Sparse Coding for Sound Event Classification

Mingming Zhang[1,2], Weifeng Li[1,2], Longbiao Wang[3], Jianguo Wei[4], Zhiyong Wu[1,2], Qingmin Liao[1,2]
[1]Shenzhen Key Lab. of Information Sci&Tech/Shenzhen Engineering Lab. of IS&DRM
[2]Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China
[3]Nagaoka University of Technology, Japan
[4]School of Computer Science and Technology, Tianjin University, China

Abstract—**Generally sound event classification algorithms are always based on speech recognition methods: feature-extraction and model-training. In order to improve the classification performance, researchers always pay much attention to find more effective sound features or classifiers, which is obviously difficult. In recent years, sparse coding provides a class of effective algorithms to capture the high-level representation features of the input data. In this paper, we present a sound event classification method based on sparse coding and supervised learning model. Sparse coding coefficients will be used as the sound event features to train the classification model. Experiment results demonstrate an obvious improvement in sound event classification.**

I. Introduction

The non-speech sound event classification has a wide use in many important applications, such as music genre classification [1-4], security surveillance [5], environment detection [6-7], health care and so on. Generally, the system of sound event detection and classification always uses the methods derived from speech recognition, which in general contains two steps: first, the sound event features are extracted from labeled training sound, such as MFCC (Mel-Frequency Cepstrum Coefficient, MFCC), PLP (Perceptual Linear Predictive, PLP); second, the classifier is trained with extracted features, such as SVM (Support Vector Machine, SVM), GMM (Gaussian Mixture Model, GMM) and HMM (Hidden Markov Model, HMM). A lot of related work has been done in last twenty years. [8] relied on the use of Wavelet transform technique for detection and on an unsupervised order estimation of GMM. The basic idea of [9] was to embed probabilistic distances into classical SVM to classify the sound events. [10] presented an efficient robust sound classification algorithm based on hidden Markov models. While the literature [11] proposed a novel method for feature extraction with spectrogram image feature. The most difference between aforementioned methods is the different combination of general features and classifiers.

Sparse coding is algorithm trying to find a high-level representation of the input signal, which first introduced by Olshausen [12]. It has to learn a dictionary called “basis functions”, and the input signal can be represented by the linear combination of the basis functions while the coefficient vector is sparse. In recent years, sparse coding is paid more and more attention in many research fields, especially image processing such as image noise reduction, image restoration, image classification and face recognition [13-14].

In audio signal processing, sparse coding can be used in speaker Identification [15], speech recognition [16-17], speech enhancement [18] and so on. Comparing with the image processing, sparse coding has got less attention on the use of audio signal processing, especially sound event classification. [19] proposed a joint sparsity classification method to exploit the inner correlation between observations for acoustic signal classification. [20] presented an algorithm for computing shift-invariant sparse coding (SISC) solutions and applied it to audio classification. [21] employed the sparse coding of auditory temporal modulations in music genre classification. Sparse coding can represent each example using a few non-zero coefficients and obtain a high-level representation of the example, therefore the sparse coefficients can be used as the new feature of sound event for sound events classification with supervised learning.

In this paper, we propose to lean a high-level representation of the input sound event features via sparse coding, and then to train a supervised classifier for our classification task.

This paper is organized as follows: In section 2, the general sparse coding algorithm is presented. In section 3, we give the proposed method. Section 4 is our detailed experiment results and the results analysis. Finally, we draw our conclusions in section 5.

II. Sparse Coding

In this section, we will give a simple description of sparse coding algorithm, including coefficient learning and dictionary learning.

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

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

The sparse representation $s \in R^{n\times 1}$ of x can be estimated by the following method:

$$\begin{aligned} &\min_{(D,s)} \; \|x - D\bullet s\|_2^2 \quad s.t. \;\; \phi(s) < \sigma \\ &s.t. \;\; \sum_j B_{i,j}^2 \le c \quad \forall j = 1\cdots n \end{aligned} \tag{2}$$

where $D=[d_1\ d_2\ \cdots\ d_n]$ is the dictionary with column vector d_j of the j^{th} atom, and s is the coefficient vector. Therefore the above sparse coding problem can be seen as a optimization problem with constrain as follows:

$$\begin{aligned} &\min_{s} \; \|x - D\bullet s\|_2^2 + \beta \sum_j \phi(s_j) \\ &s.t. \;\; \sum_j B_{i,j}^2 \le c \quad \forall j = 1\cdots n \end{aligned} \tag{3}$$

where $\|x-D\cdot s\|_2^2$ is reconstruction error, and $\beta\sum\phi(s)$ is sparsity constraint, $\phi(s)$ is penalty function such as $\|s\|_1$ (L1 penalty function), $(s_j^2+\varepsilon)^{1/2}$ (Epsilon L1 penalty function) and so on. Obviously, the optimization problem is a convex problem based on dictionary *D* or sparse coefficient *s*, but not convex based on both *D* and *s*. Therefore the general optimization method is to optimize dictionary *D* (holding *s* unchanged) and coefficient *s* (holding *D* unchanged).

A. Base Learning

For Base learning, the sparse coefficients is constant, therefore the optimization objective can be described as the following equation:

$$\begin{aligned} &\min_{D} \quad \|x-D\bullet s\|_2^2 \\ &s.t. \quad \sum_j B_{i,j}^2 \le c \quad \forall j=1\cdots n \end{aligned} \qquad (4)$$

This is a least squares optimization problem. Some methods can be used to solve the optimization problem, such as K-SVD [22], and some other algorithms for base learning is proposed in [23][24]. [25] proposed an "efficient sparse coding" to learn base: Lagrange dual. In this paper, this method will be used for base learning.

B. Coefficient Learning

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

$$\min_{s} \quad \|x-D\bullet s\|_2^2+\beta\sum_j\phi(s_j) \qquad (5)$$

This is a least squares problem with regularized constraint. If $\phi(s)=\|s\|_1$, the problem becomes a L1- regularized linear least squares problem. One appealing method is basis-pursuit (BP) [26], LASSO [27], Orthogonal Matching Pursuit (OMP) [28]. [25] also proposed a algorithm, "the feature-sign search", to lean coefficient. In this paper, the feature-sign search will be used for coefficient learning.

III. THE PROPOSED CLASSIFICATION METHOD

A. General Method

Figure 1 shows the general method for sound event classification.

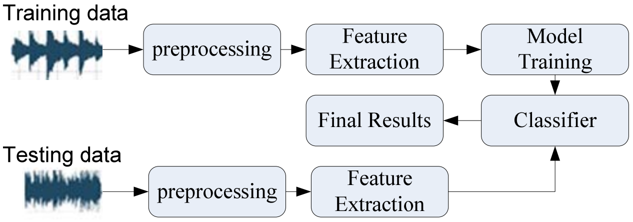

Fig. 1: Sound event classification algorithm

Preprocessing step is always noise reduction, pre-emphasis, sound segmentation and so on in order to obtain effective sound segments. Feature extraction generally includes frame feature and clip feature. In this paper, all the sound samples are clean, while most samples contains isolated sound events with long silence before and after the sound event, therefore the preprocessing step in this paper is sound segmentation.

B. The Proposed Method

Similar to the general method, the proposed method for sound event classification also needs to extract "sound features" and classifier training. Figure 2 is the whole block diagram of proposed method.

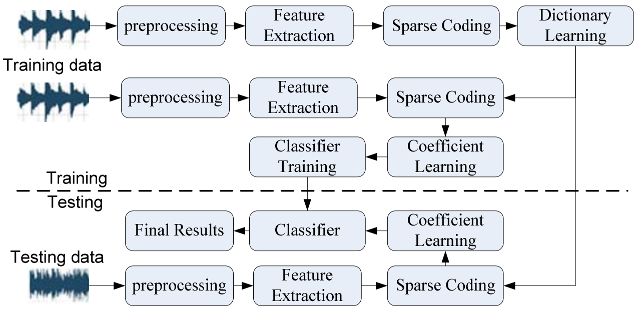

Fig. 2: Block diagram of the proposed method

In the proposed method, the "sound feature" of sound event is not the general feature such as MFCC, but the coefficients of sparse coding. Preprocessing, Feature Extraction and Model Training steps are same with the general method. By dictionary learning, the coefficients of sparse coding can well describe the properties of feature distribution. Therefore the detailed steps are presented as follows:

Step1:

- Preprocessing: in this step, the original sound samples are segmented into sound event segments and silence segments, and the silence segments are discarded while the sound event segments will be used as the effective sound clip for feature extraction..

Step2:

- Feature Extraction: this step allows us to extract sound clip features such as MFCC, PLP and so on. In this paper, we use the 39-dimension MFCC features, and also merge several frames into one clip as a clip feature for evaluation. In this step, some sound events, the length of which is too short, are discarded.

Step3:

- Dictionary learning of sparse coding: using the features learned from step2 as the training samples of the sparse coding, the dictionary is learned.

Step4:

- Coefficient learning of sparse coding: after obtaining dictionary *D*, we can obtain the high-level representation of the input feature (from step2), just as the sparse coefficients.

Step5:

- Classifier Training: the coefficients learned from step4 will be used as the new feature of the sound event to train the classifier.

Step6:

- Testing data classifying: In testing step, new feature of testing data (just as the coefficients from step4) can be learned with the learned dictionary. Then the final results can be obtained using trained classifier.

IV. EXPERIMENT

In this section, we will evaluate the performance of the proposed method, and present the experiment results.

A. Experiment Data

In this paper, our sound data are selected from an important database, the Real World Computing Partnership (RWCP) sound database produced by Mitsubishi Research Institute Inc. [29]. RWCP Sound Scene Database is a common database as a standard for objective comparison and evaluation of research results in real acoustic environment. The content includes speech measurement data collected by microphone array and dry source of non-speech sound, and all the sound event data in the experiment are from the dry source of non-speech sound.

A total of 44 sound event classes are selected from the non-speech sound database, including a wide range of sound event type, such as bells, bottles, buzzer, coin, metal, tear, whistles and so on. The sounds used in our experiment are all sampled in 16kHz.

For the sound events, 100 sound samples are contained in each class, from which 50 samples are selected for training and the rest 50 samples for testing, therefore the total samples for training and testing are respectively 2200 and 2200. Meanwhile, we consider another method for sound event classification using sparse coding: 30 samples are selected from the training data of 50 samples aforementioned for the dictionary learning of sparse coding and another 20 samples for supervised classifier training, and the testing data keep unchanged.

B. Experiment Results

To evaluate the performance of proposed method, the following methods are tested as comparisons:

1). MFCC-SVM using frame-averaged features;

2). MFCC-GMM-UBM (Gaussian Mixture Model; Universal Background Model);

3). Some special methods (such as sparse filtering [30], Spectrogram Image Feature [11]).

In this paper, all the MFCC features and sound frames in the experiment use the following parameters: the MFCC is 39 dimension feature vector, including 13-dimension coefficients and their first and second time derivatives; the frame can be obtained with hamming window of 25ms and the overlap between frames is 10ms.

The baseline method for RWCP sound event classification in our paper is Universal Background Model-Gaussian Mixture Model (UBM-GMM). UBM-GMM is an effective classifier because its generalization ability can handle “unseen” acoustic patterns.

For the UBM-GMM, we test the effect of different Gaussian mixture number on the classification results. Figure 3 shows that the classification accuracy continues to improve with the increase of Gaussian mixture number and decrease when larger than a threshold.

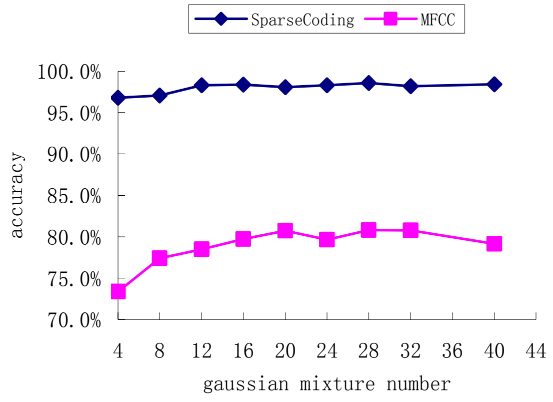

Fig. 3: The accuracy curve based on Gaussian mixture number. Blue: using MFCC feature and UBM-GMM classifier; Red: using MFCC feature, the proposed sparse coding algorithm and UBM-GMM classifier.

For the proposed sparse coding method, Figure 4 shows the results of different numbers of dictionary basis functions.

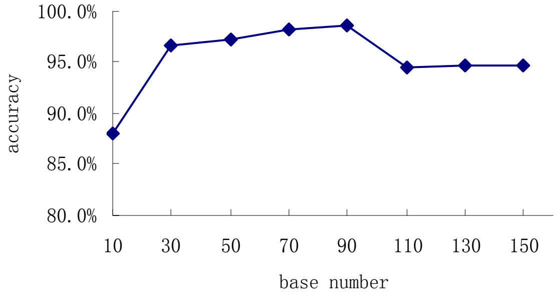

Fig. 4: The result of different base number of sparse coding dictionary (the dictionary size is feature-dimension*base-number. In this paper, feature-dimension is 39).

For the sparse coding learning with supervised model, several questions should be considered: One sparse coding dictionary is learned with all classes of data or several dictionaries for each class? The data for dictionary learning and model training use the same data or different data? Table 1 shows the four conditions. In this experiment, UBM-GMM classifier is used, and the Gaussian mixture number is 16, and the number of basis functions is 90.

According to above results, Table 2 shows the sound event classification results of different algorithm. In the experiment, the number of Gaussian mixture is 16, and the feature is based on frame other than clip, and number of basis functions is 90.

TABLE 2: CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHM ON DATABASE RWCP. 16 GAUSSIAN MIXTURE ARE USED FOR THE CLASSIFIER UBM-GMM. THE PROPOSED SPARSE CODING ALGORITHM PRESENTS TWO METHODS FOR COEFFICIENT LEARNING. THE DICTIONARY LEARNING USES THE LAGRANGE DUAL ALGORITHM [25], AND THE NUMBER OF BASIS FUNCTIONS IS 90.

UBM-GMM			SVM [31]	
MFCC	Sparse Filtering [30]	Sparse Coding (Proposed method)	MFCC	Spectrogram Image Feature [11]
79.7%	81.8%	**97.9%**	88.9%	92.8%

What's more, SVM in this paper is from CHih-Chung and Chih-Jen Lin [31].

C. Results Analysis

1). Gaussian mixture numbers

Generally, the more Gaussian mixture numbers we use in GMM classifier, the better results we can achieve, because the more Gaussian mixture numbers can describe a better approximation of the data distribution, however, because of the limitation of the training data and computation complexity, the performance will not continue to increase obviously when the mixture numbers larger than some number. Figure 3 shows that MFCC-UBM-GMM and Sparse Coding-UBM-GMM both achieve an accuracy about 97.9% with the Gaussian mixture number 16, after which the accuracy even starts to decrease.

2). Base number of sparse coding dictionary

For learning the dictionary, the number of "basis functions" is an important parameter. Figure 4 denotes that the accuracy reaches to the largest one about 97.9% when the number of basis functions is 90.

TABLE 1: DICTIONARY LEARNING AND MODEL TRAINING. ONEDICTIONARY: LEARNING ONLY ONE DICTIONARY FOR ALL CLASSES OF SAMPLES (44 CLASSES OF DATA ARE USED); MULDICTIONARY: LEARNING ONE DICTIONARY FOR EACH CLASS (44 DICTIONARIES ARE LEARNED); SAMEDATA: DICTIONARY LEARNING AND MODEL TRAINING USE THE SAME DATA (44*50=2200 SOUND SAMPLES); DIFFERENTDATA: DICTIONARY LEARNING USES44*30=1320 SOUND SAMPLES, AND MODEL TRAINING USES 44*20=880 SOUND SAMPLES

Proposed Method	SameData	DifferentData
OneDictionary	65.6%	52.7%
MulDictionary	97.9%	95.7%

3). One dictionary *vs* multiple dictionaries

From Table 1 we can see that learning a dictionary for each class of sound event can obtain a much better result than learning only one dictionary for all classes of sound events. For the data division between learning dictionary and classifier training, SameData means more data is used for dictionary and coefficient learning, therefore it performs a little better than the DifferentData.

4). Different methods

Table 2 shows the sound event classification results of different methods. Obviously, the general methods such as MFCC-UBM-GMM and MFCC-SVM have a bad result less than 90%. Sparse filtering is another sparse coding algorithm which aims to obtain good performance with shorter time, and [11] was a novel method using Spectrogram Image feature, however these two algorithms do not obtain a very good result in our experiment for the sound event classification. Finally the proposed method achieves a good performance with the accuracy 97.9%. It shows that sparse coding algorithm can obtain the high-level representation of the input signal and can better describe the property of the sound event.

According to the detailed results analysis for the accuracy of each class, we find that most classes have a good performance (the accuracy is about 100%) while several classes always have a bad result such as crumple, punch and pan. For example, analyzing the results of Sparse Coding-UBM-GMM (97.9%) in Table 1, of all the 44 classes, 30 classes of sound can achieve the accuracy of 100%, and 36 classes are larger than 97.9%. Several classes, whose recognition rates are lower than 95%, are punch (91.6%), crumple (92.3%), and pan (94.0%).

V. CONCLUSIONS

In this paper, a proposed method for sound event classification using sparse coding is presented. We use the sparse coding to extract high-level features of the sound event for supervised learning, and obtain a good result. In the experiment, we discuss the data division strategy for dictionary learning, and also discuss the effect of dictionary number on the classification result. The experiment results show that sparse coding features can better describe the properties of the sound event feature, and learning dictionary for each class of sound can obtain a much better result than that learning one dictionary for all classes of sound. Finally, we have not yet tested the performance of sparse coding in different noise conditions, which will be our future work.

VI. ACKNOWLEDGEMENTS

This work was supported in part by Shenzhen Basic Research Grant JCYJ20120831165730913, in part by National Natural Science Foundation of China (No. 61175016) and in part by National Basic Research Program of China (No. 2013CB329305).

REFERENCES

[1] 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.

[2] Changsheng Xu, Maddage M.C., Xi Shao, Fang Cao, Qi Tian, "Musical genre classification using support vector machines," Acoustics, Speech, and Signal Processing, 2003. Proceedings.

(ICASSP '03). 2003 IEEE International Conference on , vol.5, no., pp.V,429-32 vol.5, 6-10 April 2003.

[3] Ling Chen, Phillip Wright, Wolfgang Nejdl, “Improving music genre classification using collaborative tagging data,” Proceedings of the Second ACM International Conference on Web Search and Data Mining, February 09-12, 2009, Barcelona, Spain.

[4] Chin-Chia, Michael Yeh, Yi-Hsuan Yang, “Supervised dictionary learning for music genre classification,” Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, June 05-08, 2012, Hong Kong, China.

[5] Clavel C., Ehrette T., Richard G., "Events Detection for an Audio-Based Surveillance System," Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on , vol., no., pp.1306,1309, 6-6 July 2005.

[6] Chen, S., Sun Z. P., Bridge B., "Automatic traffic monitoring by intelligent sound detection," Intelligent Transportation System, 1997. ITSC '97., IEEE Conference on , vol., no., pp.171,176, 9-12 Nov 1997.

[7] 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.

[8] Rougui J. E., Istrate D., Souidene W., "Audio sound event identification for distress situations and context awareness," Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE , vol., no., pp.3501,3504, 3-6 Sept. 2009.

[9] T.H. Dat and H. Li, "Probabilistic distance SVM with Hellinger-Exponential Kernel for sound event classification,” in Proc. ICASSP, 2011, pp.2272-2275.

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

[11] Dennis J., Huy Dat Tran, Haizhou Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," Signal Processing Letters, IEEE , vol.18, no.2, pp.130,133, Feb. 2011

[12] 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.

[13] Meng Yang, Zhang D., Jian Yang, Zhang, D., "Robust sparse coding for face recognition," Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on , vol., no., pp.625,632, 20-25 June 2011.

[14] J. Herredsvela, K. Engan, TO Gulsrud, K. Skretting, “Texture classification using sparse representations by learned compound dictionaries,” in: Proceedings of SPARS '05, Rennes, France, November 2005.

[15] Naseem I., Togneri R., Bennamoun M., "Sparse Representation for Speaker Identification," Pattern Recognition (ICPR), 2010 20th International Conference on , vol., no., pp.4460,4463, 23-26 Aug. 2010.

[16] 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 , vol., no., pp.4346,4349, 14-19 March 2010.

[17] W.J. Smit, E. Barnard, “Continuous speech recognition with sparse coding,” Computer Speech & Language, Volume 23, Issue 2, pp.200-219, 2009.

[18] 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 , vol., no., pp.4758,4761, 14-19 March 2010.

[19] Haichao Zhang, Nasrabadi N.M., Huang T.S., Yanning Zhang, "Transient acoustic signal classification using joint sparse representation," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on , vol., no., pp.2220,2223, 22-27 May 2011.

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

[21] 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.

[22] Aharon M., Elad M., Bruckstein A., "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," Signal Processing, IEEE Transactions on , vol.54, no.11, pp.4311,4322, Nov. 2006.

[23] Duc-Son Pham, Venkatesh S., "Joint learning and dictionary construction for pattern recognition," Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on , vol., no., pp.1,8, 23-28 June 2008.

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

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

[26] 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.

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

[28] 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, vol., no., pp.40-44 vol.1, 1-3 Nov 1993.

[29] 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 (2000.6).

[30] J. Ngiam, P. Koh, Z. Chen, S. Bhaskar, A.Y. Ng. “Sparse filtering.” NIPS 2011.

[31] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM : a library for support vector machines.” ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.