MEL-SCALED WAVELET-BASED FEATURES FOR SUB-TASK A AND TEXTURE FEATURES FOR SUB-TASK B OF DCASE 2020 TASK 1

Technical Report

Shefali Waldekar, A. Kishore Kumar, Goutam Saha

Electronics and Electrical Communication Engineering Dept.,
Indian Institute of Technology Kharagpur, India,
shefaliw@ece.iitkgp.ernet.in, kishore@iitkgp.ernet.in, 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) 2020 for Task 1 (Acoustic Scene Classification (ASC)), sub-task A (ASC with multiple devices) and sub-task B (low-complexity ASC). The systems exploit time-frequency representation of audio to obtain the scene labels. The system for sub-task A follows a simple pattern classification framework employing wavelet transform based mel-scaled features along with support vector machine as classifier. Texture features, namely local binary pattern, extracted from log of mel-band energies is used in a similar classification framework for sub-task B. The proposed systems outperform the deep-learning based baseline systems with the development dataset provided for the respective sub-tasks.

***Index Terms*—** Haar function, LBP, 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 systems, surveillance, and audio archiving systems.

With application point of view, it is required that the machine listening algorithms be such that they are able to work with different types of audio, that is, speech, music, as well as environmental sounds. In the systems presented in this report, we use some spectral features from audio processing fields. The motivation behind using these features, specifically, mel-frequency based features, was to be able to discriminate between acoustic scenes in a way similar to the human auditory system by the exploiting the spectral characteristics of the typical audio events that characterize the scenes. The features for both the sub-tasks are derived from log-mel band energies (LogMBE). For sub-task A, we extract mel-frequency discrete wavelet coefficients (MFDWC) [2, 3] by applying discrete wavelet transform to LogMBE matrix of an audio sample. We address the low-complexity three-class classification problem of sub-task B by analysing texture of the LogMBE matrix with the help of local binary pattern (LBP) [4]. The classifier for both the systems is a support vector machine (SVM) with intersection kernel [5].

The rest of this report is organized as follows: In Section 2, we give the description of the proposed system. Next, in Section 3 we elaborate on the formation of the system and the experimental configuration. In Section 4, we present the results. It is followed by the conclusion of the work in Section 5.

2. BASIC SYSTEM CONFIGURATION

2.1. Features

The proposed systems use the following as features.

- *Mel-frequency discrete wavelet coefficients (MFDWC)* [6]: 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). Discrete wavelet transform (DWT) applied to mel-filterbank log-energies results in MFDW coefficients. Wavelet based features are especially efficient in characterizing the impulsive parts of the audio [7]. The feature extraction scheme is same as that of MFCC, except that the DWT replaces DCT [**?**]. 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 for MFDWC. The addition of the second derivatives (i.e., double-delta or acceleration features) did not prove beneficial.
- *Texture of Log mel-band energies (LogMBE)* [8]: Local binary pattern is a theoretically and computationally simple method, most commonly used in face recognition [9]. LBP is a non-parametric approach that uses a local patch of an image and compares the magnitude of the pixels to assign the local pattern one binary code [10]. LBP has been successfully applied to spectrogram to get an anti-spoofing measure [8], to linear cepstrogram [11] and MFCC matrix [12] for environmental sound classification, and to mel-scaled filterbank energies in lung sound classification [13]. We have observed that textural properties of LogMBE can be used for general grouping of environmental audio as per their location type. Therefore, we have used histograms of uniform LBP of LogMBE as features for high-level labelling of environments, i.e. *indoor*, *outdoor* and *transportation*.

2.2. Classifier

In our system, we have used SVM with intersection kernel. This kernel uses the intersection between the features of the two classes

Table 1: Class-wise accuracy (%) of baseline and proposed system for sub-task A.

Classes	Baseline	Proposed
Airport	45.0	55.1
Bus	62.9	55.6
Metro	53.5	57.2
Metro station	53.0	51.5
Park	71.3	73.1
Public square	44.9	42.1
Shopping mall	48.3	47.1
Street, pedestrian	28.8	37.7
Street, traffic	79.9	79.8
Tram	52.2	59.7
Average	54.4	55.0

as a measure of similarity [5]. Since SVM is a binary classifier, in order to determine a decision criterion for multi-class ASC, we have combined multiple SVMs following one-versus-one approach. Thus, for N classes, $N(N-1)/2$ classifiers are made, where each one trains on data from two classes.

3. PROPOSED SYSTEM

The proposed systems follow a basic pattern classification framework. This involves feature extraction after pre-processing of raw data, followed by classifier modeling with the training data, and finally classification by supplying test data features to the trained model. In both the systems, the required features are extracted from windowed frames of pre-emphasized audio. These vectors are used to train the SVM corresponding to each feature. In the present challenge, the development data is pre-divided into one train and one test partition. It should be noted that some part of development data was neither used in train nor in test fold. The data for testing comes from the evaluation dataset and follows a path similar to that of development. However, in this case whole development data is used for training the SVMs.

3.1. Experimental Framework

For both the sub-tasks, the classification is performed by a single-feature single-classifier system. We have used the the development dataset of TAU Urban Acoustic Scenes 2020 Mobile (TAUUAS20-MD) for sub-task A and 3-class (TAUUAS19-3CD) for sub-task B [14] in our experiments. All the audio signals for sub-task A and sub-task B were framed by Hamming window of 40 ms and 20ms, respectively, with 50% overlap after pre-emphasis by a factor of 0.97. The filterbank used for MFDWC features in sub-task A had 180 filters while 60 filters were chosen for LogMBE extraction in sub-task B. Haar function was used as the mother wavelet for MFDWC extraction. Delta (Δ) features, evaluated with a 3-frame window, were appended only for this feature. For low-complexity ASC, histogram of uniform LBP analysis was given as input to SVM classifier with intersection (INT) kernel. Frame-wise mean and standard deviation of the features were used for 10-class classification with the same SVM configuration.

Table 2: Device-wise accuracy (%) of baseline and proposed system for Sub-task A.

Device	Baseline	Proposed
A	68.8	69.4
B	60.2	59.0
C	59.9	65.1
S1	50.3	55.8
S2	50.0	50.9
S3	50.9	58.2
S4	45.2	46.7
S5	44.8	47.3
S6	34.8	43.0

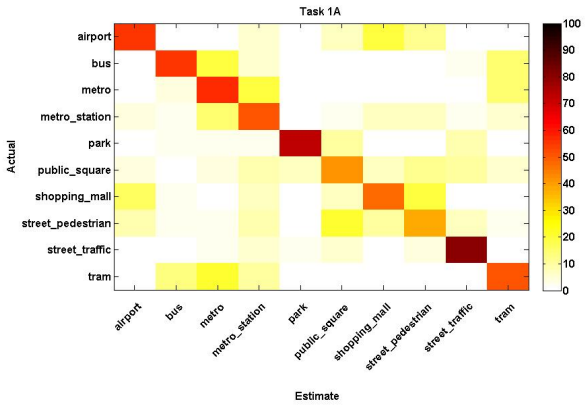

Figure 1: Confusion matrix of results of proposed MFDWC-SVM-INT system with TAU Urban Acoustic Scenes 2020 Mobile development dataset (sub-task A).

4. RESULTS

In the present challenge, class-wise mean accuracy is used as the metric. The mean accuracy of all classes reported for the openL3-DNN baseline system [15] for sub-task A is 54.1%. Thus, by obtaining a mean class-wise accuracy of 55.0%, our proposed system has outperformed the deep-learning based baseline with the development dataset of sub-task A. Class-wise performance comparison of the two systems for this sub-task is depicted in Table 1. The proposed system's results are also pictorially represented in Fig. 1. The two systems show almost equivalent accuracy for classes other than 'airport', 'bus', 'street_pedestrian' and 'tram'. Both systems' worst performance is on the 'street_pedestrian' class, although the proposed system is comparatively better. The proposed system classified the scenes from 'airport' and 'tram' classes far better than the baseline system.

From the task description of DCASE 2020's sub-task A, "This task targets generalization properties of systems across a number of different devices, and will use audio data recorded and simulated with a variety of devices". The data was recorded with three devices (A, B and C) and simulated with six devices (S1-S9). The comparative performance of the baseline and the proposed system with respect to the devices is given in Table 2. The proposed system has shown better performance for devices C, S1, S3 and S6.

The sub-task B's baseline system is similar to DCASE2019's

Table 3: Class-wise accuracy (%) of baseline and proposed system for sub-task B.

Classes	Baseline	Proposed
Indoor	82.0	86.7
Outdoor	88.5	90.5
Transportation	91.5	92.8
Average	87.3	90.0

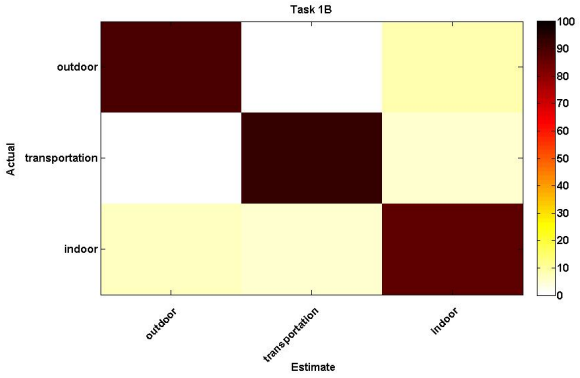

Figure 2: Confusion matrix of results of proposed LogMBE-LBP-SVM-INT system with TAU Urban Acoustic Scenes 2020 3-Class development dataset (sub-task B).

Task 1 baseline. The LogMBE-CNN based system has reported a 3-class accuracy of 87.3%. The proposed LBP-LogMBE-SVM based system has outperformed it with 90.0% accuracy. The class-wise comparison of the two systems is presented in Table 3. It can be seen that the proposed system's classification of scenes of the three classes is better than that of the baseline system, especially for 'indoor' class. It can be observed from Fig. 2, which shows the confusion matrix for the proposed system, that there is a considerable confusion between scenes from 'indoor' and 'outdoor' environments.

5. CONCLUSION

In this technical report, we have described a system for acoustic scene classification task (Task 1, sub-task A and sub-task B) of DCASE challenge 2020. The first sub-task is concerned with problem of ASC with multiple recording devices. On the other hand, sub-task B addresses a low-complexity application,in which all available data (development and evaluation) are recorded with the same device but the scenes are grouped into three high-level classes of *indoor*, *outdoor* and *transportation*. Our systems used well-known audio processing features along with SVM as classifier to produce classification better than the baseline systems for both the sub-tasks.

6. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] S. Waldekar and G. Saha, "Wavelet transform based mel-scaled features for acoustic scene classification," *Proc. Interspeech 2018*, pp. 3323–3327, 2018.

[3] ——, "Analysis and classification of acoustic scenes with wavelet transform-based mel-scaled features," *Multimedia Tools and Applications*, pp. 1–16, 2020.

[4] ——, "Texture features for high-level classification of acoustic scenes," in *2019 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2019, pp. 710–715.

[5] S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel SVMs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 66–77, 2013.

[6] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1351–1354.

[7] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on information forensics and security*, vol. 3, no. 4, pp. 763–775, 2008.

[8] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon: France (2013)*, 2013, p. 5p.

[9] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, 2011.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[11] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in *Acoustics, speech and signal processing (ICASSP), 2014 IEEE international conference on*. IEEE, 2014, pp. 3052–3056.

[12] W. Yang and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 6, pp. 1315–1321, 2017.

[13] N. Sengupta, M. Sahidullah, and G. Saha, "Lung sound classification using local binary pattern," *arXiv preprint arXiv:1710.01703*, 2017.

[14] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[15] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen and learn more: Design choices for deep audio embeddings," in *IEEE Int.˜Conf.˜on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 3852–3856. [Online]. Available: https://ieeexplore.ieee.org/document/8682475