Discrimination between Speech and Music signal

Sumit Kumar Banchhor

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. In this work we deal with the characterization of an audio signal, which is a part of a larger audio-visual system. 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.

The basic characteristics are computed in 2sec intervals. The result shows that the estimation of short time energy reflects more effectively the difference in human voice and musical instrument than zero crossing rate and spectrum flux.

Index Terms: Speech/music classification, audio segmentation, zero crossing rate, short time energy, and 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.

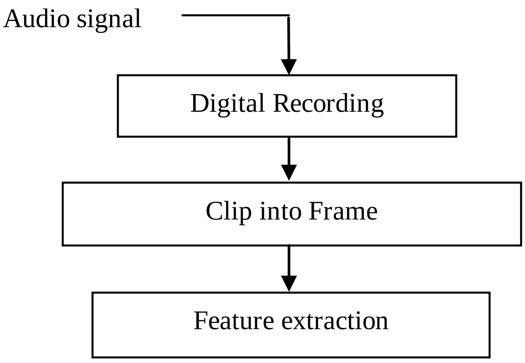

Fig.1.1. Basic processing flow of audio content analysis.

Fig. 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.

Manuscript received on July 09, 2012

Sumit Kumar Banchhor, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University, GD Rungta College of Engineering and Technology, Bhilai, India

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. 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. Each recording was of the form “Now this time”. This utterance was spoken at the habitual speaking level and most talkers repeated the phrases 10 times. For the analysis, only the manually segmented target vowel ‘now’ is used. The musical sound is recorded from flute.

IV. METHODOLOGY

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

V. EXPERIMENT AND RESULT

A. Result using zero crossing rate

It indicates the frequency of signal amplitude sign change. To some extent, it indicates the average signal frequency as:

$$ZCR = \frac{\sum_{n=1}^{N} \left|\operatorname{sgn} x(n) - \operatorname{sgn} x(n-1)\right|}{2N}$$

Where *sgn[]* is a signum function and *x(m)* is the discrete audio signal.

In mathematical terms, a "zero-crossing" is a point where the sign of a function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. The zero-crossing is important for systems which send digital data over AC circuits, such as modems, X10 home automation control systems, and Digital Command Control type systems for Lionel and other AC model trains. Counting zero-crossings is also a method used in speech processing to estimate the fundamental frequency of speech.

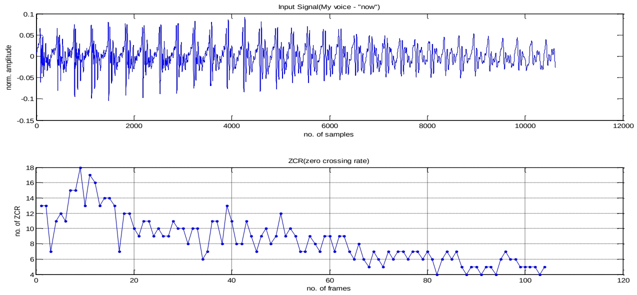

Fig.1.2. ZCR of human voice.

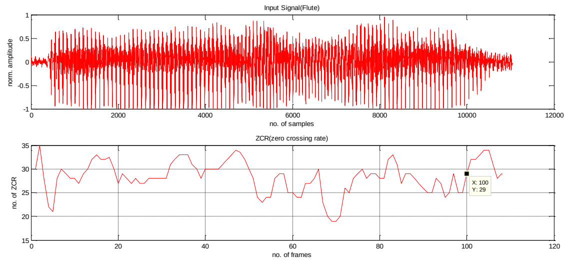

Fig.1.3. ZCR of musical instrument.

Table1.1. ZCR of human voice and musical instrument.

Frames Parameter	0	10	20	30	40	50	60	70	80	90	100
Human voice	0	13	10	10	11	12	9	6	7	5	5
Music	30	28	27	28	30	30	25	20	28	26	29

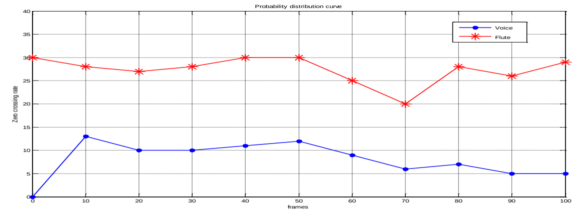

Fig.1.4. Probability distribution curves of human voice and musical instrument for ZCR.

Figure 1.4 displays the probability distribution curve of zero crossing rates of speech and music. It shows that ZCR for music is higher than speech.

B. Result using short time energy

The short-time energy (STE) measurement of a speech signal can be used to determine voiced vs. unvoiced speech. Short time energy can also be used to detect the transition from unvoiced to voiced speech and vice versa. The energy of voiced speech is much greater than the energy of unvoiced speech.

$$E_n = \left(\sum_{m=-\infty}^{\infty} x^2(m)h(n-m)\right) \qquad \text{--- (1)}$$

Eq. (1) defines the short time energy for a sampled signal where *h(n-m)* is a windowing function. For simplicity a rectangular windowing function is used as defined in eq. (2).

$$H(n) = \begin{matrix} 1 & 0 \le n \le N-1 \\ 0 & otherwise \end{matrix} \qquad \text{--- (2)}$$

N in eq. (2) is the length of the window in samples.

The selection of the window size is a compromise since a high pitched female or child’s voice may have a pitch period as small as 16 samples (at an 8 kHz sampling rate) up to 200 samples for a low pitched male voice. A window size of 160 samples or about 20 msec. is a good compromise.

We record the input signal at fs=8KHz. Now using Hamming window with the following specifications: Window size=256 samples, Window step=100 samples, Window overlap=156 samples and number of frames = (length of input – window size)/(window step), we calculate the STE for each frame using the following formula.

$$E = \sum_{m=0}^{N-1} \left| x(n)^2 / (N) \right|$$

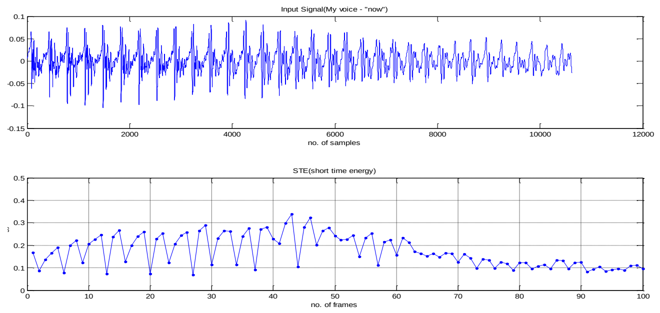

Fig.1.5. STE of human voice.

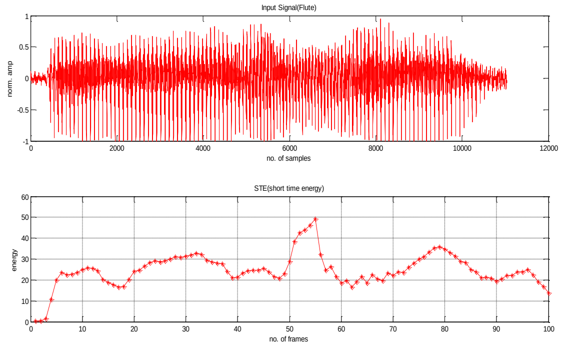

Fig.1.6. STE of musical instrument.

Table1.2. STE of human voice and musical instrument.

Frames / Parameter	0	10	20	30	40	50	60	70	80	90	100
Human voice	0.19	0.2	0.07	0.11	0.23	0.24	0.16	0.12	0.12	0.13	0.1
Music	30	28	27	28	30	30	25	20	28	26	29

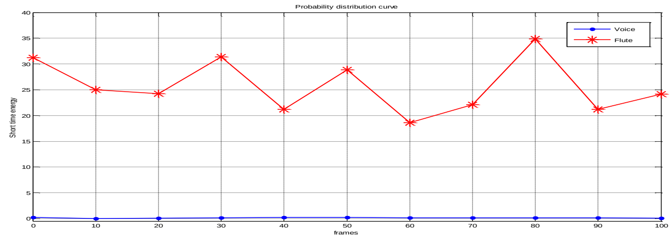

Fig.1.7. Probability distribution curves of human voice and musical instrument for STE.

Figure 1.7 displays the probability distribution curve of short time energy of speech and music. It shows that STE for music is much higher than speech.

C. Result using spectrum flux

Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. More precisely, it is usually calculated as the 2-norm (also known as the Euclidean distance) between the two normalized spectra. Calculated this way, the spectral flux is not dependent upon overall power (since the spectra are normalized), nor on phase considerations (since only the magnitudes are compared). If there is a transient or a sudden attack, the change in energy will be denoted by a jump in the difference of energy between consequent frames. It is important to note that after taking the difference in the spectrums, a positive difference value indicates a rise in energy while a negative difference value indicates a dip in energy. If this method is employed to detect transients, a threshold value should be set only for a positive difference value.

We record the input signal at fs=8 KHz. Now using Hamming window with the following specifications:

Window size=256 samples, Window step=100 samples, Window overlap=156 samples and number of frames = (length of i/p – window size) / (window step), we calculate the STE for each frame using the following formula.

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log A(n,k) - \log A(n-1,k)]^2$$

Where *A(n,k)* is the discrete Fourier transform of the nth frame of input signal.

$$A(n,k) = \sum_{m=\infty}^{\infty} x(m)w(nL-m)e^{j2\Pi km/L}$$

Where L is the window length, k is the order of DFT, and N is the total number of frames.

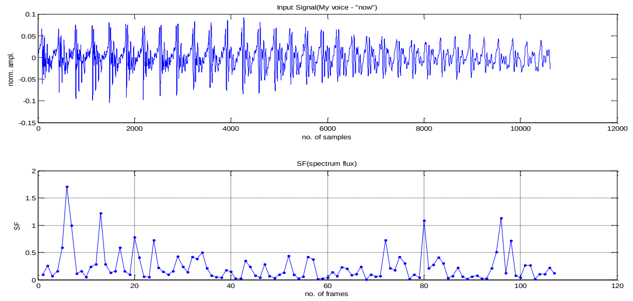

Fig.1.8. SF of human voice.

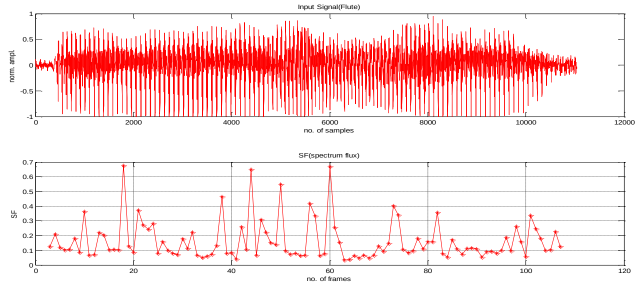

Fig.1.9. STE of musical instrument.

Table1.3. SF of human voice and musical instrument.

Frames / Parameter	0	10	20	30	40	50	60	70	80	90	100
ZCR	30	15	17	18	19	18	16	14	21	21	24
STE	31	25	24..13	31.2	21	28.7	18.4	22	34.7	21.1	24.05
SF	0.2	0.3	0.7	0.07	0.07	0.46	0.63	0.07	0.93	0.04	0.02

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

I. DISCUSSION AND CONCLUSION

Table1.4. SF of human voice and musical instrument.

Frames Parameter	0	10	20	30	40	50	60	70	80	90	100
ZCR	30	15	17	18	19	18	16	14	21	21	24
STE	31	25	24..13	31.2	21	28.7	18.4	22	34.7	21.1	24.05
SF	0.2	0.3	0.7	0.07	0.07	0.46	0.63	0.07	0.93	0.04	0.02

Figure 2.1 displays the probability distribution curve of spectrum flux of speech and music. It shows that SF for music is higher than speech.

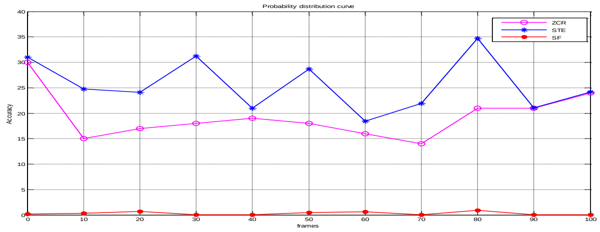

Fig.2.1 Probability distribution curves of human voice and musical instrument for SF.

Fig.2.2 Comparison of probability distribution curves of human voice and musical instrument for ZCR, STE & SF.

The In this paper, we examined the role of voice source measure in speech and musical instrument discrimination and compared the results to perceptual experiments performed on the same database. 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 zero crossing rate, short-time energy, and spectrum flux. Zero crossing rate, short-time energy and spectrum flux of musical instrument are larger than those of males. Short-time energy of musical instrument are much larger than those of human voice whereas zero crossing rate and spectrum flux of musical instrument is little larger than those of human voice. The result shows that the estimation of short time energy reflects more effectively the difference in human voice and musical instrument than zero crossing rate and spectrum flux.

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.

AUTHORS PROFILE

Sumit Kumar Banchhor received the B.E. (hons.) degree in ElectronicsandTelecommunication (2007) and M-Tech. (hons.) in Digital Electronics (2010-2011) from the University of CSVT, Bhilai, India. He has 5 international publications. 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.