Study of different Features and Feature extraction techniques used in Audio Information retrieval
Swati A. Patil

Abstract
In this paper we study different features and feature extraction techniques used in audio information retrieval. Now a days searching and retrieving exact content and user interest audio with is very challenging task due to vast amount of audio data. One common method used to facilitate the process is by using metadata. For example, the title, genre and the singer information of a song can be coded as bits sequences and embedded in the header of the song. But for many audio detail information is not provided. However, different audios have different acoustic characteristics. Here some commonly used features like MFCC, chroma, loudness, pitch, tone features are discussed. These features are used to train the audio information retrieval and recognition system Also different extraction techniques such as MFCC, LPC, LPCC, LDB, PLP are discussed.

Keywords: Chroma, MFCC, LPC, LDB, feature extraction

I. INTRODUCTION
Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications.
Digital music collections are growing ever larger, and even portable devices can store several thousand songs. As Berenzweig humorously noted in the capacities of mass storage devices are growing at a much higher rate than the amount of music, so in ten years time, a standard personal computer should be able to store all the music in the world. Already today, cell phone plans with free access to millions of songs from the Big Four (EMI, Sony BMG, Universal Music and Warner Music Group) as well as numerous smaller record companies are available on the Danish market. Accessing large music collections is thus easier than ever, but this introduces a problem that consumers have not faced to this extent before: how to find a few interesting songs among the millions available.

Some prominent task in music and audio information retrieval are genre classification, artist recognition, auto tagging, audio fingerprinting, music similarity, audio thumb nailing, rhythmic similarity. Features: Measurable properties of the observed phenomenon, usually containing information relevant for pattern recognition.

Feature extraction: Input signal is transformed into a new (smaller) space of variables that simplify analysis.

II. FEATURES USED IN AUDIO INFORMATION RETRIEVAL [4]
The features described below are divided into three groups according to [5]:

The simpler Standard low-level (SLL) features, the Frequency cepstrum Coefficients and Psychoacoustic features.

2.1 Low-Level features:
Some of the low level features are RMS, Zero Crossing Rate, Spectral Centroid, spectral Roll off, Band Energy Ratio, Flux (also called Delta Spectrum Magnitude), Bandwidth, pitch and pitch strength.

2.1.1 RMS
RMS or Root Mean Square is a measure of amplitude of a sound wave in one analysis window. This is defined as

\[ RMS = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x^2(i)} \]

where n is the number of samples within an analysis window and x is the value of the sample.

2.1.2 Zero Crossing Rate
This is a measure that counts the number of times the amplitude of the signal changes sign, i.e. crossing the x-axis, within one analysis window. The feature is defined as

\[ ZC = \frac{1}{T} \sum_{t=1}^{T} \text{func}\{s(s+t)<0\} \]

Where s is the sound signal of length T measured in time and func(A) equals 1 if A is true and 0 otherwise.

2.1.3 Spectral Centroid
This feature is effective in describing the spectral shape of the audio. The Feature is correlated with the psychoacoustic features sharpness and brightness. There are several definitions of the Spectral Centroid feature in previous work. In this study it is calculated as a weighted mean of the frequencies in the FFT transform of the signal as

\[ SC = \frac{\sum_{n=1}^{N} f(n)x(n)}{\sum_{n=1}^{N} x(n)} \]

Where x(n) represents the magnitude of bin number n, and f(n) represents the Center frequency of that bin.

2.1.4 Flux (Delta Spectrum Magnitude)
The Flux, or Delta Spectrum Magnitude, feature is a measure of the rate at which the spectral shape...
changes, or fluctuates. It is calculated by summing the squared differences of magnitude spectra of two neighbouring frames

\[ F = \sum_{k=1}^{N/2} \left| X[n] - X[n-1] \right|^2 \]

2.2 FREQUENCY CEPSTRUM COEFFICIENTS

The second group is the Frequency Cepstrum Coefficients (FCC), which includes the Mel Frequency Cepstrum Coefficients (MFCC) and the Logarithmic Frequency Cepstrum Coefficients (LFCC). These are all power spectrum representation features calculated with different frequency scales. The most frequently used coefficients for these systems are the MFCC. These are computed by taking the FFT of every analysis window, mapping the spectrum to the Mel scale, taking the base 10 logarithms of the powers and then applying a Discrete Cosine Function (DCT) to decorrelate the coefficients.

The overall performance of MFCC features was shown in [5] to be slightly better than the SLL feature. This relates to the fact that MFCC performs better at pop and rock music, but somewhat worse at classical music that contains very little vocal information.

2.3 PSYCHOACOUSTIC FEATURES

These features are more closely based on our perception of sound, and are therefore called psychoacoustic. Loudness is the sensation of signal strength, and is primarily a subjective measure for us to rank sounds from weak to strong. Loudness can be calculated (Calculated Loudness) and is then measured in Sone. One Sone is defined as the loudness of a pure 1000 Hz tone at 40 dB re 20 μPa [6].

Roughness is described in as “the perception of temporal envelope modulations in the range of about20-150Hz, maximal at 70 Hz” and is also said to be a primary component of musical dissonance. Sharpness is a measure of the high frequency energy related to the low frequency energy strength. Sounds with lots of energy in the higher frequencies, and low energy levels in the lower frequencies are considered sharp.

2.3.1 SPECTRAL ROLLOFF

As the Spectral Centroid, the Spectral Rolloff is also a representation of the spectral shape of a sound, and they are strongly correlated. It’s defined as the frequency where 85% of the energy in the spectrum is below that frequency. If K is the bin that fulfills

[\sum_{i=0}^{K} x(n) = 0.85 \sum_{n=0}^{N-1} x(n)]

Then the Spectral Roll off frequency is f(K), where x(n) represents the magnitude of bin number n, and f(n) represents the center frequency of that bin.

2.3.2 FLUX (DELTA SPECTRUM MAGNITUDE)

The Flux, or Delta Spectrum Magnitude, feature is a measure of the rate at which the spectral shape changes, or fluctuates. It is calculated by summing the squared differences of magnitude spectra of two neighbouring frames. This feature has shown good results in the SMD task in.

\[ F = \sum_{k=1}^{N/2} \left| X[n] - X[n-1] \right|^2 \]

2.3.3 LOUDNESS: can be approximated by the square root of the energy of the signal computed from the short time Fourier transform, in decibels.

2.3.4 PITCH: the Fourier transformation of a frame delivers a spectrum, from which a fundamental frequency can be computed with an approximate greatest common divisor algorithm.

2.3.5 TONE (BRIGHTNESS AND BANDWIDTH): Brightness is a measure of the higher-frequency content of the signal. Bandwidth can be computed as the magnitude weighted average of the differences between the spectral components and the centroid of the short time Fourier transform. It is zero for a single sine wave, while ideal white noise has an infinite bandwidth. [7]

2.3.6 MEL-FILTERED CEPSTRAL COEFFICIENTS (OFTEN ABBREVIATED AS MFCCS) can be computed by applying a mel-spaced set of triangular filters to the short-time Fourier transform, followed by a discrete cosine transform. The word “cepstrum” is a play on the word “spectrum” and is meant to convey that it is a transformation of the spectrum into something that better describes the sound characteristics as they are perceived by a human listener. A mel is a unit of measure for the perceived pitch of a tone. The human ear is sensitive to linear changes in frequency below 1000 Hz and logarithmic changes above. Mel filtering is a scaling of frequency that takes this fact into account.

2.3.7 DERIVATIVES: Since the dynamic behavior of sound is important, it can be helpful to calculate the instantaneous derivative (time differences) for all of the features above. [7]

3 FEATURE EXTRACTION TECHNIQUES

In music information retrieval system various features are identified and then to extract those features, many features extraction techniques are used like MFCC, LPC, PCA, PLP, RAS, LPC, RCC, PLPC, LDA, PCA etc.

3.1 MFCC (Mel-Frequency Cepstral Coefficients) is a commonly used feature extraction technique. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. But MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise. [11]

3.2 LPC (Linear Prediction Coefficient) is a common feature extraction technique used in spectral analysis.
analyzes the speech signal by estimating the formants, removing speech signal, and estimating the intensity and frequency of the remaining buzz. The process is called inverse filtering, and the remaining signal is called the residue. It is most powerful signal analysis technique. This method is robust, reliable, and accurate method for estimating the parameters.

3.3 PCA (Principal Component analysis) is a non linear feature extraction method and also known as karhunen-loeve expansion. This method is good for Gaussian data [2]

3.4 LDA (Linear Discriminant analysis) It is a nonlinear feature extraction method. It is supervised linear map, fast, eigenvector based method. Better than PCA for classification [2]

3.5 ICA (Independent Component Analysis) It is a nonlinear feature extraction method. It is linear map, iterative non-Gaussian method. The procedure used for implementation is blind course separation, used for demixing non-Gaussian distributed sources (features) .

3.6 LPCC (LPC-Derived or Linear Predictive Cepstral Coefficients) One of the common short term spectral measurements currently used are LPC derived cepstral coefficients (LPCC) and their regression coefficients [6]. LPCC shows the differences of the biological structure of human vocal tract and is computed through iteration from the LPC Parameters to the LPC Cepstrum.

3.7 PLPCC (Perceptual Linear Predictive Cepstral Coefficients) PLPCC is based on the magnitude spectrum of the speech analysis window. Like MFCC and LPC which are cepstral methods, the PLPCC is a temporal method.

3.8 (PLP) Perceptual Linear prediction The Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics of hearing. PLP discards irrelevant information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system. [3]

IV CONCLUSION
In this paper various features and feature extraction techniques are studied. In audio mining, efficient fast and exact music information retrieval technique is required for new artist in music. Various low level and high level features and their mathematical formulae are studied.

References
3. Namrata Dave, Feature Extraction Methods LPC, PLP And MFCC In Speech Recognition, International Journal For Advance Research In Engineering And Technology, Volume 1, Issue VI, July 2013
7. Rainer Typke, Frans Wiering, Remco C. Veltkamp, A SURVEY OF MUSIC INFORMATION RETRIEVAL SYSTEMS 2005