ICIC Express Letters
Part B: Applications
Volume **7**, Number **10**, October **2016**

 ISSN 2185-2766
pp. **2261–2266**

ACOUSTIC SCENE DETECTION BASED ON DEEP BELIEF NETWORK AND DYNAMIC BINARY FEATURES

Jian Wang[1,3] and Shu Xu[2]

[1]School of Equipment Engineering
Shengyang Ligong University
No. 6, Nanping Central Road, Hunnan New District, Shenyang 110159, P. R. China

[2]Xinyang Vocational and Technical College
Number 24 Avenue, Yangshan New District, Xinyang 464000, P. R. China

[3]School of Mechano-Electronics Engineering
Beijing Institute of Technology
No. 5, South Zhongguancun Street, Haidian District, Beijing 100081, P. R. China
wangjiansylu@sina.com

Received April 2016; accepted July 2016

Abstract. *In this paper we study the automatic acoustic scene classification problem using deep belief network. Target detection from acoustic signal is a difficult problem due to the noise and interference. First, we propose to use a novel spectrum based feature to analyze the audio scene. Time-frequency unit is separated from the spectrum and dynamic threshold features are extracted. Second, deep neural network structure is adopted to train the acoustic scene classifier. Finally, we verify the proposed framework on several acoustic scenes. Experimental results show that the proposed method is effective and it is suitable for practical environment with complex interference.*
Keywords: Acoustic feature, Time-frequency unit, Deep neural network, Acoustic scene analysis

1. **Introduction.** Acoustic features are key to analyse the content of audio data [1, 2]. Mel-frequency cepstral coefficient (MFCC) can be used to classify vocal signals, such as speech, cry, and laughter [3, 4]. Spectral features can be used to analyse various sounds other than speech, such as music, traffic sound, and explosion. Rakotomamonjy and Gasso [5] used histogram of gradients to construct the classification features. Their results showed that the features were effective on modelling structures in time-frequency domain. However, only linear support vector machines were combined with these features and the further extension on non-linear deep learning algorithm would be interesting. Spectral temporal modulations were introduced by Chakrabarty and Elhilali [6]. In their study, the spectral features were used and Gaussian mixture model (GMM) and hidden Markov model (HMM) were adopted to verify the proposed audio features. However, the auditory perception characters were not discussed and these features could be important for vocal sound classification.

Traditional machine learning algorithm can be applied to audio classification problem [7, 8]. Dhanalakshmi et al. [9] proposed to use neural network with radius-based function to classify scenes. The neural network only had few layers and the complicated structures in audio scene data could be better modelled with deeper layers. The recent development in representation learning has shown a promising future in neural networks that have deep structures. Cornu and Milner [10] used convolutional neural network (CNN) to model the audio scenes. The results showed that the deep neural network had outperformed the traditional algorithms. However, in their study the scene classification was limited to vocal sounds and more types of natural scenes should be considered.

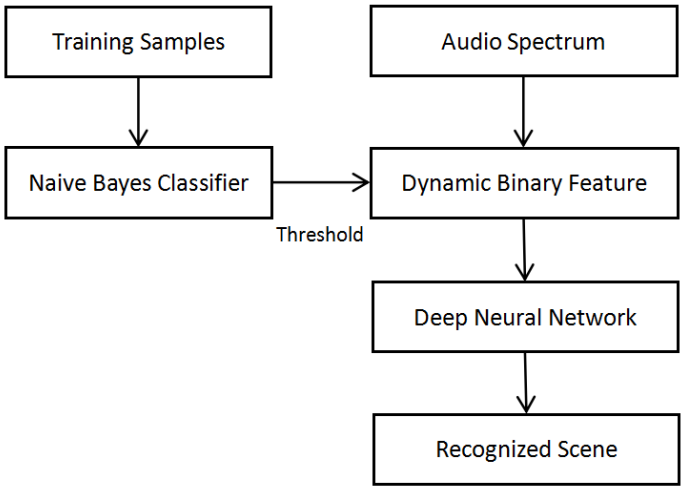


FIGURE 1. Overall flow chart of the proposed audio analysis system

In this paper, we propose to use a supervised feature extraction method to classify various audio scenes that are modelled by deep neural network (DNN). As shown in Figure 1, the training samples are modelled by Naive Bayes Classifier to classify the dynamic features. The audio spectrum is processed with binary features for feature representation. The neural network model and scene analysis are then adopted. In the feature analysis, the time-frequency unit is used for auditory perception analysis. Each time-frequency unit is preprocessed by simple threshold based function and binary features are constructed for efficient modelling. Deep neural work structure may improve the feature classification result. Compared with shallow neural network structure, it can deal with more complex data.

The rest of the paper is organized as follows: Section 2 proposed a novel feature construction method; Section 3 describes the neural network used in our system; Section 4 provides the experimental results, and finally, conclusions are given in Section 5.

2. Dynamic Binary Features for Time-Frequency Unit.

In this section we propose a novel feature construction method based on the spectrum of the sound, namely the dynamic binary features (DBF). The philosophy behind this method is the common belief that different sound sources can be separated in the time-frequency (TF) domain. Using a proper threshold we can classify each time-frequency unit into corresponding sound types.

The time-frequency unit is represented as:

$$U_{t,f} = FT(x(t_{n+1} - t_n)) \tag{1}$$

where t and f are the time and frequency, and $FT()$ stands for the Fourier Transform. n is the frame index and x is the time domain signal.

Since the human auditory system is not sensitive to phase changes, taking the modulus form we can get the amplitude $|U_{t,f}|$ for each TF unit.

A fast binary feature construction method is designed for each TF unit. The binary feature is denoted as B. We have: $B = 0$, when $|U_{t,f}| \leq Th_{t,f}$; $B = 1$, when $|U_{t,f}| > Th_{t,f}$.

The dynamic threshold is dependent on the auditory perception character. When the frequency is very high or very low, the threshold is higher, as shown in Figure 2. When the neighbouring TF units are similar to each other, the threshold is also higher. Therefore, Th can be represented as:

$$Th_{t_n,f_m} = \Psi(f_m, |U_{t_{n-\tau},f_m}|) \tag{2}$$

where m is the discrete frequency index. τ is the order of the dependency, and it usually takes the value from 1 to 3. The hidden Markov assumption is often adopted in speech signal modelling. In this case, when Markov relation is assumed, the threshold is calculated according to the neighbouring frames.

The optimal solution of Ψ can be measured in auditory perception experiment. However, in our work, we take a different approach that uses supervised learning algorithm to learn the dynamic threshold Th. First, we collect data cohorts from various audio scenes. Second, we train the threshold model using Naive Bayes framework and we have:

$$P(Th_{f_m}) = \sum_{t_n} \log(P(Th_{t_n,f_m}|B_{t_n,f_m}, U_{t_n,f_m}, \ldots, U_{t_n-\tau,f_m})) \tag{3}$$

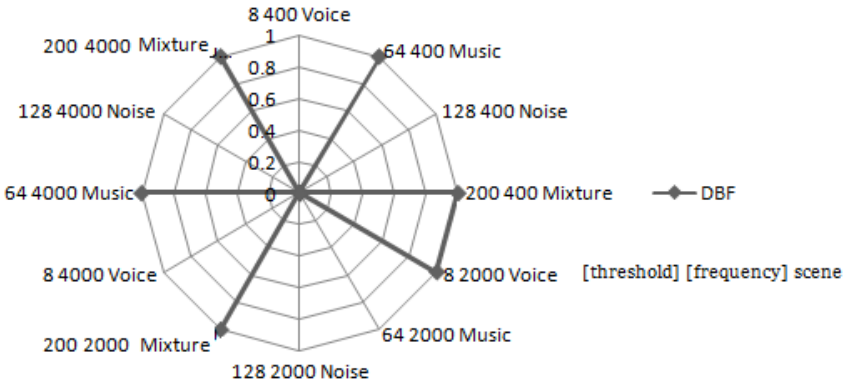


FIGURE 2. A depiction of the proposed binary feature for different frequencies and scenes

3. Supervised Classification Methods.

3.1. Gaussian mixture model.

Gaussian mixture model is a widely used algorithm that has strong data modelling ability. Theoretically, it can model any probability distribution. The setting of Gaussian mixture number is important, and it can be empirically set in different applications.

The definition of Gaussian mixture model is:

$$p\left(\mathbf{X}_t|\lambda\right) = \sum_{i=1}^{M} a_i b_i\left(\mathbf{X}_t\right) \tag{4}$$

where $\mathbf{X}_t$ is a D-dimension random vector, $b_i\left(\mathbf{X}_t\right)$ is the ith member of Gaussian distribution, t is the index of utterance, a_i is the mixture weight, and M is the number of Gaussian mixture members. Each member is a D-dimension variable following the Gaussian distribution with the mean $\mathbf{U}_i$ and the covariance $\mathbf{\Sigma}_i$:

$$\begin{aligned} b_i\left(\mathbf{X}_t\right) &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} \\ &\exp\left\{-\frac{1}{2}\left(\mathbf{X}_t - \mathbf{U}_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\mathbf{X}_t - \mathbf{U}_i\right)\right\} \end{aligned} \tag{5}$$

Note that

$$\sum_{i=1}^{M} a_i = 1 \tag{6}$$

Expectation-maximization (EM) algorithm is then used for the estimation of GMM parameters.

3.2. **Supervised classification using deep structures.** Traditional neural network uses shallow structure to learn from limited data. Deep learning uses deep network structure to learn from big data. It has outperformed many of the previous state-of-the-art machine learning algorithms in computer vision and speech recognition [13]. A desired property of DNN is that it can represent the high level feature in a supervised way. It is less likely to be influenced by the variance in the raw data [12].

DNN is rooted from the traditional multilayer neural network. However, it contains many hidden layers. DNN belongs to the family of supervised learning and it models the posterior probability $p_{y|\boldsymbol{x}}(y = s|\boldsymbol{x})$, where s stands for a class, $\boldsymbol{x}$ denotes an observed vector and y is the output of the neural network.

We can write the first L layers as vectors $\boldsymbol{v}^l$, where l is the index of nodes. The hidden binary vectors are represented as $\boldsymbol{h}^l$. Let $\boldsymbol{h}_m^l$ be the hidden units, and m is the index of the unit. The total number of hidden units is N^l. The posterior probability can be written as [13]:

$$p^l\left(\boldsymbol{h}^l|v^l\right) = \prod_{m=1}^{N^l} \frac{e^{z_m^l\left(\boldsymbol{v}^l\right)\boldsymbol{h}_m^l}}{e^{z_m^l\left(\boldsymbol{v}^l\right)} + 1} \tag{7}$$

where $z^l\left(\boldsymbol{v}^l\right) = \left(\boldsymbol{W}^l\right)^T \boldsymbol{v}^l + \boldsymbol{a}^l$. $\boldsymbol{W}$ is the weight and $\boldsymbol{a}$ is the bias vector. The output layer computes the posterior probabilities as:

$$p_{y|\boldsymbol{x}}(y = s|\boldsymbol{x}) = \frac{e^{z_s^L\left(\boldsymbol{v}^L\right)}}{\sum_{s'} e^{z_{s'}^L\left(\boldsymbol{v}^L\right)}} \tag{8}$$

4. **Experimental Result.** The dataset used in this experiment is collected locally in our lab. Sound signal is recorded using profession sound card and Adobe Audition software. The sampling rate is converted to 16K and stored in PCM coded WAV files.

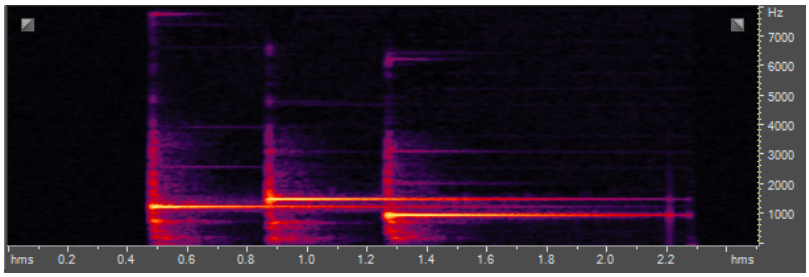
The examples of different audio scenes are demonstrated in Figure 3. We can see that the music spectrum has a periodic property. The voice signal is unstationary and formant frequencies are visible. The noise spectrum seems random and no formants are clearly seen. The mixture of those sound sources is more difficult to recognize. The proposed DBF analysis may provide useful insights on the classification of sound spectrum images.

In the experiment we compare two types of features, denoted as MFCC and DBF. We also compare two types of classification algorithms, denoted as GMM and DNN. Recognition performances are provided in Table 1. We can see that using the traditional MFCC and GMM methods, the recognition rates are lower. Using the proposed feature extraction algorithm DBF and deep neural network, the recognition rates are constantly improved over all five audio scene types.

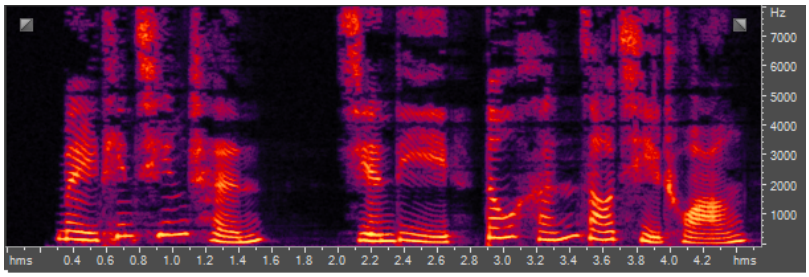
The parameter setting is further explored. The Gaussian mixture model is sensitive to mixture numbers and we set the mixture numbers to 12 for the best results. The iteration number is also important in EM algorithm and we set the iteration number to 40 for the best results. Nevertheless, the deep neural network structure still outperforms the traditional GMM mehtod.

TABLE 1. Classification performance using the proposed algorithm and traditional algorithms

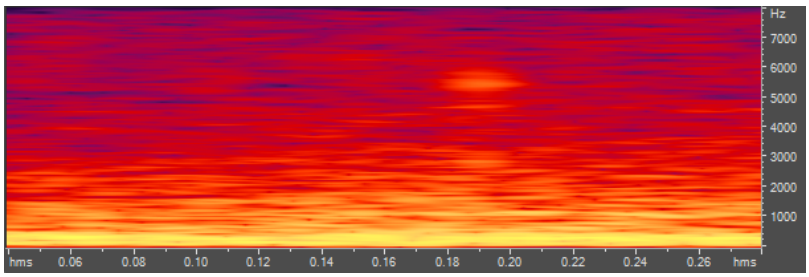
Audio scene types	MFCC + GMM	DBF + GMM	DBF + DNN
Music	88.1%	89.1%	90.1%
Voice	84.3%	87.1%	88.9%
Noise	77.8%	80.0%	83.3%
Music & Voice	75.5%	78.7%	81.5%
Noise & Voice	72.3%	77.8%	80.2%



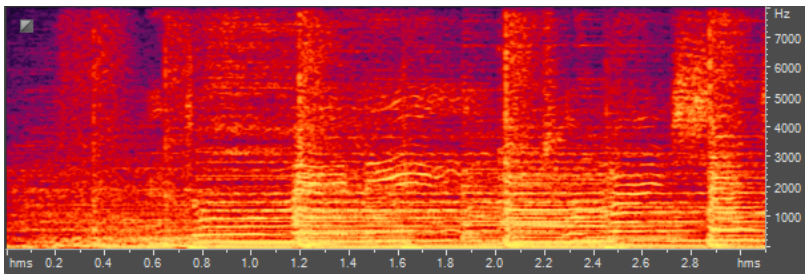
(a)



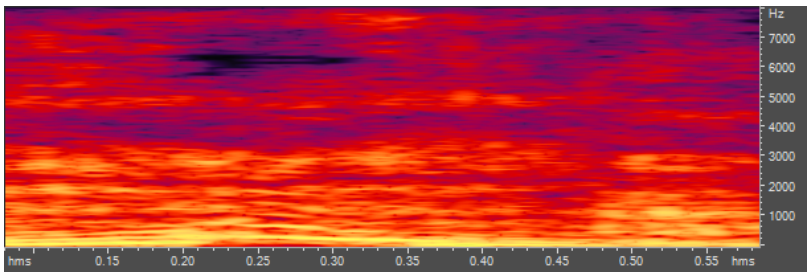
(b)



(c)



(d)



(e)

FIGURE 3. Spectrum of example audio scenes: (a) music, (b) voice, (c) noise, (d) mixture (music and voice) and (e) mixture (noise and voice)

5. **Conclusions.** In this paper we study the automatic scene classification from sound spectrum. Dynamic threshold is learned from training samples in a supervised way. The proposed feature extraction method is efficient and it is based on the human auditory perception character. Each time-frequency unit is converted into binary features and

parametric modelling algorithm is then investigated. The deep neural structure enables good feature representation and the recognition rate is constantly improved compared with traditional methods. In future work, we will study how to combine the audio scene classifier with speech recognition engine and improve the speech recognition performance under complex acoustic environment.

REFERENCES

[1] C. Wu, C. Huang and H. Chen, Automatic recognition of emotions and actions in bi-modal video analysis, *Internet of Vehicles-Safe and Intelligent Mobility*, Springer International Publishing, pp.427-438, 2015.

[2] Y. Q. Bao, L. Zhao and C. Huang, Speech emotion recognition using multiple discriminant analysis and Gaussian mixture model, *Applied Mechanics and Materials*, vol.380, pp.3530-3533, 2013.

[3] S. Chu, S. Narayanan and C. C. J. Kuo, Environmental sound recognition using MP-based features, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1-4, 2008.

[4] J. T. Geiger, B. Schuller and G. Rigoll, Large-scale audio feature extraction and SVM for acoustic scene classification, *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.1-4, 2013.

[5] A. Rakotomamonjy and G. Gasso, Histogram of gradients of time-frequency representations for audio scene classification, *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol.23, no.1, pp.142-153, 2015.

[6] D. Chakrabarty and M. Elhilali, Exploring the role of temporal dynamics in acoustic scene classification, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.1-5, 2015.

[7] D. Battaglino, A. Mesaros, L. Lepauloux, L. Pilati and N. Evans, Acoustic context recognition for mobile devices using a reduced complexity SVM, *IEEE European Signal Processing Conference*, pp.534-538, 2015.

[8] T. Lu, G. Wang and F. Su, Context-based environmental audio event recognition for scene understanding, *Multimedia Systems*, vol.21, no.5, pp.507-524, 2015.

[9] P. Dhanalakshmi, S. Palanivel and V. Ramalingam, Classification of audio signals using SVM and RBFNN, *Expert Systems with Applications*, vol.36, no.3, pp.6069-6075, 2009.

[10] T. Le Cornu and B. Milner, Voicing classification of visual speech using convolutional neural networks, *The Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, pp.10-14, 2015.

[11] C. Huang, G. Chen, H. Yu, Y. Bao and L. Zhao, Speech emotion recognition under white noise, *Archives of Acoustics*, vol.38, no.4, pp.457-463, 2013.

[12] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature*, vol.521, no.7553, pp.436-444, 2015.

[13] K. Han, D. Yu and I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, *Proc. of Annual Conference of the International Speech Communication Association*, Singapore, 2014.