ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 142

2017 International Conference on Information Technology and Intelligent Manufacturing (ITIM 2017)

Acoustic Scene Classification on Large Dataset Using Sparse Auto-encoder Based Deep Neural Network

Jianqiang Tan
Business School
Shandong Yingcai University
Jinan, Jinan250102, China

***Abstract*—In this paper we study the acoustic scene classification using a large dataset. The spectrogram of the large acoustic samples are extracted and applied with texture feature classification method. First, the acoustic scene database is built including various acoustic events. Second the image texture features on spectrogram are used to represent the acoustic samples. Third the auto-encoder is adopted to build a deep neural network classifier. Finally, we verified the proposed system on a large number of dataset and compared our results with traditional Gaussian mixture model and three-layer neural network. The experimental results show that the proposed method is effective and promising in big acoustic data classification.**

***Keywords*—acoustic scene classification; auto-encoder; deep neural network; big data**

I. Introduction

Acoustic events classification is an important topic in machine learning and data analysis[1-3]. Most of the traditional classifiers are built on a relatively small number of samples. The feature analysis is also performed on some typical acoustic events. The generalization ability of such systems are usually not well [4-5].

Acoustic features such as pitch frequency, formant frequencies, formant bandwidth, zero-cross-rate, and other spectral features can be used to analyze acoustic signal. The human voice can be detected from noise or music using pitch feature. The different environment characters can be classified by spectral features. The model used to analyze the acoustic scene is rooted from pattern recognition field[6-7]. Support vector machine is a popular classification algorithm that achieves good results on a small sample set. Neural network is also an effective learning and modeling algorithm that has many applications in machine learning and pattern classification.

In this paper, we study a novel method of construct spectrogram features using deep neural network and modeling with auto-encoder. The rest of the paper is organized as follows: Section 2 briefly introduce the database used in this paper; Section 3 provides a novel method to construct spectrogram features; Section 4 gives the details on the classification algorithm; Section 5 is the experimental results and finally conclusion is provided.

II. The database

The acoustic data is collected locally in our lab and seven categories are included: cry, music, speaking, wide band noise, other noise, explosion, and traffic. We collected real world noise from various sources and put them in the category of other noises. These scenes include supermarket, jet noise, construction noise etc. The traffic scene is a special type of category the traffic noise is treated as a scene on the road side.

The data is recorded using one microphone, 44.1k sampling rate, 16bit, single channel. The data files are saved in WAV format. All samples are labeled by human annotators and verified in a listening test by other human annotators. The basic data samples can be combined to generate different durations. They can also be split to increase the size of training and testing sample set. Examples of the data are shown in Fig. 1 and Fig. 2.

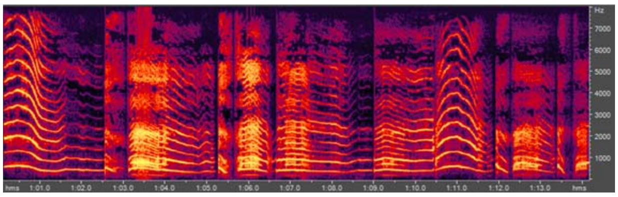

Fig. 1. Spectrogram of cry sound sample.

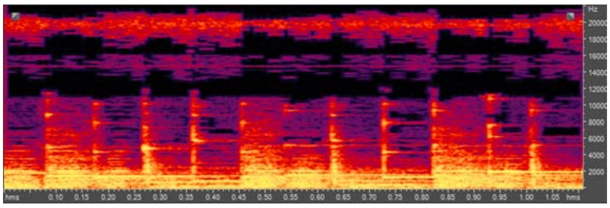

Fig. 2. Spectrogram of music scene sample.

TABLE I. ACOUSTIC SCENE DATABASE

Scene	Total Duration	Sample Size	Averaged Duration	SNR
Cry	**6120s**	**1200**	**5.1s**	**10-15dB**
Music	**9360s**	**1300**	**7.2s**	**15-30dB**
Speaking	**18900s**	**4200**	**4.5s**	**10-15dB**
Wide Band Noise	**6500s**	**1300**	**5s**	**-**
Other Noise	**7500s**	**1500**	**5s**	**-**
Explosion	**3800s**	**2000**	**1.9s**	**10-15dB**
Traffic	**11730s**	**2300**	**5.1s**	**10-15dB**

The Statistics on the big data acoustic scene database are shown in Table 1. The noise is mixed with the clean data in order to simulate the real world environment. The SNR is shown in the last column. Various SNR levels are provided in order to verify the robustness against noise in the later experiments.

III. SPECTROGRAM FEATURE EXTRACTION

Spectrogram can be combined with auto-encoding to achieve better features that used for classification.

Auto-encoder neural network is an unsupervised learning algorithm which allows the output value of the sample to be equal to the input ones. If the number of neurons in the hidden layer of neural network is much smaller than the input and output layers, the auto-encoder neural network is forced to learn the compressed representation of the input data. It is very difficult to learn the compact representation of these random data. The correlation between these associated data can be found by the auto-encoding algorithm, and the output layer reconstruction will output the input data if there are some specific structures implicated in the input data, such as some input features related to each other.

But, in contrast, the sparseness of neurons in the hidden layer can be added if the number of neurons in the hidden layer is larger or equivalent to that in the input-output layer. The correlation between the input data can be still found out by the auto-encoding neural network in this way. The correlation between the gray value of an image can be found if sparse auto-encoding neural network is used to process a spectral image.

Assuming that the neuron activation function[8] is:

$$\mathrm{y} = f(z) \tag{1}$$

and the activation value (output value) of the first neuron in the first layer of the neural network is denoted as[8]:

$$\mathrm{a}_i^l = f(z_i^l) \tag{2}$$

The sigmoid function is generally used as an activation function. The range of the sigmoid function is[8]:

$$f(z) = \frac{1}{1+e^{-z}} \tag{3}$$

In the auto-encoding neural network, the BP algorithm is used to make the unlabeled input sample equal to the output target, that is to say:

$$y^i = x^i \tag{4}$$

Meanwhile, the unlabeled sample set is expressed as:

$$\psi = \{\mathrm{x}^1, \mathrm{x}^2, ..., \mathrm{x}^m\} \tag{5}$$

and m is the number of unlabeled samples.

The output value of the j-th neuron in the hidden layer in the case of the i-th sample input is:

$$x^* = a_j^2(x^i) \tag{6}$$

The mean activation value of the j-th neuron in the hidden layer (l = 2) is expressed as:

$$\rho_{\mathrm{j}} = \frac{1}{m}\sum_{i=1}^{m} a_j^2(x^i) \tag{7}$$

IV. SPARSE AUTO-ENCODER AND DEEP NEURAL NETWORK

Its input layer and hidden layer neurons are fully connected with each other in sparse auto-encoding neural network, which is suitable for low-resolution spectrogram images. However, the network training time will be greatly extended if we want to extract 100 features from the higher resolution image, because the sparse auto-encoding neural network has 106 parameters to be learned.

For a high-resolution spectrogram, the statistical features in the local region of the image are similar to those in other parts. Therefore, there is no need to make the input layer and the hidden layer neurons fully connected. In order to reduce the parameter number of high-resolution image Neural, the convolution operation is applied to achieve the local network connection. The basic idea of local connection network is based on the fact that human visual cortical neurons only respond to stimuli in some local regions. Local connection neural networks only allow implicit layer neurons to connect with some input layer neurons. Specifically, features can be learned from a large image via a small block, which can be trained by a sparse Auto-encoder neural network, and then the features of the training image are extracted by using the convolution algorithm in turn. Finally, the convolved feature image is obtained by convolving the labeled training set and test set image with the learned feature parameters.

Structure of the classification system is shown in Fig.3. Soft-max is used for classification of the acoustic scene. The basic principle of soft-max is explained as follows.

Corresponding to logistic regression, softmax regression is an extension of logistic regression used for solving the problem of multi-classification. Suppose that the label sample set is:

$$\phi=\{(x^1,x^1),(x^2,x^2),...(x^i,x^i),...,(x^m,x^m),1<i<m\} \quad (8)$$

where the class labels of softmax regression belong to [0,1] ,the class labels of logistic regression are , where k refers to the number of classification categories. We set k = 3 in the experiment, that is, forest fire, red and red leaves. The sample set used to train the softmax regression model is the pooled feature of the labeled training set. After adding the regularized weighting attenuation term to the softmax regression model, the new cost function is a convex function, and there exists a unique minimum value. Therefore, some iterative algorithms, such as batch gradient descent, Newton method, LBFGS, are used to obtain the global optimal solution for solving the minimum cost function of the sparse auto-encoder feature learning algorithm. Finally, the accuracy of the softmax image classifier is measured. For the label test set, we first extract the convolution feature and obtain the pooled feature for getting the classification label via the trained softmax multi-classifier. If the softmax classifier outputs the same label as the test sample, it indicates that the classification result is correct, otherwise, the classification result is wrong. The total number of samples is divided by the total number of samples to obtain the correct rate for the classification of tagged test sets.

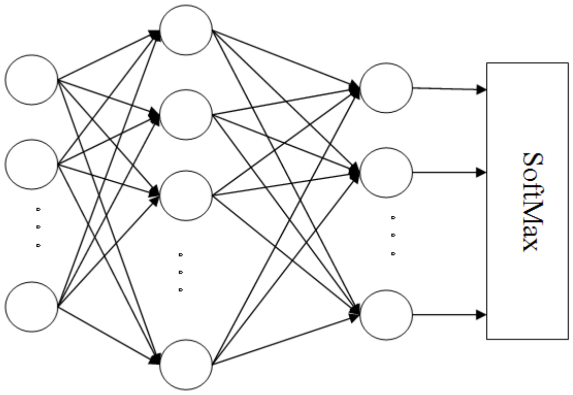

Fig. 3. Structure of auto-encoder neural network and the softmax classifier.

V. Experimental Results

In the experiment, seven classes are included for the training and recognition. They are: cry, music, speaking, wide band noise, other noise, explosion, and traffic. Their averaged recognition results are shown in Fig.4 using cross validation.

The y-axis is the recognition rate, and the x-axis is the training and testing size ratio. We can see that the music has the lowest rate and the explosion has the highest rate.

In this experiment, the feature extraction adopts a small image patch with the size of 9x9. The pooling size is 20 and the neural network hidden nodes are 200 and 400 respectively for two layers. The confusion matrix is shown in Table 2 when the training and testing ration is set to 10:1.

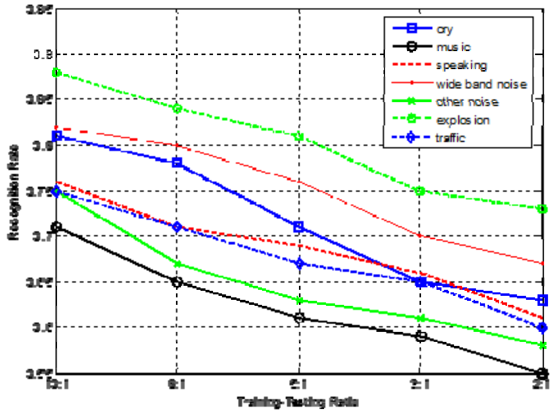

Fig. 4. Averaged Recognition Rates under Various Training-Testing Ratios.

TABLE II. Confusion Matrix of Spectrogram Recognition

	Cry	Music	Speaking	WB noise	Other noise	Explosion	Traffic
Cry	81%	5%	2%	5%	7%	0%	0%
Music	0%	71%	11%	0%	9%	0%	9%
Speaking	10%	4%	76%	0%	10%	0%	0%
WB noise	0%	6%	4%	82%	2%	3%	3%
Other noise	0%	5%	5%	2%	75%	3%	10%
Explosion	0%	0%	0%	2%	4%	88%	6%
Traffic	4%	2%	4%	5%	2%	8%	75%

A comparison with Gaussian Mixture Model (GMM) and Traditional 3-layer BP network is shown in Fig.5. GMM is initialized with K-means clustering and the number of mixtures is set to 64 for the best performance. We can see that the proposed algorithm outperforms the GMM and BP. It is more suitable for acoustic scene classification. The traditional features are confused with noisy environments and the spectrogram feature is more robust.

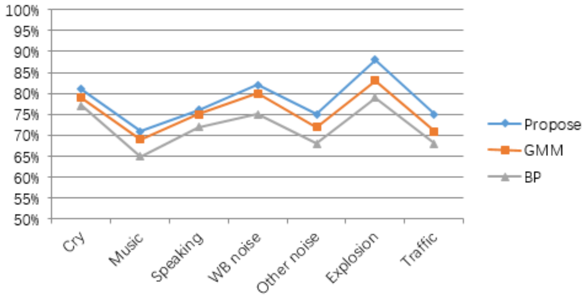

Fig. 5. Averaged recognition rate using the proposed algorithm, GMM and BP.

VI. CONCLUSIONS

In this paper a large dataset of acoustic events are introduced and deep neural network based classifier is adopted to analyze the acoustic events. Sparse auto-encoder is used to achieve a state-of-the-art performance against various traditional methods. We use a novel feature based on spectrogram analysis and achieved promising results. In the future work, the acoustic features will be further investigated and signal enhancement methods will be adopted.

References

[1] Mesaros A, Heittola T, Virtanen T. "TUT database for acoustic scene classification and sound event detection", Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016, pp.1128-1132.

[2] Giannoulis D, Stowell D, Benetos E, et al. "A database and challenge for acoustic scene classification and event detection", 21st European Signal Processing Conference (EUSIPCO 2013). IEEE, 2013, pp.1-5.

[3] Giannoulis D, Benetos E, Stowell D, et al. "Detection and classification of acoustic scenes and events: an IEEE AASP challenge", 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013, pp.1-4.

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

[5] Larsen E, Schmitz C D, Lansing C R, et al. "Acoustic scene analysis using estimated impulse responses, Signals, Systems and Computers", 2004. Conference Record of the Thirty-Seventh Asilomar Conference on. IEEE, 2003, 1, pp.725-729.

[6] Teutsch H. "Wavefield decomposition using microphone arrays and its application to acoustic scene analysis", PhD thesis, University of Erlangen-Nuremberg, 2005.

[7] Barchiesi D, Giannoulis D, Stowell D, et al. "Acoustic scene classification: Classifying environments from the sounds they produce", IEEE Signal Processing Magazine, 2015, 32(3), pp.16-34.

[8] Deng L, Seltzer M L, Yu D, et al. "Binary coding of speech spectrograms using a deepauto-encoder", Interspeech. 2010, pp.1692-1695.