Robust features for environmental sound classification

Sunit Sivasankaran, K.M.M Prabhu

Robust features for Environmental Sound Classification

Sunit Sivasankaran, K.M.M. Prabhu
Department of Electrical Engineering
Indian Institute Of Technology Madras, Chennai-36, India
Email: sunit.sivasankaran@gmail.com, prabhu@ee.iitm.ac.in

***Abstract*—In this paper we describe algorithms to classify environmental sounds with the aim of providing contextual information to devices such as hearing aids for optimum performance. We use signal sub-band energy to construct signal-dependent dictionary and matching pursuit algorithms to obtain a sparse representation of a signal. The coefficients of the sparse vector are used as weights to compute weighted features. These features, along with mel frequency cepstral coefficients (MFCC) are used as feature vectors for classification. Experimental results show that the proposed method gives a maximum accuracy of 95.6 % while classifying 14 categories of environmental sound using a gaussian mixture model (GMM).**

I. Introduction

To a careful listener, an audio recording is a rich source of information giving clues such as location, direction of vehicular movement, environmental information, speed of wind and so on. It is, therefore, only natural to ask if we could make machines imitate human listening capabilities. One step in this direction is to train a machine to automatically classify the environment based on a set of features extracted from an audio sample.

Environmental sound classification has a variety of applications. Modern hearing aids [1] consists of several programs which account for the reverberation model of environments such as meeting rooms, auditoriums and other noisy environments. Automatic recognition of the surrounding environment allows these machines to switch between programs and work with minimum user interference. Other applications include video scene analysis which is generally achieved using computationally heavy video processing techniques. Liu et al [2] have proposed a set of low level features such as frequency centroid, frequency bandwidth, energy ratio of sub-band to characterize the audio clip of a video scene and were shown to have good scene discrimination capabilities. Another interesting application of the environmental classification problem is in creating automatic diary [3] by processing the audio recorded over a certain period. Users are automatically given locations they have visited based on classifying the environments using features extracted from the audio samples.

The problem of environment classification is a sub-problem of a bigger area of research called computational auditory scene recognition (CASR) [4], where the focus is on recognizing the context. Application of CASR include providing context to devices such as hearing aids and mobile phones, enabling it to provide better service.

Previous attempts to classify environment have given rise to a new set of features. Peltonen et al [4] have used mel frequency cepstral coefficient (MFCC) as features and gaussian mixture models (GMM) and neural network as classifiers. They report an average recognition rate of only 68.4 % using MFCC as features and GMM as classifier, while classifying 17 natural sounds. Chu et al [5] have proposed to use a combination of MFCC and a set of features extracted using matching pursuit (MP) algorithm, to classify a set of 14 natural sounds. Using GMM as a preferred classifier, they have reported an accuracy of about 83.9 %. Adiloğlu et al [6] have developed a dissimilarity function to compute the distance between sounds and used support vector machine (SVM) for classification purposes.

In this paper we extend the work reported in [5] and classify the environment recording using the MP algorithm. We develop different frequency scaling methods for constructing a dictionary with an objective to capture information which are not captured by MFCC. We achieve a maximum accuracy of about 95.6 % while classifying 14 classes using a GMM classifier. The proposed algorithm constructs a dictionary using prior knowledge of the signal with a small increase in the computational cost.

The rest of the paper is organized as follows. Section II explains the feature extraction methods including the MP based feature. Section III introduces the linear piecewise model for constructing better dictionary using sub-band energy of the signal and Section IV discusses procedure to obtain weights for computing weighted features. Section V explains the experimental setup used. Results are analysed in Section VI while Section VII concludes the paper.

II. Feature extraction

A varied set of features such as zero crossing rate, MFCC, band energy ratio, spectral flux, statistical moments ([2],[4]) and features obtained using the MP algorithm and their combinations have been used to classify natural audio sounds. The best reported accuracy [5] was obtained using a combination of MFCC and MP features. MFCC is obtained by first computing the short time Fourier transform of the signal. The spectrum values of each frame are then grouped into bands using a set of triangular filters [7]. The bandwidth of the triangular filters are constant for center frequencies below 1 kHz

and increases exponentially upto 4 kHz. 13 mel frequency cepstral coefficients for each frame are obtained by taking the discrete cosine transform (DCT) of the log of the magnitude of filter outputs. Since the filter bandwidths of the filterbank are narrow below 1 kHz, MFCC can represent low frequency content of the signal adequately well.

A. Matching Pursuit for feature extraction

There are many algorithms [8] to obtain a sparse representation of a signal, given a dictionary. Commonly used algorithms are the basis pursuit (BP), orthogonal matching pursuit (OMP), iterative hard thresholding (IHT) and compressive sampling matching pursuit (CoSaMP). Sounds containing harmonic sections such as sound from an ambulance siren can also be decomposed using harmonic matching pursuit [9]. Owing to the simplicity of orthogonal matching pursuit (OMP), we use the technique to compute MP features.

Given a dictionary D of size $m \times n$ and an observed signal y, MP algorithm gives a sparse vector, x, using an iterative approach. Each iteration captures the maximum possible residual energy. The number of iteration is either fixed by predetermining the sparsity, $||x||_0 = K$ or by thresholding the residual energy, $||Dx - y||_2 \leq \epsilon$. In this paper we predetermine the value of K. Chu et al [5] have reported no significant improvement in the classification performance for $K \geq 5$ because of which we fix the value of K to five.

1) Building the dictionary : A detailed review of dictionary building techniques for matching pursuit algorithms has been described in [10]. Approaches include learned dictionaries such as K-SVD [11] which is known to represent signals better, but it is computationally intensive and data dependent. On the other hand, analytical dictionaries such as the Fourier and wavelet have fast implementations and analytic formulation with supporting proofs and error rate bounds. One such dictionary is the Gabor dictionary whose atoms are constructed using a Gabor filter which are known to have a good representation for audio signals. Gammatone filter based dictionaries which are modelled on human psychoacoustics, have also been reported to have a good audio representation [12]. We, however, use Gabor atoms to construct our dictionary since the principles developed in this paper can easily be extended to other time-frequency atoms as well.

A real discrete Gabor time-frequency atom can be represented by

$$g[k] = \frac{k_g}{\sqrt{s}} e^{-\pi(k-u)^2/s^2} \cos\left(2\pi\omega(k-u) + \theta\right), \quad (1)$$

where, the constants s, u, k_g, ω are the scale, shift, normalization factor and frequency values. Scale and shift values were set to $s = 2^p (1 \leq p \leq 8)$ and $u = \{0, 64, 128, 192\}$ respectively, to construct the dictionary. A logarithmic scale for $\omega = Ci^{2.62}$ (with $1 \leq i \leq 35, C = 0.5 \times 35^{-2.6}$) was used to accommodate finer granularity below 1000 Hz [5]. This gave a dictionary with $n = 8 \times 4 \times 35 = 1120$ atoms. Phase θ was set to zero since it has been reported to have no significant impact on the classification result.

Dictionary constructed using Gabor atoms were used by the OMP algorithm to choose five atoms which best correlate with the signal. The mean (μ) and the standard deviation (σ) of the frequency (ω_s) and scale (s_s) of the selected atoms was used as MP features. The feature set for classification is as follows:

$$[MFCC, \mu(\omega_s), \sigma(\omega_s), \mu(s_s), \sigma(s_s)].$$

We refer to the above mentioned set of features as unweighted features.

MP features depend on dictionary D, whose atoms are constructed using Gabor function. The following section describes a method to construct relevant dictionaries using energy distribution of the signal.

III. Linear piecewise scaling

In this section we introduce the piecewise method of frequency scaling to build a dictionary using apriori knowledge of the signal. Here, the sub-band energy ratio, which is the normalized energy distribution in sub-bands, is used to determine the number of atoms to be allocated per frequency band. Since MFCC has a good representation of the signal in low frequency region (≤ 1 kHz), we pass the signal through a high pass filter having a cut-off frequency of 1 kHz. This is done to ensure that the MP features do not capture the information which is already been captured by the MFCC. Now the j^{th} sub-band energy, $E_{sb}(j)$, is obtained by,

$$E_{sb}(j) = \sum_{P \in sb(j)} |X(P)|^2 \quad j = 1, 2, \ldots, N,$$

where $X(P)$ is the discrete Fourier transform (DFT) of the signal and N is the total number of sub-bands.

We then normalize the energy to obtain a distribution function as follows:

$$E_{sb_n}(j) = \frac{E_{sb}(j)}{\sum_{i=1}^{N} E_{sb}(i)}, \quad j = 1, 2, \ldots, N.$$

The product $E_{sb_n}(j) \times n_f$, rounded off to the nearest integer, decides the number of frequency elements to be allocated to the j^{th} sub-band and is denoted by $n_{sb}(j)$.

$$n_{sb}(j) = round(E_{sb_n}(j) \times n_f), \quad j = 1, 2, \ldots, N. \quad (2)$$

here $round(.)$ denotes the rounding off operator, n_f is the total number of frequency elements. In our experiment, we set $n_f = 35$.

A linear piecewise model for the j^{th} sub-band is then constructed by dividing the straight line joining the frequency boundaries of the sub-band into $n_{sb}(j)$ equi-spaced points. The corresponding frequency points are used to construct the dictionary D using Gabor atoms (1) for OMP. Fig. 1 shows the frequency allocation for ocean and casino sounds using the algorithm. An ocean sound has higher energy in the low frequency region because of which the algorithm adaptively allocates more atoms in the lower frequency band. Similarly, higher number of atoms were allocated to high frequency region in case of

a casino sound due to higher energy in the high frequency band. In one such example, 19 atoms were allocated for an ocean sound in the frequency range, $\omega < 0.1\pi$ as compared to 6 for a casino sound (Fig. 1).

IV. Computing Weighted features

After obtaining the dictionary D, OMP algorithm is used to find atoms which correlate well with the signal. Each of the selected atoms has a different value of correlation coefficient with respect to the signal. To capture this variation, we propose to use weighted mean and deviation.

If d_s are the atoms selected by orthogonal matching pursuit and r_i is the residue after i^{th} iteration of the algorithm, the inner product $x_s(i) = d_s(i)^T r_i$, for $i = 1, 2, \ldots, K$, are the non zero components of the sparse vector x and are used as weights while computing the weighted mean (μ_w),

$$\mu_w(w_s) = \frac{\sum_{i=1}^{K} abs(x_s(i)) \times w_s(i)}{\sum_{i=1}^{K} abs(x_s(i))}. \quad (3)$$

The unbiased estimator for weighted standard deviation is computed by,

$$\sigma_w(w_s) = \sqrt{\frac{V_1}{V_1^2 - V_2} \sum_{i=1}^{K} abs(x_s(i)) \times (w_s(i) - \mu_w(w_s))^2}, \quad (4)$$

where, $V_1 = \sum_{i=1}^{K} abs(x_s(i))$ and $V_2 = \sum_{i=1}^{K} x_s^2(i)$. We similarly compute the weighted mean and standard deviation of the scale parameter. The new set of features are:

$$[MFCC, \mu_w(\omega_s), \sigma_w(\omega_s), \mu_w(s_s), \sigma_w(s_s)].$$

The summary of the algorithm is detailed below:

Algorithm (Feature Extraction)

INPUT: Audio segment $y[n]$, number of Sub-bands(N), total number of frequency elements (n_f), scale range($\{s\}$), shift range ($\{u\}$)

step 1: Find the support ω.

$\rightarrow$ Pass the signal $y[n]$ through a high pass filter having a cut-off of 1 kHz.

$\rightarrow$ Divide the spectrum $Y(e^{j\omega})$ into N sub-bands.

$\rightarrow$ Compute the energy in each sub-band $E_{sb}(j) = \sum_{P \in sb(j)} |Y(P)|^2$.

$\rightarrow$ Find the number of atoms to be allocated to each sub-band, n_{sb} (Eqn.(2)).

$\rightarrow$ Find the plausible set of frequency elements $\omega = \{\omega_1, \omega_2, \ldots \omega_N\}$.

step 2: Extract feature vectors using OMP.

$\rightarrow$ Construct a Gabor dictionary D using ω as frequency elements.

$\rightarrow$ Do an OMP using D as dictionary.

$\rightarrow$ Init: $\omega_s = \emptyset, s_s = \emptyset, x_s = \emptyset, \Omega = \emptyset$, residual $r_o = y$ and counter $c = 1$.

WHILE: $c \leq 5$

$\rightarrow$ Find the column d_k of D which correlates the most with the residue:

$$k \in \arg\max_j |\langle r_{c-1}, d_j \rangle|$$
$$\Omega_k = \Omega_{k-1} \cup \{d_k\}$$
$$\omega_s = \omega_s \cup \text{ freq}\{d_k\}$$
$$s_s = s_s \cup \text{ scale}\{d_k\}$$

$\rightarrow$ Find the best coefficients

$$x_s = \arg\min_\theta ||y - D_{\Omega_k}\theta||_2$$

$\rightarrow$ update the residual:

$$r_c = y - D_{\Omega_k} x_s.$$

$\rightarrow$ c:=c+1

END WHILE:

$\rightarrow$ Find the weighted mean (3) and standard deviation (4) of ω and scale s.

OUTPUT: Weighted features - $[\mu_\omega(\omega_s), \sigma_w(\omega_s), \mu_w(s_s), \sigma_w(s_s)]$

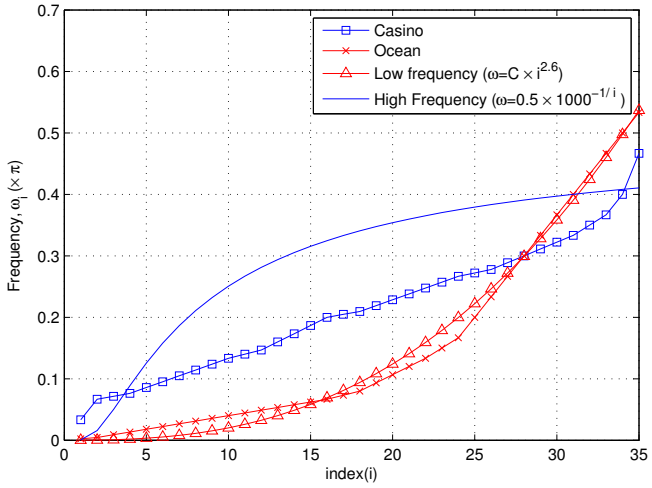


Fig. 1. Frequency allocation using the piece-wise model for casino and ocean sounds and its comparison with low and high frequency emphasized scaling model (Section VI)

V. Experiments

A. Dataset

The results presented in this paper are based on a set of data collected from an online sound repository at *freesound.org* [13]. The collection method is similar to that mentioned in [5]. In order to compare our results with those presented in [5], we have used the data set collected from [13] and applied them to our algorithm and those presented in [5]. A total of 14 audio recordings, namely, *Nature in daytime, Inside Vehicle, Restaurant, Casino, Nature at Night, Bell, Playgrounds, Street Traffic, Thundering, Train, Rain, Stream, Ocean and Street with Ambulance*, representing different environments were collected. Each one of these environments are now referred to as a class. However, no preprocessing was done on the collected data.

B. Method

For each of the 14 classes, a minimum of 4 unique recordings were collected. All collected data had a two channel recording (stereo) out of which only one channel was used. This is to avoid duplication. All the files were collected as uncompressed wav file and had a sampling rate of 44.1 kHz, but of varying time durations (from 30 secs to 8 mins). We divide each recording into segments of 4 seconds. 75 % of all the collected segments were used for training and 25% for testing. Features for both training and testing were computed on these segments. MFCC was computed by dividing the 4 sec segments into blocks of 20ms using a Hamming window with 50 % overlap. The same blocks were used to compute MP features.

To classify the segments, we build Gaussian mixture model (GMM) for each class, which is described by,

$$p = \sum_{k=1}^{N_g} \alpha_k \mathcal{N}(\mu_k, \Sigma_k),$$

where $\mathcal{N}(.)$ is the normal distribution function. α_k is the weight, μ_k and Σ_k are the mean and variance of the k^{th} mixture. N_g is the number of mixtures in the GMM. α_k, μ_k and Σ_k are obtained from the features extracted from the training data, using the standard expectation maximization (EM) algorithm.

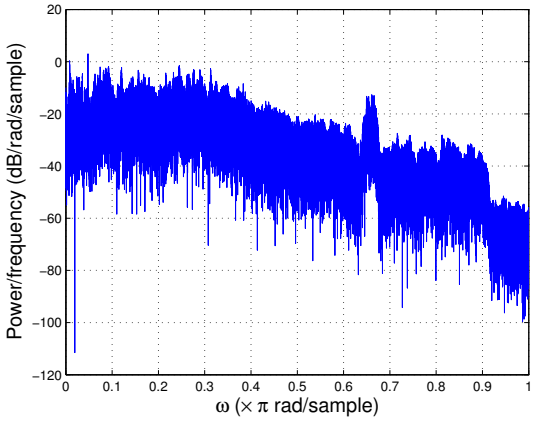
To classify a segment s, the posterior probability $p(s|\mu_i, \Sigma_i)$ is computed for each frame of the segment. A segment is assigned to k^{th} class, if the sum of posterior probabilities of all frames for the k^{th} class is maximum. If there are N_s frame in a segment, the segment is allocated to class c_k, if

$$k = \arg\max_i \sum_{j=1}^{N_s} \log\{p(s_j|\mu_i, \Sigma_i)\}.$$

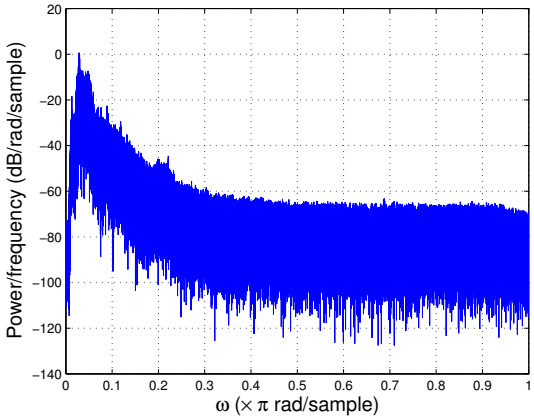
Different values of N_g were tried in [5]. Best results were obtained using a GMM having 5 mixture components. Henceforth, we construct 5 mixture GMM for all classes in our experiment. Confusion matrix was constructed and accuracy values were computed as the ratio of sum of diagonal values to the total sum of all elements in the matrix. All results reported in this paper are the average of accuracy values obtained using ten fold cross validation.

VI. Results and Analysis

We implement the algorithm detailed in [5] on our dataset and obtain an accuracy of 83.2 % while classifying the audio segments without preprocessing them. We use this as a base to compare the performance of our algorithm. On studying the periodogram of environmental sounds (Fig. 2), it was found that certain signals such as audio recordings from casino have substantial energy in the higher frequency region. To understand the impact of high frequency on the classification, we construct a dictionary using a high frequency emphasized scaling function $\omega = 0.5a^{-1/i}$, as against the low frequency emphasized scaling function $\omega = C \times i^{2.6}$ advocated in [5]. A value of $a = 1000$ was chosen for a smoother ascend towards high frequency while still distributing enough atoms in the higher part of the frequency spectrum. The variation of ω as a function of i for $a = 1000$ is shown in Fig. 1. We obtain an accuracy of 84.5 % using weighted features and high frequency emphasized scaling function, an improvement of 1.56 % over the low frequency emphasized scaling function. The increase in accuracy is due to better representation of the high frequency region in the MP features. We would like to reiterate the fact that, the low frequency region is adequately represented with MFCC. Class-wise accuracy comparison for low frequency and high frequency emphasised scaling functions using unweighted features are shown in Fig. 3. We observe that the high frequency emphasised dictionary outperforms its low frequency counterpart, while classifying sounds from nature at day time, inside vehicle, casino and nature at night time whose periodogram showed high energy in the high frequency region, while low frequency emphasised dictionary performed better in classifying sounds from ocean, rain and train which has higher energy content in the lower frequency range.



(a) Casino Recording



(b) Ambulance Recording

Fig. 2. Variation in periodograms of environmental sounds

A. Effect of using sub-band information

The results obtained using high frequency emphasised dictionary showed the need to construct signal specific

dictionaries based on energy distribution across frequencies. To do so, we divide the spectrum into N equispaced sub-bands and construct Gabor dictionaries from scaling functions obtained using sub-band energy distribution, as outlined in Section III. The results obtained are shown in Fig. 4. A weighted 4 sub-band dictionary has given an accuracy of 83.79 %. As the value of N was increased, we observe an increase in the accuracy with an exception for $N = 6$ and $N = 14$ where a dip was seen. Accuracy peaked at $N = 15$ and further increase in N showed no improvement. The class wise accuracy comparison for $N = 6, 14$ and 15 sub-band weighted dictionaries are summarised in Table I.

B. Weighted vs Unweighted

As a rule, weighted features performed better than their unweighted counterparts with notable exceptions for $N = 4$ and $N = 14$. Fig. 5 shows two such accuracy comparison for high frequency emphasized dictionary and 15 sub-band piecewise dictionary .

Our algorithm gives the best performance for $N = 15$ with an accuracy of 95.6%, an overall improvement of about 13.95 % compared to the work reported in [5] and 13.14 % over the high frequency emphasized dictionary.

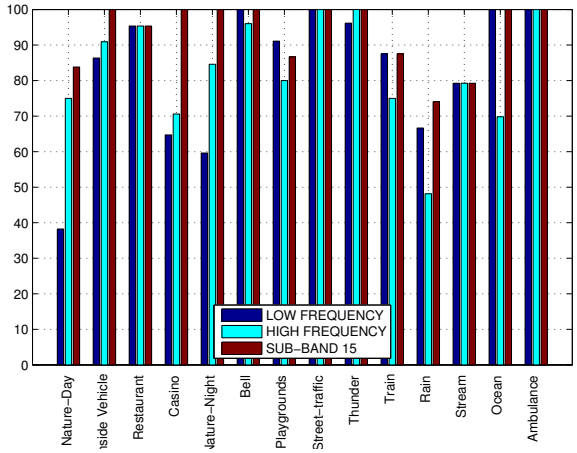


Fig. 3. Accuracy comparison of low frequency [5], high frequency and 15 band scaling (unweighted) for each class

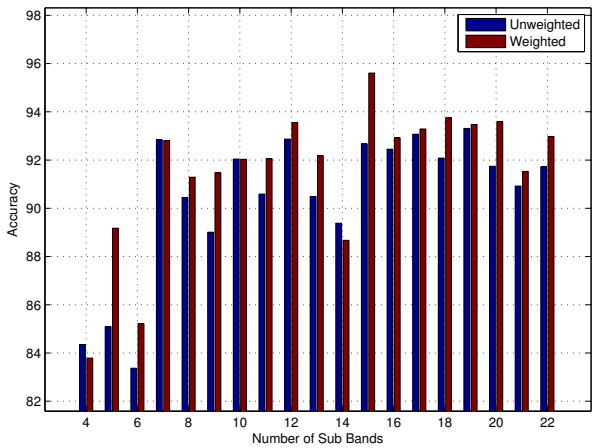


Fig. 4. Effect of increasing the number of subbands on accuracy

TABLE I
CLASS-WISE ACCURACY (IN PERCENTAGE) FOR WEIGHTED SUB BAND DICTIONARIES

	Sub-band 6	Sub-band 14	Sub-band 15
Nature-Day	83.8	82.4	83.8
Inside Vehicle	90.9	90.9	100.0
Restaurant	95.3	95.3	100.0
Casino	100.0	100.0	100.0
Nature-Night	80.8	100.0	100.0
Bell	100.0	100.0	100.0
Playground	82.2	80.0	86.7
Street Traffic	100.0	100.0	100.0
Thunder	100.0	96.2	100.0
Train	81.2	93.8	100.0
Rain	74.1	66.7	100.0
Stream	32.1	100.0	79.2
Ocean	96.2	50.9	100.0
Ambulance	100.0	100.0	100.0

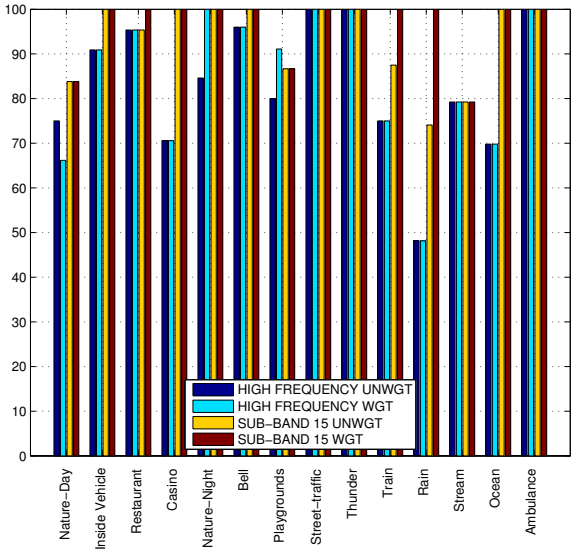


Fig. 5. Class-wise comparison of weighted and unweighted high frequency and 15 band piecewise scaling

VII. CONCLUSION

In this paper, we have proposed an algorithm to compute features for classifying environmental sounds. Features were obtained using a frequency scaling function which utilizes apriori knowledge of the signal to construct a Gabor dictionary. The sparse coefficients obtained using the OMP technique was used as weights to calculate the weighted mean and deviation. These were used as features to classify environmental sounds. Compared to [5], the proposed algorithm has a small increase in computational cost of the order of $\mathcal{O}(n \log n)$. This is incurred while computing the DFT of the signal. Results show the proposed algorithm outperforms the state of the art in environmental sound classification with a significant improvement in accuracy rate.

REFERENCES

[1] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2991–3002, Jan. 2005.

[2] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 20, no. 1-2, pp. 61–79, 1998.

[3] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, ser. CARPE'04. New York, NY, USA: ACM, 2004, pp. 39–47.

[4] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, May 2002, pp. II–1941 –II–1944.

[5] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental Sound Recognition With Time Frequency Audio Features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1142–1158, 2009.

[6] K. Adiloglu, R. Anni, E. Wahlen, H. Purwins, and K. Obermayer, "A Graphical Representation and Dissimilarity Measure for Basic Everyday Sound Events," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1542–1552, 2012.

[7] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recogntion*. Prentice Hall, 1993.

[8] J. A. Tropp and S. J. Wright, "Computational Methods for Sparse Solution of Linear Inverse Problems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, Jun. 2010.

[9] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *Signal Processing, IEEE Transactions on*, vol. 51, no. 1, pp. 101–111, 2003.

[10] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for Sparse Representation Modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[11] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: Design of dictionaries for sparse representation," in *Signal Processing, IEEE Transactions on*, 2005, pp. 9–12.

[12] S. Strahl and A. Mertins, "Analysis and design of gammatone signal models," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2379–2389, 2009.

[13] "The Freesound project." [Online]. Available: http://www.freesound.org/