M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Submitted November2004

AUDIO SIGNAL CLASSIFICATION

Hariharan Subramanian (Roll No: 04307909)
Supervisors: Prof. Preeti Rao and **Dr. Sumantra. D. Roy**

Abstract

Audio signal classification system analyzes the input audio signal and creates a label that describes the signal at the output. These are used to characterize both music and speech signals. The categorization can be done on the basis of pitch, music content, music tempo and rhythm. The signal classifier analyzes the content of the audio format thereby extracting information about the content from the audio data. This is also called audio content analysis, which extends to retrieval of content information from signals. In this report the implementation of the audio signal classification is presented. A number of features such as pitch, timbral, rhythmic features have been discussed with reference to their ability to distinguish the different audio formats. The selection of the important features as well as the common techniques used for classification has been explained. Finally an approach called the confusion matrix has been studied in order to evaluate performance of the classification system.

1. Introduction

An audio signal classification system should be able to categorize different audio input formats. Particularly, detecting the audio type of a signal (speech, background noise, and musical genres) allows such new applications as automatic organization of audio databases, segmentation of audio streams, intelligent signal analysis, intelligent audio coding, automatic bandwidth allocation, automatic equalization, automatic control of sound dynamics etc. Audio signal classification finds its utility in many research fields such as audio content analysis, broadcast browsing, and information retrieval. Recently its demand is increasing in the information retrieval field as a new approach of Query By Humming has been invented; in which the user has to hum a tune and the song that corresponds to that tune is returned. All classification systems employ the extraction of a set of features from the input signal. Each of these features represents an element of the feature vector in the feature space. The dimension of the feature space is equal to the number of extracted features. These features are given to a classifier that employs certain rules to assign a class to the incoming vector. Fig.1 shows the block diagram, which is self-explanatory.

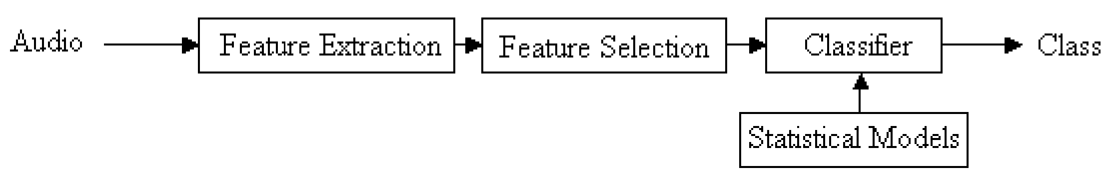

Fig.1. Block diagram of an audio signal classification system.

2. FEATURE EXTRACTION

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. Feature extraction involves the analysis of the input of the audio signal. The feature extraction techniques can be classified as temporal analysis and spectral analysis technique. Temporal analysis uses the waveform of the audio signal itself for analysis. Spectral analysis utilizes spectral representation of the audio signal for analysis.

All audio features are extracted by breaking the input signal into a succession of analysis windows or frames, each of around 10-40-ms length, and computing one feature value for each of the windows. One approach is to take the values of all features for a given analysis window to form the feature vector for the classification decision, so that class assignments can be obtained almost in real time, thus realizing a real-time classifier.

Another approach is to use the texture window, in which the long-term characteristics of the signal are extracted and the variation in time of each feature is measured, that often provides a better description of the signal than the feature itself. A texture window is a long-term segment in the range of seconds containing a number of analysis windows. In the texture-based approach only one feature vector for each texture window is generated. The features are not directly obtained in each analysis window, but statistical measures of the values are obtained for all analysis windows within the current texture window. Therefore in this case real-time classification is not possible, since at least one whole texture window has to be processed to obtain a class decision.

Since the analyzed audio files are supposed to contain only one type of audio, a single class decision is made for each type of audio, which can be derived following one of two possible approaches. The first approach is the single vector mode, which consists of taking the whole file length as the texture window. In this way, each file is represented by a single feature vector, which in turn is subjected only once to classification. The second approach is the texture window mode, which consists of defining shorter texture windows and making several class decisions along each file, one for each texture window. At the end of the file the decisions are averaged to obtain a final class decision. This average computation is weighted by the certainty of each class decision.

As discussed previously feature extraction plays an important role in classification of an audio signal. Hence it becomes all the more important to select those features that help the classification process more efficient. There are different types of features, such as the pitch, timbral features, rhythm features etc that are explained below.

2.1 Pitch:

The sound that comes through vocal tract starts from the larynx where vocal cords are situated and ends at mouth. The vibration of the vocal cords and the shape of the vocal tract are controlled by nerves from brain. The sound, which we produce, could be categorized into voiced and unvoiced sounds. During the production of unvoiced sounds the vocal cords do not vibrate and stay open whereas during voiced sounds they vibrate and produce what is known as glottal pulse. A pulse is a summation of a sinusoidal wave of fundamental frequency and its harmonics (Amplitude decreases as frequency increases). The fundamental frequency of glottal pulse is known as the pitch.

In music, the position of a tone in the musical scale is designated by a letter name and determined by the frequency of vibration of the source of the tone. Pitch is an attribute of every musical tone. The fundamental or first harmonic of any tone is perceived as its pitch. Absolute pitch is the position of a tone in the musical scale determined according to its number of vibrations per second, irrespective of other tones. The term also denotes the capacity to identify any tone upon hearing it sounded alone or to sing any specified tone. For example pitch helps the human ear to distinguish between string instruments, wind instruments and percussion instruments such as the drums, tabla etc.

After the voiced parts of the sound are selected the pitch has to be determined. There are several algorithms currently in use for accomplishing this task. These could be categorized into Time-domain and Frequency-domain analysis. In time domain analysis the pitch could be estimated by using the peaks, but due to the presence of formant frequencies (harmonics) this method could give a wrong estimation. So the formant frequencies are filtered out using a low pass filter and then zero crossing methods or any other suitable method is used to determine the pitch. The speech signal is also passed through a low pass filter in the frequency domain analysis and then the pitch is determined by analyzing the spectrum.

2.2 Timbral features:

Sound "quality" or "timbre" describes those characteristics of sound, which allow the ear to distinguish sounds that have the same pitch and loudness. Timbre is then a general term for the distinguishable characteristics of a tone. Timbre is mainly determined by the harmonic content of a sound and the dynamic characteristics of the sound such as vibrato and tremolo. In music timbre is the quality of a musical note that distinguishes different types of musical instrument. Each note produced by a musical instrument is made of a number of distinct frequencies, measured in hertz (Hz). The lowest frequency is called the fundamental and the pitch produced by this frequency is used to name the note. However, the richness of the sound is produced by the combination of this fundamental with a series of harmonics and/or partials (also collectively called overtones). Most western instruments produce harmonic sounds, and these can be calculated by multiplying the fundamental by an increasing series of numbers - x2, x3, x4, etc (whole number multiples). However many instruments produce inharmonic tones, and may contain overtones which are not whole number multiples, these being the partials. Therefore, when the orchestral tuning note is played, the sound is a combination of 440 Hz, 880 Hz, 1320 Hz, 1760 Hz and so on. The balance of the amplitudes of the different frequencies is responsible for giving each instrument its characteristic sound.

The ordinary definition of vibrato is periodic changes in the pitch of the tone, and the term tremolo is used to indicate periodic changes in the amplitude or loudness of the tone. So vibrato could be called FM (frequency modulation) and tremolo could be called AM (amplitude modulation) of the tone. Actually, in the voice or the sound of a musical instrument both are usually present to some extent. Vibrato is considered to be a desirable characteristic of the human voice if it is not excessive. It can be used for expression and adds richness to the voice. If the harmonic content of a sustained sound from a voice or wind instrument is reproduced precisely, the ear can readily detect the difference in timbre because of the absence of vibrato.

In the following equations the *r* indicates the number of the current frame, $x_r[n]$ denotes the frame in the time domain, where *n* is the time index, and $X_r[k]$ denotes the short-time Fourier transform (STFT) of that frame, where *k* is the frequency coefficient or bin index. The following are some of the timbral features,

2.2.1 Zero crossings

The zero crossings feature counts the number of times that the sign of the signal amplitude changes in the time domain in one frame. For single-voiced signals, zero crossings are used to make a rough estimation of the fundamental-frequency. For complex signals it is a simple measure of noisiness.

2.2.2 Centroid

The spectral centroid is defined as the center of gravity of the spectrum [1].

$$C_r = \frac{\sum_{k=1}^{N/2} f[k] \left| X_r[k] \right|}{\sum_{k=1}^{N/2} \left| X_r[k] \right|}$$

where *f* [k] is the frequency at bin k.

The centroid is the measure of the spectral shape and higher centroid values correspond to brighter textures with more high frequencies. Centroid models the sound sharpness. Sharpness is related to the high-frequency content of the spectrum. Higher centroid values correspond to spectra in the range of higher frequencies. Due to its effectiveness to describe spectral shape, centroid measures are used in audio classification tasks.

2.2.3 Rolloff

The rolloff is defined as the frequency below which 85% of the magnitude distibution of the spectrum is concentrated. Like the centroid, it is also a measure of spectral shape and yields higher values for high frequencies. Therefore it can be said that there exists a strong correlation between both the features. The equation for rolloff is

$$\sum_{k=1}^{M} \left| X_r[k] \right| = 0.85 \sum_{k=1}^{N/2} \left| X_r[k] \right|$$

If M is the largest value of k for which this equation is satisfied then this frequency M is the rolloff [1].

2.2.4 Flux

The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions that correspond to successive signal frames. The equation for flux is

$$F_r = \sum_{k=1}^{N/2} (|X_r[k]| - |X_{r-1}[k]|)^2$$

Flux is an important feature for the separation of music from speech [1].

2.2.5 Mel frequency cepstral coefficients (MFCC's)

MFCCs are a compact representation of the spectrum of an audio signal taking into account the nonlinear human perception of pitch, as described by the mel scale. They are one of the most used features in speech recognition and have recently been proposed to analyze and represent musical signals. MFCCs are computed by grouping the Short Time Fourier Transform (STFT) coefficients of each frame into a set of 40 coefficients, using a set of 40 weighting curves that simulate the frequency perception of the human hearing system. Then the logarithm of the coefficients is taken, and a discrete cosine transform (DCT) is applied to decorrelate them. Normally the five first coefficients are taken as features.

2.3 Rhythm features:

These are features that exhibit regularity or the structure of the audio signal. They define the characteristic of the audio signal because they follow a particular pattern. These features are rhythmical structure and beat strength. For better classification purposes it is more interesting to extract information about these features. An example is the regularity of the beats, which is expected to be higher in rock and pop. Beat strength also seems to be a valuable feature. For instance, it is likely to be higher in techno music than in jazz. A beat histogram is a curve describing beat strength as a function of a range of beats per minute values, and allows the extraction of the properties mentioned. Peaks on the histogram correspond to the main beat and other subbeats. The result is a curve describing beat strength as a function of the beat per minute (bpm) values. The high peaks in the beat histogram denote a high overall beat strength. High peaks correspond to high beat strength and peaks separated by integer bpm multiples denote rhythm regularity. All rhythm features are extracted from the beat histograms.

2.3.1 Beat strength

Statistical measures of the histogram such as mean, standard deviation, mean of the derivative, standard deviation of the derivative, third- and fourth-order central moments (called skewness and kurtosis, respectively), and entropy are evaluated to obtain an overall measure of beat strength. All these measures are computed in the "beat domain".

2.3.2 Rhythmic regularity

A beat histogram in which there is periodic spacing the peaks denotes high rhythmic regularity. This can be measured by the normalized autocorrelation function of the beat histogram. It will contain clear peaks for rhythmically regular music examples and it will be the more linear if the regularity is weaker. To reduce this to a scalar measure of rhythm regularity, the mean across the lags of the difference between the autocorrelations and the linear function is computed.

Although the computation is performed on a frame-by-frame basis, histograms are obtained in long-term intervals given by the texture windows. Hence all of the features related to the beat histogram are single-valued features to which the time domain mean and standard deviation subfeatures will not be applicable.

2.4 MPEG-7 features:

Moving Pictures Experts Group (MPEG) has defined an international standard defining a set of techniques for analyzing and describing raw data in terms of certain features. They are a subset of the features that have been discussed so far. It is an attempt to standardize the features that are used in audio signal classification. It deals the content-based description so that data can be described in terms of features.

2.4.1 Audio spectrum centroid (ASC)

A perceptually adapted definition of the centroid, which introduces a logarithmic frequency scaling centered at 1 kHz,

$$ASC_r = \frac{\sum_{k=1}^{N/2} \log_2(f[k]/1000) P_r[k]}{\sum_{k=1}^{N/2} P_r[k]}$$

where P_r is the power spectrum of the frame r [1].

2.4.2 Audio spectrum spread (ASS)

It describes concentration of the spectrum around the centroid and is defined as

$$ASS_r = \sqrt{\frac{\sum_{k=1}^{N/2} [\log_2(f[k]/1000) - ASC_r]^2 P_r[k]}{\sum_{k=1}^{N/2} P_r[k]}}$$

Lower spread values would mean that the spectrum is highly concentrated near the centroid and higher values mean that it is distributed across a wider range at both sides of the centroid [1].

2.4.3 Audio spectrum flatness (ASF)

It can be defined as the deviation of the spectral form from that of a flat spectrum. Flat spectra correspond to noise or impulse-like signals hence high flatness values indicate noisiness. Low flatness values generally indicate the presence of harmonic components. Instead of calculating one flatness value for the whole spectrum, a separation in frequency bands is performed, resulting in one vector of flatness values per time frame. The flatness of a band is defined as the ratio of the geometric and the arithmetic means of the power spectrum coefficients within that band. Each vector is reduced to a scalar by computing the mean value across the bands for each given frame, thus obtaining a scalar feature that describes the overall flatness.

2.4.4 Harmonic ratio (HR)

A measure of the proportion of harmonic components within the spectrum, defined as the maximum value of the autocorrelation (AC) of each frame [1],

$$HR_r = \max_{\tau=1, N=1} \{R_r[\tau]\}$$

2.5 Other Features

The features grouped in this last section describe the signal regarding its dynamic properties, its statistical behavior, and its predictability.

2.5.1 Root mean square

It can be defined as the root mean square (rms) energy of each signal frame.

2.5.2 Time envelope

It is the measure of the maximum of absolute amplitude in each frame.

2.5.3 Low energy rate

It can be expressed as the percentage of frames within a file that have root mean square (rms) energy lower than the mean rms energy in that file. Apart from the beat-histogram-based features, this is the only feature that is not computed on a frame-by-frame basis, but on a texture window basis.

2.5.4 Loudness

The previously discussed features were dynamic-related and are based on physical measures such as amplitude or energy. A better adaptation to the human ear perception of sound dynamics is provided by the measurement of loudness.

3. IMPLEMENTATION OF SPECIFIC FEATURES

After the brief overview of the different features it is important to figure out the important features so that these can be dealt in detail. These are

3.1 Mel Frequency Cepstrum Coefficients (MFCC's)

MFCC's employ the mel scale which is a scale of pitches which are equal in distance from one another. The normal frequency *f* hertz can be converted to the mel range by the following equation [8],

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

A cepstrum is the result of taking the Fourier transform of the decibel spectrum (power spectrum) as if it were a signal. There is a complex cepstrum and a real cepstrum. The cepstrum can be defined mathematically as

cepstrum of a signal = FT(log(FT(the signal))) where FT indicates Fourier Transform.

The real cepstrum uses the logarithm function defined for real values, while the complex cepstrum uses the complex logarithm function defined for complex values. The complex cepstrum holds information about magnitude and phase of the initial spectrum, allowing the reconstruction of the signal. The real cepstrum only uses the information of the magnitude of the spectrum. The cepstrum can be seen as information about rate of change in the different spectrum bands. Usually the spectrum is first transformed using the mel frequency bands. The result is called the MFCC's, which are used for voice identification, pitch detection and much more. This is a result of the cepstrum separating the energy resulting from vocal cord vibration from the "distorted" signal formed by the rest of the vocal tract.

The human ear exhibits a nonlinear characteristic when it comes to the perception of pitch. Hence the mel scale takes into the account of this property. Below 500Hz the frequency and the mel scales coincide and above that larger and larger intervals produce equal pitch increments. As a result, four octaves on the hertz scale above 500Hz are judged to comprise about two octaves on the mel scale. After the translation to the mel frequency scale the coefficients can be evaluated. Normally the computation of MFCC’s involves the windowing of the incoming audio signal. The log of the spectrum is computed and another transform is applied in order to obtain the cepstrum coefficients. This can be explained from Fig.2 as follows

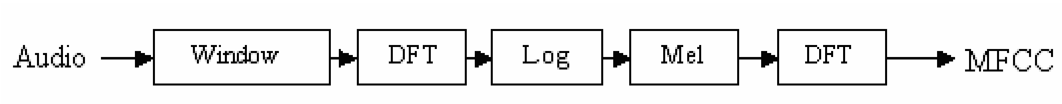

Fig.2. Block diagram to compute MFCC’s [4].

First the audio is hamming windowed in overlapping steps. For each window, the log of the power spectrum is computed using a DFT. A nonlinear map of the frequency scale perceptually weights the log spectral coefficients. This operation called the mel scaling, emphasizes mid frequency bands in proportion to their perceptual importance. At the final stage the mel weighted spectrum is transformed into cepstral coefficients using another DFT. This results in features that are dimensionally uncorrelated. Thus MFCC’s provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first few coefficients [4].

MFCC’s were originally invented for characterizing the seismic echoes resulting from earthquakes and bomb explosions. It is now used as an excellent feature vector for representing the human voice and musical signals.

3.2 Beat Strength

The feature beat strength is computed by detecting the beat of the audio signal. The main beat can be loosely defined as the regular periodic sequence of pulses corresponding to where a human would tap his foot while listening to the music. In automatic beat detection algorithms, the beat is characterized by its frequency (tempo), phase (accent locations) and a confidence measure about its detection. Beat detection can be broadly classified into two categories, event based and self-similarity based. In event based algorithms, transient events such as note onsets or percussion hits are detected and their Inter Onset Arrival Intervals (IOI) are used to estimate the main tempo. In self-similarity based algorithms, the periodicity of amplitude envelopes usually of multiple bands is calculated and used to detect the tempo [3].

Beat Strength can be defined as the rhythmic characteristic that allows discriminating between two pieces of music having the same tempo. Characteristics related to beat strength are implicitly used in automatic beat detection algorithms and known to be as important as tempo information for music classification and retrieval. The perception of beat strength and its measurement is based on the calculation of beat histograms, which are a global representation of musical rhythm based on self-similarity.

Before the beat histogram can be evaluated the occurrence of a beat has to be detected first. Most automatic beat detection systems provide a running estimate of the main beat and an estimate of its strength. One of the common automatic beat detector structures consists of filter-bank decomposition, followed by an envelope extraction step and finally a periodicity detection algorithm which is used to detect the lag at which the signal's envelope is most similar to itself. The process of automatic beat detection resembles pitch detection with larger periods. The window that is used for the filter should be larger so that capturing the signal repetitions at the beat and subbeat levels can be done [2].

The resulting histogram has bins corresponding to tempos in beats per minute (bpm) and the amplitude of each bin corresponds to the strength of repetition of the amplitude envelopes of each channel for that particular tempo. Two measures of beat strength can be derived from the beat histogram. The first measure is the sum of all histogram bins. Because of the autocorrelation calculation used for periodicity detection in the beat histogram this measure indicates how strong the self-similarity of the signal is at various tempos. The second measure is the ratio of the amplitude of the highest peak of the beat histogram to the average amplitude (peak) and indicates how dominant the main beat is. In addition to these features in order to characterize musical genres more information about the rhythmic content of a piece can be utilized. The regularity of the rhythm, the relation of the main beat to the sub-beats, and the relative strength of sub-beats to the main beat are some of the important features that can be represented as feature vectors [3].

4. FEATURE SELECTION

From a large set of features it is important to select particular set of features that would determine the nature and hence the class of the audio signal. These features determine the dimensionality in the feature space. It is important therefore to select an optimum number of features that not only keeps accordance with the accuracy and the level of performance but also reduces the computation costs. Thus there is no point in increasing the number of features as it would not have a drastic impact on the accuracy but would pave for more complexities in computation. Therefore a selected feature must have the following properties,

1) Invariance to irrelevancies: Any good feature should exhibit invariance to irrelevancies such as noise, bandwidth or the amplitude scaling of the signal. It is also upon the classification system to consider such variations as irrelevant to achieve better classification across a wide range of audio formats.
2) Discriminative Power: The purpose of feature selection is to achieve discrimination among different classes of audio patterns. Therefore a feature must take round about similar values within the same class but different values across different classes.
3) Uncorrelated to other features: It is very important that there are no redundancies in the feature space. Each new feature that is selected must give altogether different information about the signal as possible. This helps in better computation efficiency, improved performance and optimization of cost [1].

5. CLASSIFICATION

After the feature selection process it is important to classify the signal. Classification is the process by which a particular label is assigned to a particular audio format. It is this label that would define the signal and its origin. A classifier defines decision boundaries in the feature space (ie. mean vs. maximum), which separate different sample classes from each other. Classifiers are categorized by their real time capabilities, on the basis of the approach and their character.

On the basis of their real time capabilities, there are real time classifiers and non-real time classifiers. Real time classifiers can update classification results in time intervals of milliseconds. Hence their application comes of importance in the areas where the input signal consists of a sequence of different types of audio and it is absolutely necessary to keep updating, for class detection. In case of the non-real time classifiers, they analyze a longer fragment of the signal before they provide a classification result. Accuracy in this case is more than real time classifiers because they analyze a longer fragment of the incoming signal, which plays a prominent role to describe the signal.

By character classifiers can be broadly divided as taxonomic and clustering classifiers. Taxonomic classifiers make use of supervised techniques implement the category that is defined beforehand by the user or by the implementation. Clustering classifiers rely on the separation algorithm that groups the audio samples according to some similarity in them. Therefore a classifier is a set of decision rules that are used to assign a class to the unknown input signal.

On the basis of approach classifiers can be split into two types as flat or direct approach and hierarchical approach. In direct approach classifiers it is a case of single stage classification where in audio classes are decided directly by using all the features in one single step. In case of the hierarchical approach the genre dependency of features is used to suggest a hierarchical scheme so that, at each step only the features that are most appropriate to distinguish between the sub-features are used. This approach accounts for the class dependency of the features. The errors are more acceptable in this case than the direct classification techniques. It takes into consideration the future expansions of the system. If in a direct classifier the addition of a new class would mean that the feature selection algorithm would have to be modified and would have to be run with all the training samples. But in the hierarchical classifier to add a new class only a separate genre branch would be modified with respect to feature training and the selection while rest of the model remains unchanged. The hierarchical classifier is more complicated and its implementation is computationally expensive because more classification decisions have to be made and more features have to be computed [1].

In case of all the different types of classifiers it is important use an efficient algorithm that would classify the different audio inputs with less of computational complexities. At the same time the accuracy must be preserved. Normally the two most commonly used methods of computation are k-nearest neighbour and the Gaussian mixture model classifier.

5.1 k-nearest neighbour (kNN) classifier

The nearest neighbour method consists of assigning to the unlabelled feature vector the label of the training vector that is nearest to it in the feature space. In kNN, a training set T is used to determine the class of a previously unseen sample X. First, we determine the mean and maximum values in T, and similarly, for the unseen sample X. Then a suitable distance measure in the feature space is used to determine k elements in T closest to X. If most of these k nearest neighbours contain similar values, then X gets classified accordingly. This classification scheme clearly defines nonlinear decision boundaries and thus improves the performance. Furthermore, the feature distribution suggests that the number of data-points used in the example set T can be considerably reduced for faster processing; only those examples that are close to the decision boundary are actually required [6]. This can be explained by the following example.

Referring to Fig.3 each of the samples (marked by stars) have been labeled either A or B, except for the sample x. This needs to be labeled, the k NN classifier takes the k nearest, i.e. the closest, neighbours around the sample x and uses them to assign a label. This is usually done by a majority-voting rule, which states that the label assigned should be the one, which occurs most among the neighbours.

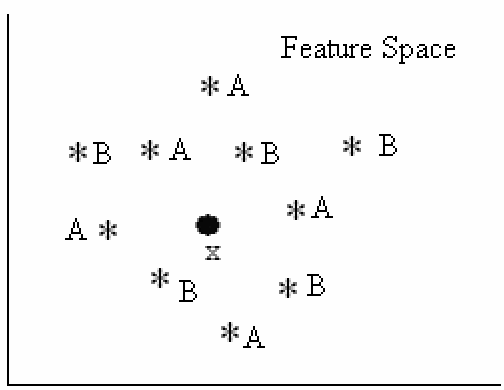

Fig.3. Example of k nearest neighbour rule [8].

For example if k=1, the sample that is nearest to the sample x is sample B. Hence the unknown sample x is assigned as B. But if k=7, then there are four samples of A and three samples of B that are closer to the sample x. Now the sample x gets assigned as A by virtue of majority. Hence it can be inferred that the value of k is critical in order to assign an unknown sample by its nearest neighbours.

As it can be seen that for k=7 sample x gets assigned to A on the basis of majority polling. It is also important to note that the k neighbours have been assumed to have equal influence on predictions irrespective of their relative distance from the query point. Since the three samples of B are closer than the four samples of A on the basis of distance, the former shows to have a greater influence on sample x even though the samples A are in majority. Therefore to increase the efficiency, it is equally important to pay attention to the relative distance of the k nearest samples to the query point in order that the unknown sample gets assigned to the sample that has greater influence on it.

Therefore there exist two main problems in this classifier that has to be addressed. The first is to find a suitable distance measure that which sample is closer to the sample to the sample x.

There are many distance metrics that are used to calculate the distance between the samples.

1) Euclidian distance:
Given two samples x and y the Euclidian distance between the samples is defined as

$$|x-y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where n is the number of features describing x and y [8].

2) City-block distance:
The city-block distance can be defined as

$$D_{city-block}(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

where n is the number of features describing x and y [8].

The second problem is the choice of k, choosing k large generally results in a linear classifier whereas small k results in nonlinear ones. This influences the generalization capability of the k NN classifier. The optimal k can be found by using for instance the leave out one method on the training set. A disadvantage of this method is its large computing power requirement, since for classifying an object its distance to all the objects in the learning set has to be calculated.

5.2 Gaussian Mixture Model (GMM) classifier

Gaussian mixture model is a weighted sum of Gaussian probability density functions, which are referred to as Gaussian components of the mixture model describing a class. The Gaussian probability density function in one dimension is a bell shaped curve defined by two parameters, mean and variance. The Gaussian distribution is usually quite good approximation for a class model shape in a suitably selected feature space. It is a mathematically sound function and extends easily to multiple dimensions. In the Gaussian distribution lies an assumption that the class model is truly a model of one basic class. If the actual model, the actual probability density function, is multimodal, it fails. Gaussian mixture model (GMM) is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class. The probability density function is defined as a weighted sum of Gaussians. The GMM classifier models each class as a linear combination of Gaussian or normal densities that is, each class k is represented by the multidimensional conditional density.

$$p(x|\omega_k) = \sum_{m=1}^{M} w_{km} p_{km}(x)$$

where ω_k is the event that belongs to class *k*, ***x*** denotes feature vector, $\boldsymbol{w_{km}}$ are the weights of the mixture, M is the total number of densities (components) in the mixture and $\boldsymbol{p_{km}}$ is normal density [1].

In a particular case when M=1, each class gets modelled by normal distribution and the classifier simplifies to a simple Gaussian classifier. Estimation of the Gaussian mixture parameters for one class can be considered as unsupervised learning of the case where samples are generated by individual components of the mixture distribution and without the knowledge of which sample was generated by which component. Clustering usually tries to identify the exact components, but Gaussian mixtures can also be used as an approximation of an arbitrary distribution.

The expectation maximization (EM) algorithm is an iterative method used to handle cases where an analytical approach for maximum likelihood estimation is infeasible, such as Gaussian mixtures with unknown and unrestricted covariance matrices and means. The values of $\boldsymbol{w_{km}}$, the mean vectors and the covariance matrices for each component in a particular class, which are the parameters for that class, are estimated using expectation maximization (EM) algorithm only. When an input vector has to be classified its conditional density in each of the classes is computed using the estimated parameters. The class for which the density value is highest becomes the class that is chosen for that vector. This decision rule is called as the maximum likelihood condition. This rule can be applied if the different classes are equally probable.

The GMM classifier has to only store the set of estimated parameters for each class while a kNN classifier needs to store all the training vectors in order to compute the distances to the input feature vector. Also the number of features that are required to attain the same level of accuracy is more in the case of kNN classifier as compared to GMM classifier. Therefore these features make the GMM more computationally optimal but the kNN classifier is still an efficient classifier that is very simple in methodology.

6. EVALUATION

After the detection of features in the audio taxanomy and its subsequent class detection it is also important to evaluate the accuracy of the output, that is the final class of the audio. This evaluation also gives an idea about the performance of the system, which in turn gives the detail about the efficiency of the different algorithms. This narrows down to one aspect i.e. the performance of the classification used. This is normally evaluated using the confusion matrix.

A confusion matrix contains information about actual and predicted classifications done by a classification system. It shows the error in classification of a particular class if that classs had been wrongly classified as another one. This in turn helps in understanding and analyzing the performance of any classifier. Performance of such systems is commonly evaluated using the data in the matrix [5]. Fig.4 shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning; a is the number of correct predictions that an instance is negative, b is the number of incorrect predictions that an instance is positive, c is the number of incorrect of predictions that an instance negative, and d is the number of correct predictions that an instance is positive.

Therefore the confusion matrix can be constructed as follows:

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Fig.4. Construction of a 2x2 classification matrix [7].

Therefore the classification matrix gives a general idea as to how the classification has performed. It is also important to note the efficiency of the confusion matrix. The most important property that describes this efficiency is the accuracy of the confusion matrix. The accuracy is the proportion of the total number of predictions that were correct. It can be defined by the equation (a + d) / (a +b +c + d) [7].

To illustrate this same feature with reference to studying the performances of the classifier consider the given example in Fig.5. The confusion matrix in the figure is a 4x4 matrix that contains the predicted as well as the actual class of an audio signal. The abbreviations Met, Ro, Ja and Cla stand for Metal, Rock, Jazz and Classical respectively, which denote the different classes of the music.

		Predicted Class			
		Met	Ro	Ja	Cla
Actual Class	Met	86	10	4	---
	Ro	8	90	2	---
	Ja	1	4	94	---
	Cla	---	---	1	99

Fig.5. Example of a confusion matrix to evaluate the classification.

From above matrix the entry of an element say X in the matrix element depicting the rows as i and the columns as j means that X% of the test samples of class i were classified as belonging to class j. Therefore looking from the matrix it can be said that while 86% of metal was correctly classified as metal, 10% was wrongly classified as rock and 4% of it was improperly classified as jazz. So was the case when 99% of classical music was correctly classified as classical while 1% of it got wrongly classified as jazz. Thus the confusion matrix is a highly efficient method to evaluate the classifier performance.

7. CONCLUSION

It can be inferred that the incoming unknown audio signal needs to be classified as speech or music signal as per its nature. This can be done only by analyzing its properties, which means that its features that define its nature have to be extracted. The different features have been studied and only those features are given the priority of selection, which give a better description of the signal. MFCC's and beat strength are used because they give information abut the pitch and the rhythmic regularity respectively that aid in classification giving better classification results than the other features. Care must also be taken to optimize the number of features selected because each feature represents a dimension in the feature space. Therefore reducing number of features reduces the computational costs and at the same time maintains the accuracy levels. The subsequent process to the feature extraction is the classification process. It is to the classifier to accurately label the signal using the features selected so that the nature of the unknown audio taxonomy is known and it is classified under a known class of audio signals. The requirements for a classifier are that it must be computationally efficient with less complexity in its algorithm that economizes its cost. Among the direct approach and hierarchical classifiers the latter has the advantage of having flexibility in structure when future expansions are considered but the drawback being that it is complicated and expensive. The two classifier algorithms knn and the Gaussian classifier have been explained. While the knn classifier is relatively simpler the Gaussian classifier uses lesser number of features to obtain a similar level of performance accuracy thereby making it computationally less expensive. Finally it is most important to evaluate whether the entire audio classification system has been faithful, efficient and accurate. This evaluation is done using confusion matrix method that is simple and efficient.

Acknowledgements

I wish to express my gratitude to Prof. Preeti Rao and Dr. Sumantra. D. Roy for their constant guidance and support, which enabled me not only to learn a lot about the topic but also helped to organize the seminar work systematically.

References

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," J. Audio Eng. Soc, Vol. 52, pp. 724-739, July/August 2004.

[2] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech and AudioProcess*, vol. 10, pp. 293–302, 2002 July.

[3] G. Tzanetakis and P. Cook, "Human Perception And Computer Extraction Of Musical Beat Strength," *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hanburg, Germany, 2002 September.

[4] J. Foote, "A Similarity Measure for Automatic Audio Classification," *Proc. AAAI 1997 Spring Symp. on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora,* Stanford, CA, 1997.

[5] A. Eronen, "Comparison Of Features For Musical Instrument Recognition," New Paltz, New York, 2001 October.

[6] Finnish Meteorological Institute, www.geo.fmi.fi/~syrjasuo/Analysis/node6.html, October 2004.

[7] Department of Computer Science, University of Regina, www.cs.uregina.ca, October 2004.

[8] The Free Dictionary.Com by Farlex, encyclopedia.thefreedictionary.com, October 2004.