels are assigned to audio streams according to the environments they represent. These environments could be indoor, outdoor, or a moving vehicle. Applications of ASC can be in context-aware and intelligent wearable devices, hearing-aids, robotic navigation, surveillance and audio archive management systems.

Any signal is generated when a system is excited by a source. In case of a speech signal, the excitation is from a person's vocal chords that provide input to the vocal tract which is working as the transformation system. On the other hand, audio signals recorded from dynamic real-life environments are a superposition of various audio events occurring simultaneously. In other words, multiple audio sources overlap to make an acoustic scene. These signals are less structured than speech/music signals. Efficient representation of such multifaceted signals can be expected from features that capture local information in both time and frequency domains. In this report, we present an ASC system employing a time-frequency feature which has been successful in different audio processing fields, and a discriminative classifier at its core. The feature is called *mel-frequency discrete wavelet coefficients (MFDWC)* [2]. In the present work, we investigate the feature's performance in classification of data from the same device as the available training data (sub-task A) and classification of data recorded with devices other than the training data (sub-task B). The classifier employed is a support vector machine (SVM) with intersection kernel [3].

The rest of this report is organized in the following way: In Section 2, we give the description of the elements of the proposed system and elaborate on the experimental framework. In Section 3, we present the results. It is followed by the conclusion of the work in Section 4.

2. PROPOSED SYSTEM CONFIGURATION

The proposed ASC system incorporates the general pattern classification framework as shown in Fig. 1. All the incoming audio signals go through pre-processing and feature extraction. Models are built from training data and then employed for classification of the test samples.

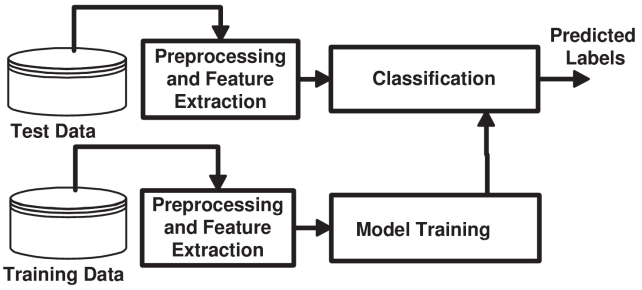

Figure 1: Schematic of general pattern classification system.

2.1. Features

2.1.1. Mel-frequency discrete wavelet coefficients (MFDWC)

In all fields of speech processing, mel-frequency cepstral coefficients (MFCC) are the most exploited features. One of the important steps in MFCC extraction is discrete cosine transform (DCT). The basis vectors of DCT span the whole frequency range of the signal. As a result, corruption of a band due to noise affects all the coefficients. Also, the DCT basis vectors have fixed time-resolution for all frequencies. Use of discrete wavelet transform (DWT) instead can deal with these issues because it has better time and frequency localization capacity, while simultaneously serving the need of DCT [4]. Unlike Fourier based transforms, wavelet transform uses short basis functions for high-frequency content and long basis functions for low-frequency content of a signal. Environmental audio data mostly carries short high-frequency transients and long-lasting low-frequency background noise at the same time [5]. The low-frequency components get subsumed by the first few filters of the mel scaled filterbank imparting high values to the coefficients

Figure 2: MFDWC feature extraction scheme.

in this region. The high-frequency transients correspond to the audio events typical to a class of environment. High values in the latter coefficients mark the presence of these scene distinguishing events. Wavelet based features are especially efficient in characterizing the impulsive parts of the audio [5]. The time-frequency localization property therefore enables wavelets to capture such diverse frequency elements of environmental audio. Discrete wavelet transform (DWT) applied to mel-filterbank log-energies results in MFDW coefficients [4]. The feature extraction scheme is same as that of MFCC, except that the DWT replaces DCT, as shown in Fig. 2 [2].

In many speech processing applications, dynamic coefficients, that is, discrete-time derivatives of features computed from local frames are used as features. We observed in our experiments that the first derivatives (i.e., delta or velocity features) improved the performance [2]. The addition of the second derivatives (i.e., double-delta or acceleration features) did not prove beneficial.

2.2. Classifier

2.2.1. Support Vector Machine (SVM)

In our system, we have used SVM with intersection kernel. This kernel uses the intersection between the features of the two classes as a measure of similarity [3]. SVM is inherently a binary classifier. For multi-class classification problem one can go for either one-vs-one or for one-vs-all approach. The classifiers obtained by the first method are typically smaller and require fewer resources than the second method. Moreover, it has also been shown that the former are marginally more accurate than the latter on standard classification tasks [6]. Therefore, we have used the one-vs-one approach, consequently training $N(N-1)/2$ classifiers for N classes. SVM requires that each data sample is represented as a vector. The mean and standard deviation of a feature matrix can act as its vector representation [7].

2.3. Experimental Framework

We have used the development dataset of TAU Urban Acoustic Scenes 2019 (TAUUAS19D) for sub-task A (basic ASC) and TAU Urban Acoustic Scenes 2019 mobile (TAUUASM19D) for sub-task B (device mismatch ASC) [8] in our experiments. In both the development datasets, training and testing portions were provided. The audio of DCASE challenge recorded with Device A is in binaural format [8], One possible way of working with such data is to first convert the audio to monophonic by averaging the two channels [7]. All the audio signals, either after binaural-to-mono conversion for device A or as they are for other devices, were framed by Hamming window of 40ms with 50% overlap after pre-emphasis by a factor of 0.97. A mel-scaled filterbank with 100 triangular filters was used for feature extraction. The choice of the number of filters is based on the results shown in Table 1. In general, the mother wavelet is chosen such that its shape is similar to the signal which is to be decomposed. In case of acoustic scenes signals it is difficult to fix a particular wavelet that is compatible with the myriad sound events. Haar function was used as the mother wavelet for DWT. Delta features were extracted with 3-frame windows. Frame-wise mean and standard deviation of the features was given as input to SVM classifier with intersection kernel.

Table 1: Class-wise mean accuracy (%) for different number of filters on sub-task A development dataset. Bold-face: Maximum mean accuracy in the column.

Number of filters	Accuracy (%)
40	63.18
60	66.48
80	66.66
100	**67.39**
120	67.12

3. RESULTS

The results of the two sub-tasks on their respective datasets are shown in Table 2. In the present challenge, class-wise mean accuracy is used as the performance metric. The mean accuracy of all classes reported for the logMBE-CNN baseline system for sub-task A is 62.5%. Thus, by obtaining a mean class-wise accuracy of 67.39%, our proposed system has relatively outperformed the deep-learning based baseline by nearly 8% on the development dataset of sub-task A. Class-wise performance comparison of the two systems for this sub-task is depicted in Fig. 3(a). The darker shades in the diagonal of the proposed system's confusion matrix exhibit its good ability to classify all scene classes with 'park' and 'street_traffic' category doing the best.

The performance of the proposed system for the three devices of sub-task B is also shown in Table 2. It can be seen that the proposed system performed better than the baseline for all three devices. According to the rules of the challenge for this sub-task, the ranking of the systems will be done by the average performance with devices B and C only. The reported baseline accuracy in this case is 41.4%. Our proposed framework achieved 52.32%, which is more than 26% relatively better. The pictorially represented confusion matrix in Fig. 3(b) is showing the average performance with device B and device C. Here we see that 'airport', 'public square' and 'tram' suffer the most mis-classifications. Many samples from 'airport' class are wrongly labeled as another *indoor* class, 'shopping mall'. The system has found it challenging to recognize the data from 'public square' class and distributed it among not only *outdoor* classes but also *indoor* classes. Most erroneous classifications of 'tram' samples is into other *vehicle* classes.

For both the sub-tasks, the proposed system seems to be finding it more difficult to distinguish among *vehicle* classes than the classes belonging to other types of environments. Another general observation is that result-improvement on data recorded with the