Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Signal Processing in Acoustics (others) : Paper ICA2016-307

Investigation on suitable acoustic features in deep neural network for environmental sound discrimination

Sakiko Mishima[(a)], Tomoyuki Mizuno[(b)], Takahiro Fukumori[(c)], Masato Nakayama[(d)], Takanobu Nishiura[(e)]

[(a)(b)] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan,
[(a)] is0188if@ed.ritsumei.ac.jp, [(b)] is0140kk@ed.ritsumei.ac.jp
[(c)(d)(e)] College of Information Science and Engineering, Ritsumeikan University, Japan,
[(c)] fukumori@fc.ritsumei.ac.jp, [(d)] mnaka@fc.ritsumei.ac.jp, [(e)] nishiura@is.ritsumei.ac.jp

Abstract

Surveillance systems have been utilized for the safety of the elder people who live alone. Most of them have been utilized for detecting hazardous situations with a video camera. However, a video camera has a problem that it is difficult to monitor the dark and blind areas. In order to solve this problem, methods for hazardous sound detection have been proposed by using environmental sounds which consist of various sounds in daily life. It is important to improve the discrimination accuracy of environmental sounds in order to accurately monitor the situation. Conventional acoustic models have been realized on the basis of a hidden Markov model (HMM) with mel-frequency cepstrum coefficients (MFCCs). However, it is difficult to discriminate the environmental sound because the conventional method for constructing the acoustic model is insufficient to express acoustic features. Deep neural network (DNN) can extract complex features from input signals. High-dimensional input features are effective to input to DNN because the environmental sound has various features. However, suitable input features are required in order to reduce the computation cost. Therefore, we investigate suitable acoustic features for DNN. We employed MFCCs, mel-filter bank, and linear-filter bank as acoustic features. From evaluation experiments, we confirmed the performance of environmental sound discrimination in each acoustic feature.

Keywords: Acoustic features, Deep neural network, Environmental sound discrimination

Investigation on suitable acoustic features in deep neural network for environmental sound discrimination

1 Introduction

In recent years, the elderly people who live alone have become increasing in the country with the aging society. Domestic accidents of the elderly people have been serious in the aging society with the nuclear family. Therefore, it is necessary to monitor the safety of them. Surveillance systems with a video camera have been utilized for monitoring the elderly people [1]. However, the systems have a problem that it is difficult to monitor the situations in dark and blind areas. In order to solve this problem, the surveillance system based on an environmental sound detection has been proposed [2]. The environmental sound consists of various sounds in daily life, for example, foot step sounds, ringtones, broken sounds of a window glass, and so on. A detection system for the environmental sound is possible to identify the hazardous situations in dark and blind areas. Thus, in this system, it is important to improve the discrimination accuracy of environmental sounds in order to accurately monitor the situations.

In the previous researches, acoustic models have been constructed on the basis of hidden Markov models (HMMs) [3] with the mel-frequency cepstrum coefficients (MFCCs) [4]. HMMs have been utilized to statistically construct the acoustic models. MFCCs have been extracted from the environmental sound as the acoustic features because MFCCs are possible to reduce dimensions of the features on the basis of auditory characteristics of human. However, it is difficult to discriminate the environmental sound because the conventional method is insufficient to express acoustic features for constructing the acoustic model. Therefore, it is necessary to construct the acoustic models using by many features which is possible to express the difference of environmental sounds.

In the recent researches, the deep neural network (DNN) has been paid attention as an effective method to construct the acoustic models [5]. DNN is possible to extract the complex features by training the complex network with input signals, which are high-dimensional features. DNN is ideal candidate for constructing the acoustic models because the environmental sound has various acoustic features, for example, duration time, frequency characteristic and generation environment of the sound. However, extracting high-dimensional features such as various acoustic features takes a high cost for training the model. Therefore, the selection of suitable input features is required in order to reduce the computational cost while keeping the higher identification accuracy of the environmental sound. In this paper, we therefore propose the method which utilizes DNN as acoustic model to discriminate the environmental sound. In order to find out the suitable acoustic features for DNN, we investigate the MFCCs, the mel-filter bank, and the linear-filter bank.

2 Conventional method for environmental sound discrimination

For discriminating the environmental sound, the system based on HMMs with MFCCs has been proposed as the conventional method [6]. Figure 1 shows an overview of the conventional method. As shown in Fig. 1, MFCCs have been extracted from the environmental sound as the acoustic features because they are possible to reduce dimensions of the features based on auditory characteristics of human. Especially, MFCCs are superior to express the sound which has dominant power in lower frequency band such as voice. Then, HMMs have been utilized to statistically construct the acoustic models. HMMs have plural states, output probabilities, and state transition probabilities. Output probabilities are approximated by Gaussian mixture models (GMMs) which are constructed plural weighted Gaussian distribution. State transition probabilities correspond to the flexibility of time. Therefore, HMMs approximate the distribution of MFCCs by the superposition of Gaussian distribution and the connection of plural states. The acoustic models are constructed by each sound class as shown in Fig. 1.

However, it is difficult to discriminate the environmental sound because the conventional method is insufficient to express acoustic features for constructing the acoustic model. The environmental sound is composed of many kind of sound which have various features, for example, duration time, frequency characteristic and environment. Thus, it is especially difficult to extract the features which are possible to properly express the environmental sound. In order to discriminate the environmental sound with high accuracy, it is necessary to construct the acoustic models with various features which can express the difference of environmental sounds.

3 Investigation on suitable acoustic features in deep neural network for environmental sound discrimination

We propose the method to discriminate the environmental sound with DNN which is utilized as acoustic model. Figure 2 shows an overview of the proposed method. As shown in Fig. 2, acoustic features are extracted from the environmental sound to input them to DNN. In this paper, in order to find out the suitable acoustic features for DNN, we investigate the MFCCs, the mel-filter bank, and the linear-filter bank. Then, DNN is trained with the acoustic feature to construct the acoustic model. Trained DNN is utilized as the acoustic models for environmental sound discrimination. To discriminate the sound, the acoustic features are inputted to DNN. DNN outputs the sound class of the environmental sound as a discrimination result.

3.1 Environmental sound discrimination based on deep neural network (DNN)

In the recent researches, DNN has been paid attention as an effective method to construct the acoustic models [5]. DNN is possible to realize the discrimination with high accuracy because it nonlinearly converts the input signals. DNN has deep layer structure which consists of many nodes as shown in Fig. 2. At the network, a node receives the weighted signals which are outputted from nodes at front layer. The node then outputs the signals for nodes at next layer

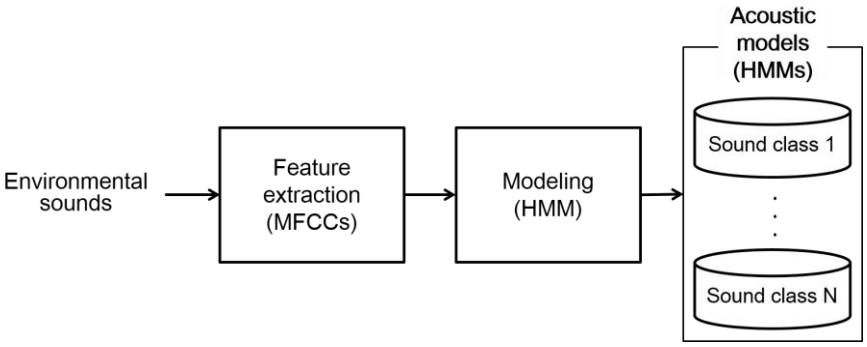


Figure 1: Overview of the conventional method

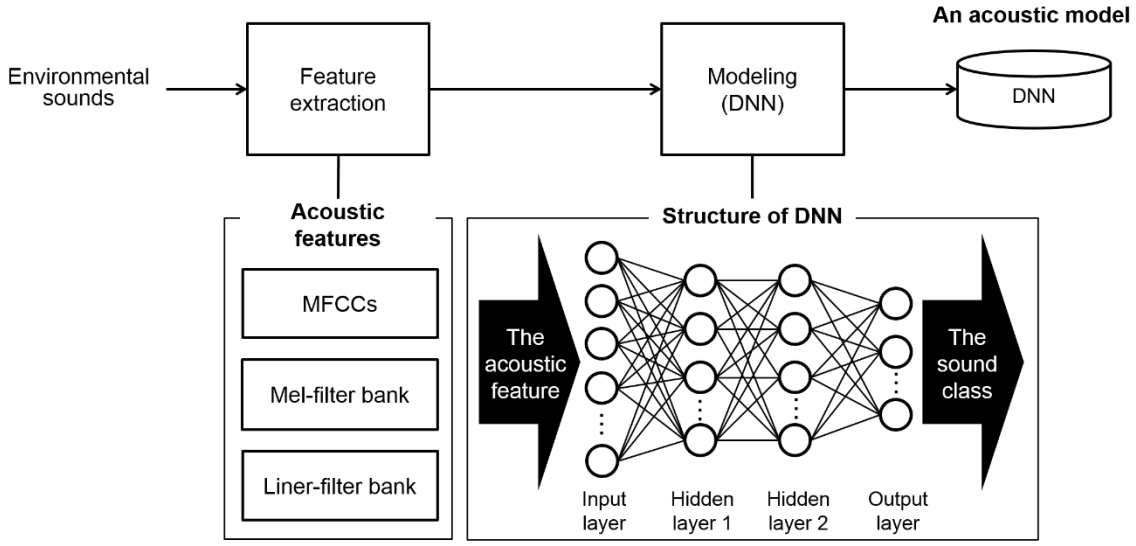


Figure 2: Overview of the proposed method

when the sum total of weighted signals exceeds the threshold. Therefore, DNN is possible to nonlinearly convert the input signals by using the nodes trained the coefficients of coupling weights. It is difficult to train the all coefficients of coupling weights at the same time because DNN is a multilayer structure. Therefore, the training process divides into pre-training and fine-training. In pre-training, hidden layers are trained from the near input layer. The auto encoder is employed as the pre-training algorithm [7]. After the pre-training, the all weights of network are tuned by fine-training. The stochastic gradient descent is employed as the fine-training algorithm [8].

3.2 Investigation on suitable acoustic features in deep neural network for environmental sound discrimination

DNN has the property that it is possible to extract the complex features with training of DNN from input signals, which are high-dimensional features. DNN is ideal candidate for constructing the acoustic models because the environmental sound has various acoustic features, for example, duration time, frequency characteristic and generation environment of the sound. Therefore, it is required to select the suitable acoustic features which are possible to accurately express the environmental sound. The suitable acoustic features contribute the reduction of the computational cost to train DNN while keeping the higher identification accuracy of the environmental sound. In order to find out the suitable acoustic features for DNN, we investigate the MFCCs, the mel-filter bank, and the linear-filter bank. Each acoustic feature is described in below.

- **Mel-frequency cepstrum coefficients (MFCCs)**

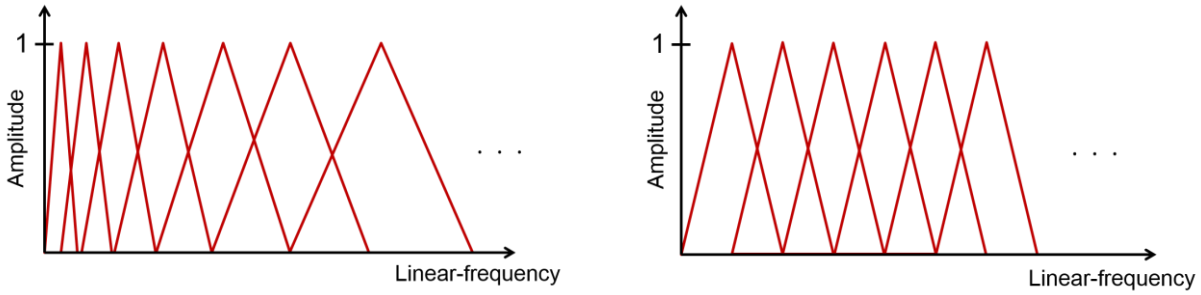
MFCCs are utilized as the acoustic feature in order to reduce dimentions from frequency information. They are calculated by the following steps. First, mel-frequency spectra are calculated by applying the mel-filter bank to power spectra at each frame and integrating the power in each filter bank. Mel-filter bank are designed on the basis of mel-frequency. Mel-frequency $Mel(f)$ is calculated from frequency f by using the mel-scale as Eq. (1).

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

Next, mel-frequency cepstra are calcurated by the discrete cosine transform (DCT) of the mel-frecuency spectrum. Finally, lower dimensions of them are extracted as MFCCs. MFCCs are able to reduce the dimensions based on auditory characteristics of human. MFCCs are especially superior to express the sound which has dominant power in lower frequency band.

- **Mel-filter bank**

Filter bank is utilized to reduce the dimensions. It consists of some bandpass filters as shown in Fig. 3. Mel-filter bank is one of the filter banks. In the mel-filter bank, bandpass filters are put at equal interbals in a mel-frequency axis. Therefore, it has a high resolution in the lower frequency. Figure 3(a) shows an image of the mel-filter bank. In this paper, the feature of mel-filter bank is defined as the feature which is calculated by applying the mel-filter bank to log-power spectra. Mel-scale is configured to approximate the auditory characteristics of human. Therefore, mel-filter bank is superior to reduce the dimension of the sound features on the basis of the auditory characteristics of human.



(a) Mel-filter bank **(b) Linear-filter bank**

Figure 3: Image of the filter bank

- **Linear-filter bank**

Linear scale filter bank is also one of the filter banks. Bandpass filters are put at equal interbals in a linear frequency axis. Therefore, bandwidths of bandpass filter are equal in a linear frequency axis. Figure 3(b) shows the linear-filter bank. The feature of linear-filter bank is defined as the feature which is calculated by applying the linear-filter bank to log-power spectra as same as mel-filter bank. In the linear-filter bank, the resolutions of frequency are equal in the all frequency. Since the environmental sound has the various features, it is conceivable that there are suitable features to discriminate the environmental sound in high frequency. Therefore, the feature of linear filter bank is suitable for the environmental sound discrimination.

4 Evaluation experiment

Evaluation experiments were conducted to compare the discrimination accuracy of the conventional and proposed methods. We compared the discrimination accuracy of multi-condition models constructed on the basis of respective methods.

4.1 Experimental conditions

We tried to construct multi-condition models by using the environmental sounds which are convoluted the impulse responses recorded at plural different indoor-environments. It is difficult to prepare the acoustic model matched the real environment. Therefore, it is sufficient to construct the multi-condition models. Table 1 shows the index of environmental sound. We utilize the environmental sounds in the database recorded by real world computing partnership (RWCP-DB) [9]. These sounds are recorded in an anechoic room. Table 2 shows the environment recorded the impulse responses. The anechoic sounds are supposed as the anechoic environment. We employ the test data which are unused samples to train the models. Table 3 shows the datasets for evaluation experiment. Discrimination accuracy in the each method is calculated with Eq. (2).

Table 1: Index of the environmental sound

Index	Environmental sound	Index	Environmental sound
1	Wood collision sounds	5	Rubbing woods sounds
2	Metal collision sounds	6	Clap sounds
3	Glass collision sounds	7	Whistle sounds
4	Jetting gas sounds	8	Electric sounds

Table 2: Environment recorded impulse responses

Environment	Reverberation time	Environment	Reverberation time
Japanese-style room	$T_{60} = 400$ [ms]	Lift station	$T_{60} = 750$ [ms]
Living room	$T_{60} = 550$ [ms]	Standard stairs	$T_{60} = 900$ [ms]
Prefabricated bath	$T_{60} = 650$ [ms]		

Table 3: Datasets for evaluation experiment

Database	RWCP-DB [9]
Training data	6720 samples (140 samples × 8 classes × 6 environments)
Open test data	2880 samples (60 samples × 8 classes × 6 environments)

$$A = 100 \times \frac{1}{N}\sum_{n=0}^{N-1} G(n), \quad G(n) = \begin{cases} 1 & (correct) \\ 0 & (incorrect) \end{cases} \tag{2}$$

where A is the identification accuracy, n is the index of an environmental sound, N is the total number of environmental sounds, and $G(n)(n = 1, \cdots, N)$ is also a function which returns 0 or 1 according to correct or not on the discrimination result.

In the conventional method, the acoustic models are constructed for 8 indoor-environmental classes and a silent class based on HMMs which have 3 states. MFCCs are utilized to extract 39 dimension features for every frame (12 dimension MFCC and 1 dimension power, deltas of them and accelerations of them). On the other hand, in the proposed method, the acoustic model is constructed based on DNN with each acoustic feature. Table 4 shows the acoustic features and the number of integrating frame for DNN. To confirm the effect of the dynamic features of MFCCs, MFCC39 include the delta and acceleration coefficients. The acoustic feature is composed by using 11 frames of MFCC39 to input the DNN because MFCC39 have delta and acceleration coefficients. On the other hand, the delta and acceleration coefficients are not utilized in MFCC13, LFBANK and MFBANK. Therefore, they are integrated by 19 frames.

Table 4: Acoustic features to construct the acoustic model with DNN

Feature name	Features per frame (Dimension)	Number of integrate frame	Dimension
MFCC39	MFCCs(12) + Power(1) + ΔMFCC(12) + ΔPower(1) + ΔΔMFCCs(12) + ΔΔPower(1)	11	429
MFCC13	MFCCs(12) + Power(1)	19	247
MFBANK	Mel-filter bank(26)	19	494
LFBANK	Linear-filter bank(26)	19	494

The network has 2 hidden layers with 500 nodes per layer and 9 nodes as output.

4.2 Experimental results

Figure 4 shows the discrimination accuracy of the conventional and proposed methods for all samples. As a result, the discrimination accuracy of DNN with MFCC39 is lower than that of HMM. In addition, the accuracy of DNN with MFCC39 has widest variance. On the other hand, the discrimination accuracy of DNN with MFCC13 is higher than that of HMM. The discrimination accuracies of DNN with the other features are also higher than that of HMM.

Figure 5 shows the itemized accuracy of the conventional method and proposed methods with MFCC39 and MFCC13. In DNN with MFCC39, some classes of discrimination accuracy is lower than that of HMM, for example, wood collision, glass collision and clap sounds. On the other hand, the discrimination accuracies of DNN with MFCC39 are comparable or higher than that of HMM. The sound of wood collision, glass collision and clap have short duration time. Therefore, the delta and acceleration coefficients are insufficient to discriminate the sound which have short duration time with DNN. The discrimination accuracy of MFCC13 also has narrower variance than that of MFCC39, as shown in Fig. 4. Therefore, the dynamic features are insufficient for DNN.

Figure 6 shows the itemized accuracy of the proposed methods with MFCC39, MFBANK, and LFBANK. The discrimination accuracy of DNN with MFCC13 fluctuates in each reverberation time of the environmental sound. On the other hand, the results of both LFBANK and MFBANK have high discrimination accuracies in each reverberation time as shown in Fig. 6. There are little difference between LFBANK and MFBANK. It is difficult to conclude the suitable acoustic feature for DNN. However, the result shows that the high-dimensional feature is sufficient to construct the acoustic model with DNN. From the above, we confirmed the effectiveness of the proposed method because the discrimination accuracy of the proposed method was higher than that of the conventional method. We also confirmed that with high-dimensional features such as MFBANK and LFBANK are effective to construct the acoustic model with DNN.

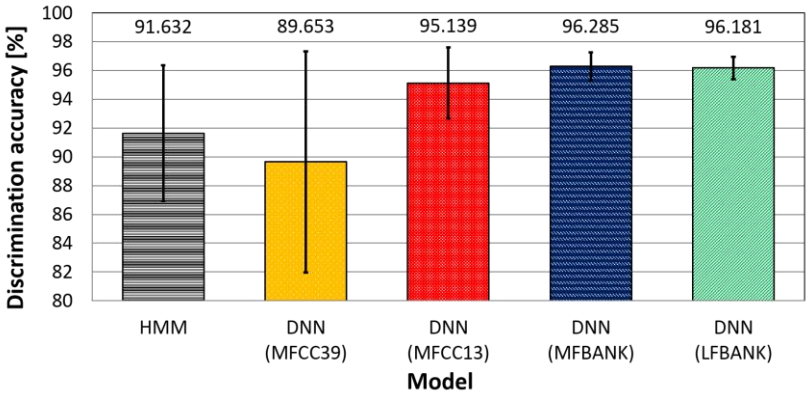


Figure 4: Discrimination accuracy of the conventional and proposed methods

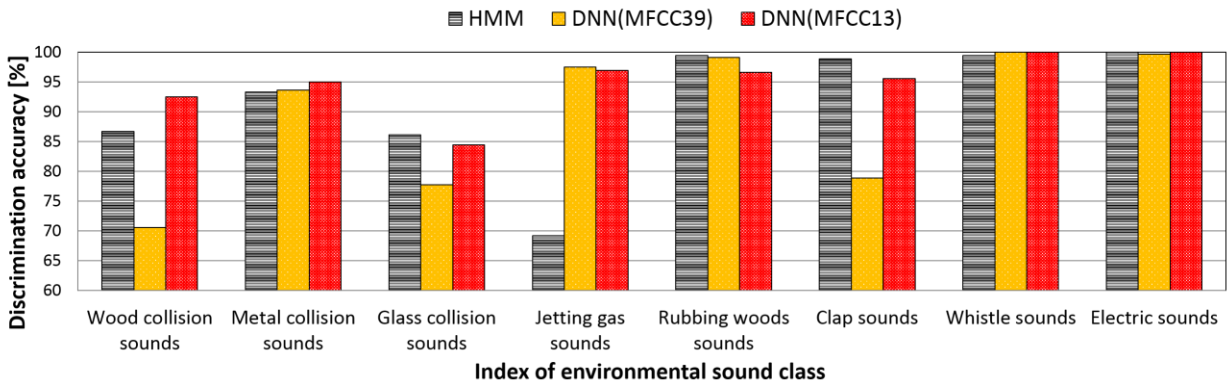


Figure 5: Discrimination accuracy of the conventional method and proposed methods with MFCC39 and MFCC13

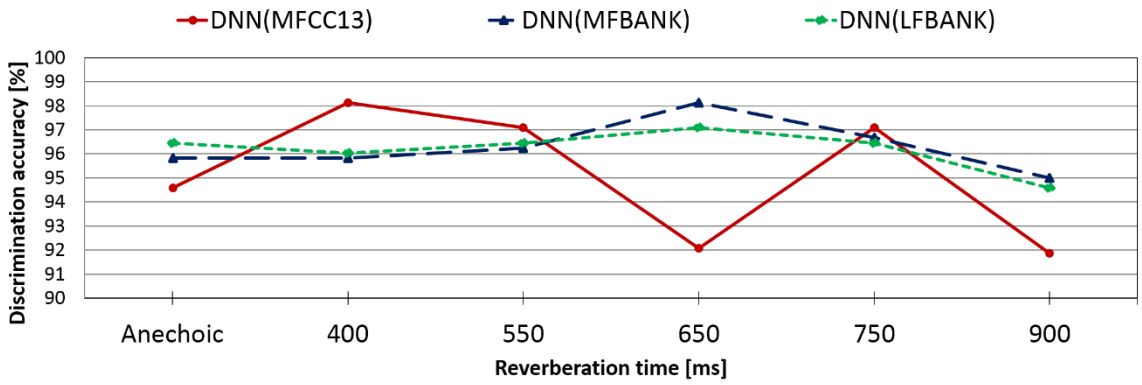


Figure 6: Discrimination accuracy of the proposed methods with MFCC13, MFBANK, and LFBANK

5 Conclusions

In this paper, we proposed the method to discriminate the environmental sound with DNN which is utilized as acoustic model. We also investigated plural acoustic features for DNN. To confirm the performance of environmental sound discrimination in each acoustic feature, we carried out the evaluation experiment. The result of evaluation experiment shows that the dynamic features are insufficient for DNN. We also confirmed that the high-dimensional features such as LFBANK and MFBANK are sufficient to construct the acoustic model with DNN. In the future works, we will utilize the other high-dimensional features for environmental discrimination, for example, the log power spectrum and the waveform, and so on.

Acknowledgments

This work was party supported by JSPS KAKENHI Grant Numbers JP26280065, JP15K16030, JP16K16094, and R-GIRO (Ritsumeikan Global Innovation Research Organization) funded by Ritsumeikan University.

References

[1] Kumar, K. S.; Prasad, S.; Saroj, P. K.; Tripathi, R. C. Multiple Cameras Using Real Time Object Tracking for Surveillance and Security System, *Emerging Trends in Engineering and Technology (ICETET)*, Goa, India, November 19-21, 2010, pp 213-218.

[2] Kawamoto, M.; Asano, F.; Kurumatani, K.; Yingbo H. A System for Detecting Unusual Sounds from Sound Environment Observed by Microphone Arrays, *Information Assurance and Security*, Xian, China, August 18-20, 2009, Vol 1, 2009, pp 729-732.

[3] Elliott, R. J.; Aggoun, L.; Moore, J. B. *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York (USA), 1st edition, 1995.

[4] Logan, B. Mel Frequency Cepstral Coefficients for Music Modeling, *International Symposium/Conference on Music Information Retrieval (ISMIR)*, Massachusetts, USA, October 23-25, 2000, In CD-ROM.

[5] Hinton, G.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Processing Magazine*, Vol 29 (6), 2012, pp 82-97.

[6] Elo, J. P.; et al. Non-speech audio event detection. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,Taipei, China, April 19-24, 2009, pp 1973-1976.

[7] Hinton, G. E.; Simon O.; Teh, Y. A Fast Learning Algorithm for Deep Belief Nets, *Neural computation*, Vol 18 (7), 2006, pp 1527-1554.

[8] Bottou, L. Large-scale Machine Learning with Stochastic Gradient Descent, *International Conference on Computational Statistics (COMPSTAT)*, Paris, France, August 22-27, 2010, pp 177-186, 2010.

[9] Nakamura, S.; Hiyane, K.; Asano, F.; Nishiura, T.; Yamada, T.; Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition. *International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, May 31 - June 2, 2000, In CD-ROM.