International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Study paper for Timbre identification in Sound

Abhilasha
Cummins College of
Engineering, Pune

Preety Goswami
Cummins College of
Engineering, Pune

Prof. Makarand Velankar
Cummins College of
Engineering,Pune

Abstract

Acoustically, all kind of sounds are similar, but they are fundamentally different. This is partly because they encode information in fundamentally different ways. Timbre is the quality of sound which differentiates different sounds. These differences are related to some interesting but difficult questions about timbre. Study has been done to distinguish between sounds of different musicals instruments and other sounds. This requires analysis of fundamental properties of sound .We have done exploration about timbre identification methods used by researchers in music information retrieval. We have observed the major method used by many as MFCC analysis. We have covered two methods i.e. MFCC and Formant for timbre analysis in more details and done comparison of them. Our study on timbre identification method will be useful to researchers in acoustics and music analysis domain.

Keywords:-Formant, MFCC, Ceptrum, Fourier transform, Filter function, Relative power

1. Introduction

Automatic timbre identification has been in the limelight of music researchers since last decade. Timbre is defined by many people in different ways. It is the property of sound for identification of source of sound. We distinguish between different musical instruments using timbre analysis.

In case of vocal music or speech, vocal track plays important role. The vocal tract with its constituents the larynx and the vocal cord within the throat, the pharynx, the tongue, the teeth and the nose represent a biological marvel of phonation. The vibrations that generate a series of overtones are originated by the vocal folds, but the spectra of the formants are shaped to a great extent by a filtration mechanism of great plasticity, which operates by the movement of the mouth from almost closed to wide open and the movement of the tongue along the horizontal and vertical axes. Well specified positions of the mouth and the tongue can produce the standardized sounds of the speech vowels, and are also capable of imitating many natural sounds.

This paper attempts to analyze different methods for timbre identification and detail study for two methods is done. MFCC (Mel Frequency Cepstrum Coefficients) and Formant analysis are the two main methods focused on in this paper. Comparative study of these methods throws more light on the possible outcomes using them. We have also explored formant analysis using tool PRAAT and its possible use for timbre identification.

2. Properties of sounds

Sound has different properties such as frequency, pitch, power, envelope, amplitude, timbre, etc. We have presented introduction about them with more exploration on timbre in the next section [4].

2.1 Frequency

Frequency describes the number of waves that pass a fixed place in a given amount of time. So if the time it takes for a wave to pass is 1/2 second, the frequency is 2 per second. If it takes 1/100 of an hour, the frequency is 100 per hour.

2.2 Power

Power is the total amount of kinetic energy contained on the sphere's surface. It is a measurement of amplitude over time. The unit of measurement for power is the watt, named after James Watt. 1 watt = 1 Newton of work per second.

2.3 Pitch

The sensation of a frequency is commonly referred to as the pitch of a sound. A high pitch sound corresponds to a high frequency sound wave and a low pitch sound corresponds to a low frequency sound wave

2.4 Envelopes / Articulation

Dynamic envelope refers to the amplitude change over time of a sound event (usually a short one, such as an instrumental or synthesized note). As a very simple example, a note can have an initial attack characterized by the amount of time it takes to change from no sound to a maximum level, a decay phase, whereby the amplitude decreases to a steady-state sustain level, followed by a decay phase, characterized by the time it take the amplitude to change from the sustain level to 0 as described in paper[4].

3. Timbre

The main attribute that distinguishes musical instruments from each other is timbre. Timbre defines the identity and the expression of a musical sound. It is separated from the expression attributes pitch, loudness, and length of a sound. Other perceptive attributes, such as brightness and roughness, can also be helpful in understanding the dimensions of timbre. The timbre of a sound depends on its wave form, which varies with the number of overtones, or harmonics that are present, their frequencies, and their relative intensities. The determination of timbre by the waveform constitutes one of the main relations among sound attributes and relates to our perception of complex sounds. This relation is one of the most difficult to describe, since both timbre and waveform are two complex quantities. All complex sounds, such as musical instruments' sounds, are a combination of different frequencies which are multiples of the fundamental frequency. This property is referred to as "harmonicity" and the individual frequencies as harmonics.

In music timbre is the characteristic tone colour of an instrument or voice, arising from reinforcement by individual singers or instruments of different harmonics, or overtones, of a fundamental pitch. Extremely nasal timbre thus stresses different overtones than mellow timbre. The timbre of the tuning fork and of the stopped diapason organ pipe is clear and pure because the sound they produce is almost without overtones. Timbre depends on the absolute frequencies and relative amplitudes of pure tone components varying in musical instruments from dull or mellow (strong lower harmonics) to sharp and penetrating (strong higher harmonics) as mentioned in [8].

Timbre is determined by an instrument's by the frequency range within which the instrument can produce overtones, and by the envelope of the instrument's sound. The timbre of spoken vowels or of a singing voice is modified by constricting or opening various parts of the vocal tract, such as the lips, tongue, or throat.

4. Timbre Analysis Methods:-

Different methods have been proposed to analyze Timbre those are mentioned in paper [4] and [9]. In our paper we are basically focusing on using MFCC and Formants to analyze Timbre.

4.1 MFCC:-

Many different approaches have been made by different researchers to classify the vocal and non-vocal parts. Basically most of them have applied statistical classifier by modeling the vocal and non-vocal part separately. For classification, two main components are required which are (1) features and (2) classifiers. Different features have been explored for singing voice detection. These features are Mel-frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPCs), Perceptual Linear Prediction Coefficients (PLPs) and the Harmonic Coefficients. MFCC (it features extraction process.) is widely used for general sound classification tasks and they are called short term features because they are calculated in short time frames.

The first step in calculating MFCC is converting the windowed frame of speech into the frequency domain by using the Discrete Fourier Transform (DFT). This is usually done by using Fast Fourier Transform (FFT) algorithm. Then the logarithm magnitude of complex Fourier coefficients is calculated as mentioned in paper [10].

MFCCs constitute Mel Frequency Cepstrum (MFC) is used for the representation of audio signal. In MFC, the frequency and are equally spaced on Mel scale, compared with frequency bands in the normal cepstrum C. MFCC is one of the popular features used for signal processing and it is used to represent tumbrel information of signal.

As MFCC provides spectral related information about the signal, it will be useful to find the variation among different audio files. The MFCC feature extraction process involves these steps as mentioned in paper [5]: Initially the audio signal x (n) is passed onto the high-pass filter in order the preserve the high frequency component which is being suppressed during the sound production mechanism of humans and pre-emphasized obtained.

The output obtained from the high pass filtering of original signal is given by,

$$x_2(n) = x(n) - a * x(n-1)' \qquad (1)$$

Where $x_2(n)$ given in equation (1) is the output signal and the value of a is usually between 0.9 and 1.0

The audio signal is divided into frames of 20 30 ms with optional overlap of 1/3 ½ of the frame size. Then each frame has to be multiplied with the hamming window in order to reduce discontinuity near the boundary of audio segments. Apply Fourier transform to each frame and obtain magnitude frequency response of each frame.
The cepstrum computed after a non-linear frequency warping onto a perceptual frequency scale as mentioned in [11], the Mel-frequency scale. The cepstrum is the inverse Fourier transform of the log-spectrum log S

$$c_n = \frac{1}{2\pi}\int_{\omega=-\pi}^{\omega=\pi} \log S(\omega) \exp j\omega n d\omega \qquad (2)$$

The C_n in equation (2) are called Mel frequency cepstrum coefficients (MFCC). Cepstrum coefficients provide a low-dimensional, smoothed version of the log spectrum, and thus are a good and compact representation of the spectral shape. They are widely used as feature for speech recognition, and have also proved useful in musical instrument recognition. Fig. 1 shows the steps for feature extraction using MFCC as mentioned in paper [5].

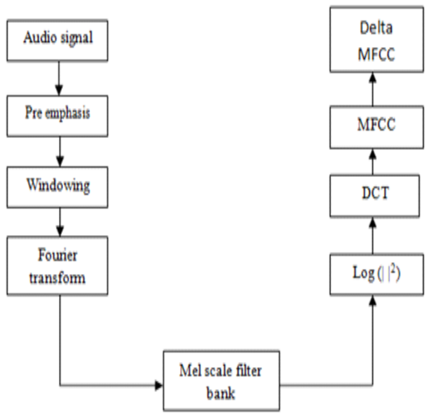

Fig. 1 MFCC feature extraction process

Mel filter bank, usually triangular band pass filter is taken and multiplies it with the magnitude frequency response of each frame and energies of each band pass filter on Mel scale.

The vocal/non-vocal classifier unit shown in the Fig .l consists of a front-end signal processor that converts digital waveforms into spectrum-based feature vectors, and a backend statistical processor that performs modeling, matching and decision making using MFCC as mentioned in paper [9].

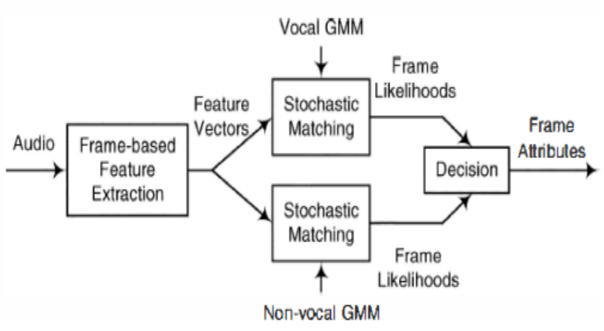

Fig.2 Front-end signal processor

4.2 Formant:-

One of the elements of an acoustic analysis is the measurement and comparison of formants. Formants are defined as peaks in the energy spectrum of vocalic sounds which correspond to the resonant frequencies of the vocal tract. According to the scientific analysis of voice in speech and singing, which is summarized in standard textbooks of acoustics, each vowel is characterized by a set of formants; a formant being a certain range of frequency with high acoustical energy emission. In the current scientific analysis of speech, generally four formants are used to identify a vowel, the first two being the most important. The numerical frequency values of these formants are extracted from the power spectra at a point of steady state by the algorithm of linear predictive coding (LPC) as mentioned in [1]. The frequencies of the resonances characterize vowel quality (i.e. vowel height and vowel frontness). The formant with the lowest frequency is labeled as the first formant (F1) and is inversely related to vowel height. The formant with the next highest frequency is labeled as the second formant (F2) and is directly related to vowel frontness. The third formant (F3) is considered to remain relatively constant for individuals as mentioned in [12]. The vocal tract with its constituents the larynx and the vocal cord within the throat, the pharynx, the tongue, the teeth and the nose represent a biological marvel of phonation. The vibrations that generate a series of overtones are originated by the vocal folds, but the spectra of the formants are shaped to a great extent by a filtration mechanism of great plasticity, which operates by the following means:

1. by the movement of the mouth from almost closed to wide open, while also changing the shape of the opening from a rounded to a horizontally stretched one; and

2. By the movement of the tongue along the horizontal and vertical axes, from the front to the back, and from down to up. Well specified positions of the mouth and the tongue can produce the standardized sounds of the speech vowels, and are also capable of imitating many natural sounds.

In study done by joseph nagyvary in [1] violin can be compared with the human voice using formant and similarity can be obtained. The sound generating mechanism of some music instrument like the violin is much more restricted by its rigidity, and it is superior only in the wider frequency range of its four strings as verified by, which can also provide a more secure pitch than the human voice. In contrast to the voice, the timbre of the violin tone is due to its greater variety of vibrating elements that are selectively brought into resonance with the appropriate vibrating modes of the violin string. The violin does not have any visibly moving filtration mechanism to shape the power spectra of the notes, although some of its parts could selectively absorb energy without radiating it. The special tone color of each violin note comes about mainly by the enhancement of certain string modes by the corresponding resonance modes of the belly: the radiation from the back and the ribs are of lesser significance.

According to Philip Harrison [12] for the purposes of phonetic analysis, formants are generally represented by their centre frequency which corresponds to the local frequency at which the energy level is the highest. Formants can be visualized and measured in several different ways. Probably the most common way of visualizing formants is through the generation of spectrograms.

4.3 Spectrogram

A spectrogram, or sonogram, is a visual representation of the spectrum of frequencies in a sound. Spectrograms are sometimes called spectral waterfalls, voiceprints, or voice grams. Spectrograms can be used to identify spoken words phonetically, and to analyze the various calls of animals. They are used extensively in the development of the fields of music, sonar, radar, and speech processing, seismology, etc. The instrument that generates a spectrogram is called a spectrograph. The sample outputs on the right show a select block of frequencies going up the vertical axis, and time on he horizontal axis as mentioned by Tong Zhang [3].

5. Timber Analysis using Formant

5.1 Fourier analysis of complex sounds:-

Relative Power (dB) = **P*n***= 10 log10 [(**A*n*/A1**)2] = 20 log10 [(**A*n*/A1)**] (3)

	1	2	3	4	5	6	7	8	9
Frequency(Hz)	440	880	1320	1760	2200	2640	3080	3520	3960
Harmonic number=fn/f1	1	2	3	4	5	6	7	8	9
Amplitude(volt)	2.000	0.00	0.667	0.000	0.400	0.000	0.286	0.000	0.222
Relative Amplitude=An/A1	1.000	0.00	0.333	0.000	0.200	0.000	0.143	0.000	0.111
Relative Power=(An/A1)^2	1.0000	0.0000	0.1111	0.0000	0.0400	0.0000	0.0204	0.0000	0.0123
Relative Power(db)	0.0	Error	-9.5	Error	-14.0	Error	-16.9	error	-19.1

Table 1.Spectrum of squarewave.

Amplitudes can be used to represent the different waves. In table1 there is representation of Spectrum of square wave which is referred from paper [13].

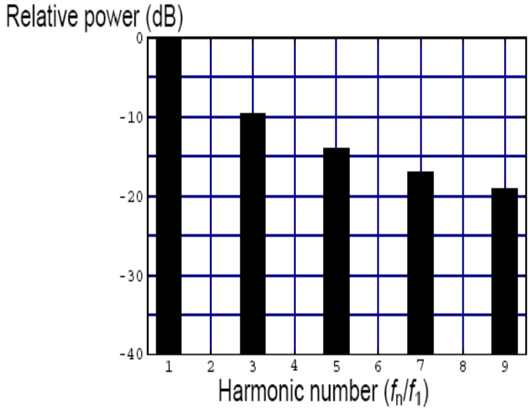

Fig 3 Relative power referred from [13].

Bars for the even harmonics are missing since these harmonic are "missing" from a square wave.

5.2 Speech analysis

The features that distinguish specific **vowels** are called **formants**.

The first two formants are particularly important in speech recognition. The frequency of the first formant increases as we open our mouth wider and lower the tongue

a) **Identify** on the graph the position of the peak corresponding to the fundamental (hint: it is the biggest peak on the graph);

b) Now label the other peaks

c) Generally, the lowest harmonics have higher power

Since the logarithm of zero is negative infinity, the Relative Power (dB) for the even harmonics is "Error".

Most **consonants** do not have harmonic frequency spectra. The features that distinguish consonants are periods of silence, voice bars, noise, and the consonant’s effects on the frequency spectra of adjacent vowels. Consonants are classified by:

-manner of articulation
-place of articulation, and
-as voiced / unvoiced.

As mentioned in paper [14] the consonant types, classified by manner of articulation, include:

a) **Plosive or stop** (p, b, t, d, k, g) – produced by blocking the flow of air somewhere in the vocal tracts

b) **Fricative** (f, s, sh, h, v, th, z) – produced by constricting the air flow to produce a turbulence

c). **Nasal** (m, n, ng) – produced by lowering the soft palate

d) **Liquid** (l, r) – generated by raising the tip of the tongue

e) **Semi-vowel** (w, y), always followed by a vowel

Filter function (dB) = Observed Power (dB) – Source Power (dB) (4)

The filter function given in equation (4) changes as we shape our mouth, adjust tongue position etc to pronounce a given sound. In particular, a sound corresponding to a specific vowel is characterized by a specific filter function that changes little from one individual to another. A filter function for a given vowel, plotted as a function of frequency, exhibits characteristic maxima, called **formants**.

Harmonic no.	1	2	3	4	5	6	7	8	9	10	11
Frequency(Hz)	105	210	315	420	525	630	735	840	945	1050	1155
Power(db)	39.6	34.5	31.0	31.6	33.0	37.2	38.4	32.0	31.6	36.1	28.1
Relative Filter =Pn-P1	0.0	-5.1	-8.6	-8.0	-6.6	-2.4	-1.2	-7.6	-8.0	-3.5	-11.5
Approx Formant Position(peak)							F1			F2	

Table 2.Formnat for spoken “aw” sound

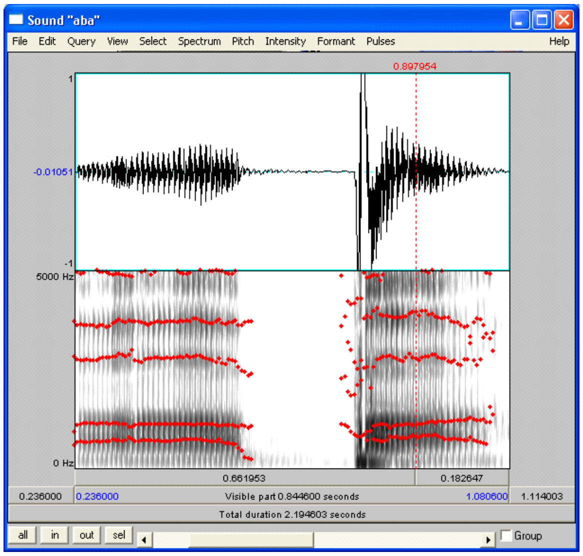

Fig 4 Frequency spectrogram with formants overlaid of the vowel-consonant combination "/aba/".

The first "a" vowel is clearly seen, then a period of silence with a faint voice bar, a sharp burst, and then the second "a" vowel.

5.3 PRAAT:-

Praat is a wonderful software package written and maintained by Paul Boersma and David Weenink of the University of Amsterdam. Available for free with open source code, there is simply no better package for linguists to use in analyzing speech. It is a computer program with which phoneticians can analyze, synthesize, and manipulate speech, and create high-quality pictures for their articles and thesis as referred in paper [6].

Following analysis can be done:-

1. Spectral analysis
2. Pitch analysis
3. Formant analysis
4. Intensity analysis
5. Manipulation

6. Conclusion and future directions

Based on our study we can use formant analysis for timbre identification. Although formant analysis is primarily used for vowel identification in speech recognition, we can use the same for timbre identification. We are analyzing different sound tracks of music and human speech. For that we have used Praat as tool that provides the analysis result of timbre.

We propose to develop a software tool that can be used for automatic identification of timbre. Automatic timbre identification has wide applications in speech and music domain such as automatic audio classification, audio clustering, speech based security applications, music search etc.

7. References

[1] A Comparative Study of Power Spectra and Vowels in Guarneri Violins and Operatic Singing by Joseph Nagyvary.

[2] The temporal character of timbre (Diploma Thesis) by Miha Ciglar.

[3] Instrument Classification in Polyphonic Music Based on Timbre Analysis by Tong Zhang.

[4] Timbre Models of Musical Sounds by Rapport .

[5] Audio Retrieval using Timbral Feature R.Christopher Praveen Kumar, D.Abraham Chandy

[6] Using Praat for Linguistic Research by Will Styler

[7] New techniques for improving the practicality of an svm-based Speech/music classifier by chungsoo lim, seong-ro lee and yeon-woo lee

[8] Multiscale fractal analysis of musical instrument Signals with application to recognition by athanasia zlatintsi.

[9] Signal processing for segmentation of vocal and Non-vocal regions in songs: a review by A. bonjyotsna & M. bhuyan

[10] Fast and scalable system for automatic artist identification by Sajad shirali-shahreza

[11] Improving Timbre Similarity : How high's the sky ? by Jean-Julien Aucouturier Francois Pachet

[12] Variability of formant measurements by Philip Harrison

[13]Fourier analysis of Complex sounds Amplitude, loudness, and decibels by Prof. Larry Kesmodel

[14]Speech analysis by Prof. Larry Kesmodel.