Novel Audio Feature Set for Monophonic Musical Instrument Classification

Shubham Bahre
Electronics and Telecommunication Engineering
College of Engineering, Pune
Pune, India
bahres15.extc@coep.ac.in

Dr. Shrinivas P. Mahajan
Electronics and Telecommunication Engineering
College of Engineering, Pune
Pune, India
spm.extc@coep.ac.in

Rohan T. Pillai
Electronics and Telecommunication Engineering
College of Engineering, Pune
Pune, India
rohant1991@gmail.com

Abstract— **This paper proposes a novel set of parameters for musical instrument classification of three different instrument classes of the instrument from audio recordings of monophonic musical sounds notes. The proposed method extracted three features: attack slope, constant Q transform and cepstral coefficients. The algorithm consisted of feature extraction and classifier learning steps. Thus, this system depends upon low-level features. A maximum accuracy of 87% was obtained when tested on the database obtained from University of IOWA Electronic Music Studios. Confusion matrix and subsequent significant classification metrics have been computed. This classification scheme finds application in audio indexing, content based retrieval, genre identification of instrumental music.**

Keywords—Music information retrieval, instrument classification, feature extraction, musical instrument identifier, constant Q transform, cepstral coefficients, MIRtoolbox

I. INTRODUCTION

Music information retrieval has been a dominating field of audio processing. Musical data is being produced in volumes every day because of easy availability of handheld devices and music producing application. Hence, classification of data for easy retrieval and quick access is needed. Each musical instrument sounds different to our ears even if same note is played. Relevant musical parameter or features are needed to quantify the distinct characteristic of each instrument. A combination of features is required when boundaries of any one of these parameters overlap.

In scientific pitch notation, whole audio frequency range can be divided into octaves. An octave is a range of frequencies between two frequencies in which one frequency is either half or double of another frequency. There are seven basic notes in an octave: C, D, E, F, G, A and B. Other than E and B, all five notes have two pitch positions. This makes the total count to 12. So by using a concept called the enharmonic equivalence [1], notes in equal tempered can be listed as in Table I. In Table I, the octave is following semitone spacing, i.e., each note is geometrically spaced from its previous or next note by twelfth root of 2.

The manuscript is arranged in the following way. First, the previous work their approach and shortcomings in instrument identification have been discussed Section II. The features used in this work are attack slope, constant Q transform and cepstral coefficients. They have been explained in detail with their specific utility to instrument characteristic in Section III. Section IV explains results and observation. Section V discusses conclusion and states the future scope of improvement.

TABLE I. OCTAVE INDEPENDENT PITCH CLASSES

Note Position	1	2	3	4	5	6	7	8	9	10	11	12
Notation	C	C#	D	D#	E	F	F#	G	G#	A	A#	B

II. RELATED WORK

In [2], the authors have proposed a hierarchical method for automatic instrument identification. They have chosen lower level MFCC like features. In step1, attack, decay, sustain and release have been calculated on wavelet decomposed audio signal. Step 2 calculates Mel Frequency Cepstrum Coefficients. RANSAC, used for robust estimation and SVM, as an example of the supervised algorithm have been used. They are compared and are giving an equivalently same performance in classification. Maximum Accuracy of 92.67% using RANSAC (with default parameter setting) and 90.69% using SVM was achieved in stage 1 and stage 2 respectively.

In [3], the authors have classified monophonic musical instrument sounds using six feature set vector that incorporated k-nearest neighbour classifier (kNNC). The features chosen in classification were constant Q transform, spectral centroid, vibrato, RMS amplitude envelope, MSA trajectories and cepstral coefficients. As a further

ISBN 978-1-5090-4760-4/17/$31.00©2017 IEEE

improvement to their work, they have used wavelet transform in Kaminskyj [4]. They showed that the accuracy has shot up if two more features wavelet features were added to the feature set.

These results show a great progress but downsides still remain. (1) Since most of the instruments sounds can be produced from synthesizer, some features specific to synthesizer must be considered. (2) Features that preserve instrument characteristics must be considered. (3) A minimum number of relevant features need to be combined to attain maximum accuracy must be chosen so as to reduce the computations.

III. Feature Extraction

Fig 1. and fig 2 show the audio waveform of C4 note of piano and violin respectively. Instrument notes have envelopes that have four basic characteristics: Attack, Decay, Sustain and Release (ADSR). It describes how the sound gets louder and softer over time. Different instruments have different values for these 4 parameters. Synthesizers model the instruments sound using the properties of the envelope. However, practical acoustic sounds do not strictly follow it. Release part is omitted if the sampler sounds are truncated [5].

As can be seen from fig 3 and comparing it with fig 1 and fig 2, the presence of ADSR is evident and their values are highly dependent on the type of instrument. A string instrument like a violin has a much lower attack and much higher sustain as compared to

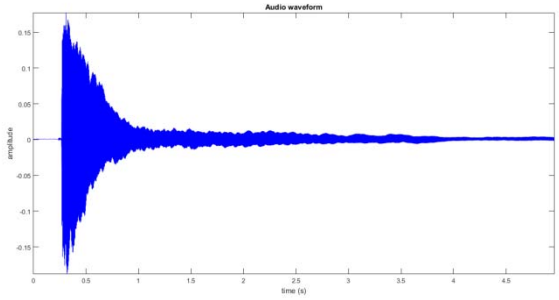

Fig 1. Waveform of Piano C4 Note

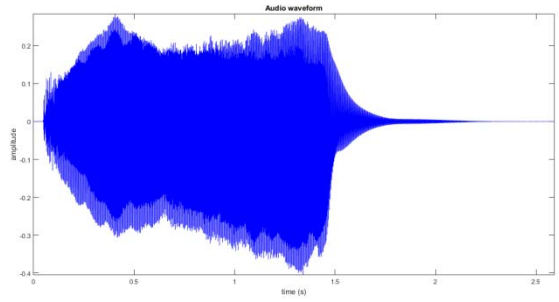

Fig 2. Waveform of Violin C4 note

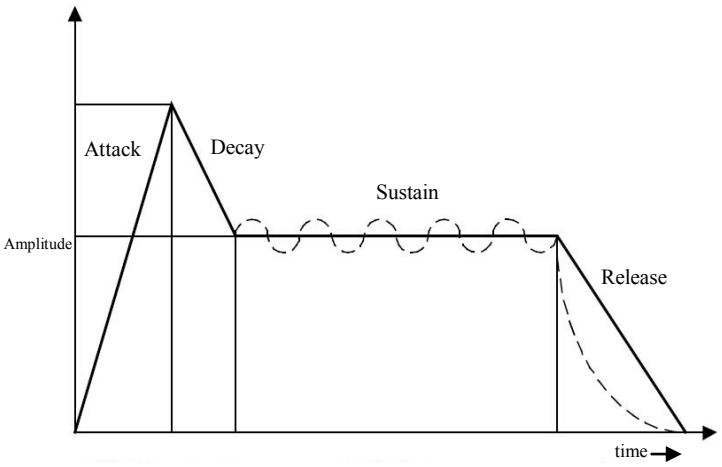

Fig 3. Audio envelope and its characteristics

The cepstrum can be defined as the ("inverse") Fourier transform of the logarithm of the absolute value of the spectrum as shown in fig 4. So the horizontal axis is neither frequency nor is it equivalent to time. Hence, the term quefrency has been suggested by Bogert [6].

Fig 5 shows the cepstrum of C4 note of Violin.

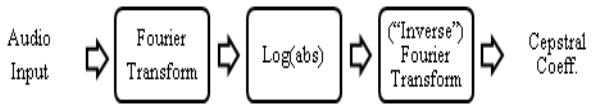

Fig 4. Cepstral coefficient calculation

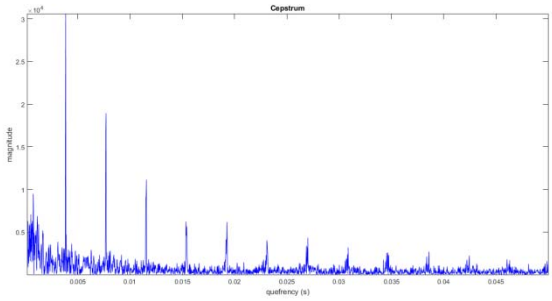

Fig 5. Cepstrum of C4 note of Violin

In scientific pitch representation [7], the standard semitone spaced equal tempered scale obeys

$$f_n = f_0 * a^n \quad (1)$$

where f_0 is the frequency of one fixed note which must be defined. Most commonly f_0 is taken 440Hz,

'n' is the number of steps you are away from f_0. For a higher note than f_0, n is positive. For a note lower than f_0, n is negative, and

f_n is the frequency of the note n half steps away. It is equal to the twelfth root of 2. The number 12 corresponds to semitone spacing.

Conventional Fourier Transform suffers from two inherent problems: (1) It does not give proper resolution for the frequency of different intervals. Musical notes are arranged in octaves. These octaves are closer in lower frequencies and sparser at higher frequencies. So higher resolution is needed at lower frequencies and lower resolution is needed at higher frequencies. For example, to resolve C2 note whose corresponding frequency is 65.41Hz, a window size of 30.8 ms or more is needed. C1 note corresponds to 32.7 Hz, a 62 ms or more is required. So if we consider a bigger frame size it gives much more resolution to higher frequencies than needed. (2) Implementing variable window size would be tedious in Fourier analysis. Hence Constant Q transform has been used. CQT offers variable window size as well as it appropriate window size for each frequency.

The difference between any two successive frequencies is given by:

$$f_{n-1} - f_n = (2^{1/B} - 1)f_n \quad (2)$$

where B is the equal tempered spacing between the frequencies. B has a value of 12 for semitone spacing and 24 for quartertone spacing. f_n is the frequency whose CQT has to be calculated. From eq. (2),

$$\frac{f_n}{f_{n-1}-f_n} = \frac{1}{2^{1/B}-1} \quad (3)$$

$$Q = \frac{f_n}{f_{n-1}-f_n} \quad (4)$$

Hence, Q-factor is defined as the ratio of a frequency being analyzed to its difference from the adjacent frequency.

CQT is based upon the Q factor value calculated in (4) and the CQT equation [3] is given by:

$$X(k) = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x(n)e^{-j\frac{2\pi Q}{N[k]}n} \quad (5)$$

W [k,n] is a window, (preferably Hanning) and N[k] represents the variable window length given by:

$$N[k] = \frac{Sampling\ Rate}{f_n} Q \quad (6)$$

A. Toolbox

Features have been calculated in MATLAB R2017 using MIRtoolbox[8]. MIRtoolbox is a MATLAB toolbox for audio feature extraction of audio that assists higher feature extraction task in musical analysis.

B. Classification

Classification is an important step after extraction of features. A number of unsupervised and supervised techniques are available. Multiclass SVM [9] has been used in this paper. Models a given training set with a corresponding group vector and classifies a given test set using an SVM classifier according to a one vs. all relation. SVM can nowadays be considered a standard tool in music genre classification [1].

IV. Results and Observations

The results obtained using the parameters described in the above section is discussed herein. The observations made from the results, the database used and the number of coefficients chosen has also been presented in the section. Based on the combination of the parameters used, i.e., one Attack slope coefficient, 11 Cepstrum coefficients and 249 CQT coefficients, the input audio signal obtained from a certain musical instrument poses a 261-dimensional problem. As the total number of instruments to be classified is 5, a 5-class SVM is employed for classification. The sample recordings used for testing the algorithm were obtained from one of the largest openly available musical instrument database, namely the University of IOWA Electronic Music Studios. The recordings consist of single notes played by a particular instrument ranging the octaves 3-6 and the number of such recordings obtained for the Flute, Guitar, Piano, Trumpet and Violin were 48, 32, 60, 47 and 124 respectively. The confusion matrix in Table II presents the classification results of the 5 instruments. The Classification Accuracy and other related classification metrics have been calculated based on the results of the confusion matrix.

TABLE II. Classification results for five different instruments

Predicted / Actual	Flute	Guitar	Piano	Trumpet	Violin
Flute	47				1
Guitar		24		1	7
Piano	6	5	46		3
Trumpet				46	1
Violin	13	2	2		107

From Table II , the true positives (when an input sample belongs to an instrument and it is correctly predicted), false positives (when an input sample doesn't belong to an instrument but is predicted as belonging to that instrument), true negatives (when an input sample doesn't belong to an instrument and is correctly predicted as not belonging to that

instrument) and false negatives (when an input sample belongs to an instrument but is predicted as not belonging to that instrument) for all instruments were calculated based on [10]. The instrument classification metrics such as Total Classification Accuracy, Sensitivity, False Positive Rate (FPR) and Precision have been calculated and is presented in Table III. The terms are defined as follows:

a. Total Classification Accuracy:

It is defined as the ratio of the sum of true positives and true negatives to total number of input samples.

b. Sensitivity of an instrument:

For an instrument, it is defined as the ratio of the true positives to the total input samples of that instrument.

c. False Positive Rate:

For an instrument, it is the number of false positives to the total number of input samples that do not belong to that instrument.

d. Precision:

For an instrument, it is the number of true positives to the total number of samples predicted as belonging to the instrument.

TABLE III. INSTRUMENT-WISE ACCURACY OF THE FIVE INSTRUMENTS

Parameter / Instrument	Sensitivity	False Positive Rate	Precision
Flute	0.9791	0.0722	0.7121
Guitar	0.750	0.025	0.7742
Piano	0.767	0.0079	0.9583
Trumpet	0.9787	0.0038	0.9787
Violin	0.8629	0.0642	0.8992

The total classification accuracy was calculated to be 86.81%. The results obtained above show that the Trumpet classification probability is highest because it has the highest Sensitivity and Precision, while its FPR is lowest, which is desirable. The average Sensitivity, FPR and Precision of the algorithm is 0.8675, 0.0347 and 0.8645 respectively.

V. CONCLUSION

The proposed novel algorithm for musical instrument classification based on novel acoustic features has been presented. The total classification accuracy figure of 86.81% is very encouraging and has been used to reliably classify five musical instruments. The distinct attributes of the algorithm are: 1) use of low level features for classification, 2) minimal number of features used, 3) trivial adaptation of the algorithm to synthesized instrument sounds, due to the use of attack slope feature, which is an absolute must for synthesizers, 4) the use of SVM as a classifier gives the algorithm a better classification accuracy and 5) use of CQT, a transform specifically suited for musical analysis helps improve the detection probability. Future work includes reduction in dimensionality, increase in the number of identified instruments by increasing more low-level features, identification of polyphonic sound and real-time implementation.

REFERENCES

[1] Alexander Lerch, "Tonal Analysis," in *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, 1, Wiley-IEEE Press, 2012, pp.79-117 doi: 10.1002/9781118393550.ch5

[2] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha, "Automatic Identification of Instrument Type in Music Signal Using Wavelet and MFCC," in *Computer Networks and Intelligent Computing: 5th International Conference on Information Processing, ICIP 2011, Bangalore, India, August 5-7, 2011. Proceedings*, K. R. Venugopal and L. M. Patnaik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 560–565.

[3] I. Kaminskyj and T. Czasejko, "Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC" , Journal of Intelligent Information Systems, vol. 24, pp. 199-221, March 2005.

[4] C. Pruysers, J. Schnapp and I. Kaminskyj, "Wavelet analysis in musical instrument sound classification," *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005.*, 2005, pp. 1-4.

[5] G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project," IRCAM, Project Report (CUIDADO), 2004

[6] B. P. Bogert, M. J. R. Healy, and J. W. Tukey: "The Quefrency Alanysis [sic] of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking". Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, 209-243. New York: Wiley, 1963

[7] B.H. Suits, 1998, May, "Frequencies of equal-tempered scale, A4 = 440 Hz", Available: http://www.phy.mtu.edu/~suits/notefreqs.html

[8] O. Lartillot, P. Toiviainen, T. Eerola, " A MATLAB toolbox for musical feature extraction from audio" , International Conference on Digital Audio Effects, Bordeaux, 2007.

[9] https://in.mathworks.com/matlabcentral/fileexchange/39352-multi-class-svm last accessed 30th may, 2017.

[10] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27, no. 8 (2006): 861-874.