MFCC combined with sparse coding for sound event classification under different noise environments

Jia-Min Mao†, Yun-Peng Wu, Li-Yang Liu and Wei-Feng Li
Department of Electrical Engineering/Graduate School at Shenzhen, Tsinghua University, China
Email: maojiamin123@163.com

In recent years, the most popular method for sound event classification can be classified into two types: 1) Extract MFCC or PLP, then train classifier for classification; 2) Convert sound into spectrogram, then use the method of image classification. However, the two methods have not achieved satisfied performance. In order to promote the classification performance, we present classification method for a sound event based on MFCC and sparse coding which has a good effect on capturing the high-level representation features of the input data. Then the coefficients of sparse coding will be employed as new sound event features to train the classification model. Our experimental results demonstrate the great robustness, adaptability and an obvious improvement on sound event classification.

Keywords: Sound event classification; MFCC; Sparse coding.

1. Introduction

As we all know, with the development of artificial intelligence, the sound event classification will have a wide use in many fields, such as environment detection [1-2], music genre classification [3], security surveillance [4], health care and so on. Generally, the method for sound event classification can be classified into two types: 1) Extract MFCC (Mel-Frequency Cepstrum Coefficient, MFCC), PLP (Perceptual Linear Predictive, PLP), then train classifier, such as GMM (Gaussian Mixture Model, GMM), HMM (Hidden Markov Model, HMM) and SVM (Support Vector Machine, SVM); 2) The method is firstly proposed in the literature[5], convert sound into spectrogram image, then use the method of image classification to accomplish the task, for example, the most common algorithm for pedestrian detection, HOG combined with SVM was proposed to audio scene classification [6]. However, the mentioned two methods for sound event classification have not achieved a satisfied performance especially under the noise environment.

As we all know, sparse coding is proposed by Olshausen [7], who attempts to find a high-level representation of the signal like the representation of visual cortex of animals. So for the sparse coding, we define a dictionary called "basis functions", then the signal can be described represented by the linear combination of the dictionary while the coefficient vector is sparse.

It is well known that sparse coding has wide applications in many fields in recent years, especially in the image processing, such as image classification, face recognition [8-9] and image noise reduction, in that it has great results on reducing the interference of noise. However, compared with image processing, audio processing has payed less attention on sparse coding, which has been ever applied on speech recognition [10], speaker identification [11], speech enhancement [12] and so on. Furthermore, in [13], it proposed a novel algorithm for computing SISC (shift-invariant sparse coding) aimed to implement audio classification. In the application of music genre classification，[14] used it has made improvement for auditory temporal modulations. Through the above examples, it can be summarized as sparse coding can generate a few non-zero coefficients to obtain a high-level representation of a sample, therefore we can employ the sparse coefficients as the feature in sound events classification.

In the paper, we propose to apply sparse coding to obtain a high-level representation of the sound event data, which will be used as the feature to train the classifier to accomplish our sound event classification. More especially, in this paper we deal with the sound classification task under different noise environments. We enlarge the training data in order to obtain better representation of sound events. Our experimental results show the effectiveness of our method.

This paper is organized as follows: Section 2 presents our proposed method in details. Section 3 shows our experiments on RWCP Sound Database, and Section 4 concludes the paper.

2. Algorithm

In the following, we will firstly introduce our overall framework. Then the most important step, sparse coding, will be emphatically described in detail.

2.1. Overview of framework

As we all know, the basic flow diagram for a sound event classification is shown in the Fig.1.

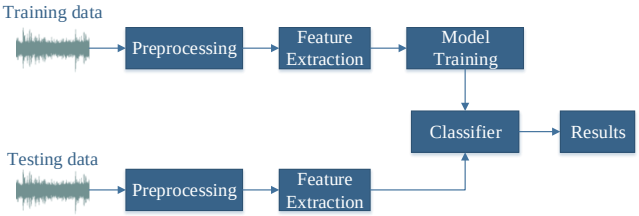

Fig.1. Basic flow diagram of sound event classification

In order to enhance the classification results, we try to extract more effective feature in our method. We use the coefficients of sparse coding as the feature instead of MFCC.

The whole flow diagram is shown in the Fig.2.

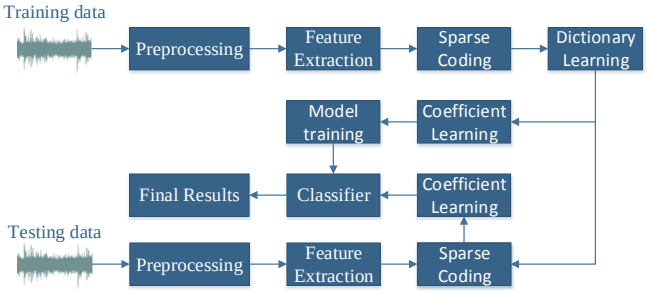

Fig.2.Flow diagram of our method

Next, we will make a brief explanation for each step in the Fig.2.

1). Preprocessing: generally, the first step of an experiment of sound event is removing the silence segments while the remaining sound event segments will be used as the effective sound clip for feature extraction.

2). Feature extraction: merging several frames into one clip firstly. Hence, it demands that the too short sound should be discarded. Then extracting features of a sound clip, while 39-dimension MFCC features will be use in this paper.

3). Dictionary learning: the features extracted from 2) will be used as the training samples of the sparse coding to learn the dictionary.

4). Coefficient learning: after learning dictionary D, we can get the sparse coefficients easily. The details will be explained in the 2.2.

5). Classifier training: we will use the above coefficients as the new features to train the classifier. In this paper, we will employ the GMM classifier.

2.2. Sparse coding

In this section, we will present the details of our core algorithm, sparse coding, which including two steps, dictionary learning and coefficient learning.

Given a sample $x \in R^{m\times 1}$, and a dictionary $D \in R^{m\times n}$ obtained by training, the signal x can be represented by a linear combination of columns of dictionary D as follows:

$$x = D \cdot s \tag{1}$$

In which $s \in R^{n\times 1}$ is called the sparse representation of x, and can be estimated by the means as follows:

$$\min_{(D,s)} \|x - D\cdot s\|_2^2 \quad s.t.\ \varphi(s) < \delta$$

$$s.t.\ \sum_j B_{i,j}^2 \le c \quad \forall j = 1\ldots n \tag{2}$$

where D is the dictionary composed of column vectors mentioned above and can be denoted as $D = [d_1 d_2 \cdots d_n]$, where d_j is the column vector of the j^{th} atom, and s is the coefficient vector. The problem (2) is equivalent to an optimization problem with constrain as follows:

$$\min_j \|x - D\cdot s\|_2^2 + \lambda \sum_j \varphi(s_j)$$

$$s.t.\ \sum_j B_{i,j}^2 \le c \quad \forall j = 1\ldots n \tag{3}$$

In which $\|x - D\cdot s\|_2^2$ is reconstruction error, while $\lambda\sum\varphi(s)$ is sparsity constraint, and $\varphi(s)$ is one of the several penalty functions, for example L1 penalty function set as $\|s\|_1$ or Epsilon L1 penalty function set as $\left(s_j^2 + \varepsilon\right)^{1/2}$ and so on. Obviously, it is a convex optimization problem based on dictionary D or sparse coefficients. Therefore, we can use the common optimization method which is holding s unchanged to optimize D and then holding D unchanged to optimize coefficient s.

2.3. Dictionary learning

When learning the dictionary D, we hold the sparse coefficients s unchanged, then the optimization objective can be simplified as the follows:

$$\min_D \|x - D\cdot s\|_2^2$$

$$s.t.\ \sum_j B_{i,j}^2 \le c \ \ \forall j = 1\ldots n \tag{4}$$

Obviously, it's a least squares optimization problem. There are many methods to solve this problem, for example, K-SVD. And there are also some other base learning algorithms proposed in [15] [16]. An "efficient sparse coding" namely Lagrange dual to learn base proposed by [17] will be used for base learning in this paper.

2.4. Coefficient learning

For coefficient learning, the dictionary D is regarded as constant, therefore the optimization objective can be described as the following equation:

$$\min_{s} \|x - D \cdot s\|_2^2 + \lambda \sum_j \varphi(s_j) \tag{5}$$

Compared with (5), a regularized constraint is added so that it can be regarded as a least squares problem with regularized constraint. If set $\varphi(s) = \|s\|_1$, it becomes a L1- regularized linear least squares problem. There have been some effective methods such as the feature-sign search [17], basis-pursuit(BP) [18], LASSO [19], Orthogonal Matching Pursuit (OMP) [20] and so on. In this paper, the feature-sign search will be used for coefficient learning.

3. Experiments

In this section, we will show some experiments with the above proposed method and analysis of the performance of results.

3.1. Experiment Data

The database we employed in this paper is a common database in real acoustic environment, called Real World Computing Partnership (RWCP) sound database, which is produced by Mitsubishi Research Institute Inc. [21]. The database consists of speech data collected by microphone array and non-speech sound collected by dry source. In our experiment, all the sound event data are from the dry source of non-speech sound (examples shown in Fig.3.), which including a wide range sound event type, such as bank, bells, coins, clock and so on. All of sounds used in our experiment are all sampled in 16kHz and all collected under 4 kinds of noise environments, the SNRs respectively are clean, 20dB, 10dB and 0dB. And according to the requirement of feature extraction, a total of 44 sound event classes are selected from the non-speech sound database.

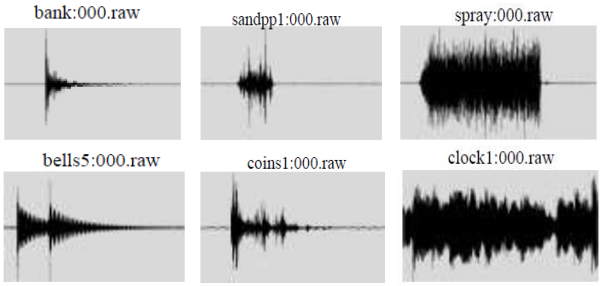

Fig.3. Waveform of samples in the RWCP Sound Database

3.2. Change the number of training samples

In the first experiment, we will observe the effect of the number of training samples of the experimental results. The database settings are as follows: for the 100 sound samples of each class, we respectively selected 20, 30, 40 or 50 samples for training and the rest 50 samples for testing. And the experiments are accomplished under the 4 kinds of noise environments respectively.

In this paper, all the parameters of the experiment are as following:

- Hamming window is used to get the frame. the length is set to 25ms and the overlap is 10ms.
- Extracted 39-dimension feature via MFCC includes 13-dimension coefficients and 13-dimension first-order difference and 13-dimension second-order difference.
- In sparse coding, the number of dictionary is set to 90 best.
- The number of Gaussian mixture is 16.

The results are shown in the Fig.4.

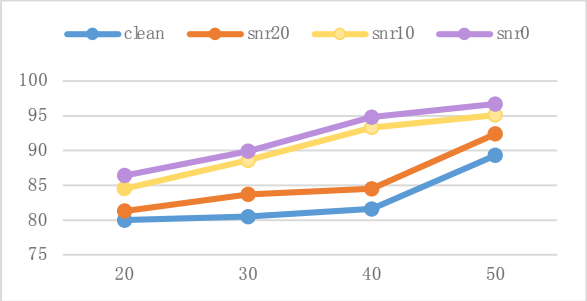

Fig.4. Classification results with different number of training samples

From the observation in the Fig. 4., it is easy to find that the performance has obvious improvement with the increasing number of training samples. Hence, in the following experiment, for each class, we select 50 training samples as the training set.

3.3. Change the training set

Obviously, in real life, we cannot know the exact SNR of a sound signal, therefore, we need to verify that the algorithm we proposed in the case of the training set is not a corresponding SNR with the testing set is still valid. Based on above viewpoints, keeping the training set the same under the 4 different noise environments is a feasible and available method.

Hence, we make the following changes of the training set and keep the testing data unchanged, meanwhile, the other parameters are also unchanged:

a) All use the clean training set;

b) All use the clean and 0db training sets;

c) Put the 4 kinds of noise environments training sets together as the training set.

The results are shown on the Fig.5.

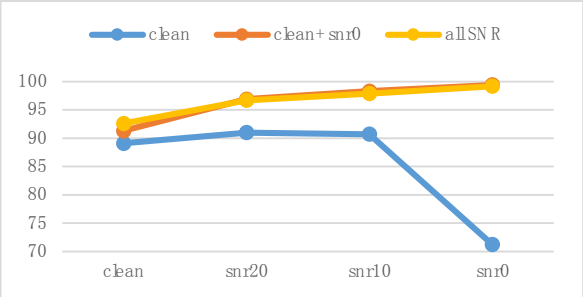

Fig.5. Classification results under different training sets

From above graph, we can conclude that our algorithm in this case is effective and even better. As a whole, the a) shows the worst results and its line appears a slump under the 0db noise environment which all because it only uses the clean samples to train; The c) has the best performance. The reason can be summarized in the two points: 1) The number of the training samples is the largest; 2) It has the samples under all the noise environments. However, the results of b) are very close to c). In the future work, we can apply the b) replacing c) to reduce the training time.

4. Conclusion

In this paper, we propose a method for sound event classification using sparse coding. The high-level features extracted from sparse coding obtain a great result. In the experiment, we mainly test the performance in the different noise conditions to demonstrates that sparse coding has great anti-interference for noise. Moreover, in order to get close to the reality, for the 4 testing sets under different noise environments, we keep the training set the same and obtain a rational and satisfied result.

References

[1] Chen, S., Sun Z. P., Bridge B., “Automatic traffic monitoring by intelligent sound detection,” Intelligent Transportation System, 1997. ITSC '97., IEEE Conference on , pp.171-176, 9- 12 Nov. 1997.

[2] Ghoraani, B., Krishnan, S., “Time–Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals,” Audio, Speech, and

Language Processing, IEEE Transactions on , vol.19, no.7, pp.2197-2209, Sept. 2011.

[3] Tzanetakis G., Cook P., "Musical genre classification of audio signals," Speech and Audio Processing, IEEE Transactions on ,vol.10, no.5, pp.293-302, Jul. 2002.

[4] Clavel C., Ehrette T., Richard G., "Events Detection for an Audio-Based Surveillance System," Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp.1306-1309, 6-6 Jul. 2005.

[5] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," J Acoustic Soc Am 115(6) (2004), pp.3033-3041.

[6] Alain Rakotomamonjy,Gilles Gasso, "Histogram of Gradients of Time–Frequency Representations for Audio Scene Classification," IEEE/ACM transactionTs on audio,speech,and language processing, vol. 23, no. 1, Jan. 2015.

[7] Olshausen B. A., Field D. J., "Emergence of simple cell receptive field properties by learning a sparse code for natural images," Nature 1996, 381:607-609.

[8] Meng Yang, Zhang D., Jian Yang, Zhang, D., "Robust sparse coding for face recognition," Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on , pp.625-632, 20-25 Jun. 2011.

[9] J. Herredsvela, K. Engan, TO Gulsrud, K. Skretting, "Texture classification using sparse representations by learned compound dictionaries," Proceedings of SPARS '05, Rennes, France, Nov. 2005.

[10] Sivaram G.S.V.S., Nemala S.K., Elhilali M., Tran T.D., Hermansky H., "Sparse coding for speech recognition," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on , pp.4346-4349, 14-19 Mar. 2010.

[11] Naseem I., Togneri R., Bennamoun M., "Sparse Representation for Speaker Identification," Pattern Recognition (ICPR), 2010 20th International Conference on , pp.4460-4463, 23- 26 Aug. 2010.

[12] Sigg C.D., Dikk T., Buhmann J.M., "Speech enhancement with sparse coding in learned dictionaries," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp.4758-4761, 14-19 March 2010.

[13] R. Grosse, R. Raina, H. Kwong, A.Y. Ng, "Shift-invariant sparse coding for audio classification," Conference on Uncertainty in Artificial Intelligence, 2007, pp. 149-158.

[14] Y. Panagakis, C. Kotropoulos, and G. R. Arce. “Music genre classification via sparse representations of auditory temporal modulations,” In Proc. European Signal Process. Conf., Glasgow, Scotland, Aug. 2009.

[15] Duc-Son Pham, Venkatesh S., “Joint learning and dictionary construction for pattern recognition,” Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on , pp.1-8, 23-28 Jun. 2008.

[16] J. Mairal, F. Bach, J. Ponce and G. Sapiro. “Online Learning for Matrix Factorization and Sparse Coding,” Journal of Machine Learning Research, vol.11, pp. 19-60. 2010.

[17] H. Lee, A. Battle, R. Raina, and A.Y. Ng, “Efficient sparse coding algorithms,” in Proc. NIPS, 2006, pp.801-808.

[18] S. S. Chen, D. L. Donoho and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM J. Sci. Comput., vol. 20, no. 1, pp. 33–61, 1998.

[19] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” J. R. Statist. Soc. Ser. B, vol. 58, no. 1, pp. 267–288, 1996.

[20] Pati Y.C., Rezaiifar R., Krishnaprasad P. S., “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” Signals, Systems and Computers. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on, pp.40-44 vol.1, 1-3 Nov.1993.

[21] S.Nakamura, K.Hiyane, F.Asano, T.Nishiura, and T.Yamada: “Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition,” 2nd International Conference on Language Resources & Evaluation, Athen ,Jun.2000.