SUBSPACE PROJECTION CEPSTRAL COEFFICIENTS FOR NOISE ROBUST ACOUSTIC EVENT RECOGNITION

Sangwook Park[1]*, Younglo Lee*[1]*, David K. Han*[2]*, and Hanseok Ko*[1]

[1]School of Electrical Engineering, Korea University, Seoul, Korea
[2]Office of Naval Research, Arlington, VA, USA

ABSTRACT

In this paper, a novel feature for noise robust sound event recognition is proposed. The proposed feature is obtained by a two-step procedure. First, a subspace bank is established via target event analysis in complex vector space. Then, by projecting observation vectors onto the subspace bank, noise effect can be reduced while generating discriminant characters originated from differing event subspaces. To demonstrate robustness of the proposed feature, experiments with several classifiers were conducted with varying SNR cases under four noisy environments. According to the experimental results, the proposed method has shown superior performance over prominent conventional methods.

***Index Terms*—** acoustic event classification, robust feature extraction, subspace learning, principal component analysis

1. INTRODUCTION

Non-linguistic sound contains rich information such as presence of humans, objects, or their activities. Acoustic Event Recognition (AER) is one of the research fields that exploits and extracts information from these sounds [1-3]. In earlier research, some features effective in speech/speaker recognition were used [1, 4]. These features were obtained by using human auditory based filter bank. However, general sounds such as breaking glass, explosion, or splashing water have different characteristics compared to vocal sound. Recently, event filter-banks were designed by expanding the acoustic spectrum beyond that of human speech [5-8]. They have shown that AER performance can be improved by designing filters that expanded its spectral range as well as their resolution beyond that required for human speech.

The AER systems using the features based on the event filter-bank have shown good performance in a clean environment. However, in many applications, the AER has to operate effectively and reliably in noisy environment. Power Normalized Cepstral Coefficients (PNCCs) and Robust Compressive Gammachirp filter bank Cepstral Coefficients (RCGCCs) have been proposed for handling noise [9, 10]. However, these features adopt noise suppression strategies that do not handle low Signal to Noise Ratio (SNR) cases due to signal distortion. Although Spectrogram Image Features (SIF) has been proposed to establish noise robust AER systems [11], it fails to deal with signal variations, even in clean environments. In [6], an event filter-bank was designed by locating filters on relatively noise-free frequency bands. However, the method may be effective only for narrow band noise. In [7, 8], an event filter-bank was composed of several spectral bases that were obtained by applying a Non-negative Matrix Factorization (NMF) to spectrogram. The spectral bases function as templates for capturing a signal envelop from noisy observation like matched filter approach. However, signal envelopes in noisy environment may be differ from the templates due to phase difference, and this difference may pose difficulties in capturing signals.

In this paper, a new feature named Subspace Projection Cepstral Coefficients (SPCCs) are proposed for achieving a noise robust AER. Under this approach, event signal is analyzed in complex vector space under an additive noise assumption and the vector space representing observations is decomposed into two orthogonal subspaces, signal and noise. When observations are projected into the signal subspace, the noise component is removed and discriminative event characteristics remain in the projection result.

The remainder of this paper is organized as follows. Section 2 explains the proposed method with its motivation. Section 3 summarizes the experimental results of the proposed feature and its comparison to several other features. After a discussion on the results, conclusions are drawn in the final section.

2. PROPOSED FEATURE

2.1. Motivation

By applying Discrete Fourier Transform (DFT) to a finite time sequence, the sequence is transformed to an observation vector lied on *K*-dimensional complex vector space. *K* is determined as a half of the number of DFT points, and the observation vector can be represented by a linear combination of basis vectors in the complex vector space. Note that the basis vector means frequency bin for

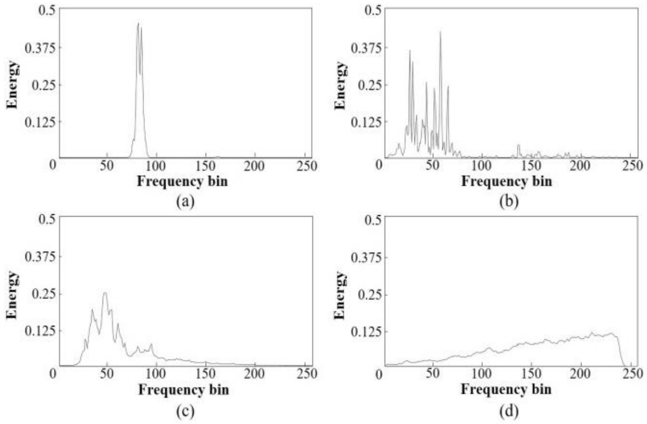

Fig. 1. Normalized energy about four types of events: (a) Whistle, (b) Piano, (c) Door opening/closing, (d) Rain sound

representing frequency domain. Fig. 1 depicts several events in terms of normalized energy versus frequency. In case of an event such as whistle that has harmonic components, energy is confined in finite frequency bins. On the other hand, energy of an event without harmonic component is spread in wide frequency bands. As shown in the case of door or rain, however, their energy distributions are different according to frequency bin. From these observations, firstly, envelopes of an event can be represented by linear combination of R_i bases ($R_i < K$) where the event subspace is composed of these bases. Secondly, event subspaces as well as the bases that compose them vary depending on each event. The second consideration is one of the key factors for successful AER.

2.2. Expected effect by projecting onto event subspace

In a scenario that assumes uncorrelated-additive noise, the covariance matrix of observation can be represented by

$$C_X = C_S + C_N \tag{1}$$

where C_X, C_S and C_N are the covariance matrices of observation, event signal, and noise, respectively. Due to the influence of noise, C_X may be a full-rank matrix although C_S is singular matrix composed of a subspace approximated to a rank of R_i. According to the assumption, the noise component must be in a null space of C_S, which implies that the two subspaces, signal and noise, are orthogonal to each other. Consequently, the noise effect can be reduced by projecting the observation onto the signal subspace.

To perform AER, an observation is projected onto each event subspace. Since each event has a different subspace, this projection will construct a signal that differs according to event. Hence, the projection result is applicable as a discriminant character for AER.

2.3. Establish subspace-bank

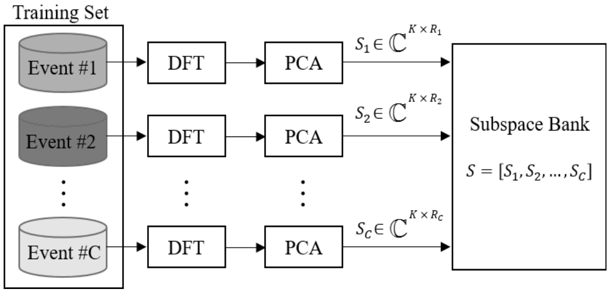
Training Set
Event #1
Event #2
Event #C
DFT
PCA
$S_1 \in \mathbb{C}^{K \times R_1}$
$S_2 \in \mathbb{C}^{K \times R_2}$
$S_C \in \mathbb{C}^{K \times R_C}$
Subspace Bank
$S = [S_1, S_2, \dots, S_C]$

Fig. 2. Block diagram for training subspace bank

For subspace learning, several algorithms using dimensional reduction approach have been introduced [12]. Among the methods, Principal Component Analysis (PCA) is a popular yet powerful method for establishing an event subspace [13]. As shown in Fig. 2, each event subspace is separately trained via PCA. The i^{th} event dataset is transformed into a K-dimensional complex vector space by DFT. Then, PCA is conducted to produce R_i-eigenvectors selected by preserving 90% of data energy to establish the i^{th} subspace S_i. Note that the data energy is a summation of all eigenvalues. By repeating this procedure for every events, a subspace bank is generated. The subspace bank $\boldsymbol{S} \in \mathbb{C}^{K \times R}$ is used in feature extraction procedure which is mentioned in the next section. Note that $\mathbb{C}$ is a symbol of matrix whose element is complex number, and R is equal to sum of each subspace dimensions. In Fig. 2, C is the number of target events.

2.4. SPCC extraction

The proposed feature extraction procedure is shown in Fig 3. First, pre-emphasis is performed in order to compensate loss of high frequency component in input sequence $x[n]$. After framing, $x_p[n]$ is transformed into a complex vector space by Short Time Fourier Transform (STFT) to $\boldsymbol{X} = \mathcal{F}\{x_p[n]\} \in \mathbb{C}^{K \times L}$ where n is time sequence index, x_p is the result of framing and windowing, and L is the number of frames.

$\boldsymbol{X}$ is projected onto each basis vector in the subspace bank $\boldsymbol{S}$. Since the subspace bank is established by concatenating each event subspace, the projection result $\boldsymbol{P} \in \mathbb{C}^{R \times L}$ is composed of concatenated vectors projected onto all event subspaces for every frame. Next, for analyzing the $\boldsymbol{P}$,

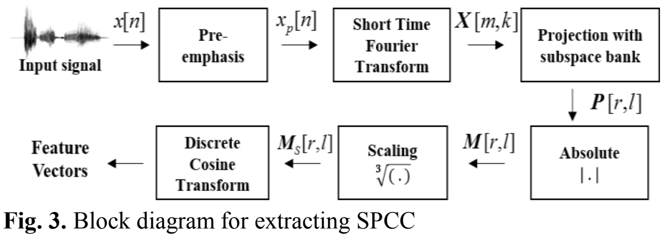

Fig. 3. Block diagram for extracting SPCC

TABLE 1
THE ACOUSTIC EVENT DATABASE AND ITS SIZE

Event	# of data	Event	# of data
Bark	318	Male scream	149
Bell	227	Female scream	207
Door op./cl.	346	Breaking glass	263
Piano	162	Siren	267
Pop music	288	Tire skid	239
Rain	283	Whistle	299
Water pouring	248		

its elements are converted to $\boldsymbol{M}\in\mathbb{R}^{R\times L}$ in real numbers by applying an absolute function. Note that the $\mathbb{R}$ is a symbol of real matrix. To reduce magnitude deviation, the $\boldsymbol{M}$ is scaled by applying 3rd order root function. For decorrelation and compression, the proposed feature, SPCC, is finally obtained by applying Discrete Cosine Transform (DCT) to $\boldsymbol{M}_S$.

3. EXPERIMENT

3.1. Database

Database as provided in Table 1 was collected in several locations by a portable recorder with an average length of 1.5 seconds. For validation of noise robustness, four noisy environments (*cafeteria, office, cross-road, and pub*) were chosen from the ETSI background noise database [14]. In the experiments, the sound event database was mixed with the four types of noise at SNR in 20, 10, 5, and 0 dB by using the ADDNOISE library [15].

3.2. Experiment Setting

3.2.1. Cross-validation Test

For performance assessment, the database was categorized into 5 subsets and with a ratio of 4:1 between training and test sets. Mean average recognition rate for 13 target events and all possible combinations are provided in the recognition results.

3.2.2. Conventional Methods

In order to demonstrate effectivity of the proposed feature, prominent features, MFCC, PNCC [9], RCGCC [10], SIF [11], and Spectral Basis Vector - Cepstral Coefficient (SBVCC) [8] were considered as conventional methods. For conducting a Short Time Fourier Transform (STFT), frames were defined as 32 ms time sequences with an overlap with the next frame for 16 ms. DFT was conducted with 512-points and the Hamming window.

In the DCT step, 40 coefficients including delta coefficients were extracted to compose cepstral coefficient vectors of the considered cepstral features (e.g. MFCC, PNCC, RCGCC, and SBVCC). Then, the Gaussian Mixture Model (GMM) modelled by 32 mixture components was applied to these frame-based features [8].

In the case of SIF, parameters required for feature extraction were determined as described in [11]. Correspondingly, the classifier for SIF was configured to apply Support Vector Machine (SVM) using a quadratic polynomial kernel. Note that GMM is not appropriate for the SIF because its dimension (486-dimension) is too large to estimate GMM parameters. Additionally, Deep Belief Network (DBN) consisted of 3 hidden layers with each layer having the same number of nodes was also considered as the classifier. Experimentally, SIF performed best when the DBN structure is trained with 200-nodes for each layer.

3.2.3. Proposed Method

The proposed feature is a type of cepstral features and subsequently both parameters for feature extraction and feature dimension were same with those of cepstral features. In order to demonstrate effective of the SPCC, several classifiers were applied. Firstly, GMM and SVM trained with linear kernel were used for performance comparison to other prominent conventional features. Note that the input feature for SVM was obtained by averaging over all frames. Secondly, Deep Belief Network (DBN) was also applied to the proposed feature for comparing with the classifier based on deep learning. The DBN structure was same with one applied to SIF except the number of nodes. In this case, the number of nodes was set to 100, experimentally. The input vectors for DBN were extracted as follows. 80 features were extracted by calculating the mean and standard deviation for each of the 40 DCT coefficients over all frames [16].

3.3. Experiment Results

Table 2 shows the experiment results when 32-mixture GMM is applied to several cepstral features including SPCC. In case of clean condition, all features performed well with a recognition rate of over 90%. By adding noise, all performances were drastically degraded in low SNR conditions. Due to the signal enhancement procedures in PNCC and RCGCC, their performances are shown to be better than MFCC. However, this holds only in sufficiently high SNR conditions. Otherwise, some problems such as signal distortion and residual noise may occur. Since SBVCC is extracted based on energy detectors of each event, performance can be improved in low SNR. In SPCC, it is clearly observed that SPCC retains higher recognition rate than that of other considered methods even in low SNR cases.

Table 3 summarizes the results obtained by applying SVM or DBN to SIF and SPCC. Since the SIF is composed of statistics on each partial region of spectrogram, the feature does not cover the case when the signal is in different region. To overcome this case, the signal interval in spectrogram must be perfectly found. However, this is impractical in real noisy condition. For this reason, average recognition rate of SIF in clean condition is the worst among

TABLE 2
EXPERIMENT USING 32 MIXTURE GAUSSIAN MIXTURE MODEL RESULTS [%]

	SNR	0 dB	5 dB	10 dB	20 dB	Clean		SNR	0 dB	5 dB	10 dB	20 dB	Clean
cafeteria	MFCC	45.42	57.95	70.60	87.76	91.59	*Pub*	MFCC	33.13	45.78	61.88	80.58	91.59
	PNCC	57.01	72.08	82.99	90.30	91.40		PNCC	39.82	58.43	69.77	88.74	91.40
	RCGCC	52.40	66.37	79.66	90.31	91.62		RCGCC	30.49	53.89	69.16	87.90	91.62
	SBVCC	59.63	71.74	83.21	90.82	91.82		SBVCC	39.75	53.61	70.98	90.21	91.82
	SPCC	**80.31**	**80.31**	**88.07**	93.34	94.95		SPCC	48.33	**72.92**	**86.18**	93.58	94.95
office	MFCC	46.50	58.48	70.68	89.08	91.59	*cross-road*	MFCC	47.02	64.36	78.59	89.75	91.59
	PNCC	47.50	65.47	79.10	89.75	91.40		PNCC	78.38	86.78	89.88	91.16	91.40
	RCGCC	54.97	68.46	81.55	91.04	91.62		RCGCC	65.67	79.90	88.20	91.35	91.62
	SBVCC	59.98	74.47	85.22	91.56	91.82		SBVCC	71.65	83.42	89.76	91.46	91.82
	SPCC	**76.26**	**85.30**	**90.96**	94.22	94.95		SPCC	**88.01**	**92.24**	93.76	94.46	94.95

TABLE 3
EXPERIMENT USING SUPPORT VECTOR MACHINE OR DEEP BELIEF NETWORK RESULTS [%]

	SNR	0 dB	5 dB	10 dB	20 dB	Clean		SNR	0 dB	5 dB	10 dB	20 dB	Clean
SIF-SVM	*Cafeteria*	51.31	63.01	68.76	70.71	70.45	SIF-DBN	*Cafeteria*	63.75	75.50	80.63	82.60	82.82
	Office	58.83	67.75	70.28	70.80	70.45		*Office*	70.75	78.48	81.58	82.81	82.82
	Pub	42.39	60.90	69.19	70.59	70.45		*Pub*	51.04	70.89	79.47	82.75	82.82
	Cross-road	70.04	70.78	70.50	70.37	70.45		*Cross-road*	80.86	82.05	82.57	82.85	82.82
SPCC-SVM	*Cafeteria*	51.05	68.35	85.45	**97.75**	**99.11**	SPCC-DBN	*Cafeteria*	70.02	77.74	84.54	91.09	93.27
	Office	55.30	74.09	90.72	**98.39**	**99.11**		*Office*	71.85	80.66	86.06	91.79	93.27
	Pub	41.20	52.37	71.15	**95.77**	**99.11**		*Pub*	**58.49**	70.26	78.79	87.86	93.27
	Cross-road	77.01	90.18	**96.77**	**98.96**	**99.11**		*Cross-road*	80.22	85.38	89.13	92.99	93.27

all considered features although its degradation according to SNR is less than other features. Otherwise, in cases both SVM and DBN, SPCC generally outperforms SIF in clean as well as noisy condition.

4. DISCUSSION

As previously mentioned, main idea of the SPCC is to reduce noise effect and extract discriminant character by projection. By comparing to other results, these effects can be identified. From overall results, SPCC has the best performance. On recognition performance using GMM, the performance of SPCC is improved on the average by 17.21%, 9.26%, 11.01%, and 8.17% compared to MFCC, PNCC, RCGCC, and SBVCC, respectively. Also, its performance is improved on the average by 15.65% and 5.50% compared to SIF in recognition using SVM and DBN, respectively.

For classification of the SPCC, GMM shows better performance on average than other classifiers. In clean condition, SVM shows the best performance among the considered classifiers. However, it has the largest degradation in 0 dB cases. In the aspect of classification criteria, all frames have same weight when classification is performed with SVM. Practically, importance of each frame may be different from each other. Although DBN also has the same issue for classification, its performance is better than SVM in 0 dB cases because input feature for DBN includes not only frame-mean but also frame-variance. In particular, DBN shows the best performance in 0 dB *pub* condition. On the other hand, each frame has different weight in terms of likelihood in classification using GMM. Based on maximum likelihood criteria, GMM usually shows noise robust performance without increasing feature dimension.

5. CONCLUSIONS

In this paper, SPCC is proposed as a noise robust AER feature. SPCC can be obtained by projecting the observation onto an event subspace-bank. By means of the projection procedure, noise components can be suppressed while extracting discriminant character generated by differing event subspaces. For performance assessment, prominent conventional features such as MFCC, PNCC, RCGCC, SIF and SBVCC were considered. And several classifiers such as GMM, SVM, and DBN were considered and the experiments were conducted with varying SNR cases under four noisy environments such as *cafeteria, office, pub*, and *cross-road*. In reference to the experimental results, the proposed feature, SPCC-GMM, has shown to outperform the other conventional features.

6. REFERENCES

[1] W. Choi, J. Rho, D. K. Han, and H. Ko, “Selective background adaptation based abnormal acoustic event recognition for audio surveillance”, in Proc. IEEE Conference on Advanced Video and Signal Based Surveillance, Beijing, China, pp. 118-123, Sep. 2012.

[2] K. Yamano and K. Itou, “Browsing audio life-log data using acoustic and location information,” in *Proc. IEEE Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, Sliema, Malta, pp. 96-101, Oct. 2009.

[3] M. Xu, C. Xu, L. Duan, J.S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” *ACM Trans on Multimedia Computing and Communications and Applications*, vol. 4, no. 2, pp.1-23, May 2008.

[4] W. Choi, S. Kim, M. Keum, D. K. Han, and H. Ko, "Acoustic and visual signal based context awareness system for mobile application", *IEEE Trans. on Consumer Electronics*, vol. 57, no. 2, pp.738-746, May 2011.

[5] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, “Robust sound event classification using deep neural networks,” *IEEE/ACM Trans. on Audio, speech, and Lang. Proc.*, vol. 23, no. 3, pp. 540-552, Mar. 2015.

[6] S. Park, W. Choi, D. K. Han, and H. Ko, "Acoustic event filterbank for enabling robust event recognition by cleaning robot," *IEEE Trans. on Consumer Electronics*, vol. 61, no 2, pp. 189-196, May 2015.

[7] M. K. I. Molla and K. Hirose, “Audio classification using dominant spatial patterns in time-frequency space,” in *Proc. INTERSPEECH*, Lyon, France, pp. 2915-2919, Aug. 2013.

[8] W. Choi, S. Park, D. K. Han, and H. Ko, “Acoustic event recognition using dominant spectral basis vectors,” in *Proc. INTERSPEECH*, pp. 2002–2006, Sep. 2015.

[9] C. Kim and R. M. Stern, “Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction,” in *Proc. INTERSPEECH*, Brighton, UK, pp. 28-31, Sep. 2009.

[10] M. J. Alam, P. Kenny, and D. O’Shaughnessy. “Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique,” *Digital Signal Processing*, vol 29, pp. 147-157, Jun. 2014.

[11] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions,” *IEEE Signal Processing Letters*, vol. 18, no. 2, Feb. 2011.

[12] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Overview of multilinear subspace learning,” in *Multilinear subspace learning dimensionality reduction of multidimensional data*, Boca Raton: CRC press, 2014, pp. 71-87

[13] T. Bouwmans and E. H. Zahzah, “Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance,” *Computer Vision and Image Understand.*, vol. 122, pp. 22-34, May 2014.

[14] European Telecommunications Standards Institute, “ETSI: EG 202 396-1 v1.2.2,” 2008.

[15] ITU, “Objective measurement of active speech level,” *ITU-T Recommendation* P.56, 1993.

[16] Z. Kons and O. T. Ronen, “Audio event classification using deep neural networks,” in *INTERSPEECH*, in *Proc. INTERSPEECH*, Lyon, France, pp. 1482–1486, Aug. 2013.