A Speech/Music Discriminator based on Frequency energy, Spectrogram and Autocorrelation

Sumit Kumar Banchhor, Om Prakash Sahu, Prabhakar

***Abstract*—Over the last few years major efforts have been made to develop methods for extracting information from audio-visual media, in order that they may be stored and retrieved in databases automatically, based on their content. In this work we deal with the characterization of an audio signal, which may be part of a larger audio-visual system or may be autonomous, as for example in case of an audio recording stored digitally on disk. Our goal was first to develop a system for segmentation of the audio signal, and then classify into one of two main categories: speech or music.**

Segmentation is based on mean signal amplitude distribution, whereas classification utilizes an additional characteristic related to frequency. The basic characteristics are computed in 2sec intervals, resulting in the segments' limits being specified within an accuracy of 2sec. The result shows the difference in human voice and musical instrument.

***Index Terms*—Speech/music classification, audio segmentation, zero crossing rate, short time energy, spectrum flux.**

I. INTRODUCTION

In many applications there is a strong interest in segmenting and classifying audio signals. A first content characterization could be the categorization of an audio signal as one of speech, music or silence. Hierarchically these main classes could be subdivided, for example into various music genres, or by recognition of the speaker. Audio classification can provide useful information for understanding and analysis of audio content. It is of critical importance in audio indexing. Feature analysis and extraction are the foundational steps for audio classification and identification. In the present work only the first level in the hierarchy is considered.

Revised Manuscript Received on March 2012.

Sumit Kumar Banchhor, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University, GD Rungta College of Engineering and Technology, Bhilai, India, +91 9893880318, (e-mail: sumit.9981437433@gamil.com).

Om Prakash Sahu, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University, GD Rungta College of Engineering and Technology, Bhilai, India, +91 9827889143, (e-mail: omprakashsahu@gamil.com).

Prabhakar, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University, GD Rungta College of Engineering and Technology, Bhilai, India, +91 9165860844, (e-mail: prabhuelexstar@gamil.com).

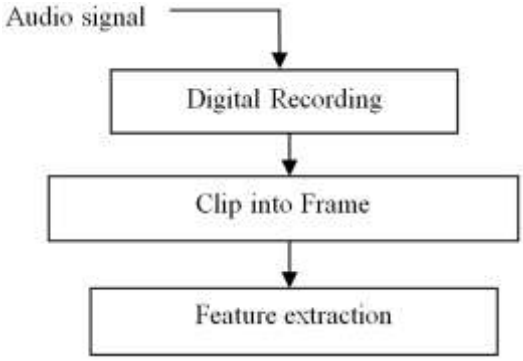

Figure 1.1 Basic processing flow of audio content analysis.

Figure 1.1 shows the basic processing flow which discriminates between speech and music signal. After feature extraction, the input digital audio stream is classified into speech, non speech and music.

II. PREVIOUS WORK

A variety of systems for audio segmentation and/or classification have been proposed and implemented in the past for the needs of various applications. We present some of them in the following paragraphs:

Saunders [4] proposed a technique for discrimination of audio as speech or music using the energy contour and the zero-crossing rate. This technique was applied to broadcast radio divided into segments of 2.4 sec which were classified using features extracted from intervals of 16 msec.

Four measures of skewness of distribution of zero-crossing rate were used with a 90% correct classification rate. When a probability measure on signal energy was added a performance of 98% is reported.

Scheirer and Slaney [5] used thirteen features, of which eight are extracted from the power spectrum density, for classifying audio segments. A correct classification percentage of 94.2% is reported for 20 msec segments and 98.6% for 2.4 sec segments. Tzanetakis and Cook [8] proposed a general framework for integrating, experimenting and evaluating different techniques of audio segmentation and classification. In addition they proposed a segmentation method based on feature change detection. For their experiments on a large data set a classifier performance of about 90% is reported.

In [9] a system for content-based classification, search and retrieval of audio signals is presented. The sound analysis uses the signal energy, pitch, central frequency, spectral bandwidth and harmonicity.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This system is applied mainly in audio data collections. In a more general framework related issues are reviewed in [1].

In [3] and [6] cepstral coefficients are used for classifying or segmenting speech and music. Moreno and Rifkin [3] model these data using Gaussian mixtures and train a support vector machine for the classification. On a set of 173 hours of audio signals collected from the WWW a performance of 81.8% is reported. In [6] Gaussian mixtures are used too, but the segmentation is obtained by the likelihood ratio. For very short (26 msec) segments a correct classification rate of 80% is reported.

A general remark concerning the above techniques is that often a large number of features are used. Furthermore the classification tests are frequently heuristic-based and not derived from an analysis of the data.

III. SPEECH DATABASE FORMULATION

Speech recording from age group, 19-25 was taken. This utterance was spoken at the habitual speaking level and most talkers repeated the phrases 10 times.

IV. METHODOLOGY

The target vowel was manually segmented using GOLDWAVE software and stored with .wav extension.

V. EXPERIMENT AND RESULT

A. *Result using frequency energy and sub bands*

The frequency spectrum is divided into four sub-bands with intervals [0, ω 0 /8], [ω 0 /8, ω 0 /4], [ω 0 /4, ω 0 /2] and [ω 0 /2, ω 0]. The ratio between sub band power and total power in a frame is defined as:

$$D = \frac{1}{FE}\int_{li}^{Hi} |F(w)|^2 \, dw \quad \text{------ (1)}$$

Here $F(\omega)$ denotes the Fast Fourier Transform (FFT) coefficients, Where *Lj* and *Hj* are lower and upper bound of sub-band j respectively. FE is an effective feature, especially for discriminating speech from music signals. In general, there are more silence frames in speech than in music; thus, the variation of FE (or E) measure will be much higher for speech than that for music.

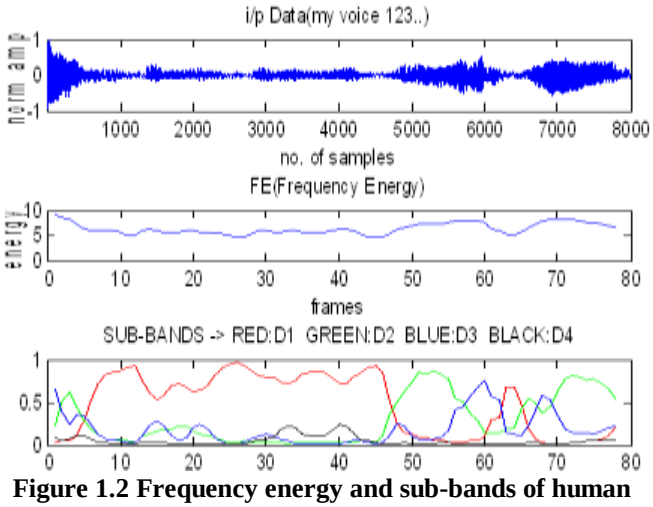

Figure 1.2 Frequency energy and sub-bands of human voice.

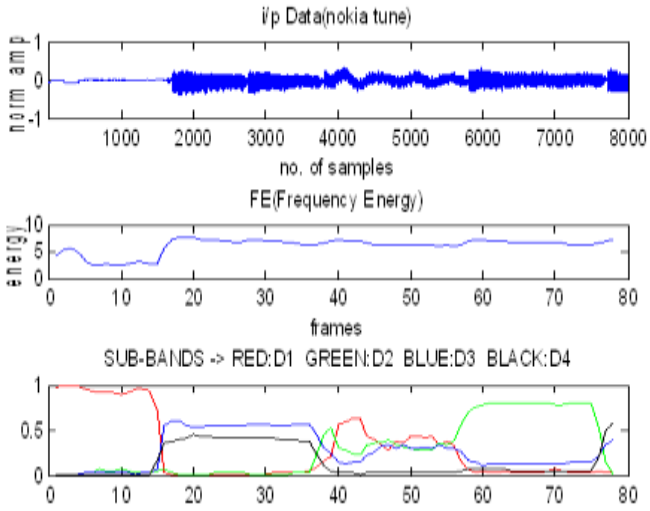

Figure 1.3 Frequency energy and sub-bands of musical instrument.

Figure 1.2 and 1.3 displays the frequency energy and sub-bands of speech and music.

B. *Result using spectrogram*

Since 1950s, many theories have promoted the development of speech recognition, such as Linear Predictive Analysis, Dynamic Time Warping, Vector Quantization, Hidden Morkov Model, and so on. Plenty of Automatic Speech Recognition (ASR) solution is applied from lab to life. The foundation of ASR is to choose speech features. Some usual features such as LPC, LPCC, MFCC and others are all based on time-domain analysis or frequency-domain analysis alone. Their respective limitations lie in: time-domain analysis doesn't reflect spectral characteristics; on the contrary, frequency-domain analysis doesn't make out the time variation. The time-frequency-domain analysis is a method combining the advantage of both parties, which shows the relationship of time, frequency, and amplitude directly. Based on this idea, people pay attention to express speech signal with spectrogram, and apply spectrogram to speech recognition [8].

In 1970s, Victor W.Zue and Ronald A.Cole pursued speech recognition based on spectrogram by spectrogram reading [9]. After 1980s, the research on spectrogram focused on how to extract feature form spectrogram. Mathew J.Palakal and Michael J.Zoran tried to pick up constant characteristics for speaker recognition using Artificial Neural Network [10]. Hideki Kawahara decomposed speech signal to the convolution of spectral parameters, which is used to form special spectrogram, and a series of pulses like VOCODER, and used the spectrogram for speech synthesis [11]. There were many applications in practice of these theories, such as the application of voiceprint recognition in financial security and Judicial verifying [12].

A series of experiments by Zue and his colleagues demonstrated that the underlying phonetic representation of an unknown utterance can be recovered almost entirely from a visual examination of the speech spectrogram [13].

The most common format is a graph with two geometric dimensions: the horizontal axis represents time; as we move right along the x-axis we shift forward in time, traversing one spectrum after another, the vertical axis is frequency and the colors represent the most important acoustic peaks for a given time frame, with red representing the highest energies, then in decreasing order of importance, orange, yellow, green, cyan, blue, and magenta, with gray areas having even less energy and white areas below a threshold decibel level.

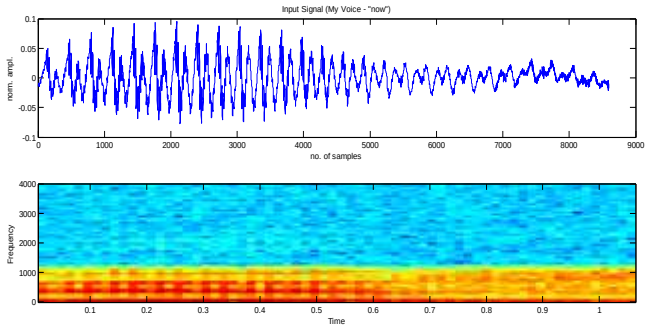

Figure 1.4 Spectrogram of human voice.

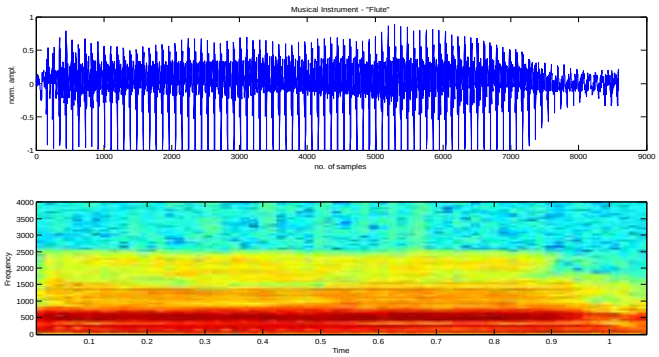

Figure 1.5 Spectrogram of musical instrument.

Figure 1.4 and 1.5 displays the spectrogram of speech and music. It shows that energy in decibel for music is much higher than speech.

C. Result using autocorrelation

Autocorrelation is the cross-correlation with itself. Informally, it is the similarity between observations as a function of the time separation between them. It is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal which has been buried under noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

In statistics, the autocorrelation of a random process describes the correlation between values of the process at different points in time, as a function of the two times or of the time difference. Let X be some repeatable process, and i be some point in time after the start of that process. (i may be an integer for a discrete-time process or a real number for a continuous-time process.) Then X_i is the value (or realization) produced by a given run of the process at time i. Suppose that the process is further known to have defined values for mean μ_i and variance σ_i^2 for all times i. Then the definition of the autocorrelation between times s and t is

$$R(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s}$$

Where "E" is the expected value operator. Note that this expression is not well-defined for all time series or processes, because the variance may be zero (for a constant process) or infinite. If the function R is well-defined, its value must lie in the range $[-1, 1]$, with 1 indicating perfect correlation and -1 indicating perfect anti-correlation.

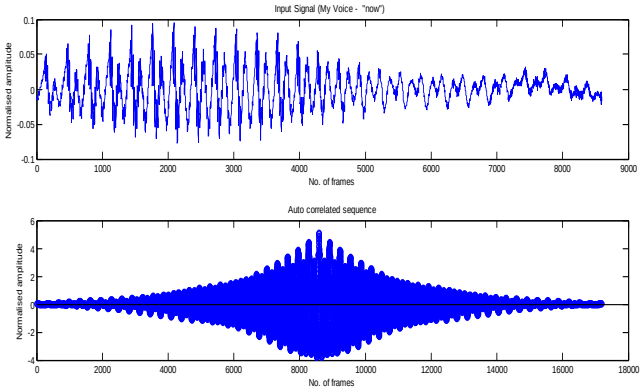

Figure 1.6 Autocorrelation of human voice.

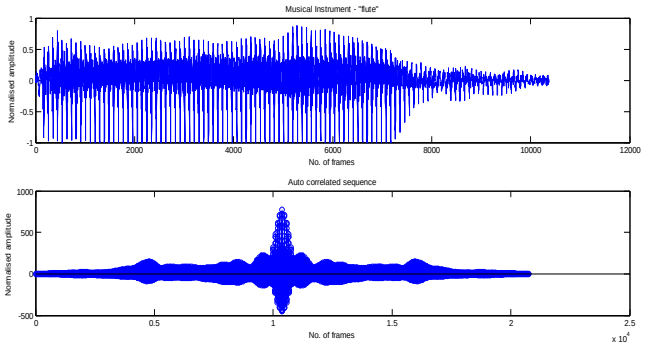

Figure 1.7 Autocorrelation of musical instrument.

Figure 1.6 and 1.7 displays the autocorrelation of speech and music. It shows that autocorrelation for music is much higher than speech.

VI. DISCUSSION AND CONCLUSION

In this paper, we examined the role of voice source measure in speech and musical instrument discrimination Voice source measures were extracted from a large database.

We used three different parameters in the analysis. From the experiments, we could observe evident results for frequency energy, spectrogram and autocorrelation. It is concluded that the frequency characteristics are very different between human voice and music apparatus, the distribution of FE of music is relatively even on each sub-band, but maximum FE is concentrated on first sub-band for speech.

Spectrogram shows that energy in decibel for music is around 2500 which is much higher than speech whose energy in decibel is around 1000. Autocorrelation for music is much higher than speech.

REFERENCES

1. J. Foote. An overview of audio information retrieval. Multimedia Systems, pages 2-10, 1999.
2. E. Scheier and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, 1997.
3. G. Tzanetakis and P. Cook. A framework for audio analysis based on classification and temporal segmentation. In Proc.25th Euromicro Conference. Workshop on Music Technology and Audio Processing, 1999.
4. E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. IEEE Multimedia Magazine, pages 27-36, 1996.
5. J. Foote. An overview of audio information retrieval. Multimedia Systems, pages 2-10, 1999.
6. P. Moreno and R. Rifkin. Using the fisher kernel method for web audio classification. In Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pages 1921{1924, 2000.
7. M. Seck, F. Bimbot, D. Zugah, and B. Delyon. Two-class signal segmentation for speech/music detection in audio tracks. In Proc. Eurospeech, pages 2801-2804, Sept. 1999.
8. Ruan boyao. The application of PCNN on speaker recognition based on spectrogram. Master Degree Dissertations of Wuyi University. 2008.
9. An expert spectrogram reader: A knowledge-based approach to speech recognition Zue, V.; Lamel, L.; Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86. Volume: 11 Publication Year: 1986 , Page(s): 1197 - 1200
10. Mathew J.Paiakal and Michael J.Zoran. Feature Extraction form Speech Spectrogram Using Multi-Layered Network Models. Tools for Artificial Intelligence, 1989. Architectures, Languages and Algorithms, IEEE International Workshop on Volume, Issue, 23-25 Oct 1989. Pages: 224 – 230.
11. Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigne. Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication. Volume 27, Issue 3-4. Apr 1999. Pages: 187 – 207.
12. Yang yang. Voiceprint Recognition Technology and Its Application in Forensic Expertise. Master Degree Dissertations of Xiamen University. 2007.
13. V.W. Zue and R.A. Cole, "Experiments on Spectrogram Reading, "*IEEE Conference Proceedingo, ICASSP*, Washington D.C.,1979, pp. 116-119

AUTHOR PROFILE

Sumit Kumar Banchhor received the B.E. (hons.) degree in Electronics and Telecommunication (2007) and M-Tech. (hons.) in Digital Electronics (2010-2011) from the University of CSVT, Bhilai, India. From 2009, he is currently Asst. Prof. in the department of ET&T, GD Rungta College of Engineering and Technology, university of CSVT, Bhilai. His current research includes speech and image processing.

Om Prakash Sahu received the B.E. degree in Electronics and Telecommunication (2005) and M-Tech. in Instrumentation and control system (2008) from the University of CSVT, Bhilai, India. He has 3 international and 1 international publications. From 2006 he is Lecturer and currently Reader & HOD in the department of ET&T, GD Rungta College of Engineering and Technology, university of CSVT, Bhilai. His research work includes biomedical instrumentation and Robotics.

Prabahkar received the B.E. degree in Electronics and Telecommunication (2005) and M-Tech. in Digital Electronics (2010-2011) from the University of CSVT, Bhilai, India. From 2009, he is currently Asst. Prof. in the department of ET&T, GD Rungta College of Engineering and Technology, university of CSVT, Bhilai. His current research includes speech and image processing.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication