IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 3, Issue 1 (Sep. – Oct. 2013), PP 22-28
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Content Based Environmental and Natural Sounds Classification Using SVM

[1]Jithina T. S.,[2]Renjith R. J.
[1]*M.Tech Student,Electronics and Communication Department, SCT College of Engineering Pappanamcode, Kerala, India*[1]
[2]*Asst.Proffessor, Electronics and Communication Department, SCT College of Engineering Pappanamcode, Kerala, India*[2]

Abstract: *Audio signal classification system analyzes the input audio signal and label the signal to a class. The categorization can be done on the basis of pitch, loudness, rms value of signal etc. The signal classifier analyzes the content of the audio format thereby extracting information about the content from the audio data. A number of spectral and temporal features and Mel Frequency Cepstral Coefficients are used for classification purposes. In this paper the implementation of the audio signal classification using Suport Vector Machine (SVM) is presented. Finally the confusion matrix and overall accuracy has been studied in order to evaluate performance of the classification system.*
Keywords**:** *Audio feature extraction, loudness, Mel Frequency Cepstral Coefficients, pitch, spectral centroid, spectral flux, spectral sparsity, and spectral roll off, SVM*

I. Introduction

Audio, which includes various types of voice, music, and various kinds of environmental and natural sounds, is an important type of media, and also a significant part of audiovisual data [1]. Compared to research done on content-based image and video database creation, classification and retrieval, very little work has been done on the audio part of the multimedia bit stream [4]. Today content based feature extraction and classification of audio used in many applications such as audio sensors in surveillance and monitoring application, speech recognition, speaker identification, music retrieval applications [2]. Traditional tasks in the area of the automatic audio classification and matching are speech/music segmentation, classification and audio retrieval [1]. Audio classification and retrieval is an important and challenging research topic and is still at a preliminary stage. Compared to speech and music very little work done on the environmental and sounds.

One of the first content-based indexing and retrieval of audio was the Muscle Fish database [9]. In this paper, a short-time analysis was performed with four spectral features and the root mean square (rms) level of the sound. Then mean, variance, and autocorrelation of each feature were stored and used for the classification of all sounds in the database. The temporal feature integration method was adopted to classify the sound events and is quite complex and high computational cost [2]. Mel Frequency Cepstral Coefficients (MFCCs) and energy as audio features and the classification procedure were also done by the Neural Network proposed by Foote proposed [3]. Experimental results showed that the Support Vector Machine (SVM) approach with various acoustic features achieved lower error rate compared to other system [6].The whole system is simple to implement and computational cost is very less with high accuracy.

Before any audio signal can be classified under a given class, the features in that audio signal are to be extracted. These features will decide the class of the signal. The block diagram of the proposed system is shown in Figure. 1. In the feature selection, various temporal and spectral features of the sound signals are selected.

II. Audio Feature Extraction

An efficient classification scheme for audio typically based upon the fundamental process of acoustic feature extraction. An audio feature is any qualitatively or quantitatively measurable aspect of a sound. The audio features can be classified into different ways. Commonly audio features can be classified as physical features and perceptual features. Physical features refer to mathematical measurements computed directly from the sound wave, such as energy function, spectrum, the harmonicity [2], the Zero Crossing Rate (ZCR) and the cepstrum. Perceptual features are related to the perception of sounds by human beings, including loudness, pitch, timbre, and rhythm.

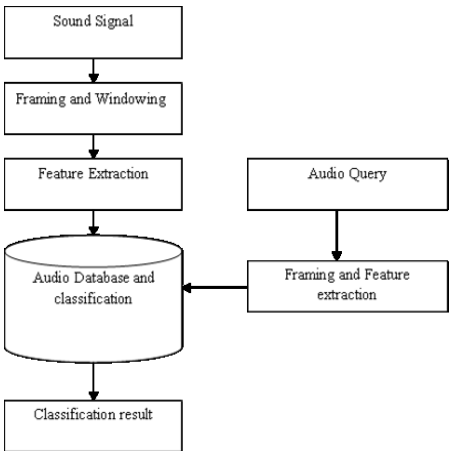

Figure1.Block diagram of the audio classification system using SVM

Alternatively the audio features can be classified as temporal features and spectral features. The temporal domains are the native domain of the sound signal and extracted directly from the audio signal without using any transformation therefore the complexities of these features are less [3].The temporal features include zero crossing rate, pitch, loudness, entropy etc. The spectral features are computed from the Short Time Fourier Transform of the signal including spectral centroid, spectral sparsity, spectral roll off etc. The figure 2 shows the block diagram of audio feature extraction [5].

2.1. Zero Crossing Rates (ZCR)

Zero Crossing Rates is the rate of sign-changes along a signal. In other words it is a measure of number of times the signal value crosses the zero axis rated by the number of values of the signal. Periodic sounds tend to have small ZCR, while noisy sounds tend to have high ZCR and are a simple measure of the frequency content of a signal. ZCR for each frame is defined as follows[9].

$$Z_n = \sum_{m=0}^{N} \left| \text{sgn}[x(m)] - \text{sgn}[x(m-1)] \right| w(n-m) \quad (1)$$

2.2. Loudness

The amplitude of a sound wave determines its loudness. There is minimum amplitude required for you to hear a sound. This varies with thespecies ofanimal. Sound loudness is related to the strength of a sound perception of the ear. It is related to the intensity of the sound signal. The loudness of the sound defines as the root mean square level in decibels of the windowed frame of the sound. The loudness for each frame is defined as follows [1].

$$\text{Loudness} = 20 \log_{10} [\text{rms (windowed frame)}] \quad (2)$$

2.3. Pitch

The perceived pitch of a sound is just the ear's response to frequency, i.e. for most practical purposes the pitch is just the frequency of the received signal. Pitch is the fundamental period of an audio waveform, and is an important parameter in the analysis and synthesis of audio and speech signals. In an audio signal, we can still use pitch as a low-level feature to characterize the periodicity of waveforms in different audio signals[9],. In time domain analysis the pitch could be estimated by using the peaks of the sound signal in another way pitch can be determined by analyzing the spectrum.

2.4. Spectral centroid

The spectral centroid is used to measure the brightness of a sound. It determines the point in the spectrum where most of the energy of the sound signal concentrated and is correlated with the dominant frequency of the sound .Spectral centroid is defined as the weighted mean of the frequencies present in the frame, determined using a Fourier transform, with their magnitudes as weights, The spectral centroid for each frame is defined as follows [9],

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n).\, X(n)}{\sum_{n=0}^{N-1} X(n)} \quad (3)$$

Where X (n) represent the weighted frequency value of binary number n and f (n) represents the centre frequency of that binary.

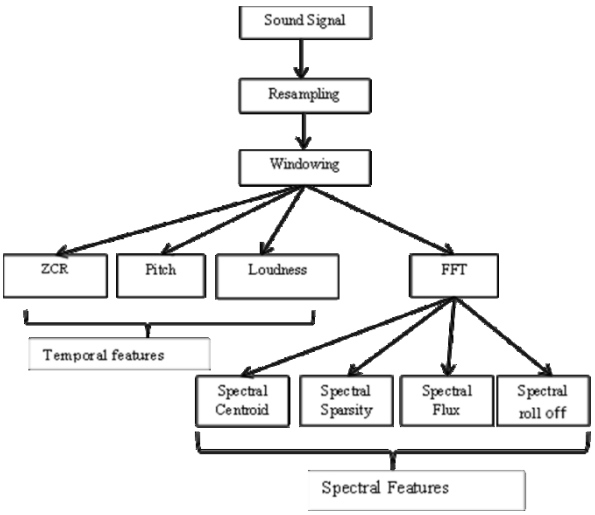

Figure.2.The block diagram of various audio feature extraction scheme

2.5. Spectral sparsity

Spectral sparsity means distributed spectral features of the sound signal. Spectral sparsity should be large for pure sine tones or bells and smaller for sounds with significant noise. It has been provide a good description for sound classification. The spectral sparsity for each frame is defined as follows [1]:

$$\text{Spectral sparsity} = \frac{\max[|X(1),X(2),\ldots\ldots\ldots\ldots X(N)|)}{\sum_{n=0}^{N-1} X(n)} \quad (4)$$

2.6. Spectral roll off

Spectral roll off is a measure of the bandwidth of the analyzed block of the audio samples. Spectral roll off is defined as the frequency binary below which the accumulated magnitudes of the spectrum of the sound signal reach the certain percentage k of the overall sum of the magnitudes. Spectral roll off is small for silences and large for noises [4]. The spectral roll off of the sound signal for each frame is defined as follows,

$$\text{Spectral roll off} = 0.85 \times \sum_{n=1}^{N} X[n] \quad (5)$$

2.7. Spectral flux

Spectral flux can be used to identify various musical sounds. Spectral flux is a measure of the change in energy between various frequencies bands in a sequence of frequency spectra measured from the audio data. The spectral flux of the sound signal for each frame is defined as follows,

$$\text{Spectral flux} = \sum_{n=1}^{N} (X[n] - X[n-1])^2 \quad (6)$$

Where X[n] and X [n-1] are the normalized magnitude of the Fourier transform at the current time t, and the previous time t-1, respectively. The spectral flux is a measure of the amount of local spectral change.

2.8. Mel Frequency Cepstral Coefficient.

Mel-Frequency Cepstral Coefficient (MFCC) are considered as the low-level audio features as they are successfully used in speech recognition, speaker identification and further proved to be efficient for environment and natural sound classification. MFCC employ Mel scale which relates the perceived frequency to measured frequency. Incorporating this scale makes the sound to match more closely what humans hear.The normal frequency f hertz can be converted to the Mel range by the following equation [3].

$$m = 1127.01048 \log (1 + f / 700) \quad (7)$$

In MFCC usually the spectrum is first transformed using the Mel frequency bands. The result is called the MFCC's, which are used for natural sound detection. The MFCC is used to represents the shape of the

spectrum with very few coefficients [3]. The cepstrum is defined as the Fourier transform of the logarithm of the spectrum of the audio signal. The Mel cepstrum is the spectrum computed on the Mel-bands instead of the Fourier spectrum. The use of Mel-cepstrum allows us to obtain better mid-frequency part of the signal.

The flow chart of the MFCCs calculation is shown in figure.2. Here first the incoming audio signal is windowed using hamming window in overlapping steps. For each window, the log of the power spectrum is computed using a Fast Fourier Transform [3]. A nonlinear map of the frequency scale perceptually weights the log spectral coefficients. This operation called the Mel scaling. At the final stage the Mel weighted spectrum is transformed into Cepstral coefficients using Discrete Cosine Transform. This results in features that are dimensionally uncorrelated. Thus MFCCs provides a compact representation of the spectral envelope.

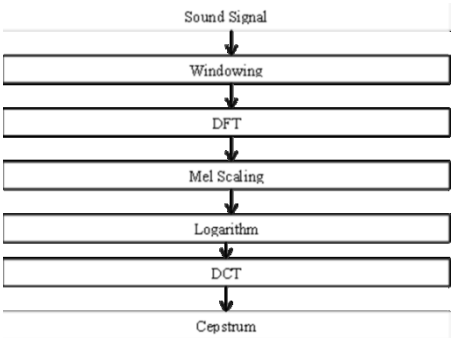

Figure.3. Flow chart of calculation of MFCCs

III. CLASSIFICATION

3.1 Support Vector Machine (SVM)

SVM is a useful technique for music, speech and sound classification. A classification task usually involves with training and testing data which have to be classified. Each instance in the training set contains one target values and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set based on the given attributes [8].

Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction. Feature selection and SVM classification together have a use even when prediction of unknown samples is not necessary [6].

The SVM theory is a new statistical technique and has drawn much attention on this topic in recent years. An SVM is a binary classifier that makes its decisions by constructing an optimal hyper plane that separates the two classes with the largest margin. It is based on the idea of structural risk minimization (SRM) induction principle [6] that aims at minimizing a bound on the generalization error, rather than minimizing the mean square error. For the optimal hyper plane $w.x + b = 0$, $w \epsilon R^N$ and $b \epsilon R$ the decision function of classifying a unknown point x is defined as [6].

$$f(x) = \text{sign}(wx + b) = \text{sign}(\sum_{i=1}^{Ns} \alpha_i m_i x_i . x) \qquad (8)$$

Where N_S is the support vector number, x_i is the support vector, α_i is the Lagrange multiplier and $m_i \varepsilon$ {-1, +1} describes which class x belongsto [6].

In most cases, searching suitable hyper plane in input space is too restrictive to be of practical use. The solution to this situation is mapping the input space into a higher dimension feature space and searching the optimal hyper plane in this feature space. Let $z = \varphi(x)$ denotes the corresponding feature space vector with a mapping φ from R^N to a feature space Z [6]. It is not necessary to know about φ from R^N . We just provide afunction called kernel which uses the points in input space to compute the dot product in feature space Z, that is

$$Z_i . Z_j = \varphi(x_i).\varphi(x_j) = K(x_i, xj) \qquad (9)$$

Finally, the decision function becomes

$$f(x) = \text{sign}(\sum_{i=1} \alpha_i m_i K(x_i . x) + b) \qquad (10)$$

Functions that satisfy Mercer's theorem can be used as kernels. Typical kernel functions include linear kernel, polynomial, radial basis kernel, Gaussian etc. Kernel parameters also have a significant effect on the decision boundary. The degree of the polynomial kernel and the width parameter of the Gaussian kernel control the flexibility of the resulting classifier .The lowest degree polynomial is the linear kernel, which is not

sufficient when a non-linear relationship between features exists. A degree-2 polynomial is already flexible enough to discriminate between the two classes with a sizable margin. The degree-5 polynomial yields a similar decision boundary with greater curvature [8].

3.2 Support Vector Machine for Multiclass Classification One against One method

In this method, SVM classifiers for all possible pairs of classes are created. Therefore, for M classes, there will be binary classifiers. The output from each classifier in the form of a class label is obtained. The class label that occurs most is assigned to that point in the data vector. In case of a tie, a tie-breaking strategy may be adopted. A common tie-breaking strategy is to randomly select one of the class labels that are tied [11].

The number of classifiers created by this method is generally much larger. However, the number of training data vectors required for each classifier is much smaller. Therefore, this method is considered more symmetric than the one against- the-rest method. Moreover, the memory required to create the kernel matrix is much smaller. However, the main disadvantage of this method is the increase in the number of classifiers as the number of class increases. For example, for 7 classes of interest, 21 classifiers will be created.

IV. EXPERIMENTS

An audio database for training consists of 12 classes of sound signals. The names of these audio classes are alarm, applause, baby cries, beep, dog barking, doorbells, laugh, knocking, motor bike running sound, organ sound, whistle and women scream. There are totally 240 sounds in this database. Each sound signals are resampled to make equal number of sample for each sound. Frame based acoustic feature extraction is adopted here. Here the sounds signals are resampled and windowed using 40ms hamming windows hopped at every 20ms.The reason for using frames of this length is, in general, dynamic audio sources tend to have some stationary characteristics, for instance spectral characteristics, at these time scales.

The sound signals are then windowed using 40ms hamming windows hopped at every 20ms.The various temporal features such as pitch, zero crossing rates, loudness are calculated. The Discrete Fourier Transform is applied to each frame to obtain distribution of energy into frequencies then various spectral features such as spectral centroid, spectral sparsity, spectral flux and spectral roll of are calculated. After the detection of the different features, feature vector is created of same length for training set.

Here SVM classifier is used as pattern classifier. Separate set of test and training sounds are used for classification and recognition of sounds. The radial basis kernel used as a kernel function for SVM classification. The classification accuracy was estimated using evaluation of the confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classifier.

V. RESULTS

The audio features of each test sound signals are extracted and feature vector are created. Figure 4 shows the various temporal and spectral features of the doorbell sound. To understand the classification performance, confusion matrix is shown, which allows us to observe the degree of confusion among different classes. The confusion matrix given in Table I is built from a single arbitrary trial, constructed by applying the classifier to the test set and displaying the number of correctly/incorrectly classified items. The rows of the matrix denote the environment classes we attempt to classify, and the columns depict classified results. To further understand the classifier performance no of hit and miss for each class is shown in Table II. The SVM classifier for environment and natural sounds was good with an accuracy of 91.30% and error rate is 8.7%.

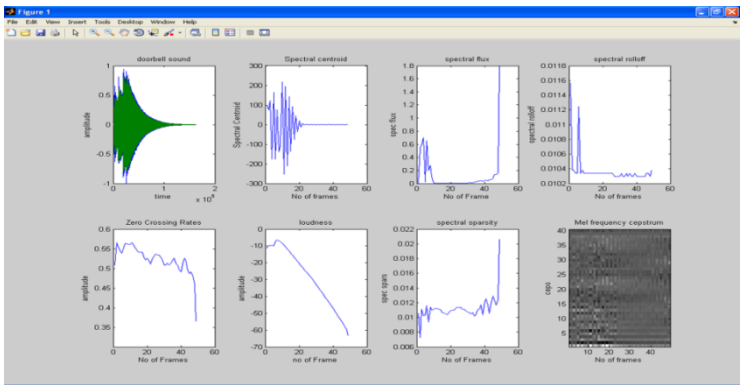

Figure.4. Different audio features of doorbell sound (A).Whistle sound,(B).Spectral centroid.(C).Spectral flux,(D)Spectral roll off,(E).Zero Crossing Rates.(F).Loudness,(G)Spectral sparsity,(H).Mel Frequency Cepstral Coefficients

VI. CONCLUSION AND FUTURE SCOPE

Content based acoustic feature extraction and classification based on Support Vector Machine for environmental and natural sounds are presented here. The different acoustic features have been studied both temporal and spectral feature, which is flexible, intuitive and physically interpretable. Therefore these features can be used for the classification and recognition purposes of the audio data. It is important to optimize the number of features selected because each feature represents a dimension in the feature space

Table I: The confusion matrix of the SVM classifier

	Alarm	Applause	Baby cries	Beep	Dog	Door bell	Knock	Laugh	Motor bike	Organ	Whistle	Women scream
Alarm	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Applause	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
Baby cries	**2**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Beep	**0**	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Dog	**0**	**0**	**0**	**0**	**4**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Door bell	**0**	**0**	**0**	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**
Knock	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**0**	**0**	**0**	**0**	**0**
Laugh	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**3**	**0**	**0**	**0**	**0**
Motor bike	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**0**	**0**	**0**
Organ	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**0**	**0**
Whistle	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**0**
Women scream	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**

Table II. The number of correct and miss classification of the classifier

Sound class used for testing	**Hit**	**Miss**	**Accuracy**
Alarm	**3**	**0**	**100%**
Applause	**3**	**1**	**75%**
Baby cries	**3**	**2**	**60%**
Beep	**3**	**0**	**100%**
Dog	**4**	**0**	**100%**
Door bell	**3**	**0**	**100%**
Knock	**4**	**0**	**100%**
Laugh	**3**	**1**	**75%**
Motor bike	**4**	**0**	**100%**
Organ	**4**	**0**	**100%**
Whistle	**4**	**0**	**100%**
Women scream	**4**	**0**	**100%**
Overall accuracy	91.30%		

The detailed study of multiclass SVM classifier is presented here which will label the sound signal using the features selected so that the nature of the unknown audio sound is known and it is classified under a known class of audio signals with an accuracy rate of 85%.Future work include expand the number of class of the sound signal and study on the classification and visualization of these sounds for various application such as surveillance system and visualizing sound events for deaf and hard of hearing.

REFERENCES

[1]. Gordon Wichern, JiachenXue, Harvey Thornburg ,Brandon Mechtley, and Andreas Spanias, "Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds," IEEE Trans. Audio, Speech, And Language Processing, Vol. 18, No. 3, March 2010

[2]. Stavros Ntalampiras, IlyasPotamitis and Nikos Fakotakis" Sound Classification based on Temporal Feature Integration" 4th International Symposium on Communications,Control and Signal Processing, ISCCSP 2010, Limassol, cyprus, 3-5 March 2010.

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

[4]. Lu, G., & Hankinson, T." A technique towards automatic audio classification and retrieval". In 4th International Conference on Signal Processing. Beijing. (Retrieved November 3), 2002,

[5]. E. Wold, T. Blum, D. Keislar, and J. Wheaton. "Content-based classification, search, and retrieval of audio". Multimedi IEEE, Trans 27–36, 1996.
[6]. Jia-Ching Wang, Jhing-Fa Wang, Cai-Bei Lin, Kun-Ting Jian, and Wai-He Kuok"Content-Based Audio Classification Using Support Vector Machines and Independent Component Analysis"IEEE Trans. The 18th International Conference on Pattern Recognition (ICPR'06)
[7]. A. Eronen," Comparison of Features for Musical Instrument Recognition", New Paltz, New York, 2001.
[8]. GuodongGuo and Stan Z. Li" Content-Based Audio Classification and Retrieval by Support Vector Machines", IEEE Trans. On Neural Networks, vol. 14, no. 1, january 2003.
[9]. E. Wold, T. Blum, D. Keislar, and J. Wheaton,"Content-based classification search and retrieval ofaudio," IEEE Multimedia Magazine, vol. 3, pp. 27–36,July 1996.