ACOUSTIC SCENE CLASSIFICATION BASED ON SPECTRAL ANALYSIS AND FEATURE-LEVEL CHANNEL COMBINATION

Dinesh Vij[*], *Naveen Aggarwal*[†] *Bhaskaran Raman* *K.K.Ramakrishnan* *Divya Bansal*

UIET, Panjab University, Chandigarh
[*]vijdinesh@gmail.com
[†]navagg@gmail.com

IIT, Bombay
br@cse.iitb.ac.in

University of California, Riverside
kk@cs.ucr.edu

PEC University of Technology, Chandigarh
divya@pec.ac.in

ABSTRACT

This paper is a submission to the sub-task Acoustic Scene Classification of the IEEE Audio and Acoustic Signal Processing challenge: Detection and Classification of Acoustic Scenes and Events 2016. The aim of the sub-task is to correctly detect 15 different acoustic scenes, which consist of indoor, outdoor, and vehicle categories. This work is based on spectral analysis, feature-level channel combination, and support vector machine classifier. In this short paper, the impact of different parameters while extracting features is analyzed. The accuracy gain obtained by feature-level channel combination is then reported.

Index Terms— Acoustic Scene Classification, MFCC, Wavelet Packets, SVM, Binaural, Channel Combination

1. INTRODUCTION

This short paper describes our submission to the sub-task Acoustic Scene Classification of the Detection and Classification of Acoustic Scenes and Events 2016 (DCASE 2016) challenge. It is 2nd official IEEE Audio and Acoustic Signal Processing challenge, organized by IEEE Signal Processing Society. The field of Acoustic Scene and Event Detection is generally overshadowed by the traditional field of Automatic Speech Recognition (ASR). The features which have been widely used in speech recognition community such as Mel Frequency Cepstral Coefficients (MFCCs) are used as it is, in the field of Acoustic Scene and Event Detection. Furthermore, the parameters such as window size are also kept same while extracting these features. This short paper has analyzed the relevance of different parameters for the scene classification task, and suggested the required modifications for this task. Then the accuracy gains obtained by using Wavelet Packet Transform (WPT) features and feature-level channel combination are further presented. An overview of the whole system is shown in Figure 1. The input binaural signal is split into two channels, and signal from each channel is processed separately. The initial results reported in this paper correspond to the channel 1, which are further augmented by feature-level combination of channel 1 and channel 2 as discussed in section 4. The following subsections provide more details on each system block.

This research is supported by the CARTS Project funded by ITRA, Media Lab Asia.

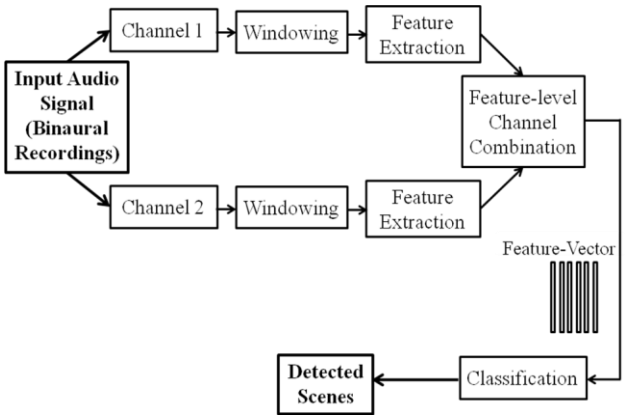

Figure 1: Basic architecture of Acoustic Scene Classification.

The paper is organized as follows: Section 2 describes the windowing process. Section 3 details the different features used for capturing the characteristics of various acoustic scenes. Feature-level channel combination and classification technique used is discussed in section 4. The evaluation results obtained after 4-fold cross-validation are presented in section 5. Finally, section 6 concludes the paper and gives future directions. In section 7, the classification results achieved on evaluation dataset (after challenge completion) are discussed.

2. WINDOWING

In speech recognition, short windows of 25ms with a step of 10 ms are used so as to correctly detect the basic unit of speech i.e. phoneme. The speech signal changes too much in large windows. Therefore, the audio signal in ASR is segmented using small windows to capture the pattern of different phonemes. In case of acoustic scenes, there is no such fixed small pattern. In case we have to detect events within a scene, then the abrupt changes (which can be captured by small windows) needs to be taken care of. But for cumulative acoustic scene detection, larger window sizes are more appropriate because the spectral characteristics of cumulative acoustic scenes do not change significantly over the short time spans. Therefore, primary analysis windows of size 40ms (DCASE baseline default), 125ms, 250ms, 500ms, 1s, 2s, 3s, 4s, and 5s are tested in this work. The win-

dows are then shifted over time in a way, so that the new analysis window overlaps with the previous one. A window overlap of 0% to 90% is tested to obtain a time series of feature vectors. Although the different classes resolute at different window sizes, but an overall window size of 2s with 50% overlap is found to be appropriate for all the classes.

3. FEATURE EXTRACTION

Four different types of features are extracted from the 2s windows mentioned above, to capture the distinctive acoustic signatures of various scenes. The extracted features are then normalized to the same scale using z-score normalization.

3.1. Mel Frequency Cepstral Coefficients (MFCCs)

MFCC features have been widely used for speech recognition. Also, it has been suggested in the literature that higher MFCC coefficients represent fast changes in the filterbank energies and these fast changes degrade ASR performance. Therefore only first 13 MFCC coefficients have traditionally been used by most of the researchers. Cumulative acoustic scenes change even more slowly than the phoneme change of ASR. So, we proposed that similar detection accuracy can be achieved by using lesser number of MFCC coefficients. Experimental results revealed that different classes showed best results at different number of MFCC dimensions, but overall best results were obtained with 9-d MFCCs. This technique proves out to be a very simple dimensionality reduction technique based on the characteristics of cumulative acoustic scenes. Rastamat library [1] is used for extracting MFCC features. The 0^{th} coefficient of 9-d MFCC is replaced with true log energy. Then, mean (9-d) along with the standard deviation (9-d) of the MFCC features extracted from multiple windows is calculated, to obtain a single feature vector for each recording. Overall accuracy obtained by 18-d MFCC features with parameter tuning is 71.61%, which proves out to be better than DCASE baseline accuracy. Our argument was further supported by the fact when we applied pre-emphasis filter, and it had no effect at all in increasing the accuracy of scene detection. Pre-emphasis is used in speech recognition to compensate for the rapid decaying spectrum of speech. But in case of cumulative acoustic scenes, the spectrum decays slowly. Although, few acoustic scenes such as "train" and "tram" showed better results at higher MFCC dimensions, probably due to high frequencies present in these scenes, yet overall best results for the 15 different acoustic scenes were obtained with 9-d MFCCs.

3.2. Spectral Centroid

It measures the "center of mass" of the spectrum or brightness of sound. Different acoustic scenes have different center of mass. Therefore, 1-d spectral centroid feature is calculated for each 2s window. Then, mean (1-d) and standard deviation (1-d; to measure the spread of spectrum around the mean of spectral centroid) of the features extracted from multiple windows is calculated, to obtain a single feature vector for each recording. MFCC combined with Spectral Centroid improved the classification accuracy from 71.61% to 72.68%.

3.3. Spectral Flux

It measures the rate of change of local spectral information i.e. the squared difference of the power spectra between two adjacent frames. Power spectra of different acoustic scenes may vary differently in their local neighborhoods. Therefore, 1-d spectral flux is calculated for each window. Then, mean of the features extracted from multiple windows is calculated, to obtain a single feature vector for each recording. Addition of Spectral Flux feature improved the classification accuracy from 72.68% to 74.74%.

3.4. Wavelet Packet Transform (WPT)

The spectral structure of acoustic scenes is different from that of speech. Even the properties of one type of acoustic scene differ from another. Few scenes consist of only low, mid, or high frequencies, while frequency spectra of others constitute a mix range of frequencies. Therefore, conventional speech features such as MFCC have a limited power in recognizing different scenes. Wavelets can be used in combination with MFCCs for a comprehensive description of the spectrum. Such a representation is successful at capturing differences among different classes. WPT filters a signal into equal-width subbands at each level, and partitions the signal's energy among the subbands. Daubechies4 (db4) wavelet with 5-level wavelet packet decomposition is used in this work. Log Root Mean Square features from the last level nodes (wavelet packets) of complete wavelet packet tree without node selection are used [2]. Addition of WPT features further increased the classification accuracy from 74.74% to 76.16%.

4. FEATURE-LEVEL CHANNEL COMBINATION AND SCENE CLASSIFICATION

To take advantage of the additional cues embedded in the binaural recordings of the dataset, features are extracted from both the channels separately. Then feature-level channel combination is performed so as to obtain a single feature vector for each recording. Combination of features from both the channels increased the detection accuracy of the system (79.63%) as compared to a single channel system (76.16%). For the scene classification task, our system follows a standard SVM-based approach using an RBF kernel. LIBSVM library [3] is used for modeling the final feature vectors with SVM.

5. EVALUATION

We evaluated our approach using a 4-fold cross-validation on the evaluation setup files (both training & testing) provided by the DCASE 2016 organizers. Average classification accuracy over the folds is then calculated. Development dataset is used for this purpose. It consists of 78 segments (30-sec each, totaling 39 min of audio) for each acoustic scene [4].

Final classification accuracy obtained after 4-folds cross-validation is 74.08% using the provided development dataset. Scene-wise classification results for the same are shown in Table 1. The final proposed model is then applied on Evaluation dataset, which consists of 26 segments (30-sec each, totaling 13

min of audio) for each acoustic scene, and the results are submitted to the challenge.

Acoustic Scene	Baseline System	Our Approach
Beach (outdoor)	71.9	73.94
Bus (vehicle)	62.0	69.61
Café (indoor)	83.9	81.40
Car (vehicle)	75.7	75.92
City center (outdoor)	85.6	83.57
Forest path (outdoor)	65.9	78.18
Grocery store (indoor)	76.6	88.29
Home (indoor)	79.4	62.78
Library (indoor)	61.3	72.28
Metro station (indoor)	85.2	97.55
Office (indoor)	96.1	90.88
Park (outdoor)	24.4	39.03
Residential (outdoor)	75.4	60.72
Train (vehicle)	36.7	63.22
Tram (vehicle)	89.5	73.89
Overall accuracy	**71.3**	**74.08**

Table 1: Scene-wise classification accuracy in % obtained after 4-fold cross-validation.

6. CONCLUSION AND FUTURE DIRECTIONS

Overall accuracy reported by the baseline system is 71.3%. Compared to the baseline system, our system provides a 2.78% absolute gain in the classification accuracy. Also, our system shows an improvement in accuracy in 9 of 15 acoustic scenes. The least accurate results correspond to the park scene (39.03%). Our system outperforms the baseline system for outdoor and vehicle categories, but higher detection accuracies are observed for indoor scenes in general. Future directions include representing acoustic scenes in terms of component acoustic events, and inferring important properties of these events which would aid in detecting the acoustic scenes associated with them. To differentiate between two scenes having common type of events, repetition frequency of individual events within each scene can be useful. Also, two scenes can be matched for some pre-defined sequence of events, as this sequence might be different for different scenes. Additional binaural features could also be considered to extract important information from the two channels of binaural recordings. Future work would also focus on early detection where minimum amount of recording needed by the classifier to correctly identify a scene would be tested.

7. RESULTS ON EVALUATION DATASET

Although our system outperforms the baseline system on the development dataset, similar results are not achieved on evaluation dataset. Baseline system achieved a better scene detection accuracy of 77.2% as compared to 74.4% of our proposed system. Nevertheless, our system achieved consistent detection accuracy on both the development (74.1%) and evaluation (74.4%) datasets, which proves the reliability of our approach.

8. REFERENCES

[1] http://labrosa.ee.columbia.edu/matlab/rastamat/.

[2] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Trans. Biomedical Engg.*, vol. 58, no. 1, pp. 121-131, Jan. 2011.

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th EUSIPCO*, Budapest, Hungary, 2016.