Musical Instrument Recognition using Spectrogram and Autocorrelation

Sumit Kumar Banchhor, Arif Khan

Abstract*— Traditionally, musical instrument recognition is mainly based on frequency domain analysis (sinusoidal analysis, cepstral coefficients) and shape analysis to extract a set of various features. Instruments are usually classified using k-NN classifiers, HMM, Kohonen SOM and Neural Networks. Recognition of musical instruments in multi-instrumental, polyphonic music is a difficult challenge which is yet far from being solved. Successful instrument recognition techniques in solos (monophonic or polyphonic recordings of single instruments) can help to deal with this task.

We introduce an instrument recognition process in solo recordings of a set of instruments (flute, guitar and harmonium), which yields a high recognition rate. A large solo database is used in order to encompass the different sound possibilities of each instrument and evaluate the generalization abilities of the classification process. The basic characteristics are computed in 1sec interval and result shows that the estimation of spectrogram and autocorrelation reflects more effectively the difference in musical instruments.

***Index Terms*— Speech/music classification, audio segmentation, spectrogram, autocorrelation.**

I. INTRODUCTION

Recognizing objects in the environment from the sounds they produce is arguably the primary function of the auditory system. An organism that can sense a threat at a distance has a competitive advantage (in the evolutionary sense) over one that cannot. Recognition is possible, in part, because acoustic features of sounds often betray physical properties of their sources. As a simple example, large objects tend to produce sound energy at frequencies lower than those produced by small objects. If an organism's goal is to recognize sounds as arising from particular source classes, recognition should be based on those acoustic features that are invariant across the sounds within each class yet distinguish between the sounds of different classes. For many classes of sound sources, acoustic characteristics that correlate with physical or behavioral properties are examples of such highly discriminatory features. Successful automatic classification of musical sounds is useful in many applications –classification of audio files scattered on the Internet, automatic scoring of recorded music, automatic indexing of recordings, multimedia labeling and many others. Computational auditory scene analysis (CASA), automatic music transcription frameworks and content-based search systems, all find such a capability to be extremely helpful. However, musical instrument recognition has not received as much research interest as, for instance, speech and speaker recognition, even though both the amateur music lover and the professional musician would benefit from such systems. The challenge of automatic classification of musical sounds poses many questions: Accuracy - is it possible to distinguish among virtually identical sounds coming from different instruments, for example certain sounds of Viola and Violin? Taxonomy - what should be the classes? Should sounds recorded in different environments using different instruments and playing techniques, classified in the same class? e.g. when classifying into musical instruments, should recordings of a string ensemble in a noisy environment and a pizzicato sound of a single violin recorded in an anechoic chamber considered the same class? Which instruments should be classified in the same classes when categorizing samples into instrument families? Generality - which are the common qualities of sounds of a specific class (e.g. the sounds of a classical guitar) which separate them from other classes, regardless of the sound database being used and the recording conditions? Validity of data - are the sound databases consistent? Do they contain "bad" or misclassified samples? One of the broad goals of computational auditory scene analysis research is to create computer systems that can learn to recognize the sound sources in a complex auditory environment.

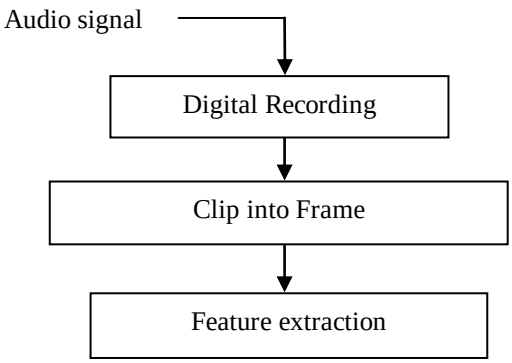

Figure 1.1 Basic processing flow of audio content analysis.

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.

Manuscript Received February 09, 2011.

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).

Arif Khan, Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University, GD Rungta College of Engineering and Technology, Bhilai, India, +91 8103195321, (e-mail: princearif.arif@gmail.com).

II. PREVIOUS WORK

Many attempts in music instrument recognition have taken place in the last thirty years. Most of them have focused on single, isolated notes (either synthesized or natural) and tones taken from professional sound data-bases [1]. Recent works have operated on real-world recordings, polyphonic or monophonic, multi-instrumental or solo [2]. However, the issue is yet far from being solved. The work on recognition from separate notes still remains crucial, since it can lead to further optimization of the methods used and to insights on the recognition of multi instrumental, commercial recordings.

The majority of the recognition systems used so far concentrate on the timbral-spectral characteristics of the notes. Discrimination is based on features such as pitch, spectral centroid, energy ratios, spectral envelopes and mel frequency cepstral coefficients [3, 4]. Temporal features, other than attack, duration and tremolo, are seldom taken into account. Classification is done using k-NN classifiers, HMM, Kohonen SOM and Neural Networks [5, 6]. A limitation of such methods is that in real instruments the spectral features of the sound are never constant. Even when the same note is being played, the spectral components change. One has to take into consideration many timbral components and the way they can vary, which is often rather random, in order to develop a robust recognition system.

III. METHODOLOGY

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

IV. EXPERIMENT AND RESULT

A. 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.

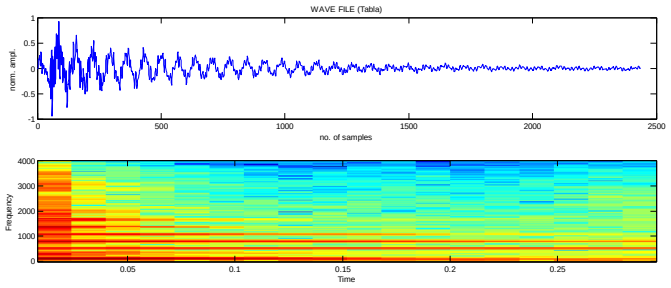

Figure 1.2 Spectrogram of tabla.

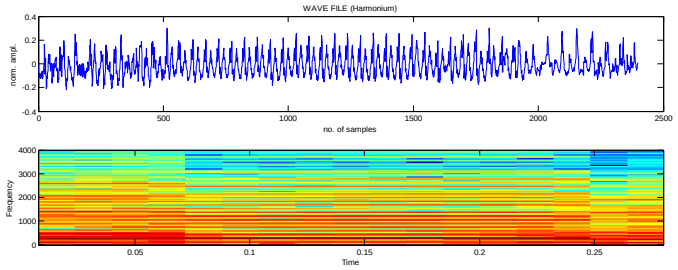

Figure 1.3 Spectrogram of harmonium.

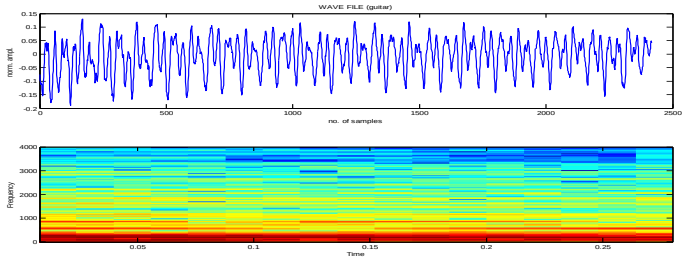

Figure 1.4 Spectrogram of guitar.

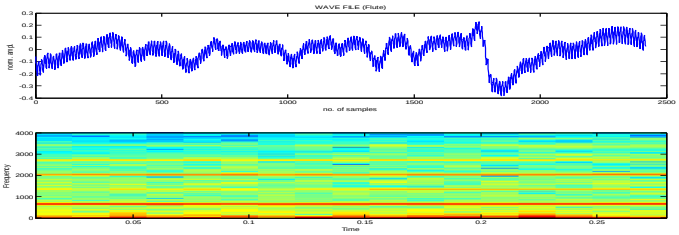

Figure 1.5 Spectrogram of flute.

Figure 1.2, 1.3, 1.4 and 1.5 displays the spectrogram of table, harmonium, guitar, and flute.

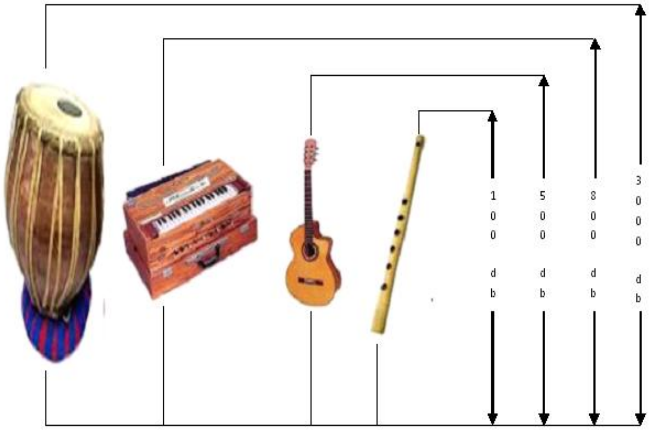

Figure 1.6 Pictorial representation of spectrogram of table, harmonium, guitar, and flute.

B. 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.

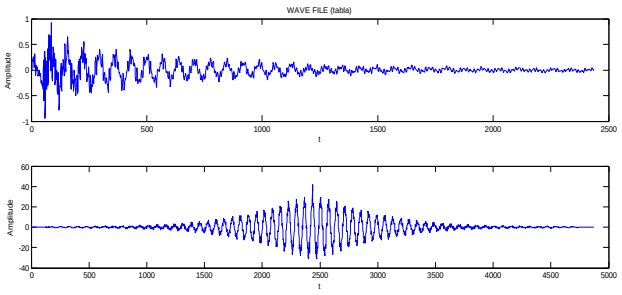

Figure 1.7 Autocorrelation of tabla.

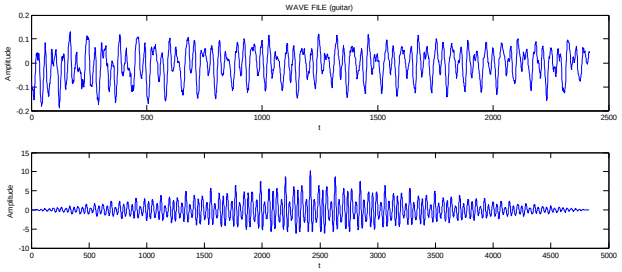

Figure 1.8 Autocorrelation of harmonium.

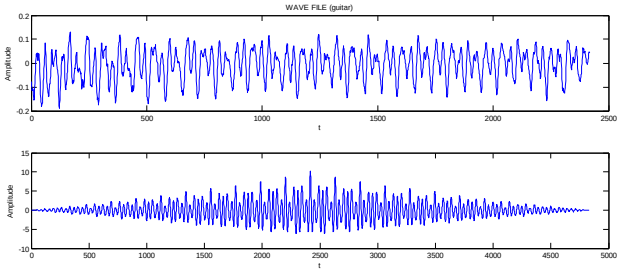

Figure 1.9 Autocorrelation of guitar.

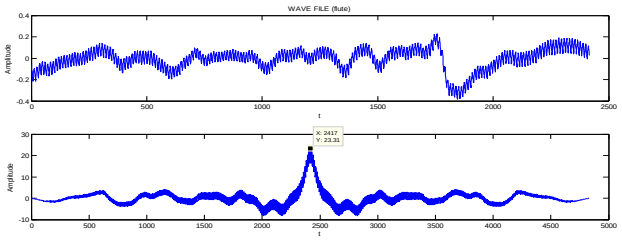

Figure 1.10 Autocorrelation of flute.

Figure 1.7, 1.8, 1.9 and 1.10 displays the autocorrelation of table, harmonium, guitar and flute.

Figure 1.11 Pictorial representation of autocorrelation of table, harmonium, guitar, and flute.

V. DISCUSSION AND CONCLUSION

In this paper, we dealt with recognition of sound samples and presented several methods to improve classification results. Tones are extracted from a large database of four musical instruments (table, harmonium, flute and guitar).

We use two different parameters in the analysis. From the experiments, we could observe evident results for spectrogram and autocorrelation. Maximum and minimum values of amplitude for autocorrelation for all musical instruments have different ranges. Spectrogram of tabla is much larger than those of harmonium, guitar and flute.

Result shows that the estimation of spectrogram and autocorrelation reflects more effectively the difference in musical instrument.

REFERENCES

1. K.D. Martin: Sound-Source Recognition: A Theory and Computational Model, Ph.D. thesis, MIT, 1999
2. A. Livshin, X. Rodet: Musical Instrument Identification in Continuous Recordings, Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX-04), Naples, Italy, October 5-8, 2004
3. A. Eronen, A. Klapuri: Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2000, pp. 753-756
4. T. Kitahara, M. Goto, H. Okuno: Musical Instrument Identification Based on F0-Dependent Multivariate Normal Distribution, Proc. of the 2003 IEEE Int'l Conf. on Acoustic, Speech and Signal Processing (ICASSP '03), Vol.V, pp.421-424, Apr. 2003
5. A. Eronen: Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs, Proc. of the Seventh International Symposium on Signal Processing and its Applications, ISSPA 2003, Paris, France, 1-4 July 2003, pp. 133-136
6. G. De Poli, P. Prandoni: Sonological Models for Timbre Characterization, Journal of New Music Research, Vol 26 (1997), pp. 170-197, 1997

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.

Arif Khan received the B.E. degree in Electronics and Telecommunication (2008) and is persuing M-Tech. in Digital Electronic from the University of CSVT, Bhilai, India. From 2010, 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.