Journal of Computer Applications (JCA)
ISSN: 0974-1925, Volume V, Issue 1, 2012

Audio Classification Using Support Vector Machines and Independent Component Analysis

V. Elaiyaraja[a,*], P. Meenakshi sundaram [b,1]

Abstract - **In this paper, we present a new audio classification system. First, a frame-based multiclass support vector machine (SVM) for audio classification is proposed. The accuracy rate has significant improvements over conventional file-based SVM audio classifier. In feature selection, this study transforms the log powers of the critical-band filters based on independent component analysis (ICA). This new audio feature is combined with linear prediction coefficients (LPC) – derived cepstrum coefficients (LPCC), Mel Frequency Cepstral coefficients(MFCCs) perceptual features to form an audio feature set. The superiority of the proposed system has been demonstrated via a 6-class sound database with a 91.7% accuracy rate.**

Index Terms **– support vector machine (SVM), linear prediction coefficients (LPC), Mel Frequency Cepstral coefficients(MFCCs), independent component analysis (ICA).**

I. INTRODUCTION

Audio classification, especially the speech or music discrimination, has been becoming a focus in the research of audio processing and pattern recognition. It can be used in a lot of areas and applications, such as audio segmentation, audio indexing and retrieval, automatic speech recognition and etc. Audio is usually treated as an opaque collection of bytes with only the most primitive fields attached such as file name, file format, sampling rates, etc. Users accustomed to searching, scanning, and retrieving text data may be frustrated by the inability to look inside the audio objects. Generally speaking, audio classification is a pattern classification problem. Firstly, the needed features are extracted from the audio signals. Then audio snippets are classified using some learning algorithms. The efficacy of an audio classification or categorization system depends on the ability to capture proper audio features and to accurately classify each feature set corresponding to its own class.

A closely related area, research on audio classification is relatively new. Wold *et al*. [4] presented an audio retrieval system named Music Fish based on audio classification. This work is a milestone about audio retrieval because of the content based analysis which distinguishes it from previous works. In this system, pitch, harmonicity, loudness, brightness and bandwidth were used as the audio features. The nearest neighbor NN) rule was adopted to classify the query audio into one the defined audio classes.

Foote [1] proposed the use of mel-frequency cepstral coefficients (MFCCs) plus energy as audio features. The classification procedure was also done by the NN rule. Pfeiffer *et al*. [2] adopted a filter bank consisting of 256 phase-compensated gamma phone filters proposed to extract audio features. More recently, Li [3] concatenated the perceptual and cepstral feature sets for audio classification. A new classifier name nearest feature line (NFL) for audio classification was also presented and produced better results than the NN-based and other conventional methods. This study was improved by the author's later work [5]. With the same feature set as [3], Guo and Li used support vector machines (SVMs) with a binary tree structure to tackle the audio classification problem. Experimental results showed that the SVM approach with perceptual and cepstral feature sets achieved lower error rate than Music Fish system and NFL approach. This SVM classifier was also adopted by Lin *et al*. [6] for audio classification. With similar feature set as that of [5], Lin *et al*. applied wavelet to extract subband power and pitch information. Also, the MFCCs are replaced bfrequency cepstral coefficients.

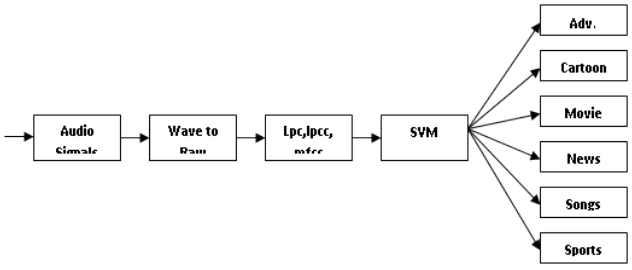


Figure 1. Block diagram of the proposed audio classification system.

In the works of [4-6], the means and standard deviations of all individual features in an audio file were computed over the nonsilent frames to form a feature vector of an audio file. In this paper, we present a frame-based audio feature set and a frame-based multiclass SVM for audio classification. The block diagram of the proposed system is depicted in Fig. 1. In accordance with our experiments, the presented In the works of [4-6], the means and standard deviations of all individual features in an audio file were computed over the nonsilent frames to form a feature vector of an audio file. In this paper, we present a

Manuscript received 20/Feb/2012.

V. Elaiyaraja[a,*],
Assitant Proffessor,Arasu Engg.college,Kumbakonam
E-mail: rajaelaiya@gmail.com.
P.Meenakshi sundaram [b,1],
Assitant Proffessor,
Mohamed sathak college of arts & Science, Chennai,
E-mail: sundaramjkm@gmail.com

frame-based audio feature set and a frame-based multiclass SVM for audio classification. The block diagram of the proposed system is depicted in Fig. 1. In accordance with our experiments, the presented frame-based strategy significantly outperforms the file-based approaches adopted by [5], [6]. In the feature selection, besides total spectrum power, subband powers, brightness, bandwith, pitch and MFCCs, a new audio feature based on independent component analysis (ICA) is also presented. For computing MFCCs, the log powers of the critical-band filters are obtained first. Discrete cosine transform (DCT) is then applied to the log powers to remove some of their correlations. Instead of DCT, ICA transform is applied to the log powers to generate a new audio feature, mel-frequency ICA-based feature. The prominence of the proposed audio feature is the theoretical maximization of the statistical independence among all the critical-band log powers

II. FEATURE EXTRACTION

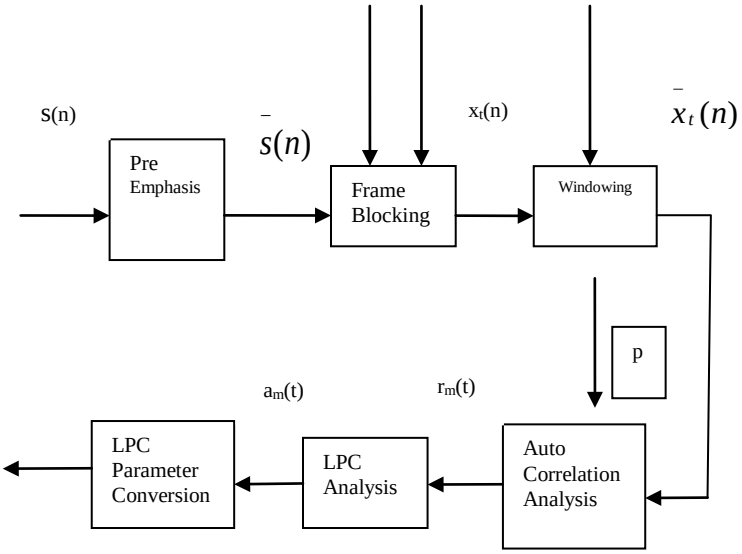
2.1 Linear Prediction Coefficients

The theory of linear prediction (LP) is closely linked to modelling of the vocal tract system, and relies upon the fact that a particular speech samples may be predicted by a linear weighted sum of the previous samples. The number of previous samples used for prediction is known as the order of predictions. The weights applied to each of the previous speech samples are known as linear prediction coefficients (LPC). They are calculated so as to minimize the prediction error.

2.2 Linear Prediction Cepstral Coefficients

In many applications, Euclidean distance is used as a measure of similarity or dissimilarity between feature vectors. The sharp peaks of the LP spectrum may produce large errors in a similarity test, even for a slight shift in the position of the peaks. Hence, linear prediction coefficients are converted in to cepstral coefficients using a recursive relation. Cepstral coefficients represent the log magnitude spectrum, and the first few coefficients model the smooth envelope of the log spectrum [1]. These coefficients can be obtained either from linear prediction coefficients or from the inverse discrete fourier transform (IDFT) of log magnitude spectrum of the speech signal. In both cases, the process results in estimating vocal tract system characteristics from the speech signal [3]. Atal explored the LP derived cepstral coefficients, and proved their effectiveness over LPC and other features such as pitch and energy contours.

2.3 LPC Processor for Speech Signal Classification



The fig.3.1 shows the block diagram of LPC processor. The basic steps in the processing include the following:

2.2.1. Preemphasis

The digitized speech signal, s(n), is put through a low order digital system (typically a first order FIR filter), to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The digital system used in the preemphasizer is either fixed or slowly adaptive (eg to average transmission condition, noise background, or even to average signal spectral). Perhaps most widely used emphasis network is the fixed first order system.

$$H(z) = 1 - \bar{a}z^{-1}, \quad .9 \le a \le 1.0 \qquad (3.1)$$

In this case, the output of the pre-emphasis network, $\bar{S}(n)$, is related to the input network s(n), by the difference equation.

$$\bar{s}(n) = s(n) - \bar{a}s(n-1) \qquad (3.2)$$

The most common value for $\bar{a}$ is around 0.95 (for fixed-point implementations a value of $\bar{a} = 15/16 = 0.9375$)

A simple example of a first order adaptive preemphasizer is the transfer function.

$$H(z) = 1 - \bar{a}_n z^{-1} \qquad (3.3)$$

Where $\bar{a}_n$ changes with time(n) according to the chosen adaption criterion. One possibility is to choose $\bar{a}_n = r_n(1)/r_n(0)$. The magnitude characteristic of $H(e^{j\omega})$ For the value $\bar{a} = 0.95$. It can be seen that at $\omega = \pi$ (half sampling rate) there is a 32 dB boost in the magnitude over that at $\omega = 0$.

2.2.2. Frame Blocking

In this step that preemphasized speech signal, $\bar{s}(n)$, is blocked in to frames of N samples, with adjacent frames being separated by M samples. Fig.3.1 illustrates the blocking into frames for the case in which M = (1/3)N. The first illustrated frame consists of the first N speech samples.

The second frame begins M samples after the first frame, and overlaps it by N-M samples. Similarly the third frame begins 2M samples after the first (or M samples after the second frame) and overlaps it by N-2M samples. This process continues until all the speech is accounted for within one or more frames. It's easy to see that if M = N, then adjacent frames overlap, and the resulting LPC spectral estimates will be correlated from frame to frame. If M << N then LPC spectral estimates from frame to frame quite smooth. On the other hand if M > N. there will be no overlap between adjacent frames. In fact some of the speech signal will be totally lost (i.e., never appear in any analysis frame) and the correlation between the resulting LPC spectral estimates of adjacent frames will contain a noisy component whose magnitude increases as M increases. This situation intolerable in any practical LPC analysis for speech recognition .if we denote the ℓ^{th} frame of speech by $x_l(n)$ and there are L frames within the entire speech signals, then

$$x_1(n) = \bar{s}(M\ell + n), n = 0,1,..., N-1, \ \ell = 0,1,..., L-1 \quad (3.4)$$

That is the first frame of speech, $x_0(n)$, encompasses speech samples

$$\bar{s}(0), \bar{s}(1),...., \bar{s}(N-1) \quad (3.5)$$

The second frame of speech, $x_1(n)$, encompasses speech sample

$$\bar{s}(M), \bar{s}(M+1),...., \bar{s}(M+N-1) \quad (3.6)$$

And the L^{th} frame of speech, $x_{L-1}(n)$, encompasses speech samples

$$\bar{s}(M(L-1)), \bar{s}(M(L-1)+1),...., \bar{s}(M(L-1)+N-1) \quad (3.7)$$

Typical values for N and M are 300 and 100 when the sampling rate of speech is 6.67kHz.These corresponds to 45-msec frames, separated by 15msec, or a 66.7Hz frame rate.

2.2.3. Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is identical to the one discussed with regard to the frequency domain interpretation of the short-time spectrum, to use the window to taper the signal to 0 at the beginning and end of each frame. If we define the window as w(n), 0 = n = N-1, then the result of windowing is the signal

$$\bar{x}_\ell(n) = x_\ell(n)w(n), \qquad 0 \le n \le N-1 \quad (3.8)$$

A "typical" window used for the autocorrelation method of LPC (the method most widely used for recognizing system) is the Hamming window which has the form

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \ 0 \le n \le N-1 \quad (3.9)$$

2.2.4. Autocorrelation Analysis

Each frame of windowed signal is next auto correlated to give

$$r_\ell(m) = \sum_{n=0}^{N-1-m} \bar{x}_\ell(n)\bar{x}_\ell(n+m), \qquad m = 0,1,...., p \quad (3.10)$$

Where the highest autocorrelation value p, is the order of the LPC analysis. Typically values of p from 8 to 16 have been used, with p=8 being the value used for most systems to be described in the book. A side benefit of the autocorrelation analysis is that the 0^{th} autocorrelation, $R_l(0)$, is the energy of the l^{th} frame. The frame energy is an important parameter for speech-detection systems.

2.2.5. LPC Analysis

The next processing step is the LPC analysis, which converts each frame of p+1 autocorrelation into an "LPC parameter set" in which the set might be the LPC coefficients, the reflection (or PARCOR) coefficients, the log area ration coefficients, the cepstral coefficients, or any desired transformation of the above sets. The formal method for converting from autocorrelation coefficients to an LPC Parameter set (for the LPC autocorrelation method) is known as Durbin's method and can formally be given as the following algorithm (for convenience, we will omit the subscript l on $r_l(m)$):

$$E^{(0)} = r(0) \quad (3.11)$$

$$k_i = \left\{ ri) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right\} / E^{(i-1)} \quad (3.12)$$

$$\alpha_i^{(t)} = k_i \quad (3.13)$$

$$\alpha_j^{(t)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-j)} \quad (3.14)$$

$$E^{(t)} = (1 - k_i^2) E^{(i-1)} \quad (3.15)$$

Where the summation in Eq. (3.12) is the omitted for i=1. the set of equations (3.11 - 3.15) are solved recursively for i=1,2…..,p and the final solutions is given as

a_m = LPC coefficients = $\alpha_m^{(p)}$, $1 \le m \le p$ (3.16)

k_m = PARCOR coefficients (3.17)

g_m = log area ratio coefficients = $\log\left(\frac{1-k_m}{1+k_m}\right)$ (3.18)

2.2.6. LPC Parameter Conversion to Cepstral Coefficients

A very important LPC parameter set, which can be derived directly from the LPC coefficients set, is the LPC cepstral coeffients, c(m). The Recursive relation between LP coefficients and LPC coefficients is given by

$$c_0 = \ln \sigma^2 \quad (3.19)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \quad (3.20)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \qquad m > p \quad (3.21)$$

Where σ^2 is the gain term in the LPC model. The cepstral coefficients, which are the coefficients of the Fourier transform representation of the log magnitude spectrum, have been shown to be a more robust, reliable feature set for speech recognition than the LPC coefficients, the PARCOR coefficients, or the log area ratio coefficients. Generally, a cepstral representation with q > p coefficients is used, where

$$q \approx \left(\frac{3}{2}\right)p \qquad (3.22)$$

2.3 Mel Frequency Cepstral Coefficients

MFCCs have been widely used in the field of speech recognition and are able to represent the dynamic features of a signal as they extract both linear and non-linear properties. MFCC can be a useful tool of feature extraction in vibration signals as vibrations contain both linear and non-linear features. The Mel-frequency Cepstral Coefficients (MFCC) is a type of wavelet in which frequency scales are placed on a linear scale for frequencies less than 1 kHz and on a log scale for frequencies above 1 kHz. The complex cepstral coefficients obtained from this scale are called the MFCC. The MFCC contain both time and frequency information of the signal and this makes them useful for feature extraction.

III. Experimental Results

3.1 Summary of the Work

In this project, we have taken four different categories of audio signals namely news, advertisement, sports and music. We have extracted the LPC, LPCC & MFCC features for these signals and are given as input to SVM. The SVM is used as a modelling technique in this work for audio classification. The SVM in multi-class mode is used for audio classification. The experimental results shown that the method achieves an accuracy of about 94%.

Category	LPC	LPCC	MFCC
Advt.	42.50	82.50	55%
Cartoon	30.00	47.50	48%
Movie	37.50	55.00	60%
News	65.00	95.00	60%
Songs	80.00	47.50	38%
Sports	82.50	90.00	58%

Samples	LPC	LPCC	MFCC
1 Sec	43 %	65 %	24 %
2 Sec	60 %	67 %	50 %
5 Sec	62 %	69 %	62 %
10 Sec	60 %	77 %	76 %

IV. Conclusion

This study has presented an audio classification system based on frame-based multiclass SVM. Compared with conventional file-based SVM classifier, the proposed audio classifier has considerable improvement averagely 13.9%. This study also presents a new feature set including LPC, LPCC, MFCCs ICA-based feature, MFCCs and perceptual features. With the proposed audio classifier and feature set, the accuracy rate can achieve 9 1.7%. Future works include the expansion of audio classes and the study of audio segmentation which is a preprocess for audio classification in many related applications.

References

[1] T. Foote, “Content-based retrieval of music and audio,” Multimedia Storage and Archiving Systems II, Proc. of SPIE, vol. 3229, pp. 138–147, 1997.

[2] S. Pfeiffer, S. Fischer, and W. E. Elsberg, “Automatic Audio Content Analysis,” Tech. Rep. 96-008, Univ. Mannheim, Mannheim, Germany, Apr. 1996.

[3] S. Z. Li, “Content-based audio classification and retrieval using the nearest feature line method”, IEEE Transactions on Speech and Audio Processing, vol. 8, no. 5, pp. 619-625 Sept. 2000.

[4] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-based classification search and retrieval of audio,” IEEE Multimedia Magazine, vol. 3, pp. 27–36, July 1996.

[5] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” IEEE Transactions on Neural Networks, vol. 14, no. 1, pp. 209-2 15, Jan. 2003.

[6] C. C. Lin, S. H. Chen, T. K. Truong, and Y. Chang, “Audio classification and categorization based on wavelets and support vector machine,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 644-651, Sept. 2005.

[7] Guodong Guo and Stan Z. Li, “Content-Based Audio Classification and Retrieval by Support Vector Machines”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 14, NO. 1, JANUARY 2003.

[8] S. Esmaili, S. Krishnan and K. Raahemifar, “Content Based Audio Classification And Retrieval Using Joint Time-Frequency Analysis”, IEEE,ICASSP 2004, VOL. 665.

[9] Pawan Lingms, Cory Butz, “Interval Set Classifiers using Support Vector Machines*”, IEEE.2004, Vol. 0-7803-8376-1/04/.

[10] Stan Z. Li_ GuoDong Guo, “Content-Based Audio Classification and Retrieval Using SVM Learning”, Microsoft Research China, 5/F Beijing Sigma Center, Beijing 100080, China.

[11] Stan Z. Li, “Content-Based Audio Classification and Retrieval Using the Nearest Feature Line Method”, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 5, September 2000

Biography

V.Elaiyaraja is currently working as Assistant professor in the Department of Information Technology, Arasu Engineering College, Kumbakonam, TamilNadu. He obtained his B.Tech in Information Technology from SCAD College of Engineering and Technology, Tirunelveli in 2005 and his M.E degree in Computer Science and Engineering from Annamalai University in 2008. He has about 2 years of Industrial experience in the field of Mobile Communications and around 4 years of teaching Experience. He has published research papers in International Journals and National Conference. He has coordinated and organized two national conferences and one National level Technical Symposium. His areas of Interest are Mobile Computing, Networks and Algorithms.

P.Meenachi Sundaram working as Assistant professor in computer science department, Mohamed Sathak Group of Institutions in Chennai. He obtained in MCA from Thanthai Hans Rover college, Permabulur and his M.Phil., Degree from Annamalai University, Chidambaram. He has submitted Ph.D., thesis in Dravidan University, Kuppam, Andra Pradesh. He has above five years of Teaching Experience in software engineering. He has published research papers in International Journals of Action Research Engineering and Eradicate poverty.