ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Automatic Classification of Instrumental Music & Human Voice Using Formant Analysis

[I]Diksha Raina, [II]Sangita Chakraborty, [III]M.R Velankar
[I,II]Dept. of Information Technology, Cummins College of Engineering, Pune, India
[III]Asst. Professor, Dept. of Information Technology, Cummins College of Engineering, Pune, India

Abstract
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 used formant analysis for distinguishing between instrumental music & human voice. With automatic identification of instrumental music & human voice, we can avoid manual changing of channels which is generally done by drivers while driving the vehicle which is risky. The variation of formants can successfully identify instrumental and human voice files. After analyzing all the four formants i.e. F1, F2, F3, F4, we found out that 1st formant (F1) of an instrumental and a speech file will be specific to find a distinguishable pattern. Comparing the average value of F1 with the individual F1 values given at a particular time, we came to the conclusion that the deviation in speech is more than that of instrumental file .The improvised version of the algorithm can be used to sort song collections based on artists & instruments. Combining this algorithm of differentiation of instrumental/human voice files with various implementations of transistors, we can do automatic switching of FM stations to a station playing music from a station where RJ is speaking or advertisement is going on.

Keywords
Music information retrieval, sound, spectrogram, formant, MFCC.

I. Background

Acoustically, all kind of sounds are similar, but they are fundamentally different. This is partly because they encode information in fundamentally different ways. Music and speech both use sound features, and so are received and analyzed by the same organs. Many of their similar acoustical features are used in different ways. They are saved as audio files without any distinction. One purpose of this project is to compare and differentiate them [1].

Our Domain Music Information Retrieval (MIR) is a multidisciplinary research field that includes wide range of disciplines. An incomplete listing of these disciplines includes acoustics, psychoacoustics, signal processing, computer science, musicology, library science, informatics, and machine learning, etc [1].

Music is a complex amalgam of acoustic, rhythmic, harmonic, structural, and cultural phenomena. The grand challenge of MIR research is the development of retrieval systems that deal with music on its own terms. The main problem addressed by us was to analyze & develop the feature of automatic identification of instrumental music/human voice in PRAAT.

II. Introduction

Approaches to solve the problem:-

A. Mfcc

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum") [2].

MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken on a telephone. MFCCs also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc.

B. Formant

Formants are defined by Gunnar Fant as 'the spectral peaks of the sound spectrum of the voice', also used to distinguish specific vowels. It is often measured as an amplitude peak in the frequency spectrum of the sound.One of the elements of an acoustic analysis is the measurement and comparison of formants. According to the scientific analysis of voice in speech and singing, 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. Based on our study we can use formant analysis for timbre identification [3].

III. Related Work

A. Mfcc

MFCCs have been used once for speech recognition by Young, Woodland and Byrne (1993), their success has been due to their ability to represent the speech amplitude spectrum in a compact form.

Steps for MFCC:

1. Divide the speech signal into frames. Usually by applying window function," Hamming window", which removes edge effects at fixed intervals. Here we need to model small (typically 20ms) sections of the signal that are statistically stationary. We generate a cepstral feature vector for each frame.
2. Take the Discrete Fourier Transform (DFT) of each frame. We then retain only the logarithm of the amplitude spectrum. Discard phase information because the amplitude of the spectrum is much more important than the phase. Then, take the logarithm of the amplitude spectrum because the perceived loudness of a signal has been found to be approximately logarithmic.

3. Smooth the spectrum and emphasize perceptually meaningful frequencies.
4. The Mel scale is based on a mapping between actual frequency and perceived pitch as apparently the human auditory system does not perceive pitch in a linear manner. The mapping is approximately linear below 1 kHz and logarithmic above. The components of the Mel-spectral vectors calculated for each frame are highly correlated [2].

B. Formant

The main aim was to compare recognition results using formant features for describing spectral details as compared to conventional mel-cepstrum representation. In order to directly assess the usefulness of the formants, the same total number of features was used for both representations, and exactly the same low-order cepstrum features were used for describing general spectral shape. Thus the only difference was in the use of formants versus higher cepstral coefficients for representing detailed spectrum shape [4].

Formants are the distinguishing or meaningful frequency components of human speech and of singing. By definition, the information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. In speech, these are the characteristic partials that identify vowels to the listener. Most of these formants are produced by tube and chamber resonance, but a few whistle tones derive from periodic collapse of Venturi effect low-pressure zones. The formant with the lowest frequency is called f1, the second f2, the third f3 and the fourth f4.

Table 1: List of first 4 formants for Instrumental file.

f1	f2	f3	**f4**
576	1240	1195	2862
633	948	1881	3209
621	1694	2243	3262
612	1517	2315	3711
624	1654	2199	2392
610	1511	2333	3900
591	1503	2351	3950

The first two formants determine the quality of vowels in terms of the open/close and front/back dimensions. The first formant f1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant f2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]). Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. However, the first two formants are most important in determining vowel quality, and this is often displayed in terms of a plot of the first formant against the second formant [3].

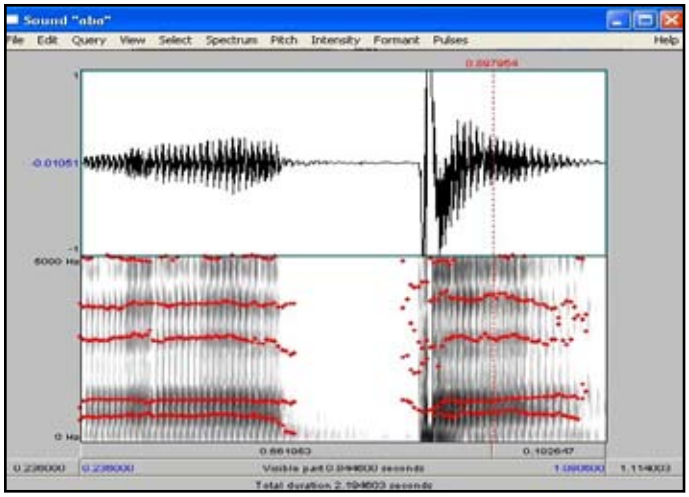

Fig. 1: Spectrogram with formants overlaid of the vowel-consonant combination "aba".

Fig.1 shows spectrogram for the human voice recorded and use of formants for vowel-consonant combination. The first "a" vowel is clearly seen, then for "b", a period of silence with a faint voice bar, a sharp burst, and then the second "a" vowel.

IV. Our Approach

A. Spectrogram

The spectrogram is a spectro-temporal representation of the sound. The horizontal direction of the spectrogram represents time, the vertical direction represents frequency. The time scale of the spectrogram is the same as that of the waveform, so the spectrogram reacts to the zooming and scrolling. To the left of the spectrogram, the frequency scale is seen. The frequency at the bottom of the spectrogram is usually 0Hz (hertz, cps, cycles per second), and a common value for the frequency at the top is 5000 Hz.

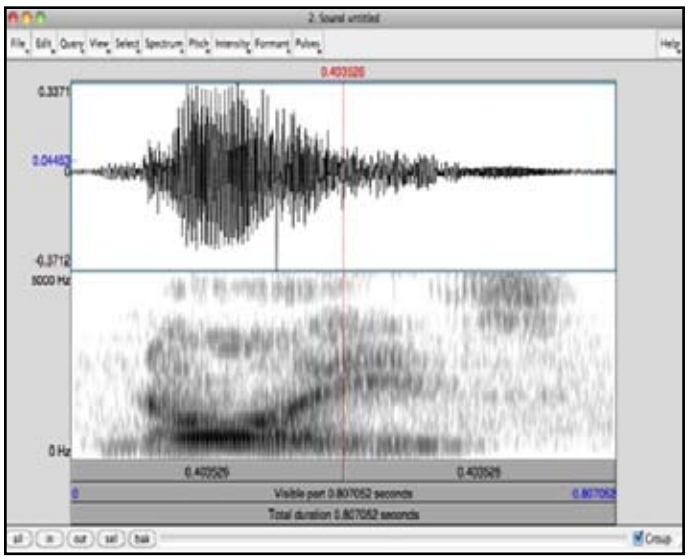

Fig. 2: Spectrogram of instrumental file

Darker parts of the spectrogram mean higher energy densities; lighter parts mean lower energy densities. If the spectrogram has a dark area around a time of 1.2 seconds and a frequency of 4000 Hz, this means that the sound has lots of energy for those high frequencies at that time. We have observed that spectrogram

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

analysis is not a correct approach for the problem to be solved.

B. Formant

In Praat, the inbuilt feature of formant allows us to view the formant analysis of a particular selected area. We get four formants for particular points which are f1, f2, f3, f4. Analyzing all the formants we found out that there is no distinguishable pattern found out for all of them together, but if we look at the f1 formant of an instrumental and a speech file, we are finding a specific distinguishable pattern. The deviation is speech is more than that of instrumental file. We have tested different sound files for our proposed approach.

The details of same are as under.

1. Total Speech clips tested: 37 (Male and Female voice samples).
2. Total Instrumental clips tested: 63 (Violin, Guitar, Piano, and Flute)

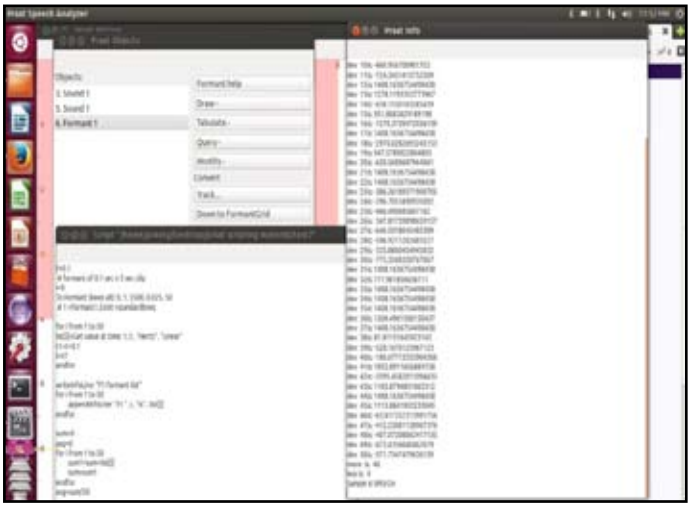

Fig. 3: Screenshot of analysed human voice file in Praat

Above data is for the human voice sample of 5 seconds and following are few important data observations.

1. Average(f1)=683
2. Deviation= Avg-f1i
3. D1=683-633= 50 (where i is ith sample and D is deviation)
4. Deviation from100(compared with average value) 70%-90%
5. Deviation from150(compared with average value) 50%-60%

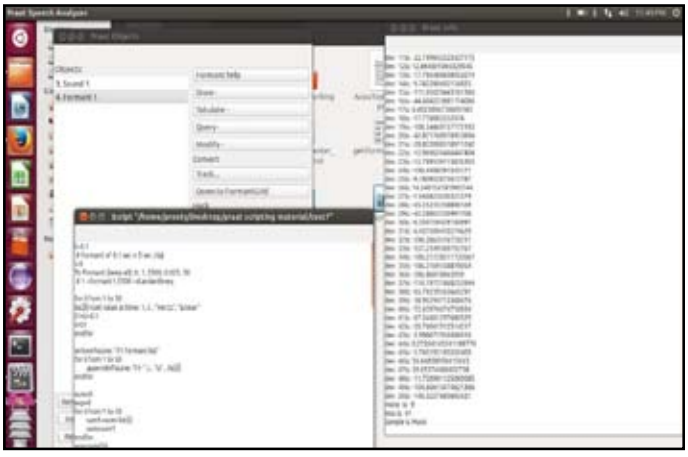

Fig. 4: Screenshot of analysed instrumental musical file in Praat.

Above data is for the instrumental sample of 5 seconds and following are few important data observations.

1. Average(f1)=626
2. Deviation= Avg-f1i
3. D1=626-576= 109 (where i is ith sample and D is deviation)
4. Deviation from100(compared from average value)20%-33%
5. Deviation from150(compared from average value)15%-30%

Data observations from Fig. 3 and 4 i.e. human voice sample and instrumental sample has given us the proposed approach for distinguishing them using formant f1 and has given us reasonably good results.

V. Algorithm For Identification

We have taken a sample of 5 sec with time difference of 0.1 sec which makes use of 50 formant values. So, count <50.

Where,

T = formant of 0.1 sec n 5 sec clip

T1= variable (T1=T+0.1)

Dev = Deviation with respect to formant f1

M = More deviation

L = Less deviation

M1, L1 = Variables

Count = Counting the no of formants

Average = Average of f1

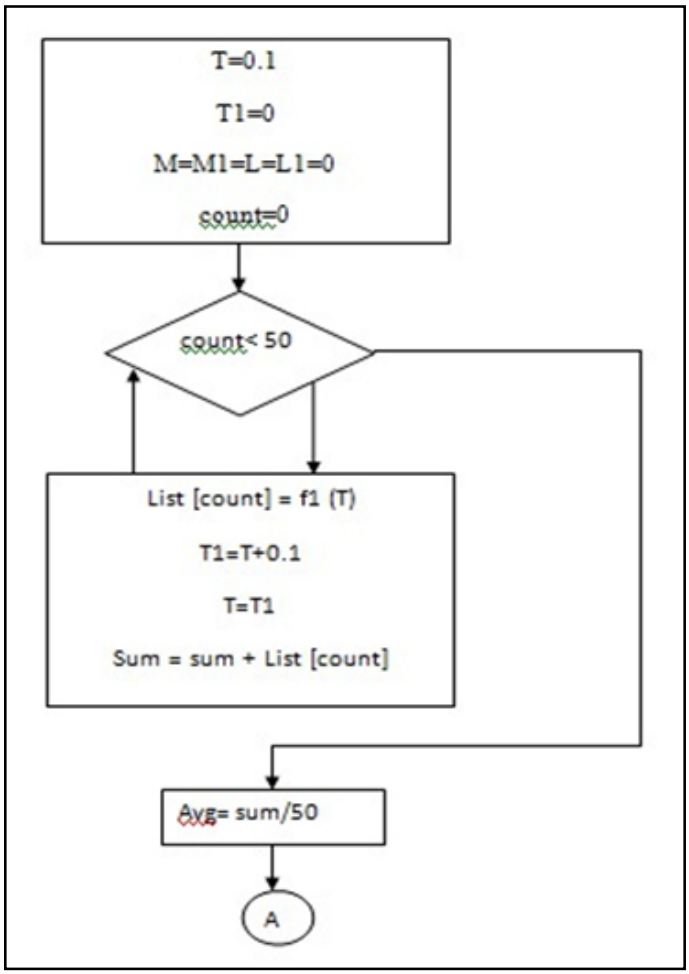

Fig. 5: Flowchart classification of instrumental music & human voice using formant analysis (PART A)

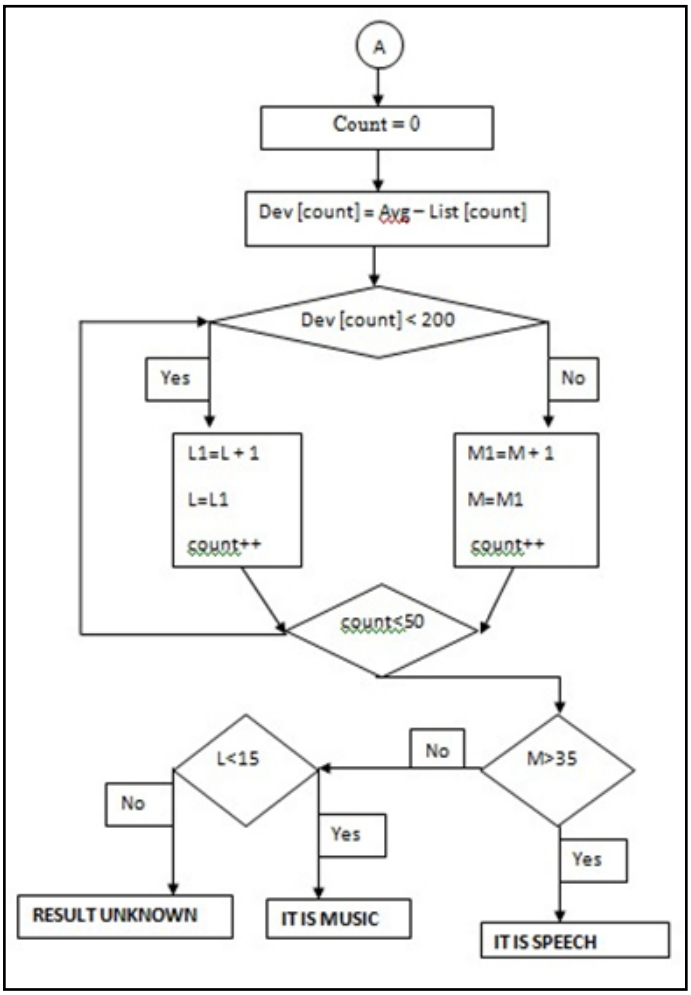

Fig. 6: Flowchart classification of instrumental music & human voice using formant analysis (PART B)

VI. Results

- For Human voice: out of 37 samples, 32 clips were identified correctly according to our experiments and algorithm used.
- For Instrumental music: out of 63 samples, 45 clips gave the correct result according to our experiments and algorithm used.

Table 2: Test results of instrumental music files

Samples tested	Total Samples	Correct identification	Accuracy of our algorithm in %
Instrumental -Guitar	22	18	82
Instrumental -Piano	10	7	70
Instrumental -Violin	12	8	67
Instrumental– Flute	19	12	63
Overall Results	63	45	72

It has been observed that for different instruments the accuracy of our algorithm is different for the tested samples. It is likely that the possible less accuracy for violin and flute is due to more close resemblance of them to human voice than other instruments. This is a likely outcome and we need to study timber characteristics of instruments and there closeness to human voice in more depth to support the argument.

Table 3: Test results of human voice files

Samples tested	Total Samples	Correct identification	Accuracy of our algorithm in %
Male voice	10	9	19
Female voice	27	23	85
Overall Results	37	32	86.4

It has been observed that accuracy for male voice is better than female voice. We need to work more on different parameters and aspects to fine tune our algorithm and draw conclusions. Our initial results are encouraging and further study and improvisation in algorithm can produce better results.

VII. Conclusion/ Future Scope

The variation of formants can successfully identify instrumental and speech files. Future scope based on our algorithm can be differentiation of singing, recitation, various different instruments, male/female/child voice classification, chorus voice identification, classification of different melody. Some of the clippings (may be instrumental or speech) giving ambiguous results may be identified correctly if we can combine the MFCC and formant together. The accuracy of our algorithm should be improvised and the consideration of differentiation of human voices (male/female/child) and mixed voices (chorus) should be taken care of. The improvised version of the algorithm can be used to sort song collections based on artists & instruments. Combining this algorithm of differentiation of instrumental/speech files with various implementations of transistors, we can do automatic switching of FM stations to a station playing music from a station where RJ is speaking or advertisement going on.

References

[1]. Downie, J. Stephen, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research" [Page no-1].

[2]. Logan Beth, Cambridge research laboratory, "Mel Frequency cepstrum coefficient for music modeling" [Page 4-7].

[3]. Wikipedia contributors, "Formant," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Formant&oldid=606192133 (accessed April 30, 2014).

[4]. Kuhn Michael, Computer Engineering and Networks Laboratory, " Social Audio Features for Advanced Music Retrieval Interfaces"

[5]. Yading Song, Simon Dixon, Marcus Pearce, Centre for Digital Music, Queen Mary University of London, "Evaluation of musical features for emotion classification".

[6]. Meinard Muller, Verena Konz, Peter Grosche, Saarland University, Max-Planck-Institute f¨ur Informatik Campus E1 4, 66123 Saarbr¨ucken, Germany, "Music Information Retrieval".

[7]. Wu Chou and Liang Gu ,Bell Laboratories, Lucent Technologies, Murry Hill, NJ 07974, USA"Robust Singing

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Detection in Speech/Music Discriminator Design"

[8]. John N. Holmes (1), Wendy J. Holmes (2) and Philip N. Garner (2) (1) Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH, U.K. (2) Speech Research Unit, DRA Malvern, St. Andrews Road, Malvern, Worcs., WR14 3PS, U.K. "USING FORMANT FREQUENCIES IN SPEECH RECOGNITION" [page 3-5].

[9]. Wikipedia contributors, "Mel-frequency cepstrum," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Melfrequency_cepstrum&oldid=600281256 (accessed April 30, 2014).

[10]. Abhilasha, Preety Goswami, Prof. Makarand Velankar, "Study paper for Timbre identification in Sound" International Journal of Engineering Research & Technology Vol.2 - Issue 10 (October - 2013)