Musical Instrument Extraction through Timbre Classification

Sang Hyun Park
NVIDIA Corporation
Santa Clara, CA 95050
Email: andyp@nvidia.com

***Abstract*—Contemporary technological advancement of internet and online servers allows many musical pieces to be readily available to the users to enjoy. The users may listen to the music, share with friends, or create another musical piece by either remixing or sampling. One may desire to simply play the music as it is or sample just one instrument out of the music, however, this task can be challenging due to the complexity of the modern musical pieces. A musical piece may contain multiple musical instruments and this require the user to distinguish the instrument from the others in order to play the correct part of the music.**

In this paper, a machine learning approach is presented to extract a musical instrument from a complex music using timbre classification.

***Index Terms*—Timbre recognition, timbre classification, machine learning, instrument recognition, instrument extraction.**

I. Introduction

Human ears possess an ability to distinguish musical colors. We can easily distinguish the sound of piano from the sound of guitar because they have different feeling or color in their sound. When required training is processed, human ears can extract a certain instrument using the knowledge of timbre of the sound of the specific instrument. Various audio features are used in recent researches in order to achieve automatic timbre recognition system using machine learning algorithms. Recognizing audio features can be a challenging problem since it is a continuous task unlike descrete data, image or text recognition. Acquisition and recognition need to be in sync of time frame in order to achieve correct predictions.

In order to achieve a correct set up for the recognition, we need to explore more about the musical charactersitics, timbre. Then, we represents the audio features that we use in this project.

II. Timbre

Timbre can be a set of subjective opinions of individuals toward a sound that is independent from the frequency (pitch) or the amplitude (loudness). Timbre is also known as color or tone of sounds. Unlike frequency to pitch or amplitude to loudness, there is no dominant attributes to timbre and this limitation makes the definition of timbre to vary and subjective.

The American Standards Association definition 12.9 of timbre describes it as, "...that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar," and a note to this definition adds that, "Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus." (American Standards Association 1960, 45).

Due to many attributes of sound are required to recognize timbre, it is necessary to acquire features that are objective rather than subjective. J. F. Schouten suggested the following five acoustic parameters to be considered related to timbre.

1) The range between tonal and noiselike character
2) The spectral envelope
3) The time envelope in terms of rise, duration, and decay (ADSR-attack, decay, sustain, release)
4) The changes both of spectral envelope (formant-glide) and fundamental frequency (micro-intonation)
5) The prefix, or onset of a sound, quite dissimilar to the ensuing lasting vibration

Using the five parameters presented above R. Erickson presented a table of subjective experience to objective characteristics as shown on TABLE 1.

Although Timbre is decided with multiple aspects of acoustic featurs, spectrum of audio seems to be most affected feature.

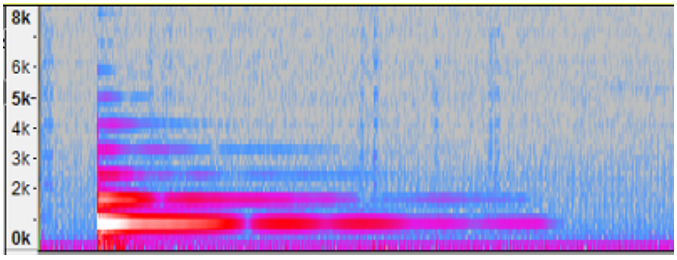

Fig. 1. Octave 5G on Piano

Figure 1 shows the spectrum of a sound of Piano at the frequency of Octave 5G. The color shows the amplitude of the the corresponding frequency. White is the strongest, red, blue, gray are in order.

We can see how the different timbre shows different spectrum when we compare Piano, Guitar, and Vibe. Features related to spectrum should be sufficient to recognize different timbre to a certain extent.

TABLE 1

Subjective Experience to Objective Characteristics of Timbre

Subjective Experience	**Objective Characteristics**
Tonal character, usually pitched	Periodic sound
Noisy, with or without some tonal character, including rustle noise	Noise, including random pulses characterized by the rustle time (the mean interval between pulses)
Coloration	Spectral envelope
Beginning/ending	Physical rise and decay time
Coloration glide or formant glide	Change of spectral envelope
Microintonation	Small change (one up and down) in frequency
Vibrato	Frequency modulation
Tremolo	Amplitude modulation
Attack	Prefix
Final sound	Suffix

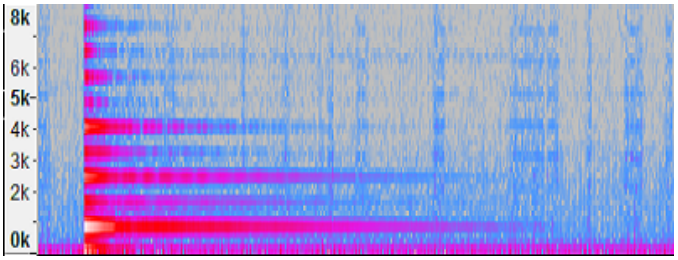

Fig. 2. Octave 5G on Guitar

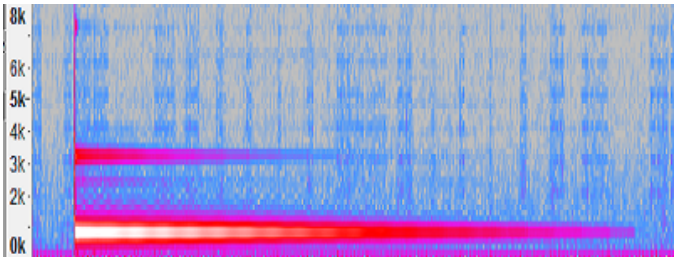

Fig. 3. Octave 5G on Vibe

III. Features Extraction

Features are collected using YAAFE (Yet Another Audio Feature Extractor). The software produces many acoustic features in easy and efficient ways. Although there are 27 features that we can get with the current version of YAAFE, we are only using few of the features in order to simplify the beginning research process. We choose some of the features that are related to spectrum. Features that are acquired and used for this project are described in the following subsections.

A. MFCC

Mel-Frequency Ceptrum Coefficients feature is used broadly in acoustic, sound, and speech related research areas due to it's compactibility to represent MFC which becomes the short span of spectrum of an audio frame. In this research, we used 13 ceptral coefficients to represents MFC, Hanning weighting window to apply before FFT, 40 Mel Filter Banks of 130 6854 Hz, 1KB block size and 512B step size.

B. Spectral Shape Statistics

This statistics includes the following four separate attributes: centroid, spread, skewness, and kurtosis.

$$
\begin{aligned}
\mu_i &= \frac{\sum_{n=1}^{N} f_k^i * a_k}{\sum_{n=1}^{N} a_k} \\
centroid &= \mu_1 \\
spread &= \sqrt{\mu_2 - \mu_1^2} \\
skewness &= \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu 3}{S_w^3} \\
kurtosis &= \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{S_w^4} - 3
\end{aligned}
$$

This features uses four consecutive frames of audio data to achieve the above characteristics. In this research, we used Hanning weighting window to apply before FFT, 1KB block size and 512B step size.

IV. Data Set

Audio data set for this project is produced by me using the electronic keyboard which can make piano sound and some other instruments such as guitar, vibe, etc. I was able to achieve all scale of notes for each instruments but this method is limited due to the limitation of the number of instruments that I can produce with my keyboard.

V. Machine Learning Algorithm

A. SVM

Since we have the label specifying if the given audio is piano or not, we can use supervised learning on this. We used SVM with linear kernel to simplify the work for now.

B. KNN

Even if we have label, we can disregard the label and make group of similar sounds using KNN. Then, label the group using the majority label for the corresponding group. This may not be allowing wrong placements of dataset but it is similar to how human would process.

VI. Results

Using separately recorded notes as test sets, we were able to achieve the below results for deciding if a note is played by the piano or others.

ML Tool	Accuracy
SVM	73.54
KKN	75.12

VII. Conclusion

We were able to achieve some functionality to distinguish piano sound from the other sounds using the machine learning algorithm. However, the accuracy is not good enough to be used in real life yet. There are two potential optimizations to be made in order to increase the accuracy.

A. Features

The usage of the extracted features were not clear. In order to make sense of the features and correctly use them, knowledge of acoustic and speech studies seems to be required. This project can be rerun with the full-fleged feature list and usage in future.

B. Machine Learning Algorithm

Due to the time and member limitation, we did not extensively use SVM and other algorithms. This is a must for next engagement with this project in order to use more advanced and known good algorithms.

C. Dataset

Since the dataset I acquired from my electronic keyboard may suffice for current limitation in the project but this is still limited since the world has so much more timbre to consider. Next time this project should be done with more various sound sources.

D. Future Work

This project has a small scope where we only find a short frame of sound is from piano or not. The first goal will be properly achieve the mentioned task. Then, we can diverge to three different paths. One is to distinguish more instruments from the given sound. Another is to detect piano in a frame where more instruments than piano are played at the same time. The other is to implement continuous recognition and detection in a long audio stream. When these three can be achieved, we can finally meet the last goal which is to extract a specified instrument from music. This project will be the starting point to achieve the goal.

References

[1] T. H. Park, "Towards Automatic Musical Instrument Timbre Recognition", Ph.D. Thesis, Princeton University, 2004.

[2] X. Zhang, Z. W. Ras, "Analysis of Sound Features for Music Timbre Recognition", *IEEE CS International Conference on Multimedia and Ubiquitous Engineering* (MUE 2007). April 26-28, Seoul, Korea, pp. 3-8.

[3] B. Mathieu, S. Essid, T. Fillon, J.Prado, and G. Richard, "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software", *Proc. of the 11th ISMIR conference*, Utrecht, Netherlands, 2010.

[4] D. K. Fragoulis, J. N. Avaritsiotis, and C. N. Papaodysseus, "Timbre recognition of single notes using an ARTMAP neural network", *Proc. of the 6th IEEE International Conference on Electronics, Circuits and Systems*, Paphos, Cyprus, 1999.

[5] R. Loughran, J. Walker, M. O'Neill, M. O'Farrell, "Musical Instrument Identification Using Principal Component Analysis and Multi-Layered Perceptrons", *IEEE International Conference on Audio Language and Image Processing*, Shanghai, pp. 643-648, 2008.

[6] S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357-366, 1980.

[7] O. Gillet, and G. Richard, "Automatic Transcription of Drum Loops", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.