Interspeech 2018
2-6 September 2018, Hyderabad, India

Wavelet Transform based Mel-scaled Features for Acoustic Scene Classification

Shefali Waldekar, Goutam Saha

Dept of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India
shefaliw@ece.iitkgp.ernet.in, gsaha@ece.iitkgp.ernet.in

Abstract

Acoustic scene classification (ASC) is an audio signal processing task where mel-scaled spectral features are widely used by researchers. These features, considered de facto baseline in speech processing, traditionally employ Fourier based transforms. Unlike speech, environmental audio spans a larger range of audible frequency and might contain short high-frequency transients and continuous low-frequency background noise, simultaneously. Wavelets, with a better time-frequency localization capacity, can be considered more suitable for dealing with such signals. This paper attempts ASC by a novel use of wavelet transform based mel-scaled features. The proposed features are shown to possess better discriminative properties than other spectral features while using a similar classification framework. The experiments are performed on two datasets, similar in scene classes but differing by dataset size and length of the audio samples. When compared with two benchmark systems, one based on mel-frequency cepstral coefficients and Gaussian mixture models, and the other based on log mel-band energies and multi-layer perceptron, the proposed system performed considerably better on the test data.

Index Terms: environmental sounds, cepstral features, Haar wavelet

1. Introduction

Assigning a textual label to an audio signal based on the general acoustic characteristics of recording location or surroundings is referred to as acoustic scene classification (ASC) [1]. This field of audio signal processing has gained popularity because of the importance of information obtained from environmental sounds in applications like surveillance, smart devices, robotics, data archiving, and hearing aids.

In the popular 'bag-of-frames' (BoF) approach for soundscape (audio equivalent of landscape) classification, an audio stream is represented by a long-term statistical distribution (e.g. Gaussian mixture models (GMMs)) of some set of short-term spectral features (e.g. mel frequency cepstral coefficients (MFCCs)) [2]. However, the much simpler one-point average approach was shown to be better than the BoF paradigm when evaluated on larger audio scene datasets with less within-class variability [3]. Nevertheless, the baseline system for ASC provided with the first two challenges on detection and classification of acoustic scenes and events (DCASE) was the BoF system [4, 5]. Following the current trend of use of deep learning in all machine learning applications and by its success in DCASE 2016 [6], the baseline system of DCASE 2017 was a multi-layer perceptron (MLP) trained on log mel-filter bank energies (MBE) [7].

Audio signals captured from acoustically dynamic surroundings cover almost the entire audio frequency range of 20Hz to 20 kHz. Moreover, unlike in the case of speech signals where one excitation source and one transformation system are active at a time, sounds generated from multiple audio sources overlap to form an acoustic scene. Features that can capture local information in both time and frequency domains would provide better representation of such multifaceted signals. However, according to the Heisenberg-Gabor limit, a simultaneous sharp localization in both the complementary domains is not possible [8]. The wavelet-based transforms have basis functions that are more concentrated in both the domains than the Fourier based transformations [9]. Consequently, better time-frequency localization ability can be expected from features that use wavelet transform in their extraction process.

Although spectral features dominate the research in this field, other possibilities like time-frequency features obtained from matching pursuit algorithm [10] and time-frequency features based on the histogram of gradients [11] have also shown considerable potential in environmental audio classification. In this paper, we propose the use of a variant of MFCC for ASC. The feature, mel-frequency discrete wavelet coefficients (MFDWC) makes use of wavelet transform in place of the conventional cosine transform in MFCC extraction [12]. The classifier employed is support vector machine (SVM) with radial basis function (RBF) kernel. The rest of the paper is organized in the following way. The evolution of the wavelet-based features from the conventional MFCC is discussed in Section 2. The proposed system architecture and the evaluation setup used in this work are described in Section 3. In Section 4, we present the results obtained on multiple datasets and discussions on the same. It is followed by the conclusions drawn from this work in Section 5.

2. Evolution of MFDWC

In all fields of audio signal processing, the most exploited features are mel-scale based. The inspiration behind using mel-scaled filterbank is the logarithmic sensitivity of the human hearing system to the frequency of audio signals. Two such features, namely MFCC and log MBE, were also used in the systems considered as the baseline in this work [5, 7]. Conventionally, MFCC are obtained when discrete cosine transform (DCT) is applied to log MBE. If x_n represents the log-energy of the n^{th} filter of N filters, then $MFCC_k$, the k^{th} MFCC coefficient ($k = 1, \cdots, M$, $M < N$), is given by

$$MFCC_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi k}{N}(n + \frac{1}{2})\right) \quad (1)$$

The purpose of DCT is three-fold. First, it replaces the inverse Fourier transform needed to get the cepstrum, second, it performs decorrelation since the filters of the mel-scaled filterbank are overlapping, and third, it brings more information to the lower frequencies which in turn allows the use of fewer

10.21437/Interspeech.2018-2083

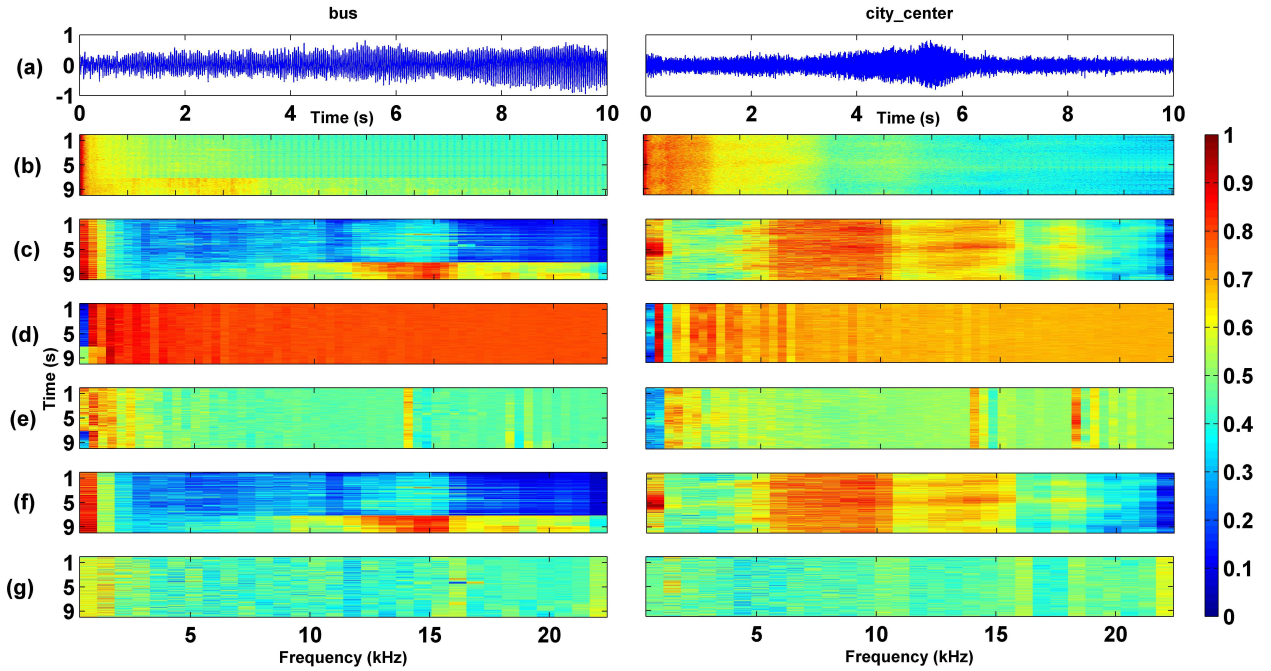

Figure 1: *Comparison of mel-scale based features. (a) samples from two different classes; (b) spectrograms of the samples; pictorial representation of features (c) log-MBE; (d) MFCC; (e) NOBTC; (f) and (g) approximation and detail coefficients, respectively, of MFDWC*

coefficients than the filters. The feature extraction scheme in [13] captures speech information in a more efficient manner than the standard MFCC because it applies DCT in blocks based on dominant formant frequency zones. These block-based MFCC were shown to be better for speaker recognition. Non-overlapped block transform coefficients (NOBTC), obtained when the DCT blocks do not overlap, have reportedly outperformed MFCC in speech/music discrimination [14] and acoustic scene classification [15, 16].

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. With the block-DCT approach, many, if not all, coefficients might get corrupted in the presence of a band-limited noise because the number of blocks is two or three. Also, the DCT basis vectors have fixed time-resolution for all frequencies. Use of discrete wavelet transform (DWT) instead can overcome these shortcomings because it has better time and frequency localization capacity, while simultaneously filling the need of DCT [12]. 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. This property makes wavelets more suitable for working with environmental audio data which might carry brief high-frequency transients and long-lasting low-frequency background noise at the same time [17].

DWT of a signal $x[n]$ is defined by the equation

$$W(j,k) = \sum_j \sum_k x[k] 2^{-j/2} \psi(2^{-j} n - k) \tag{2}$$

where $\psi(t)$ is called the mother wavelet and is a fast decaying time function with finite energy. All our experimental results shown in this work are by using the Haar function as the mother wavelet, which is the oldest, simplest and a compactly supported wavelet. It is given by

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & otherwise \end{cases} \tag{3}$$

One level of wavelet decomposition of a time-domain signal is equivalent to passing it through a low-pass filter $g[n]$ and a high-pass filter $h[n]$ simultaneously. The output of the filters is downsampled by a factor of two because their frequency content is half of that of the original signal. These two filtered signals represent the approximate and the detail information of the signal and can be expressed as

$$y_a[k] = \textstyle\sum_n x[n] g[2k - n] \tag{4a}$$
$$y_d[k] = \textstyle\sum_n x[n] h[2k - n] \tag{4b}$$

For more details on wavelets, the interested reader may refer to [9] and for multiresolution signal analysis to [18].

When DWT replaces DCT in the MFCC extraction scheme, the new set of features is called mel-frequency discrete wavelet coefficients (MFDWC) which has proven its mettle in speech recognition [12] and audio event detection [17]. In order to exhibit the superior discriminative capacity of the wavelet-based features, we present a portrayal of different mel-scaled features in Fig 1. Two randomly picked samples from the DCASE 2017 ASC dataset, which turned out to be from 'bus' and 'city_center' classes, their spectrograms, corresponding log MBE, MFCC, NOBTC, and approximation and detail parts of MFDWC are shown in Fig 1(a), (b), (c), (d), (e), (f) and (g) respectively. To enable proper visual comparison, all the matrices were put through min-max normalization so that the values are in the range [0, 1]. It should be noted here that the approximation and detail coefficients' dimensions are half of that of the other

three features. The figure shows that the change in the signal pattern in time (after 7 sec for 'bus' sample and between 4-6 sec for 'city_center' sample) is captured by all the features, but it is done vividly and with fewer components by MFDWC. In case of 'bus', it was a high-frequency (13-15 kHz) sound of a sudden gush of wind (probably due to door opening or another vehicle passing by). The class 'city_center' represents a busy road with multiple vehicles moving simultaneously, which is clearly seen by the high energy in 5-15 kHz band in the logMBE and MFDWC approximation component.

3. Proposed system

The proposed ASC system incorporates the general pattern classification framework. All the incoming audio signals go through pre-processing and feature extraction processes. Since the DCASE data is in binaural stereo format (i.e. the two channels carry different values), in our system, the first pre-processing step is to convert the data samples to monophonic audio by averaging the two channels. Monophonic audio frames of 20 ms having 50% overlap were chunked by applying Hamming window. Pre-emphasis factor of 0.97 was used to emphasize the high-frequency content. The feature extraction block diagram is elaborated in Fig. 2. For MFDWC extraction, a filterbank of 60 mel-scaled triangular filters was employed. The choice of the number of filters is motivated by the fact that we are dealing with audio signals that can go up to a frequency of 20 kHz. Only one level of decomposition was performed with Haar wavelet and both the approximation and detail coefficients of the DWT were retained. Discrete-time derivatives or delta (Δ) features, evaluated with a 3-frame window, were appended only for approximation coefficients of MFDWC.

Models were built from features of training data and then employed for classification of the test samples. In our system, frame-wise mean and standard deviation of the features were given as input to an SVM classifier with RBF kernel. Most of the submissions of DCASE 2016 employed fusion based or deep-learning based classification [6]. Nonetheless, SVM's presence in the top ten indicated it as a viable choice over baseline systems' classifiers, given that discriminative classifiers are more suited for ASC task than the generative ones [1, 19] and deep-learning based ones need more resources. Since SVM is a binary classifier, for our multi-class classification problem we have used the one-vs-one approach, consequently training $N(N-1)/2$ classifiers for N classes. The cost parameter (C) of the SVM and the γ parameter of the Gaussian RBF were empirically determined by applying grid-search on DCASE 2013 ASC dataset [4].

3.1. Evaluation Setup

We have used the development dataset of TUT Acoustic Scenes 2016 [5] and TUT Acoustic Scenes 2017 [7] in our experiments, henceforth referred to as TUTAS16D and TUTAS17D respectively. The two datasets differ from each other in the length of the audio streams and amount of data as given in Table 1. According to the DCASE challenges' ASC task setup, development data is partitioned into k folds, where k=4 for both the datasets. Fold-wise mean classification accuracy is used as the performance metric during development. For testing the proposed system, the evaluation datasets of DCASE 2016 (TUTAS16E) (390 samples of 30sec each) and DCASE 2017 (TUTAS17E), (1620 samples of 10sec each) are used. The corresponding development datasets are used as training data during testing. For performance comparison, we used the baseline systems of the DCASE challenges of 2016 (Baseline 1) and 2017 (Baseline 2), which are MFCC-GMM based and log MBE-MLP based respectively.

Table 1: *Development data description*

Name	**TUTAS16**	**TUTAS17**
Duration per audio	30 sec	10 sec
Number of files	1170	4680
Number of classes	15	
Classes	lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tram	
Data format	44.1 kHz, 16 bit, binaural stereo wav files	
Location	Finland	

4. Results and Discussion

Wavelet (W) decomposition of a signal gives approximation (WA) and detail (WD) coefficients. For an efficient information representation, we first evaluated different configurations of MFDWC. These configurations are listed in Table 2 along with their corresponding results. We observed that the WD coefficients are better than the WA coefficients when they are considered individually. However, when appended with their respective velocity or delta (Δ) coefficients, the performance of the latter surpasses that of the former. This shows that ΔD coefficients do not contribute any useful information. Besides appending the WA and WD coefficients and their respective Δs that form different feature-fusion configurations, we also used score-fusion of the features to find an optimum feature configuration. In the table, feature-level and score-level fusion are indicated by "_" and "+" respectively. We used FoCal Multiclass toolkit [20] to find weights for the score fusion. From the mean accuracy obtained on the two datasets, the best configuration for TUTAS16D is the score-level fusion of WA and WD appended with their respective Δs, while for TUTAS17D it is feature-level fusion of WA, ΔA and WD coefficients. Finally, we de-

Table 2: *Fold-wise mean accuracy and standard deviation (%) for different combinations of MFDWC. W: Full decomposition; WA: Approximate coefficients; WD: Detail coefficients; Δ: Velocity coefficients; ΔA: Approximation coefficients' velocity; ΔD: Detail coefficients' velocity; _: Feature-level fusion; +: Score-level fusion.* ***Bold-face****: Maximum mean accuracy in the column.*

Feature	**Dim.**	**TUTAS16D**	**TUTAS17D**
WA	30	62.62±4.31	69.66±2.23
WA_AΔ	60	76.74±2.91	78.23±2.70
WD	30	68.71±2.50	74.24±1.25
WD_DΔ	60	68.45±1.05	73.22±1.12
W	60	76.74±3.02	79.56±2.54
W_Δ	120	78.53±3.96	80.63±3.44
WA_AΔ_WD	90	78.62±2.43	**80.84±3.31**
WA+WD	30,30	76.67±3.77	79.43±1.89
WA+WD_DΔ	30,60	77.44±2.97	79.49±1.86
WA_AΔ+WD	60,30	78.47±3.53	80.45±2.01
WA_AΔ+WD_DΔ	60,60	**78.90±3.74**	80.50±2.24

Figure 2: *Block diagram of MFDWC extraction*

Table 3: *Fold-wise mean accuracy and standard deviation (%) for different mel-scale based features. NA: Not applicable.* ***Bold-face****: Maximum mean accuracy in the column.*

Feature	TUTAS16D	TUTAS17D	TUTAS16E	TUTAS17E
LogMBE-SVM	65.37±3.87	70.98±2.08	67.18	59.88
MFCC_Δ_ΔΔ-SVM	71.93±5.72	74.46±2.17	73.85	60.99
NOBTC_Δ-SVM	75.13±2.24	78.58±3.67	78.97	62.59
MFDWC_O-SVM	**78.62±2.43**	**80.84±3.31**	**81.79**	**69.88**
MFCC_Δ_ΔΔ-GMM (Baseline 1)	71.29±3.81	NA	77.20	NA
LogMBE-MLP (Baseline 2)	NA	75.12±1.62	NA	61.00

cided upon WA_ΔA_WD as the optimum feature configuration because the second dataset is bigger and more complex and its result on the first dataset is very close to the maximum value. Henceforth, this configuration is referred to as MFDWC_O in this paper.

Figure 3 shows a pictorial representation of confusion matrices obtained by Baseline 2 and proposed framework when evaluated on the TUTAS17D dataset. From the pictures, one can observe that more than any other class, the proposed system has the upper hand on 'library' and 'train' classes. It is noteworthy that 'office' class is better classified by Baseline 2. Overall, the misclassification pattern is more or less similar in both cases owing to the use of the same type of features (i.e. mel-scaled) by both the systems.

The superiority of MFDWC over other mel-scale based features was demonstrated in Fig 1. We further reinforce the fact in Table 3 with the help of the classification accuracy obtained on the development and evaluation datasets by the application of these features to SVM classifier. We also show in the table the performance of the two baseline systems on their respective datasets. The proposed feature performed above 10% better than Baseline 1 on TUTAS16D, while the relative improvement was close to 8% compared to Baseline 2 on TUTAS17D. On the respective test data, however, the proposed system showed around 6% and above 14% better results. Note that the performance of all features during development is better for 2017 than for 2016. But the former's evaluation dataset was seen to be a tougher nut to crack for all. Although the proposed system managed to stay ahead in all conditions shown here, the results could be improved with advanced techniques and fusion based approaches to come in the range of the best performing systems of the two challenges.

5. Conclusions

In this paper, we discussed the usefulness of the time-frequency localization property of wavelets while working with environmental audio. We presented a novel use of wavelet-based mel-scaled features for acoustic scene classification. The feature MFDWC, where DCT from the conventional MFCC extraction was replaced by DWT, displayed more discriminative power than other mel-scale based features. Accompanied by SVM as the classifier, the feature outperformed the MFCC-GMM based and log MBE-MLP based baseline systems.

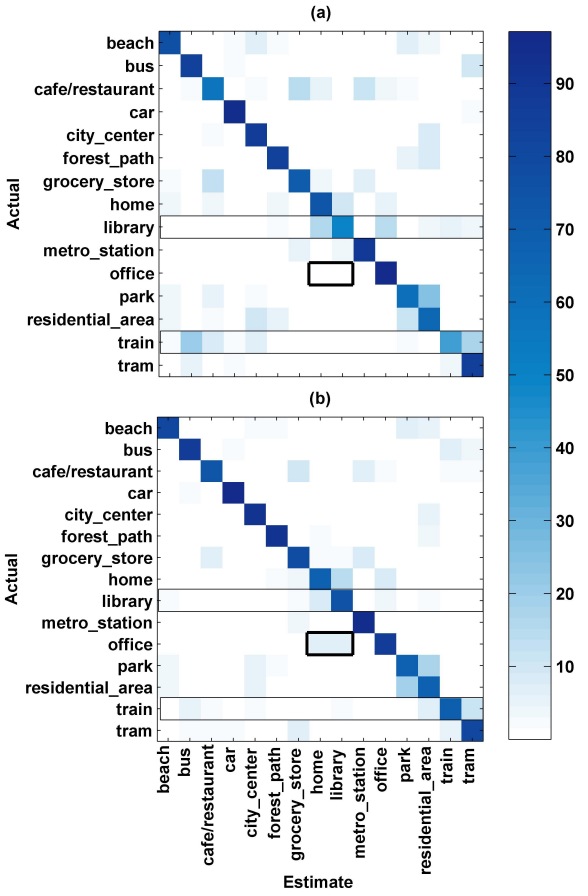

Figure 3: *Pictorially depicted confusion matrices of (a) log MBE-MLP based Baseline 2, and (b) MFDWC_O-SVM based proposed system, both evaluated on TUTAS17D dataset. Examples of better performance of (b) are marked in thin-edged boxes, while for (a) it is in thick-edged box.*

Researchers in machine learning are now exploring possibilities with deep-learning architectures. The ASC system presented here is comparatively less resource hungry. This work shows that wavelet-based features in ASC task hold promise.

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] J.-J. Aucouturier, B. Défréville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[3] M. Lagrange, G. Lafay, B. Défréville, and J.-J. Aucouturier, "The bag-of-frames approach: A not so sufficient model for urban soundscapes," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. EL487–EL492, 2015.

[4] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

[6] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[7] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[8] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[9] I. Daubechies, *Ten lectures on wavelets*. Siam, 1992, vol. 61.

[10] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[11] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 142–153, 2015.

[12] 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.

[13] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2012.

[14] V. Ghodasara, D. S. Naser, S. Waldekar, and G. Saha, "Speech/music classification using block based MFCC features," *Music Information Retrieval Evaluation eXchange (MIREX)*.

[15] V. Ghodasara, S. Waldekar, D. Paul, and G. Saha, "Acoustic scene classification using block-based MFCC features," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016), Budapest, Hungary, Tech. Rep*, 2016.

[16] S. Waldekar and G. Saha, "Classification of audio scenes with novel features in a fused system framework," *Digital Signal Processing*, 2018.

[17] 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.

[18] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[19] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[20] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scorestutorial and user manual," *Software available at http://sites. google. com/site/nikobrummer/focalmulticlass*, 2007.