International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 10, No. 3 (2017), pp. 41-52
http://dx.doi.org/10.14257/ijsip.2017.10.3.05

Complex Sound Recognition Method Based on Cluster Labels

Cuiqing Jiang*, Peng Fan, Kun Liang and Zhao Wang

School of Management, Hefei University of Technology,
AnHui, Hefei, 230009, China
fanpeng812@163.com

Abstract

Complex sound is the kind of sound which contains numerous types of sounds and the boundaries between them are indistinguishable. Aiming at the problem of poor identification accuracy of identifying complex sound in noisy environment, this paper proposes a mixed feature representation of complex sound, which is based on the combination of Mel Frequency Cepstral Coefficients and Short-Term Energy, by extracting and analyzing features of complex sound. And a training samples selection algorithm based on double cluster labels is proposed to select representative training sets. Then we train and identify complex sound by employing the Hidden Markov Model. Experiments show that using the mixed feature parameter and the proposed algorithm can greatly improve the identification accuracy of complex sound in noisy environment.

Keywords*: Complex sound; Mel Frequency Cepstral Coefficients; Hidden Markov Model; Cluster labels*

1. Introduction

Real-live sounds are basically aliasing and complex, existing systems identify these sounds by separating sound sources[1]-[2]. Be different with these aliasing sounds, complex sound of this study is the kind of sound which contains numerous types of sounds and the boundaries between them are indistinguishable. We just identify this type of sounds integrally and we don't need to separate these sound sources. For example, the vehicles whistles in communities are different because of the different types of vehicles and different distances from the sound sources to the microphone, so when we identify them, we just need to identify whether they belong to whistle without knowing their specific whistle ways. For another example, train whistle under the railway environment contains wind whistle and electric whistle, and is along with various sounds of the wind, grinding sound of the track, brakes and the environmental sounds, *etc*. There are also many such sounds in natural environment, such as the sound of thunder and lightning, the sound of water and wind, *etc*., with different times, the characteristics are different.

Complex sound recognition is the key issue in the fields of intelligent transportation and ambient intelligence, *etc*. In recent years, researches of speech recognition have made significant breakthroughs [3]-[4]. However, researches of non-speech recognition such as Environmental Sound Recognition (ESR) can't meet the application requirements [5]-[6]. Complex sound recognition is a kind of classification problem. Existing studies mainly use feature extraction methods and classification methods of speech recognition. In terms of feature selection and extraction, MFCC is most commonly used. Nevertheless, there is various and unpredictable noise in real and complex environment. The classification effect of MFCC declines significantly when noise occurs, so the traditional MFCC feature can't meet the needs of practical applications [7]-[8]. In order to solve this problem, this paper proposes a mixed feature parameter which contains MFCC and STE to represent the features of complex sound by analyzing complex sound.

ISSN: 2005-4254 IJSIP

Common classifiers for classifying sound include Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Support Vector Machine (SVM), *etc*. Studies show that sample-based learning methods are the most effective to design classifiers. Thus the quantity and quality of training samples are two key factors which affect the performance of the classifiers [9]. But in the traditional training process, there are following problems of training samples: firstly, when the number of training samples is large enough, the statistic-based learning methods for classification can obtain the classifiers which have good generalization abilities, but the computational cost of training classifiers also increases. Secondly, there are some redundant samples in a lot of sample libraries, in fact, similar samples don't need repeatedly training [10]-[11]. Thirdly, in practice, complex sound contains many types of sounds and they alternately occur. Therefore, if we mark them manually, the cost is very high. And marking some types of sounds may also need expertise, for example, train whistle contains wind whistle and electric whistle. If we choose some samples which are more useful for classification to label, we can use less manual annotation to get higher quality of training data sets[12]-[13]. So this paper proposes an algorithm based on double cluster labels for selecting training samples to improve training phase of HMM classifier.

The rest of this paper is organized as follows. In Section II, we will first introduce the theories of some sound features, and then give the reasons of selecting the proposed mixed features. In Section III, we will give the details of the proposed training samples selection algorithm. Experimental results and analyses will be presented in Section IV. Finally, Section V concludes the work.

2. Feature Extraction

2.1. Mel Frequency Cepstral Coefficients

The human auditory system is a very good sound recognition system within the audible range of the human ear. MFCC is the commonly used feature in sound recognition, because it uses a non-linear frequency scale to simulate the auditory system and its performance is relatively stable. The transform relation of Mel frequency Mel (f) and actual frequency f is given in formula (1).

$$\mathrm{Mel(f)} = 2593\,\mathrm{lg}(1 + \mathrm{f}/700), \tag{1}$$

The steps of extracting MFCC feature parameters as follows.

(1) Firstly, we transform the framed and windowed sound signals by Discrete Fourier Transform (DFT) to get power spectrum on the frequency spectrum.

(2) Then, we convolve the power spectrum through a set of band-pass filters which contain M triangle filters, and get logarithmic power spectrum of Mel frequency by calculating logarithmic energy of filters.

(3) Finally, we transform the logarithmic power spectrum by Discrete Cosine Transform (DCT) to get MFCC coefficients.

$$\mathrm{C_t(n)} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \mathrm{S_t(m)} \cos\left(\frac{\pi \mathrm{n(m-0.5)}}{\mathrm{M}}\right), 0 \le \mathrm{n} < \mathrm{M}, \tag{2}$$

Where n is the number of MFCC, and $\mathrm{C_t(n)}$ is the $\mathrm{n^{th}}$ MFCC parameter of the $\mathrm{t^{th}}$ frame, $\mathrm{C_t(0)}$ is deleted because it is DC component and MFCC coefficients finally contain $\mathrm{C_t(1)}, \cdots, \mathrm{C_t(12)}$. $\mathrm{S_t(m)}$ is the logarithmic power spectrum of sound signals. M is the number of triangle filters and M values 12 in this paper.

2.2. Differential Features

Standard MFCC parameter only reflects the static characteristics of sound, and the dynamic characteristics can be described by the difference coefficients of MFCC. Experiments have shown that the combination of dynamic features and static features can effectively improve the recognition performance of the system. We calculate the first order difference coefficients of MFCC by the formula (3).

$$D_n = \frac{\sum_{k=1}^{K}(C_{n+k} - C_{n-k})}{\sqrt{2\sum_{k=1}^{K} k^2}} \tag{3}$$

Where D_n represents the n^{th} first order difference coefficients of MFCC. C_n is the n^{th} cepstral coefficient. K represents the time difference of the first derivative and K values 2 in this paper.

2.3. Short-Term Energy

Short-term energy is the measurement of the intensity of the sound, and is a kind of important time domain feature of sound signals. It is denoted by STE and defined as the formula (4).

$$E_t = \sum_{i=0}^{N-1} x_t^2(i), \tag{4}$$

Where E_t is the STE of the t^{th} frame. N is the frame length and values 1024 in this paper, i is the sample point within a frame.

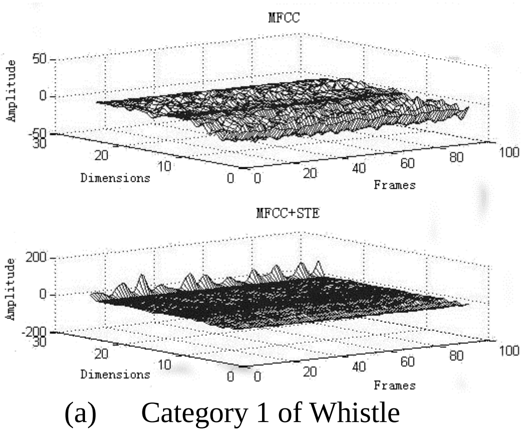
2.4. Mixed Features

This paper proposes a mixed feature parameter which contains MFCC (including the first order differential coefficients) and STE on the following bases.

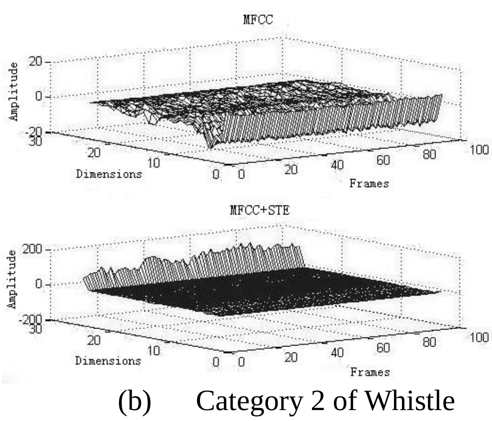
(1) STE is the reaction of time domain features, and has little correlation with MFCC parameters. The combination of them can reflect the features of complex sounds better.

(2) STE is a scalar value. Combining with MFCC parameters doesn't significantly increase the dimensions of feature components and the computational complexity is small.

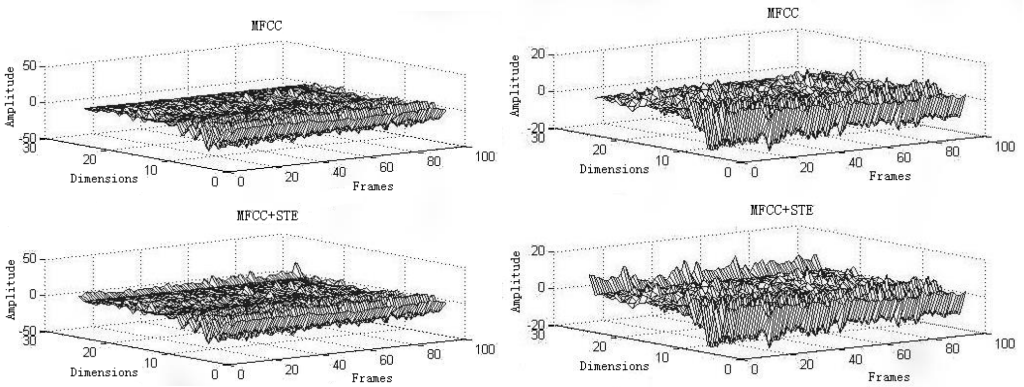
(3) We take the train whistle detection for example, and analyze some sound features of the running trains. As Figure 2 shows, the mixed feature parameter which contains MFCC and STE can reflect the differences between different sound signals better.



(a) Category 1 of Whistle



(b) Category 2 of Whistle



(c) Category 1 of Non-Whistle (d) Category 2 of Non-Whistle

Figure 1. Figures of MFCC Features and Mixed Features which Contain MFCC and Short-Term Energy

The following conclusions can be obtained from Figure 1.

(1) Through comparison between Figure 1(a) and Figure 1(b), we can find that although category 1 of whistle and category 2 of whistle are both whistle, their features are significantly different. Similarly, through comparison between Figure 1(c) and Figure 1(d), we also can find that although category 1 of non-whistle and category 2 of non-whistle are both non-whistle, their features are also significantly different. The results show that even if they belong to the same category of complex sound, the features may be significantly different.

(2) Through comparison between Figure 1(a) and Figure 1(c), we find that the MFCC parameters of category 1 of whistle and category 2 of whistle are similar, so using only traditional MFCC can’t distinguish the whistle and non-whistle well.

(3) When we use the mixed feature parameter, the features of different sounds are significantly different. So we expect the classification effects of the mixed feature parameter would be better.

3. Proposed Training Samples Selection Algorithm

In order to classify the features of complex sound better, this paper constructs the classifier with HMM. HMM is a kind of generative model classifier, and for this type of classifier, good training sample consists on the following characteristics (1) Similarity. In a category, the samples should have some certain similarities. The samples which are away from the distribution areas of most samples may lead to a serious decline in the performance of the classifier. (2) Non-redundancy. When we select the useful samples from a category, we only need the samples which can describe the distribution of samples in the category and don’t need too many similar samples, so we need to remove redundant samples. (3) Comprehensiveness. Representative sample sets which are selected should cover every category as well as possible. Therefore, HMM is a kind of learning model based on statistics and the quality of HMM training samples directly influences the effect of the classifier.

In the complex sound environment, there are some differences in duration, background noises and amplitude of sound clips in the same category. If we train the complex sound in the same category together as one class by traditional manual annotation, the obtained model may be not convergent or the model parameters may be not representative. In order to solve the above problems, we propose an algorithm based on double cluster labels for selecting training samples to improve training phase of HMM classifier. The distance of K-Means clustering method is cosine distance because the sound signals are time series signals and the features are vectors. The algorithm based on double cluster labels for selecting training samples is shown in Figure 2.

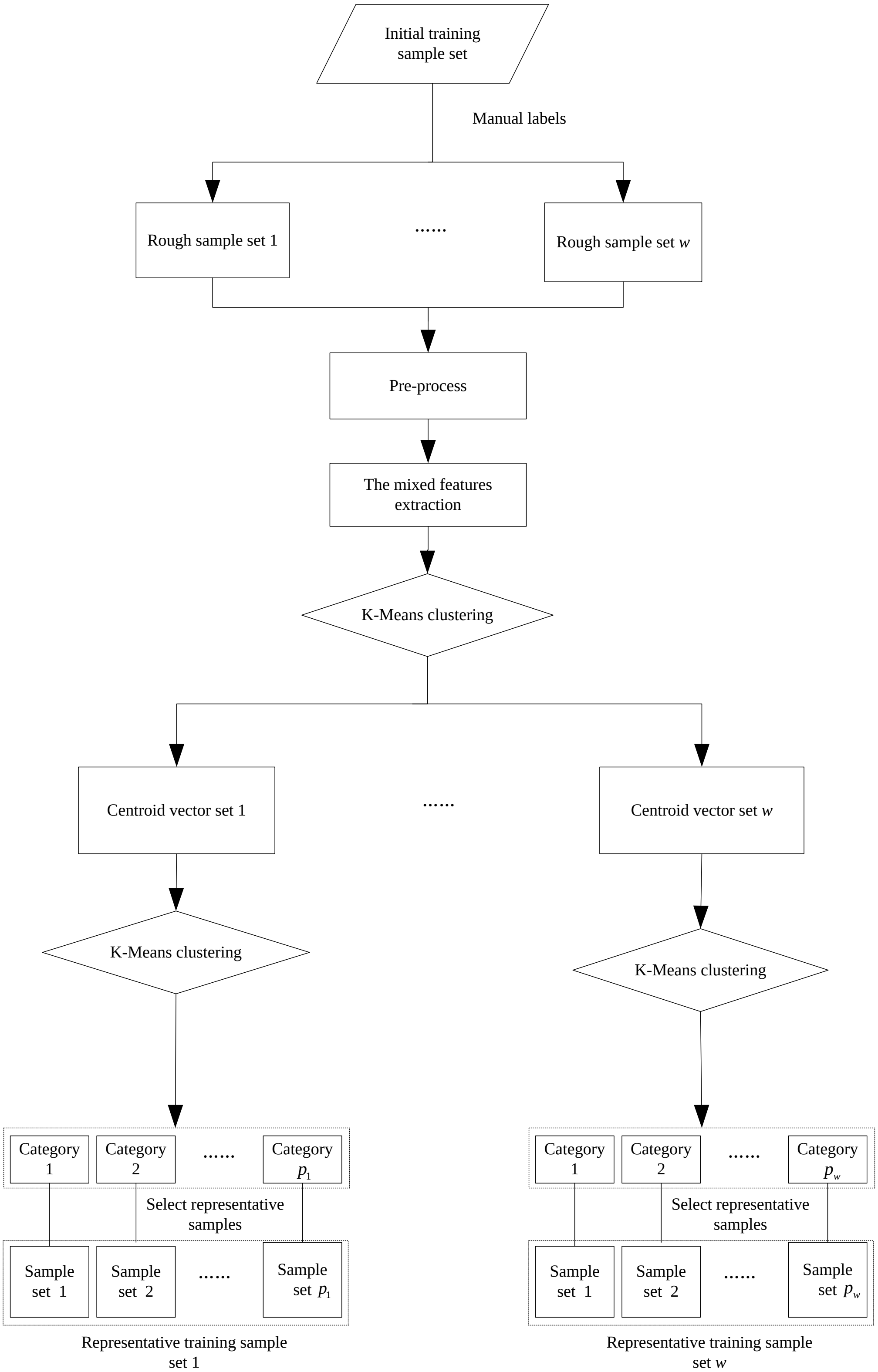


Figure 2. Algorithm Based on Double Cluster Labels for Selecting Training Samples

The algorithm for selecting representative training sample sets is as follows.

(1) Select a large number of experimental samples as the initial training sample set, and then classify them into rough sample set 1, rough sample set 2,⋯, rough sample set *w*. Each rough sample set represents a same category.

(2) Pre-process every sample of each rough sample set and extract their mixed features to get the corresponding mixed eigenvector matrix.

(3) Cluster each mixed eigenvector matrix into one group by K-Means clustering method and the distance measure is cosine similarity. Select the centroids to represent the samples and we can obtain a total of *w* centroid vector sets.

(4)Cluster the *w* centroid vector sets separately. The number of classes is decided according to the actual situation. Each centroid vector set can be clustered into P_w classes, and then select few samples as the final representative training sample sets.

In step (4), to every class, the selection method of the representative training sample sets is as follows. Firstly, we get the cosine distance from each centroid vector to the class center by K-means. Then we arrange the distances in ascending order. Finally, we select the $(1 + k \times d)$th centroid vector as representative centroid vector.

$$k = 0, 1, 2, \cdots, N - 1\text{；}\quad d = \frac{X}{N}$$

Where *X* represents the number of centroid vectors of the class. *N* is the number of centroid vectors of representative centroid vector sets.

Even if the number of the training samples is limited, we can still get better classification performance through the training sample sets which are constructed by the proposed selection algorithm.

Through the above algorithm, we can extract a plurality of more representative training sample sets from a rough sample set according to the mixed feature parameters. The advantages of these representative sample sets are as follows. First, because the sets are obtained by clustering based on feature parameters, it is possible to overcome the difficulties of subjective classification. Second, the distribution of these sets approaches the distribution of every category and they remove the redundant samples, so they ensure the quality of training samples, while reducing the number of training samples, and they greatly improve the efficiency of training and recognition accuracy.

4. Experiments

4.1. Experimental Setup

In this paper, we take the train whistle recognition for an example. The experimental goal is to identify the train whistle from the sound of trains, that is, to distinguish the whistle from the non-whistle. Since the train whistle contains wind whistle and electric whistle, *etc*., and non-train whistle contains various sounds of the wind, grinding sound of the track and brakes, *etc*., so train whistle recognition is a kind of typical complex sound recognition in noisy environments.

(a). Data Sets

The data sets of this paper were collected from the nearby train station in Hefei and Wuhu, *etc*. And we used microphones and other equipment to acquire the sound signal data of trains, the sound format is wav, the sampling rate is 48 kHz, 16 bit, mono and the encoded form is PCM.

Taking the multiple semantic characteristics of complex sound signals into account, we marked the original sound signals of trains as two categories which contain whistle and non-whistle by artificial identification. In reality, whistle can be subdivided into wind whistle and electric whistle, *etc*. Non-whistle can be subdivided into sound of the wind, grinding sound of the track and brakes, *etc*. Since these complex sound signals can’t be directly marked by artificial annotation, so we directly marked the whistle as category 1 of whistle, category 2 of whistle,⋯, *etc*., and we directly marked the non-whistle as category 1 of non-whistle, category 2 of non-whistle,⋯, *etc.*

(b). Feature extraction and classifier design

In this paper, the sound samples of the original data sets are different in lengths of time. In order to obtain the same number of features, we used the time window method to automatically cut the sound samples and we can obtain the sound clips of equal size. The length of each clip is 0.5 seconds and the length of the overlap between adjacent clips is 0.25 seconds. The features of fragments are composed by features of frames and the used frame length is 1024 points, frame shift is 480 points. The features of frames are composed by 12-dimensional MFCC parameters, their first-order differential coefficients and one-dimensional STE, a total of 25 dimensions. We employed HMM classifier as the experimental classifier. The number of initial state is set as 6 and the initial probability of state is [1,0,0,0,0,0]. Each state contains three Gaussian probability density functions and the convergence threshold of training samples is set as 5 * e^{-7}.

(c). Data Distribution

The initial training sample set of the experiment contains a total of 200 train sound samples, and the lengths of time range from 30 seconds to 180 seconds. Through the algorithm based on cluster labels, we finally selected 29 sound signals of whistle and 47 sound signals of non-whistle as the representative training sample set. Besides, we obtained test samples by collecting new samples in real time, and the test sample set contains a total of 230 samples in this paper.

(d). Evaluation Criteria

The evaluation criteria of features selection is the classification accuracy of HMM in the case of selecting different features, that is, the whistle recognition rate, the non-whistle recognition rate and the comprehensive recognition rate. They are calculated as follows.

$$whistle\ recognition\ rate = \frac{the\ number\ of\ correctly\ identified\ whistle}{the\ total\ number\ of\ whistle\ samples} \times 100\% \quad (5)$$

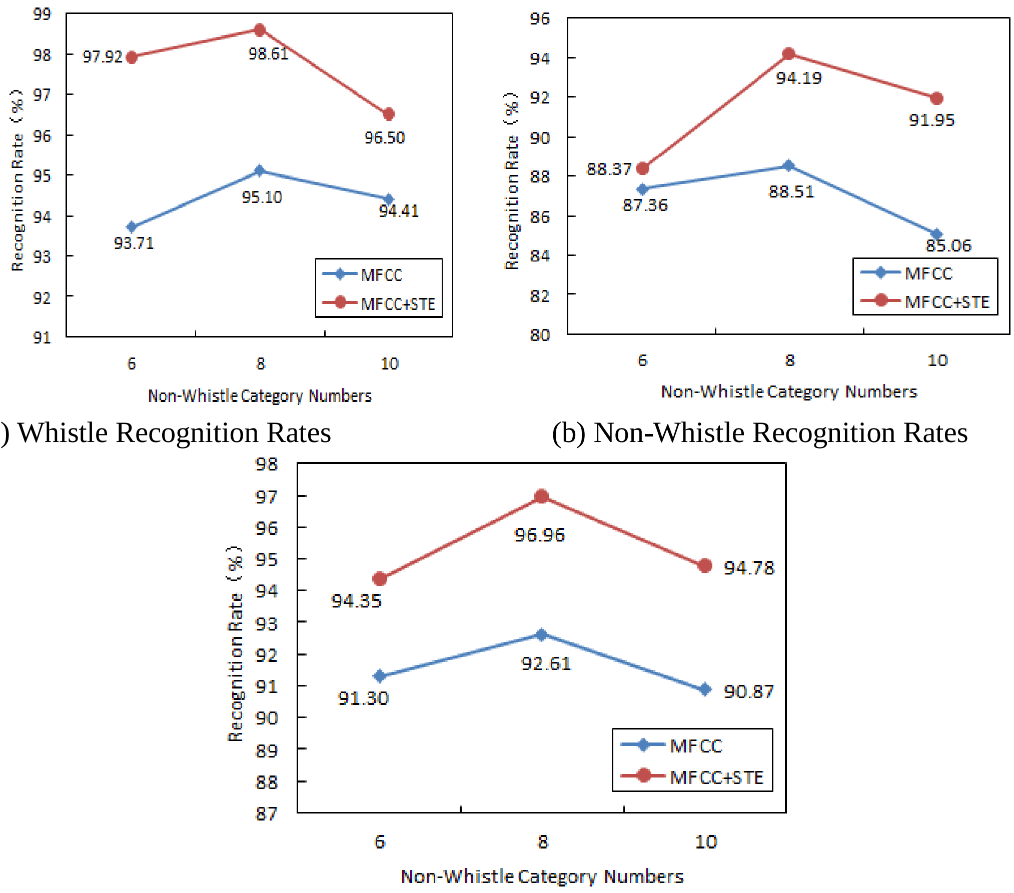
$$non-whistle\ recognition\ rate = \frac{the\ number\ of\ correctly\ identified\ non-whistle}{the\ total\ number\ of\ non-whistle\ samples} \times 100\% \quad (6)$$

$$comprehensive\ recognition\ rate = \frac{the\ number\ of\ correctly\ identified\ samples}{the\ total\ number\ of\ test\ samples} \times 100\% \quad (7)$$

In this paper, since the complexity of non-whistle is greater than whistle, so we mainly discussed how the category number of non-whistle affects training samples selection. The number of whistle categories was identified as 3 through contrast experiments. The evaluation criteria of the quality of sample selection are the recognition accuracy of HMM classifier and the time required for training and recognition when the category numbers of the obtained representative training sample set of non-whistle through cluster labels are different.

4.2. Experiment 1

In order to demonstrate the mixed feature parameters based on MFCC and STE can reflect the differences of sound samples better than traditional MFCC in complex sound environment, so we respectively experimented with traditional MFCC and the mixed features based on MFCC and STE in the case of different category numbers of the representative training sample set of non-whistle. The results are shown in Figure 3.



(a) Whistle Recognition Rates

(b) Non-Whistle Recognition Rates

(c) Comprehensive Recognition Rates

Figure 3. Recognition Rates of Traditional MFCC and the Mixed Feature under Different Category Numbers

We can conclude the following results from Figure 3.

(1) When the category numbers of training samples are the same, the whistle recognition rates of the mixed features increase at least 2.09% than traditional MFCC, the non-whistle recognition rates increase at least 1.01% and the comprehensive recognition rates increase at least 3.05%. The results show that the mixed features based on MFCC and STE indeed can represent the features of complex sound better than traditional MFCC.

(2) When we use the same feature, whether traditional MFCC or mixed features, the recognition accurate rates of non-whistle all are lower than whistle because non-whistle is more complex and has more noise.

(3) In the case of the same features, when we set the category number of the representative training sample set of non-whistle as 8, the whistle recognition rates, the non-whistle recognition rates and the comprehensive recognition rates are the highest. The results show that at this time, the obtained representative training sample set of non-whistle is the most reasonable.

(4) When we use the mixed features based on MFCC and STE, and set the category number of the representative training sample set of non-whistle as 8, the whistle recognition rate is 98.61%, the non-whistle recognition rate is 94.91% and the comprehensive recognition rate reaches to 96.96%. The three kinds of recognition rates are highest.

4.3. Experiment 2

On the basis of experiment 1, when we set the category number of the representative training sample set of non-whistle as 8, we added the experiment which selects the representative training sample sets by traditional manual annotation. Taking into account the convergence of models and the accuracy of parameters, we did many experiments and selected the best results which are shown in Figure 4.

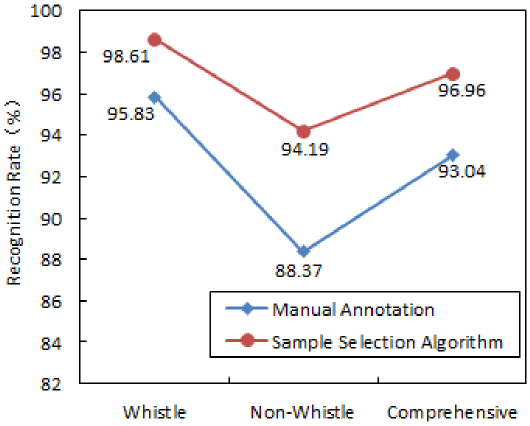


Figure 4. Manual Annotation Method and Sample Selection Algorithm

From Figure 4, we can conclude the following results

(1) In this experiment, training the samples selected by manual annotation may counter non-convergence problems, because the selection process is full of subjectivity and we may mistakenly train the two kinds of samples with quite different characteristics together. Therefore, if we want to get the right training sample sets, we have to do many experiments.

(2) Compared to manual annotation method, using the proposed training sample selection algorithm (cluster labels algorithm) to build training model libraries can significantly improve recognition accuracy. Because the manual annotation is full of subjectivity and uncertainty, the proposed algorithm is more objective and scientific.

(3) Compared cluster labels algorithm, the use of manual annotation methods generally require more training samples and more training time, and the experimental performances are worse than the cluster labels.

4.3. Contrastive Analysis of the Experimental Run Time

In the case of different category numbers of the representative training sample set of non-whistle, we respectively experimented with traditional MFCC and the mixed features based on MFCC and STE. The experimental time of training and recognition is compared in Figure 5 and Figure 6. (The experimental environment of this paper is as follows: CPU: Intel Core i5-3470; Dual-core 3.20GHz; RAM: 4GB; OS: win7 32 Bits)

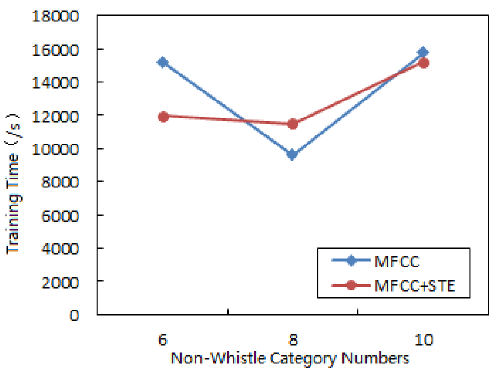


Figure 5. Comparison of Training Time

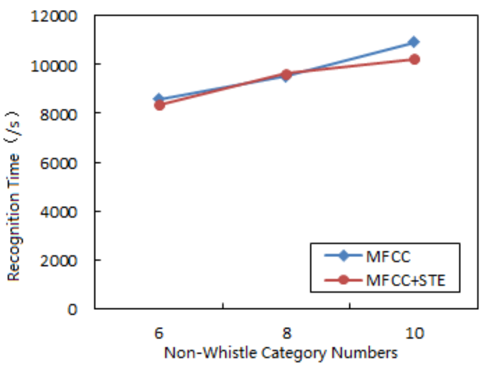


Figure 6. Comparison of Recognition Time

We can get the following results by Figure 5 and Figure 6.

(1)When the category numbers of the representative training sample set of non-whistle are the same, the run time of using the mixed features has little change in comparison with traditional MFCC. The result shows that adding STE feature can’t lead to increase operational complexity.

(2) According to the conclusions of Table I, and through Fig.4, we can find that when the category number of the representative training sample set of non-whistle is determined as 8, not only the recognition rate is the highest, but the training time is the shortest.

(3) The Fig.5 shows that with the increase of the category number of the representative training sample set of non-whistle, the recognition time will increase. These is because that with the increase of the category number, the number of obtained HMM models will increase. Thereby the time of matching models will increase in recognition phase.

5. Conclusion and Future Work

In this paper, we proposed a training samples selection algorithm based on double cluster labels to select representative training sets and selected the mixed feature parameter which contains MFCC and STE to represent the features of complex sound. Experimental results show that this method can improve the efficiency of the model and identification accuracy. MFCC and STE are basic features of sounds, so future work will consider exploring new and different features. The proposed algorithm uses K-means which is the most common clustering method and we can consider using other clustering methods and improving the clustering methods in the future work.

Acknowledgements

This research was financially supported by National Natural Science Foundation of China (Grant No. 71331002 and Grant No. 71571059) and Fund for Humanities and Social Science Fund Research Planning of the Ministry of Education (Grant No. 13YJA630037 and Grant No. 15YJA630010).

References

[1] G. Pop, A. Caranica, H. Cucu and D. Burileanu, “Sound event recognition in smart environments”, International Conference of Speech Technology and Human-Computer Dialogue, **(2015)**.

[2] S. Innami and H. Kasai, “NMF-based environmental sound source separation using time-variant gain features”, Computers &
Mathematics with Applications, vol. 64, no. 64, **(2012)**, pp. 1333-1342.

[3] G. Hinton, D. Li, Y. Dong and G. E. Dahl, “A Mohamed Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”, IEEE Signal Processing Magazine, vol. 29, no. 6, **(2012)**, pp. 82-97.

[4] M. Cutajar, E. Gatt, I. Grech and O. Casha, “Comparative study of automatic speech recognition techniques”, let Signal Processing, vol. 7, no. 1, **(2013)**, pp. 25-46.

[5] S. Chachada and C. C. J. Kuo, “Environmental sound recognition: A survey”, Apsipa Transactions on Signal & Information Processing, vol. 3, **(2014)**, pp. 1-9.

[6] P. Khunarsal, C. Lursinsap and T. Raicharoen, “Very short time environmental sound classification based on spectrogram pattern matching”, Information Sciences, vol. 243, no. 18, **(2013)**, pp. 57-74.

[7] M. L. Narayana and S. K. Kopparapu, “Effect of noise-in-speech on MFCC parameters”, Wseas International Conference on Signal, Speech and Image Processing, and, Wseas International Conference on Multimedia, Internet & Video Technologies, World Scientific and Engineering Academy and Society (WSEAS), **(2009)**.

[8] T. Kinnunen, R. Saeidi, F. Sedlak, A. L. Kong and J. Sandberg, “Low-Variance Multitaper MFCC Features: a Case Study in Robust Speaker Verification”, IEEE Transactions on Audio Speech & Language Processing, vol. 20, no. 7, **(2012)**, pp 1990-2001.

[9] J. H. Jeon and Y. Liu, “Automatic prosodic event detection using a novel labeling and selection method in co-training”, Speech Communication, vol. 54, no. 3, **(2012)**, pp. 445-458.

[10] S. Wang, Z. Li, C. Liu, X. Zhang and H. Zhang, “Training data reduction to speed up SVM training”, Applied Intelligence, vol. 41, no. 2, **(2014)**, pp. 405-420.

[11] S. Duan, J. Zhang, P. Roe and M. Towsey, “A survey of tagging techniques for music, speech and environmental sound”, Artificial Intelligence Review, vol. 42, no. 4, **(2012)**, pp. 637-661.

[12] S. Ntalampiras, “A Novel Holistic Modeling Approach for Generalized Sound Recognition”, IEEE Signal Processing Letters, vol. 20, no. 20, **(2013)**, pp. 185-188.

[13] S. Ntalampiras, I. Potamitis and N. Fakotakis, “Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions”, IEEE Transactions on Multimedia, vol.13, no. 4, **(2011)**, pp. 713-719.

Authors

Cuiqing Jiang*, he is a Professor, doctor. His main research areas: Data mining, Artificial intelligence, Business intelligence.

Peng Fan, he is a master. His main research areas: Data mining, Pattern recognition.

Kun Liang, he is a doctor. His main research areas: Data mining, business intelligence.

Zhao Wang, he is a doctor. His main research areas: Data mining, business intelligence.