Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 731 – 737

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Algorithm of Abnormal Audio Recognition Based on Improved MFCC

Chuan Xie[a],Xiaoli Cao[a],Lingling He[a]

[a]*College of computer science and information engineering, Chongqing technology and business university, Chongqing 400067, china*

Abstract

Characteristics extraction has a great effect on the audio training and recognition in the audio recognition system. MFCC algorithm is a typical characteristics extraction method with stable performance and high recognition rate. For the situation that MFCC has a large amount of computation, an improved algorithm MFCC_E is introduced. The computation of MFCC_E is reduced by 50% compared with the standard algorithm MFCC, and it make the hardware implementation is easy. The experimental result indicated that MFCC_E and MFCC have the same recognition rate roughly, yet the computational complexity of MFCC_E is much smaller.

Keywords：audio recognition; characteristics extraction; MFCC; MFCC_E; GMM

1. Introduction

In the field of target tracking, the video and the radio data are the two most important kinds of information, and in the past two decades, the video tracking has been in a dominant position, but its tracking performance will be greatly reduced when the tracking target is out of the observation range. Compared with video tracking, acoustic sensor has the advantages of low cost, small size, high efficiency, and the audio signal changes slowly over time, so the collected audio signal is stable and reliable. Therefore, the audio recognition and the audio target localization have become the research hotspots in recent years [1].

In the field of audio tracking, audio feature parameters and the choice of classifier will affect the

* Corresponding author. Tel.:+86-13368005360; fax:+86-23-62755167.
E-mail address: chuanxie@yeah.net.

doi:10.1016/j.proeng.2012.01.032

complexity and the recognition performance of the tracking system directly. The main classic algorithms of feature parameters extraction are the Linear Prediction Coefficient（LPC), the Linear Prediction Cepstrum Coefficient（LPCC) and the Mel Frequency Cepstrum Coefficient（MFCC) [2] [3] [4] etc. However, the application of all these algorithms need a amount of calculation, which will not only increase the cost, limit its application scope, and more importantly reduce the probability of its hardware implement. Based on the analysis of the standard MFCC algorithm, this paper presents an improved algorithm to extract audio characteristic parameters，which will get better recognition effect. Compared with taking the MFCC coefficient as the characteristic parameters single-handed, the new method can get the same recognition ratio, but its calculation is reduced obviously and it is more suitable for hardware implementation.

2. Audio recognition principle

Audio recognition is a pattern recognition process essentially, which includes some function modules, such as audio signal pre-processing, characteristics extraction, characteristics modeling (Reference Model Library), pattern matching etc ,and the principle of basic structure is shown in Figure 1.

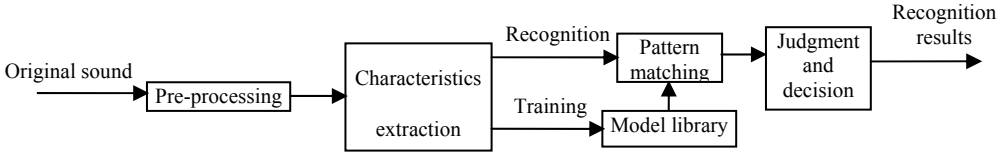

Fig.1 Block diagram of sound recognition system

A audio recognition system mainly includes two stages training and recognition. Both of them need the pretreatment to the original audio and the characteristics extraction.

2.1. pre-processing

Audio signal pre-processing includes filtering, A/D conversion, pre-emphasis, frame, endpoint detection etc. Assuming that， x(n) is the digital audio signal after the A/D conversion, the pre-processing process is shown as follows[5]:

1) Normalization process. Normalization is designed to eliminate the difference from different audio samples, and control the sample fluctuation value in the range of [- 1, + 1].

2) pre-emphasis. We usually use a digital filter with 6dB/ Octave to finish pre-emphasis, which is shown in (1). In formula (1), μ is a constant which is taken 0.97 usually.

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

3) Process the audio signal by window-framing. Although the audio signal is nonlinear and time-varying, it is steady in a short time, so its short-time characteristics can be extracted by framing with the frame-length in 10~30ms. For avoiding the big characteristic change between two frames, we take the half frame length as the frame shift, that is to say there is half overlapped data between the two adjacent frames. For the short-time analysis, we must extract audio signals through the windowing mode within the window. At the same time, the audio signal is 0 outside the window and the most widely used window function is the Hamming window. Generally it takes 256 points as a frame and the overlapped data is 128 points.

2.2. characteristics extraction

The choose of audio characteristics depends on the specific system ,and the representative characteristics are the amplitude, the zero-crossing ratio, the Linear Prediction Coefficient（LPC), the Linear Prediction Cepstrum Coefficient （LPCC) and the Mel Frequency Cepstrum Coefficient （MFCC),etc. In the characteristics extraction module, we will finish the following tasks such as analyzing and processing the acoustic signal, removing redundant information, and acquiring important information which will affect the audio recognition .With such an advantage as sorting the low-high frequency from the spectrum, Cepstrum is widely used in the field of audio recognition such as LPCC and MFCC.

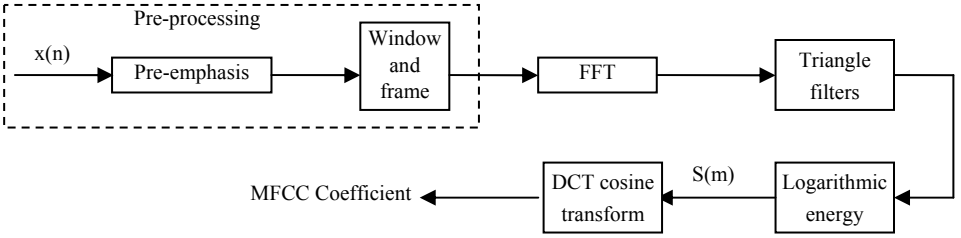

Fig.2 The MFCC coefficient extraction process

Considering the fast and instable changing of the acoustic signal in the time domain, the signal is usually transformed to the frequency domain to analyze the characteristic parameters. Mel Frequency Cepstrum Coefficient extraction process is shown in Figure 2[6]. The audio data after the pre-processing will be calculated to get the spectral parameters of every frame data through by the FFT, and then use the spectral parameters data of each frame to complete a convolution operation with the Mel frequency filters which is consists of a group of M triangular band-pass filters(M is 20~40 usually).Then process the output from each frequency band with a Log operation to calculate the Log Energy S(m)， m=1,2,3,……N. Finally calculate the N parameters with discrete cosine transform to find Mel Cepstrum coefficients as the audio characteristics, which is shown in formula (2). And in formula (2),n is the number of MFCC, Ci(n) is the n-th MFCC coefficients of the i-th frame, S(m) is the logarithmic power spectrum of the audio signal, and M is the number of triangular filters[5].

$$C_i(n)=\sum_{m=1}^{M}S(m)\cos\left[\frac{\pi n(m-0.5)}{M}\right],\quad 0\le m<M \tag{2}$$

2.3. training and recognition

Audio recognition system requires establishing an audio library to finish the pre-process and characteristics extraction to the audio sample， and the audio library will be trained by the classifiers. The widely used classifiers are the Support Vector Machine(SVM)[7], the Hidden Markov Model(HMM)[8] and the Gaussian Mixture Model(GMM)[9].

Essentially GMM is a multi-dimensional probability statistical model based on parameter estimation, it is considered that the characteristics of every audio will form a specific distribution in the characteristic space, and the distribution can be combined by a plurality of Gauss distributions. Gauss distribution combination of different parameters can be used to characterize the different voices, i.e., each kind of audio characteristic parameter is corresponding to a GMM.

GMM has been widely used in speaker recognition and speech recognition and the training process follows the method in the references [10] and the method is described as follow: extract the feature parameter vector from the training samples to train GMM, for the audio recognition system with a variety of abnormal audio , each kind of audio will be replaced by a GMM to get the model parameters of each kind of abnormal audio, and ultimately get the complete GMM formula (3)to describe each kind of abnormal audio .Among them, P_i is the Weight of the mixed component ; μ_i is the mean vector; Σ_i is covariance matrix, and N is the step of mixing.

$$\lambda = \{P_i, \mu_i, \Sigma_i\}; \quad i=1,2,\ldots\ldots N \tag{3}$$

The recognition process is that make use of the feature vector extracted from the test samples combining with the GMM classifier to get the recognition result of each type of test samples through by seeking the maximum posterior probability. Finally, we add all the recognition results of test samples of each kind together to get the general recognition ratio of each type of audio.

3. Improved algorithm

In the characteristics extraction process, the modules such as pre-emphasis, windowing, framing, FFT, filters, logarithmic energy, cosine transform all contain a lot of multiplication, and among them, the multiplication in the FFT module occupies most of the entire treating processes. Too much multiplication leads to high requirement for capability, large energy consumption, lower stability, and narrow application range to the system.

3.1. Improved characteristics extraction process

In the new arithmetic, adjust the framing module of the pre-processing after the triangle filter group, and the improved MFCC characteristics extraction process called as MFCC_E which is shown in Figure 3.

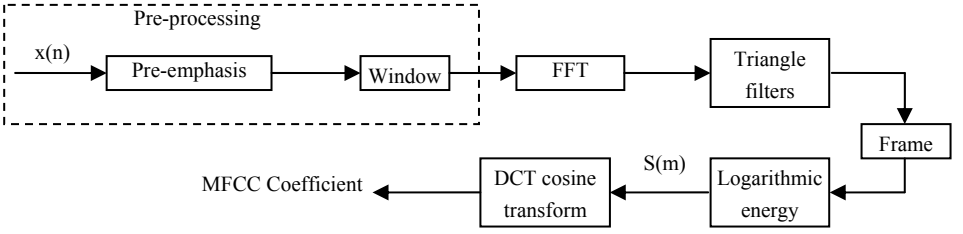

Fig.3 The MFCC_E coefficient extraction process

In essence, the windowing module itself has a framing function, because we can extract a frame audio data from each window. Set the window length to make sure that each frame of audio data contains 128 points and there is no duplicate data between the two adjacent frames, so although the frame length becomes shorter, the total frames is consistent with the original algorithm.

The improved work principle is shown in figure 4.fn, fn+1, fn+2 are the audio frames in the original algorithm with a 256-spots frame length, and there are 128 overlapped spots between two adjacent frames. sn+1, sn+2 are the output frames from the MFCC_E windowing module with a 128- points frame length, and there is not overlapped data among the frames. The outputs become Sn, k, Sn+1, k, Sn+2, k is the output after the FFT operation and the Mel filter modules, and we get the logarithm energy spectrum Fn, k, Fn+1, k with the superposition of the adjacent outputs. The FFT operation and the Mel filter method are the same with the original MFCC algorithm. However, the frame size is only the half of the original

algorithm,256 points reduce to 128 points, so the FFT module needs to process the 128-points audio frames only, that is to say the computation is the half of the original algorithm, Similarly, Mel filter's computation also is half of the original algorithm. Therefore, the computation will be reduced by 50% through by turning the framing module after the Mel filter.

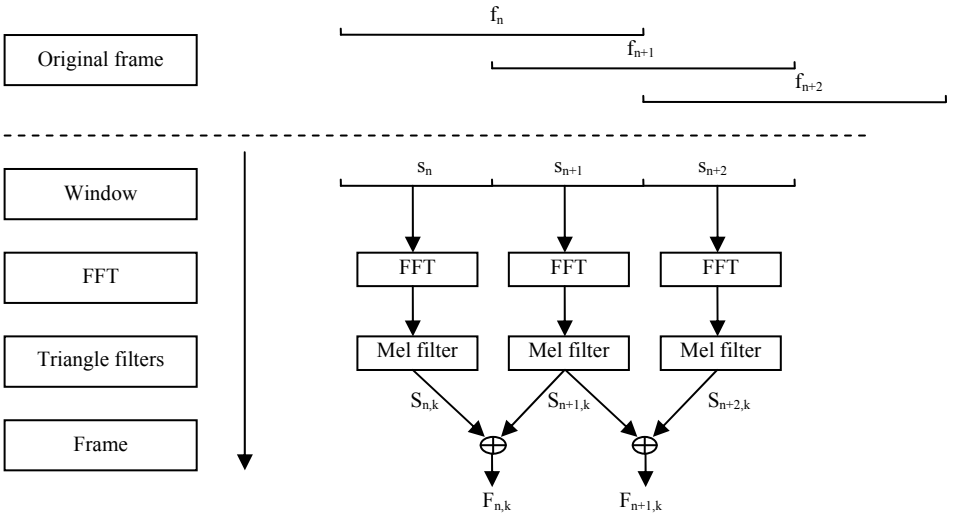

Fig.4 The framing output of MFCC_E

3.2. Improved pre-emphasis parameters

We follow formula (1) to finish the pre-emphasis module, and we take 0.97 as the value of μ. If 0.97 is replaced by 31/32, the formula (1) will evolve into formula (4).

$$H(z) = 1 - \frac{31}{32} z^{-1} = 1 - (1 - \frac{1}{32}) z^{-1} \tag{4}$$

Although there is not much difference between 31/32 and 0.97, the benefits are very obvious. Because we can get 1/32 through by shifting operations (right shift 5 bits), so the original multiplication in pre-emphasis module can be replaced by shifting and addition operations, which make the hardware implement is easy.

4. Experimental results

Apply the improved algorithm to the forest burglary protection system, choose the forest as the experiment background, and take three kinds of abnormal sounds which are easy to appear in burglary trees such as cutting down trees, sawing the tree and trees collapse as the experiment source material.

4.1. Experimental environment

The experiment run on a computer, whose basic frequency is 2.66GHz, the memory is 2GB, the operating system is Windows XP, and the simulation software is Matlab. The experiment sounds are from cutting down trees, sawing trees and trees collapse. Take 40 samples from each kind of sound，so the

sampling rate is 16khz quantified as 16bit, the original algorithm frame length is 256 spots, and the improvement algorithm frame length is 128 spots.

Take 80% from the total samples randomly as the training sample, then the remaining 20% is the recognition sample. Complete each group experiment in five times, show the average recognition ratio of each kind of sound, at last calculate the average recognition ratio with all sounds in the same GMM order.

4.2. GMM training library

Sample training process is shown in Figure 5. Extract characteristic parameters from the train sample, then carry out module train on all kinds of abnormal sound sample, finally get the three abnormal sound models.

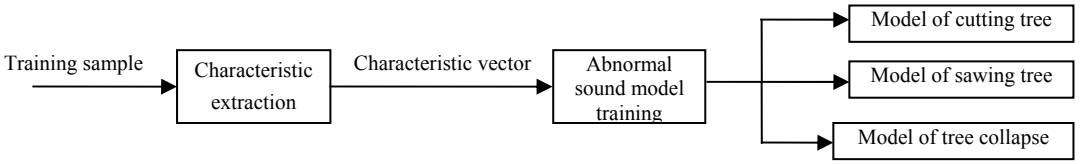

Fig.5 GMM training module

4.3. Experimental data

Table 1 and table 2 show the recognition ratio of the original MFCC algorithm and the MFCC_E to the abnormal sound respectively, and N is the GMM mixed order. The test data shows that there is not big difference between the recognition ratios of MFCC and MFCC_E, and both of them have a high recognition ratio, moreover the recognition ratio increases along with the GMM mixed order.

Tab.1 recognition ratio of the original MFCC algorithm to the unusual sound

sample	recognition ratio		
	N=12	N=16	N=20
sound of cutting tree	0.925	0.975	0.975
sound of sawing tree	0.85	0.9	0.95
sound of tree collapse	0.75	0.775	0.875
average recognition ratio	0.841	0.883	0.933

Tab.2 recognition ratio of the original MFCC_E algorithm to the abnormal sound

sample	recognition ratio		
	N=12	N=16	N=20
sound of cutting tree	0.925	0.925	0.975
sound of sawing tree	0.825	0.9	0.925
sound of tree collapse	0.75	0.8	0.9
average recognition ratio	0.833	0.875	0.933

5. Conclusion

This article introduces a kind of sound characteristic parameter extraction algorithm MFCC_E, which is applied in the GMM training and the recognition, and the experiment proved that MFCC_E algorithm can make the same effect with MFCC. The recognition rate can reach to 90% if the GMM mixed order is appropriate. The essential superiority of MFCC_E lay in reducing the complexity of the algorithm. Compared with the MFCC algorithm the computation of MFCC_E reduced by 50%, moreover the pre-emphasis module might be done by hardware directly.

The experimental result showed that the higher the GMM mixed order is, the better the recognition ratio of abnormal sound is, but the higher order causes the number of parameter increase, even lose the convergence model when training the data. So, how to choose the appropriate mixed order will be the key in the future research.

Acknowledgements

The Chongqing Science and Technology Committee Found (Grant No. CSTC,2011AC2136)..The Chongqing education committee Found（Grant No. KJ100709）

References

[1]Zajdel W, Krijnders J D, Andrnga T. Sound-video sensor fusion for aggression detection [C]. Proceedings of the 2007 IEEE International Conference on Advanced Video and Signal based Surveillance. London: IEEE Computer Society,2007: 200 - 205.

[2] ZHANG Ling-hua,ZHENG Bao-yu,YANG Zhen.. A Study of Feature Parameters Based on LPC Analysis with Applications to Speaker Identification[J]. JOURNAL OF NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS,2005, 25(6):1-6.

[3] Rong Wei,Tao Zhi, Gu Ji-hua.. Identification of Chinese whispered speech based on modified LPCCand [J].computer engineering and applications,2007,43(30):213-216.

[4]Lee C H, Chou C H, Han C C. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis[J]. Pattern Recognition Letters,2006,27(2):93-101.

[5]Lv Xiao-yun,Wang Hong-xia.Abnormal audio recognition algorithm based on MFCC and short-term energy [J]. JOURNAL OF COMPUTER APPLICATIONS,2010,30(3):796-798.

[6]Wang J C,Wang J F, Wang Y S. Chip design of MFCC extraction for speech recognition [J]. Integration,2002, 32 (1/2) : 111 - 131.

[7]Rabaou I A, Davy M, Rossignol S. Using one-class SVMs and wavelets for audio surveillance [J]. IEEE Transactions on Information Forensics and Security, 2008,3(4):763-775.

[8]Rabaou I, Lachir I Z, Ellouze N. Using HMM-based classifier adapted to background noises with improved sounds features for audio surveillance application [J]. International Journal of Signal Processing, 2008,5(1):46-55.

[9]Radhakr I R, Divakaran A, Smaragdis A. Audio analysis for surveillance applications [C]. Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Washington, DC: IEEE Computer Society, 2005:158-161.

[10]Hu Yi-ping. Research and Implementation of Speaker Recognition Based on GMM [D]. Xiamen: Xiamen university, 2007.