# ROBUST ACOUSTIC FEATURE EXTRACTION FOR SOUND CLASSIFICATION BASED ON NOISE REDUCTION

*Jiaxing Ye* *Takumi Kobayashi* *Masahiro Murakawa* *Tetsuya Higuchi*

National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Japan

## ABSTRACT

In this paper, we present a novel method for environmental sound classification in non-stationary noise environment. The proposed method mainly consists of three stages: noise source separation and acoustic feature extraction and multi-class classification. At first stage, we employ probabilistic latent component analysis (PLCA) to perform time-varying noise separation. To alleviate the artifacts introduced by source separation, a series of spectral weightings is applied to enhance reliability of audio spectra. At feature extraction stage, we extract acoustic subspace to effectively characterize temporal-spectral patterns of denoised sound spectrogram. Subsequently, regularized kernel Fisher discriminant analysis (KFDA) is adopted to conduct multi-class sound classification through exploiting class conditional distributions based on extracted acoustic subspaces (features). The proposed method is evaluated with Real World Computing Partnership (RWCP) sound scene database and experimental results demonstrate its superior performance compared to other methods.

***Index Terms***— Sound recognition, PLCA, source separation, eigen-decomposition, discriminant analysis

## 1. INTRODUCTION

Recently, research area of machine audition has gained a lot of attention [1], where the goal is to recognize real-life acoustic events by using effective audio signal processing techniques. Various potential applications promoted those studies, such as intelligent environment context recognition in robotics [2] and audio-based surveillance [3]. Initial works on acoustic event classification mostly followed speech recognition approaches, such as employing the Mel-Frequency Cepstral Coefficients (MFCCs) to describe sound pattern and performing classification through HMM modeling [4]. However, several critical defects restrained availability of speech recognition techniques for sound event recognition, i.e. the MFCCs are not robust to noise presence while Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) models are insufficient to fully characterize temporal-spectral dynamic patterns in non-speech sounds. A thorough comparison of applying conventional audio processing techniques for sound event recognition can be found in [5], which extensively investigated acoustic features of spectrograms of Short-Time Fourier Transform (STFT), Continuous/Discrete Wavelet Transform (CWT/DWT) and MFCCs together with conventional classifiers, such as Artificial Neural Network (ANN) and Learning Vector Quantization (LVQ).

Advanced classification schemes have been examined for audio classification lately, such as using SVM [6] to exploit non-linear time-frequency distributions in acoustic events. However, motivated by biological evidence that local time-frequency information contributes greatly to human auditory, more significant progress has been carried out for audio representation development. Various novel acoustic features have been proposed, such as spikegram [7], a neural-spike-like representation of sound event, likewise, another sparse audio feature is presented in [2], which is based on matching pursuit with Gabor dictionaries. Besides, based on sparse audio representation, Dennis et al. [8] further developed spike neural network (SNN) for noise robust sound classification and achieved superior results comparing to conventional methods.

It is suggested, from psychology of hearing, human beings are able to recognize noise corrupted acoustic events effortlessly, since we can concentrate on certain sound we are interested in and isolate it from background noise. This study endeavors to realize humanlike sound recognition scheme by employing advanced signal processing techniques. A three-step framework is proposed, which includes noise reduction, robust acoustic feature extraction and multi-class classification. The noise reduction stage is designated to imitate hearing concentration function of human auditory, and thereupon lays basis for noise robust audio classification. In order to tackle single channel noise source separation problem, we employ Probabilistic Latent Component Analysis (PLCA) algorithm [9], which learns dictionaries and their activation weights for representing non-stationary background noise and sound event separately, and noise source separation can be effectively performed thereafter; however, it simultaneously introduces artifacts (distortions) to the denoised sound, especially in extremely noisy environment. To alleviate such artifact interference, we develop a series of spectral weightings to enhance availability of sound spectra based on characteristic of PLCA model. After noise reduction, we

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

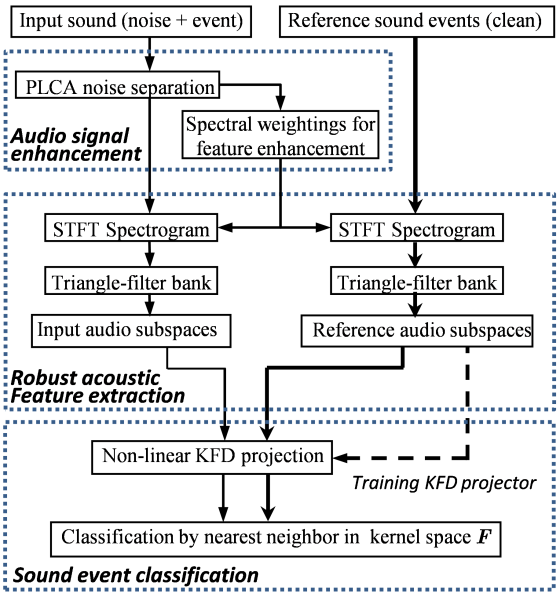


**Fig. 1**. Overview of the propose sound classification system

extract acoustic feature extraction in following steps: 1. uniformly spaced spectral triangle filter bank is applied to sound spectrogram to generate robust spectral feature with low dimensionality; 2. we extract acoustic subspace from filtered spectrogram to characterize predominant patterns of audio event within/across time and frequency domains, which is more representative comparing to conventional frame-based features. In addition, subspace method inherently provides denoising mechanism [10] and thus exhibits more robust property. At multi-class environmental sound classification stage, based on aforementioned audio feature, we adopt regularized Kernel Fisher Discriminant Analysis (KFDA) to exploit non-linear class conditional distributions of audio events, which is equivalent to the method presented in [11].

## 2. THE PROPOSED APPROACH

As presented in Fig. 1, the proposed method consists of three main stages: PLCA noise reduction, robust acoustic feature extraction and KFDA sound classification. In the following sections, we introduce details in those steps.

### 2.1. PLCA noise source separation

PLCA is an effective non-negative matrix decomposition algorithm for non-stationary source modeling [9]. The formulation of PLCA can be expressed as:

$$P(f,t) \approx \sum_{z=1}^{C} P(z)P(f|z)P(t|z) \tag{1}$$

where $P(f,t)$ denotes the (normalized) magnitude of audio spectrogram at $t$ frame and $f$ frequency band, which is regarded as a random variable. $P(f|z)$ is a multinomial distribution representing frequency basis vectors (dictionary) corresponding to the sound source, $P(z)$ is the possibility distribution of latent variable $z$ and $P(t|z)$ denotes the latent variable activations along time coordinate. The number of latent components is denoted by $C$ which is determined by user. The PLCA decomposition is carried out by minimizing KL divergence $d(P(t,f)||Q(t,f))$ between input spectra $P(t,f)$ and reconstructed spectra $Q_t(f) = \sum_z P(f|z)P(z)P(t|z)$. PLCA is efficient for modeling non-stationary sound since the learnt dictionary accommodates invariant characteristics of input sound, for this reason, linear combinations of its dictionary elements (basis) are sufficient for representing the time-varying patterns in audio signal.

In sound source separation, input spectrogram is viewed as a mixture of noise and sound event that can be written as:

$$P(f,t) \approx \sum_{z\in S} P(z)P(f|z)P(t|z) + \sum_{z\in N} P(z)P(f|z)P(t|z) \tag{2}$$

where $P(f|z)$ for $z \in N$ and $P(f|z)$ for $z \in S$ represent the noise dictionary and sound events dictionary, respectively. Noise reduction can therefore be realized by setting noise activations $P(z)$ for $z$ to zero, therefore, denoised sound spectra can be obtained by:

$$P(f,t) \approx \sum_{z\in S} P(z)P(f|z)P(t|z) \tag{3}$$

We perform noise source separation in a semi-supervised manner, in which noise dictionary $P(f|z)$ for $z \in N$ is trained by background noise collected beforehand; while sound event dictionary $P(f|z)$ for $z \in S$ and dictionary element activations $P(z)P(t|z)$ for $z \in S$ are estimated through the PLCA decomposition in (2).

Though PLCA presents state-of-the-art performance for noise reduction [12], artifacts are produced through noise source separation process, especially under extremely noisy conditions, e.g. -5dB SNR. Increasing EM iterations for KL divergence minimization and initializing more latent components in PLCA model are possible ways to suppress artifacts; however, they introduce heavy computation load which greatly degrade algorithm availability. In this work, we present an alternative solution to enhance intelligibility of denoised sound through focusing on frequency bands with fewer artifacts generated. It is grounded on the fact in human auditory that noise corrupted sound is recognized via processing the audio in local high SNR frequency bands [13]. We realize such spectral concentrating mechanism by assigning a series of spectral reliability weighting, which is derived from reconstruction error of PLCA model for background noise estimation. The frequency band-wise reconstruction error

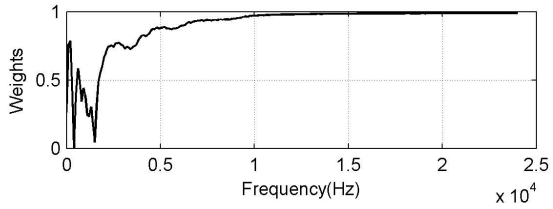


**Fig. 2**. Reliability weights for denoised sound spectra

possibility distribution is expressed as:

$$P(f|e) = \sum_t (X_{t,f} - \tilde{X}_{t,f})^2 / \sum_f \sum_t (X_{t,f} - \tilde{X}_{t,f})^2 \quad (4)$$

where $X$ is input noise spectrogram and $\tilde{X}$ is reconstructed spectrogram by PLCA model. $P(f|e)$ presents spectral error possibility distribution. Based on which, a spectral reliability measure can be derived as:

$$\omega_f = 1 - p(f|e)/max(p(f|e)) \quad (5)$$

from which lower error rate bands are assigned with higher weights, in contrast, high error possibility bands will be indexed by smaller weights to suppress the significance. A series of spectral reliability weightings for babble noise is shown in Fig.2 and it can be applied to audio spectrogram by $S_{\cdot t} = X_{\cdot t} \cdot \omega_f$, where $S_{\cdot t}$ denotes $t-$frame denoised spectrum, $X_{\cdot t}$ represents enhanced spectral feature by weightings $\omega_f$.

### 2.2. Robust feature extraction

Based on the aforementioned process, we further develop robust acoustic representation by using triangle filter bank and eigen-decomposition. Mel-filter bank is well developed for characterizing speech signal. We followed the Mel-filter banks idea and design a uniformly positioned triangle-filter bank to describe sound events, which renders three main advantages: First, the triangle-filter produces more robust audio representation by introducing fuzzy assignment over audio spectra. Second, in contrast to Mel-filter bank, which emphasizes on low band contents where speech signal mainly lays in, the uniformly positioned filters render identical resolution across all frequency bands, and thus capture richer temporal-spectral dynamics in sound event. Besides, filter bank effectively reduce the feature dimension and therefore facilitates sound classification. The filtering process is expressed as:

$$\tilde{s_n} = \sum_f \alpha_n(f) s(f) \quad (6)$$

where $s(f)$ denotes denoised audio spectum corresponding to $f$ band, $\alpha_n(f)$ represent spectral weighting coefficients of $n$-th filter. Then filtered spectrum can be written as $\tilde{S}_t = [\tilde{s}_1, ..., \tilde{s}_n]^T$.

Based on filtered audio spectrogram, $[\tilde{S}_1, ..., \tilde{S}_T], \tilde{S}_T \in \Re^{N \times 1}$, we perform eigen-decomposition to further characterize significant temporal-spectral spectral patterns in sound event. The eigen decomposition to can be written as:

$$R_{\tilde{S}_T} = U \Lambda U^T, \quad R_{\tilde{S}_T} = E\{S_t S_t'\} \quad (7)$$

where $U = [u_1 ... u_N]$ are the eigenvectors characterizing predominant patterns of sound and $\Lambda = diag(\lambda_1 ... \lambda_N)$ is diagonal eigenvalue matrix. The contribution ratio of k-th eigenvector uk is defined as:

$$\eta_k = \lambda_k / \sum_{i=1}^{N} \lambda_i \quad (8)$$

which shows its signicance in representing the audio. We select the first $K$ principle eigenvectors with highest contribution ratios $U_K = [u_1 ... u_K], K < N$ to form a sound representation. In addition, the eigenvectors in $U_K$ are normalized by their contribution ratios through computing $\tilde{u}_k = \{\lambda_k / \sum_{i=1}^{K} \lambda_i\} \cdot u_k,\ k = 1...K$ which follows a similar manner to [14]. Contribution weightings give prominence to the principle eigenvectors for describing sound event. In the end, we concatenate $K$ normalized principle eigenvectors to build acoustic feature vector.

### 2.3. Sound classification by regularized kernel Fisher discriminant analysis

We employ KFDA to perform sound classification. Let $X_i = \{x_1^i ... x_{N_i}^i\}$ be audio features from class $i$. $\phi(x)$ is nonlinear mapping of input vector $x$ into kernel feature space $\mathcal{F}$. KFDA seeks a direction $w \in \mathcal{F}$ maximizing class separability which is defined as:

$$J(w) = \frac{w^T S_B^\phi w}{w^T (S_B^\phi + \lambda\epsilon) w}, S_B^\phi = (m_1^\phi - m_2^\phi)(m_1^\phi - m_2^\phi)^T$$

$$S_W^\phi = \sum_i \sum_x (\phi(x) - m_i^\phi)(\phi(x) - m_i^\phi)^T, m_i^\phi = \frac{1}{N_i} \sum_{n=1}^{N_i} \phi(x_n^i) \quad (9)$$

where $m_i^\phi$ denotes class center, while $S_B^\phi$ and $S_W^\phi$ are within-class and between-class variances in kernel feature space. Since the within-class variance may be singular, a regularization term is added. It is typical Rayleigh quotient and solution can be found in [15].

## 3. EXPERIMENTS

### 3.1. Dataset

We evaluate the proposed sound classification method using RWCP sound scene dataset [16], in which sound files are recorded at 48 kHz sampling rate with high SNR. There are 105 categories of sounds including some duplicated classes.

| Method | Proposed | Dennis [8] | YE [11] | MF-HMM[8] | MFCC-HMM[8] |
|---|---|---|---|---|---|
| Clean | 100% | 98.5% | 100% | 95.7% | 99.0% |
| 20dB | 100% | 98.0% | 98.5% | 94.2% | 62.1% |
| 10dB | 100% | 95.3% | 98.5% | 84.7% | 34.4% |
| 0dB | 100% | 90.2% | 59.5% | 69.5% | 21.8% |
| -5dB | 99.0% | 84.6% | 30.0% | 53.8% | 19.5% |
| Avg. | 99.6% | 93.3% | 77.3% | 79.6% | 47.3% |

**Table 1**. Results comparison on 10 sound classes

such as 5 types of coin sounds. We removed duplicated classes and thereafter obtained 62 distinct sound categories with 5949 samples for evaluation. Typical non-stationary babble noise is extracted from NOISEX92 database for evaluating the robustness of the proposed method.

### 3.2. Parameter setting

We set window length to 10ms with 7.5ms overlapping in Short-time Fourier Transform (STFT). At PLCA noise reduction stage, background noise dictionary size and foreground sound event dictionary size are both fixed to 50, since it was found enough to model babble noise and sound events. The number of EM iterations is experimentally set to 100. $N = 50$ triangle-filters are used to build filter bank in (6). We concatenate $K = 3$ principle eigenvectors with highest contribution ratio to form audio representation, from which over 95% of patterns in sound are accommodated according to their contribution ratios (7). The regularization term in (8) is set to $10^{-4}$. 10 seconds babble noise are utilized for training the noise model in PLCA, which is removed at noisy audio samples generation stage to ensure there is no overlap in training and testing background sound. In KFDA, spread parameter in gauss kernel is set to 0.3.

### 3.3. Experiment 1: validation by comparison

In first experiment, we compare the proposed approach with other methods [8,11]. The evaluation protocol is similar to [8], in which 10 classes of sound events are selected from RWCP database, i.e. bells5, bottle1, buzzer, cymbals, horn, kara, metal15, phone4, ring and whistle1. For each class, 40 files are randomly selected, from which half are used for training and the other half are for testing. The babble noise is added to testing data with 20, 10, 0 and -5 dB SNRs. Final results are computed by averaging the outcomes across 5 runs of test. Meanwhile, several conventional methods such as MFCC with HMM and Mel-Frequency (MF) with HMM are also evaluated [8]. According to results shown in Table.1, proposed scheme achieved favorable classification accuracy for all noise conditions and significantly outperformed all other methods. In addition, we plot all features extracted from 10 categories of sounds corrupted with 20 to -5 dB babble noise

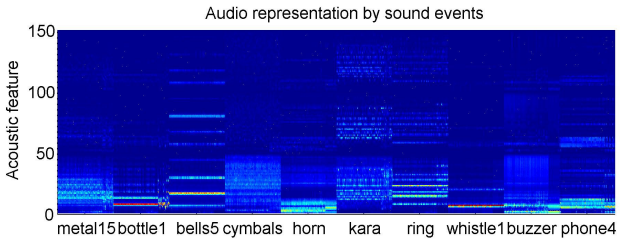


**Fig. 3**. Robust feature extracted from noise corrupted sounds

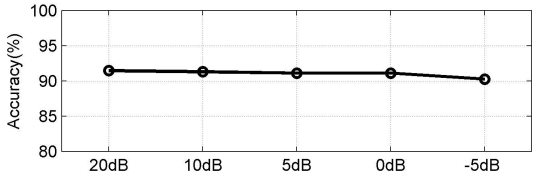


**Fig. 4**. Classification results on 62 sound classes with various noise intensities

in Fig. 3, from where clusters for each sound class can be clearly observed, which manifests robustness of the proposed feature. In addition, significance of adopting noise reduction together with spectra enhancement can be observed through comparing our result with result shown by [11], in which a similar audio (subspace) representation is employed. As a conclusion, the proposed method greatly outperform [11] under extremely noisy conditions.

### 3.4. Experiment 2: evaluation with more data

To further confirm superiority of the proposed method, we perform more extensive sound classification test using RWCP database with 62 sound classes and 5949 samples. The performance is measured by 10-fold cross validation, in which clean data is applied for training and test samples are corrupted by babble noise with 20, 10, 5, 0 and -5dB SNRs, respectively. From results in Fig.4, proposed method obtained an average accuracy of 91.04%. Even under challenging -5dB SNR condition, it achieved more than 90% classification accuracy.

## 4. CONCLUSION

This paper proposed novel noise robust sound classification approach presenting favorable performance in severe acoustic environment. The proposed method assembles advanced signal processing techniques to realize the human auditory mechanism, in which noise corrupted sound event is recognized based on part of spectral information that with high SNR. In detail, three main steps are included: noise source separation, robust acoustic feature extraction and regularized KFDA multi-class classification. We demonstrated the proposed method with RWCP sound scene database. Extensive experiments validated superiority of the proposed approach.

## 5. REFERENCE


[1] Wang, Wenwu. "Machine Audition: Principles, Algorithms and Systems." IGI Global press, 2011

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

[3] Stavros Ntalampiras, Ilyas Potamitis, Nikos Fakotakis, "On acoustic surveillance of hazardous situations", *in Acoustics, Speech and Signal Processing, 2009, IEEE International Conference on,* pp. 165-168, 2009

[4] R. Radhakrishnan and A. Divakaran, "Systematic acquisition of audio classes for elevator surveillance", *Image and Video Communications and Processing*, *vol. 5685 of Proceedings of SPIE*, pp. 64–71., 2005

[5] Cowling, M. and R. Sitte, "Comparison of techniques for environmental sound recognition", *Pattern Recognition Letters*, vol.24, no.15, pp. 2895-2907, 2003

[6] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and Thomas S. Huang, "Real-world acoustic event detection", *Pattern Recognition Letters*, vol. 31, no.12, pp.1543–1551, Sept. 2010

[7] X. Valero, F. Alías, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification", *Multimedia, IEEE Transactions on*, vol. 14, no. 6, pp. 1684-1689, 2012.

[8] J. Dennis, Q. Yu, H. Tang, H. Tran and H. Li, "Temporal coding of local spectrogram features for robust sound recognition", *in Acoustics, Speech and Signal Processing, 2013. IEEE International Conference on,* pp. 803-807, 2013

[9] Shashanka, M.V., B. Raj and P. Smaragdis, "Probabilistic Latent Variable Models as Non-Negative Factorizations", *special issue on Advances in Non-negative Matrix and Tensor Factorization, Computational Intelligence and Neuroscience Journal*, May 2008

[10] Beheshti, S. and M. A. Dahleh. "Noise variance in signal denoising", *In Acoustics, Speech, and Signal Processing, 2003, IEEE International Conference on.* p.p. 185-188, 2003

[11] Jiaxing Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Kernel discriminant analysis for environmental sound recognition", *in Acoustics, Speech and Signal Processing, 2013. IEEE International Conference on,* pp. 808-812, 2013

[12] Zhiyao Duan, G. J. Mysore and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments", in Proc. *Interspeech*, 2012.

[13] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, pp. 1562-1573, 2006

[14] Takumi Kobayashi, "Generalized Mutual Subspace Based Methods", *In Proc of Asian Conf. on Computer Vision*, 2012

[15] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Muller, "Fisher discriminant analysis with kernels", *In Proc. of Neural Networks for Signal Processing IX*, 1999, pp. 41–48, 1999

[16] Real World Computing Partnership, "RWCP Sound Scene Database", http://tosa.mri.co.jp/sounddb/index.htm

[17] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993