# ROBUST MINIMUM STATISTICS PROJECT COEFFICIENTS FEATURE FOR ACOUSTIC ENVIRONMENT RECOGNITION

*Shiwen Deng*[1,3]*, Jiqing Han*[2]*, Chaozhu Zhang*[1] *, Tieran Zheng*[2]*, Guibin Zheng*[2]

[1]College of Information and Communication Engineering, Harbin Engineering University
[2]School of Computer Science and Technology, Harbin Institute of Technology
[3]School of Mathematical Sciences, Harbin Normal University
Email: dengswen@gmail.com, {jqhan, zhengtieran, zhengguibin}@hit.edu.cn

## ABSTRACT

Acoustic environment recognition has been widely used in many applications, and is a considerable difficult problem for the real-life and complex environment. This paper proposes a novel feature, named minimum statistics project coefficients (MSPC), and intents to solve this problem. The MSPC feature is extracted from the background sound which is more robust than the foreground sound for the task of acoustic environment recognition. Experimental results show the outstanding performance of the MSPC feature compared with the conventional acoustic features, especially in very complex acoustic environments.

***Index Terms—*** Acoustic environment recognition (AER), background sound/noise, sound event, minimum statistics.

## 1. INTRODUCTION

Acoustic environment recognition (AER), which is referred to classifying different sound environments depending only on sound information and intents to answer "Where am I in the acoustic space ?", can be widely used in many application scenarios. These scenarios include audio classification and segmentation, robotic navigation [1, 2], mobile robots [3], mobile device-based services, audio retrieval [4], audio forensics [5], and other wearable and context-aware applications. Moreover, understanding the acoustic environments can provide an effective and efficient way to prune out irrelevant scenarios and sound events, which have important advantages in acoustic/sound event detection [6]. Although there are many efforts on AER, the distance is far from the practical applications in real-life.

Research in general AER has received more interest in the last years. Two main strategies are usually employed in exploring the problem of AER. The first normally focuses on the recognition of discrete sound effects or specific acoustic/sound events in an environment, by pre-extracting and modelling them [6–11]. In these studies, they believe that the acoustic environment can be characterized by the presence of individual sound events [6, 9–11], or a mixture of the key audio effects and the background sounds [7, 8]. These methods are similar to that used in the document classification by key words. Obviously, the performance of the AER is mainly depended on the recognition accuracy of the sound events or key audio effects. However, these methods suffer some drawbacks in real-life sound environments: 1) These sound events or key audio effects require to be manually defined and selected; 2) There are a large number of these sound events or key audio effects in real-life environment, and it is unrealistic to define and select all of them; 3) It is difficult to sure that some sound events or key audio effects must be emerging in a specific acoustic environment; 4) Some sound events or key audio effects can also be heard in different acoustic environment.

The second strategy intends to characterize the general acoustic environment type as a whole [12–17]. Based on the assumption that the acoustic environment can be characterized by all sounds in it, the features, therefore, are extracted from these sounds. The features or their combinations contain both conventional acoustic features used in [12–15], such as Mel-frequency Cepstral coefficients (MFCC), linear prediction coding coefficients (LPC), linear prediction Cepstral coefficients (LPCC), and more complex features in [12, 15, 17], such as matching pursuit (MP) and independent component analysis (ICA) features. Unfortunately, the above assumption is not true in the real-life acoustic environment, especially in the complex environment where a large number of and many types of sounds occur simultaneously. For instance, it is difficult to decide whether the talk of some people is belong to an office acoustic environment or not. Therefore, the second strategy also suffers the analogous drawbacks in the first strategy.

In this paper, we also consider the acoustic environment

This work was supported in part by the Major Research plan of the National Natural Science Foundation of China (No. 91120303), National Natural Science Foundation of China (No. 91220301), Academic Core Funding of Young Projects of Harbin Normal University of China (No. KGB201225), and Open Fund by Smart Education and Information Engineering (Harbin Normal University) (No. SEIE2013-01).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

type as a whole, and focus on the problem of robust feature extraction. To this end, all the sounds in a specific acoustic environment can be viewed as two parts: the foreground sounds and the background sounds or background noises. The foreground sounds are usually the dominant sounds and can be easily heard, but they are not robust for recognizing acoustic environment due to the analogous drawbacks in the first strategy. However, we find that the background sounds are very robust for AER in our work. A novel feature, which is extracted from the background sounds and named minimum statistics project coefficients (MSPC), is proposed by using the tracking minimum statistics algorithm [18–20]. Next, each acoustic environment is modelled by a Gaussian Mixture Model (GMM), and the recognition is performed based on the likelihood of each model. The experimental results show the robust performance of the proposed feature, especially in very complex acoustic environments.

## 2. MSPC FEATURE AND ACOUSTIC ENVIRONMENT RECOGNITION

### 2.1. Tracking minimum statistics of background sound spectrum

By viewing all sounds in an acoustic environment as two parts: foreground and background sounds, we consider the observed sound signal $y$ as the sum of a foreground sound $s$ and a background sound $n$, $y = s + n$. Similar to the idea of tracking noise components in [18–20], we assume that the spectrum of the background sound has the minimum of the spectrum energy in a local time-frequency window, since the foreground sound is dominant and has more spectrum energy. Tracking the minimum statistics (MS) for feature extraction from the background sound spectrum is carried out as follows.

Firstly, the observed sound signal $y$ is divided into frames by an analysis window function and is analyzed by using the short-time Fourier transform (STFT):

$$Y(l,m) = \sum_{n=0}^{N-1} y(n+lH)w(n)\exp\left(-j\frac{2\pi}{N}nm\right) \quad (1)$$

where $n$ is the sampling time index, $N$ is the size of the STFT, $m$ is the frequency bin index, $l$ is the time frame index, $w$ is an analysis window, and $H$ is the hop size.

Secondly, the smoothing is carried out in both frequency and time [18, 19], respectively. The frequency smoothing of the power spectrum in each frame is defined by

$$P_f(l,m) = \sum_{i=-B}^{B} b(i)|Y(l,m-i)|^2 \quad (2)$$

where $b$ denotes a normalized window function of length $2B+1$, i.e., $\sum_{i=-B}^{B} b(i) = 1$. The smoothing in time is performed by a first-order recursive averaging

$$P(l,m) = \alpha_s P(l-1,m) + (1-\alpha_s)P_f(l,m) \quad (3)$$

where $\alpha_s (0 < \alpha_s < 1)$ is the smoothing parameter.

Finally, the MS of the power spectrum is tracked by the following non-linear rule [20]:

$$X(l,m) = \begin{cases} \gamma X(l-1,m) + \dfrac{1-\gamma}{1-\beta}(P(l,m) - \beta P(l-1,m)), \\ \qquad \text{if } X(l-1,m) < P(l,m) \\ P(l,m), \qquad \text{if otherwise} \end{cases} \quad (4)$$

where $\beta$ and $\gamma$ are constants which are determined experimentally (we set $\beta = 0.8$ and $\gamma = 0.995$ throughout this paper). $X(l,m)$ mainly captures the information of the background sound and part information of the foreground sound.

### 2.2. Feature extraction

The MS calculated from the $l$ frame can be viewed as a vector, say $\bar{\mathbf{x}}_l \in \mathbb{R}^N$, and $\bar{\mathbf{x}}_l = [X(l,0), \cdots, X(l,N-1)]^T$. The vector $\bar{\mathbf{x}}_l$ is converted to the log-scale:

$$\hat{\mathbf{x}}_l = 10\log_{10}(\bar{\mathbf{x}}_l) \quad (5)$$

and is normalized:

$$\mathbf{x}_l = \frac{\hat{\mathbf{x}}_l}{\|\hat{\mathbf{x}}_l\|} \quad (6)$$

which is called log-scaled and normalized minimum statistics (LNMS) vector. The LNMS vectors, however, are unsuitable to be directly used as the feature in classification tasks due to its high dimensional. Hence, these vectors require to be transformed into a low-dimensional subspace to obtain more effective representations by using principal component analysis (PCA) or independent component analysis (ICA) as done in [17, 21].

Here, the eigenvalue decomposition (ED) is performed to extract the basis vectors of the subspace and to obtain the reduced-dimension features. Given the training data which contains $L$ frames, the LNMS vectors calculated from the training data can be written as a matrix, $\mathbf{X} \in \mathbb{R}^{L\times N}$ and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_L]^T$, where $L$ is the total number of frames and $N$ is the number of frequency bins. The covariance matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ of the LNMS matrix $\mathbf{X}$ is given by $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, and the ED is defined as

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \quad (7)$$

where $\mathbf{U} \in \mathbb{R}^{N\times N}$ is a unitary matrix and contains the eigenvectors as columns, $\mathbf{\Lambda}$ is the diagonal matrix containing the eigenvalues $\lambda_1, \cdots, \lambda_N$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. To perform dimensionality reduction, only the first $K$ basis vectors of $\mathbf{U}$ are retained, i.e. the first $K$ columns of $\mathbf{U}$, and

is denoted by $\mathbf{U}_K \in \mathbb{R}^{N\times K}$. Therefore, the $K$-dimensional basis project feature of LNMS vector $\mathbf{x}$, $\mathbf{z} \in \mathbb{R}^K$, can be given by

$$\mathbf{z} = \mathbf{U}_K^T \mathbf{x} \tag{8}$$

The new feature vector $\mathbf{z}$ is called minimum statistics projection coefficients (MSPC) feature, since it is the vector of the projection coefficients of the MS against the basis vectors. The MSPC feature is used for training statistical models and performing the classification task.

### 2.3. GMM classifier for recognition

We use GMM as the classifier for recognizing sound environments. The GMM with $M$ components for the $c$-th class is defined by

$$p_c(\mathbf{z}) = p(\mathbf{z}|c) = \sum_{m=1}^{M} \pi_m^c \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_m^c, \boldsymbol{\Sigma}_m^c) \tag{9}$$

where $\mathbf{z}$ is the input feature, $\boldsymbol{\mu}_m^c$ and $\boldsymbol{\Sigma}_m^c$ denote the mean and the covariance of the Gaussian distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_m^c, \boldsymbol{\Sigma}_m^c)$, and $\pi_m^c$ is the weighting parameters of the $m$-th Gaussian component. In the training phase, the model parameters for the $c$-th class are estimated by maximum likelihood estimation using Expectation-maximization (EM) algorithm on training data of the $c$-th class. In the testing phase, the likelihood of the features of the input segment of each class model is calculated by using Eq. (9), and the segment can be classified into the $c^*$-th class with maximum likelihood as follows

$$c^* = \underset{c}{\operatorname{argmax}}\ p_c(\mathbf{z}|c) \tag{10}$$

## 3. EXPERIMENTAL EVALUATION

### 3.1. Experimental Setup

To investigate the performance of the proposed MSPC feature, the empirical evaluation was performed on seventeen different types of environmental sounds. These environment types considered were *airport, basketball, beach, bus, celebration, classroom, countryside, football, highway, kitchen, market, office, party, protest, restaurant, street, train*. The environmental sounds used for the material were collected from [22]. Each of environmental types consisted of 5 to 8 recordings, and each of these recordings of varying lengths (3–6 minute long) was recorded in real-life environments from different locations. Each recording contained a lot of sound events, some of which were also contained in other recordings of different environments. For instance, speech could be clearly heard in *office, restaurant, and protest* environments. Moreover, all the recordings were converted into WAV formats and downsampled to 22050 Hz sampling rate, monochannel, and 16 bits per sample.

For the modelling and recognition of environmental sounds, all recordings were first divided up into the 2 s segments. The total number of these segments was 8408 about 4.67 hours. Then, GMM classifiers for each concept were trained on 60% of these segments, and tested on the remaining 40%, selected at random. Furthermore, all features were calculated from a rectangle window of 256 points (11.6 ms with nonoverlap), and all results showed in the following experiments were averaged over 5 trials by randomly selecting the training and testing sets.

### 3.2. Comparison of overall recognition accuracy

Several major features, such as MFCC, LPC, and LPCC, were commonly used in audio signal processing, and hence the recognition accuracy of MSPC feature was compared with these conventional features. In the current experiment, 12-dimensional MSPC, 12-dimensional MFCC, and 10-order LPC and LPCC features were used for training the GMMs, respectively. The GMMs for each class were trained with 9 mixtures.

Figure 1 shows the recognition rates of these four features across all 17 classes. Firstly, MSPC and MFCC achieve better average recognition rates ($87.1\%$ and $69.5\%$) than LPC and LPCC ($16.9\%$ and $13.4\%$). Especially, LPC and LPCC almost lose their abilities of classification in some sound environments, such as *basketball, celebration, classroom, restaurant*. The LPC and LPCC features are designed to capture the resonance of the vocal tract of a single sound source, but the acoustic environments in the current experiment contain a lot of varying sound sources, which is the important reason of the poor performances of these features. Moreover, the LPC and LPCC obtain comparatively higher recognition rates in *bus, countryside, kitchen* than in other acoustic environments. The reason is that these acoustic environments contain comparatively fewer sound sources, and the temporal overlapping of these sound sources is not often.

Secondly, figure 1 shows that MSPC achieves a better average recognition rate ($87.1\%$) than MFCC ($69.5\%$). Especially, MSPC obtains outstanding performance in acoustic environments of *celebration, office, party, protest, restaurant, street*, where MFCC shows very poor recognition performances. Obviously, these acoustic environments contain a large number of sound events which are overlapping simultaneously. The same sound events can be heard in different environments, while the different sound events can be heard in the same environment. For instance, loud voices can be heard in *celebration, office, protest, street* environments, while loud voices and police siren can be heard in *street* environments. Although the existence of sound events can not stably indicate what the sound environment is, these sound events usually have much more energy than background sounds and become the dominant sounds of the current acoustic environment. The MFCC is designed to capture the characteristics of

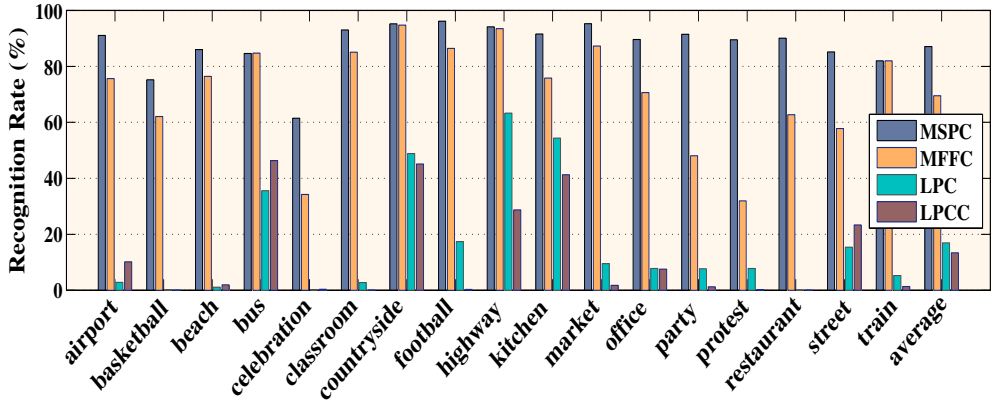


**Fig. 1**. Recognition accuracy across all 17 classes obtained and overall average accuracy with different features (MSPC, MFCC, LPC, and LPCC) using GMM with segment length of 2 s.

the spectral envelope of a sound, and therefore the MFCC in fact represents the information of the dominant sound events in sound environments. This results in the poor performance of the MFCC in some complex sound environments. In contrast with the MFCC, the MSPC can capture the information of the background sounds which are more robust than the sound events in the task of AER, and achieves better recognition rates, especially for recognizing very complex acoustic environments.

Thirdly, figure 1 shows that MSPC achieves better or slightly better recognition rates than MFCC in some comparatively simple acoustic environments, such as *beach, bus, countryside, highway, train*. These acoustic environments usually contain fewer sound events, and can be recognized by modelling these dominant sound events. This is the reason that MFCC can obtain their good performance in these acoustic environments. Although MSPC feature loses some information of the dominant sound events since it only traces the minima of the spectral components, it still outperforms MFCC in these acoustic environments. This result also shows the robustness of the MSPC feature for AER.

Moreover, figure 2 shows the average recognition accuracies of MSPC and MFCC increase with the number of mixtures of GMM. When the number of mixtures is larger than 8, the performance increase slightly for both MSPC and MFCC, MSPC outperforms MFCC in any number of mixtures, from 1 to 20. Also, figure 3 shows the average recognition accuracy of MSPC increase with the number of $K$ which is the dimension of the MSPC feature. According to this figure, a tradeoff between recognition accuracy and reduce dimension can be found.

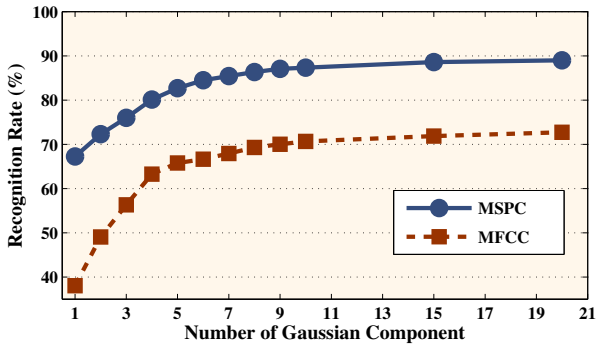


**Fig. 2**. Average recognition accuracies of MSPC and MFFC using GMM with a varing number of mixtures.

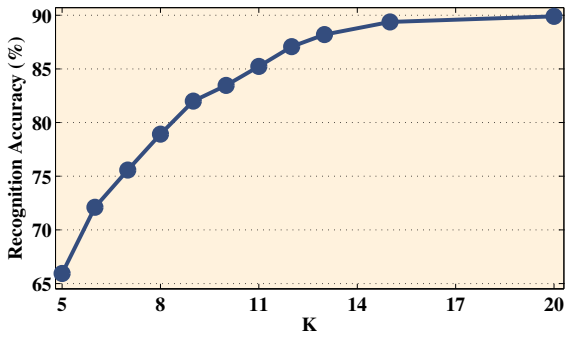


**Fig. 3**. Average recognition accuracy of MSPC with varying number of $K$.

## 4. CONCLUSION

We present in this paper a novel robust MSPC feature for AER by tracking minimum statistics of the background sound spectrum. The MSPC feature can capture most of characteristics of background sound and little information of foreground sounds. The performance of MSPC feature outperforms the performance of the conventional acoustic features in all of the seventeen types of acoustic environments, especially in very complex environments. This revealed that the background

sounds are more important and robust than the foreground sounds in the task of AER.

## 5. REFERENCES


[1] J. Pineau, M.Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Special Iss. Socially Interactive Robots, Robot., Autonomous Syst.,* vol. 42, no. 3–4, pp. 271–281, 2003.

[2] A. Kalmbach, Y. Girdhar, and G. Dudek, "Unsupervised Environment Recognition and Modeling using Sound Sensing," in Proc. Robotics and Automation, 2013. pp. 2699–2704

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in Proc. ICME, pp. 885–888, 2006.

[4] G. Chechik, E. Ie, M. Rehn, S. Bengio, and R. F. Lyon, "Large-scale content-based audio retrieval from text queries," in Proc. Multimedia Information Retrieval, pp. 105–112, 2008.

[5] G. Muhammad, K. Alghathbar, "Environment Recognition for Digital Audio Forensics Using MPEG-7 and Mel Cepstral Features," *Journal of Electrical Engineering,* vol. 62, no. 4, pp. 199–205, 2011.

[6] T. Heittola, A. Mesaros, A. Eronen and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, andMusic Processing,* 2013, doi:10.1186/1687-4722-2013-1.

[7] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 14, no. 3, pp. 1026–1039, 2006.

[8] R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for auditory scene categorization," *IEEE Trans. on Multimedia,* vol. 18, no. 6, pp. 596–606, 2008.

[9] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust., pp. 158–161, 2005.

[10] T. Heittola, A. Mesaros, A. Eronen, T. Virtanen, "Audio context recognition using audio event histograms," in Proc. European Signal Process. Audio Acoust., pp. 1272–1276, 2010.

[11] G. Roma, J. Janer, S. Kersten, M. Schirosa, "Ecological acoustics perspective for content-based retrieval of environmental sounds," *EURASIP Journal on Audio, Speech, and Music Processing,* 2010, doi:10.1155/2010/960863.

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

[13] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-Based Context Recognition," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 14, no. 1, pp. 321–329, 2006.

[14] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification, " *ACM Trans. Speech Lang. Process.,* vol. 3, no. 2, pp. 1–22, 2006.

[15] K. Lee, and D. P. W. Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 18, no. 6, pp. 1406–1416, 2010.

[16] R. Mogi and H. Kasai, "Noise-robust environmental sound classification method based on combination of ICA and MP features," *Artificial Intelligence Research,* vol. 2, no. 1, pp. 107–121, 2013.

[17] H.-G Kim, N. Moreau, T. Sikora, "Audio classification based on mpeg-7 spectral basis representations," *IEEE Trans. on circuits and systems for video technology,* vol. 14, no. 5, pp. 716–725, 2004.

[18] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing,* vol.9, no. 5, pp. 504–512, 2001.

[19] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing,* vol. 11, no. 5, pp. 466–475, 2003.

[20] S. Rangachari, P. Loizou, "A noise-estimation algorithm for highly nonstationary environments," *Speech Communication,* vol. 48, pp. 220–231, 2006.

[21] M. A. Casey, "General sound classification and similarity in MPEG-7," *Org. Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[22] Online free sound resource *freesound*, http://www.freesound.org.