FEATURE IS ALL YOU NEED

SeungHeonDoh
Graduate School of Culture Technology
KAIST
seungheondoh@kaist.ac.kr

ABSTRACT

The most basic element that composes a music is an instrument. The combination of instruments creates new music and is delivered to the user. Musical instruments are given their own characteristics according to their manufacturing methods and materials. Understanding the unique characteristics of these instruments is expected to make a significant contribution to music classification and generation models.

In this report, The automatic classification of musical instruments, using NSynth dataset, which contain 10 classes of different musical instruments, including bass, brass, flute, guitar, keyboard, mallet, organ, reed, string and vocal. We use three feature group for representing timbre features, pitch/harmony features, rhythm features. Experiments use old Machine-Learning algorithm and 5fold-cross validation. The main points of this report are using small data set(Train 1200, Test 200) and focusing on relationship between 3 feature group and Machine-Learning algorithm. Using the proposed feature group, classification of 92.5% for 10 class of musical instruments.

***K*eywords** Instrument Classification · Feature extraction · Machine Learning

1 Introduction

The most representative paper in the music classification is GTZAN paper. GTZAN used three feature group for representing timbre features, pitch/harmony features, rhythm features. But, they use KNN/GMM classifiers and accuracy is not very high.

In this report, task is changed genre classification to instrument classification. we investigate various machine learning algorithm, including Softmax Regression,Random Forest Classifier, Support Vector Machine, KNeighbors Classifier, Gaussian Navie Bayes. Also, feature-extraction make total 61 features, using 3group of feature's characteristic.

The report is structured as follows. Section 2 provides a review of data-set and task. In Section 3, we described our methods, including the machine-learning algorithm and feature-extraction. Section 4,5 provide experiment and result of experiment. We conclude with a discussion in Section 6

2 Data Set and Visualization

We use a subset of the NSynth dataset which is a large collection of musical instrument tones from the Google Magenta project. The subset has 10 classes of different musical instruments, including bass, brass, flute, guitar, keyboard, mallet, organ, reed, string and vocal. The data-set are all .wav file. The files that stored as 16000Hz, 16-bit, 4sec, mono audio file.

Through the data visualization step we have found clues to feature extraction. In Waveform, we checked the difference of envelop for each instrument. Through the spectrogram, we were able to see the main frequency and temporal change of the main frequency according to the flow of time.

- waveform: visualizes the change in amplitude as the time axis changes.
- spectrogram: Shows the difference of amplitude in color according to time axis and frequency axis change

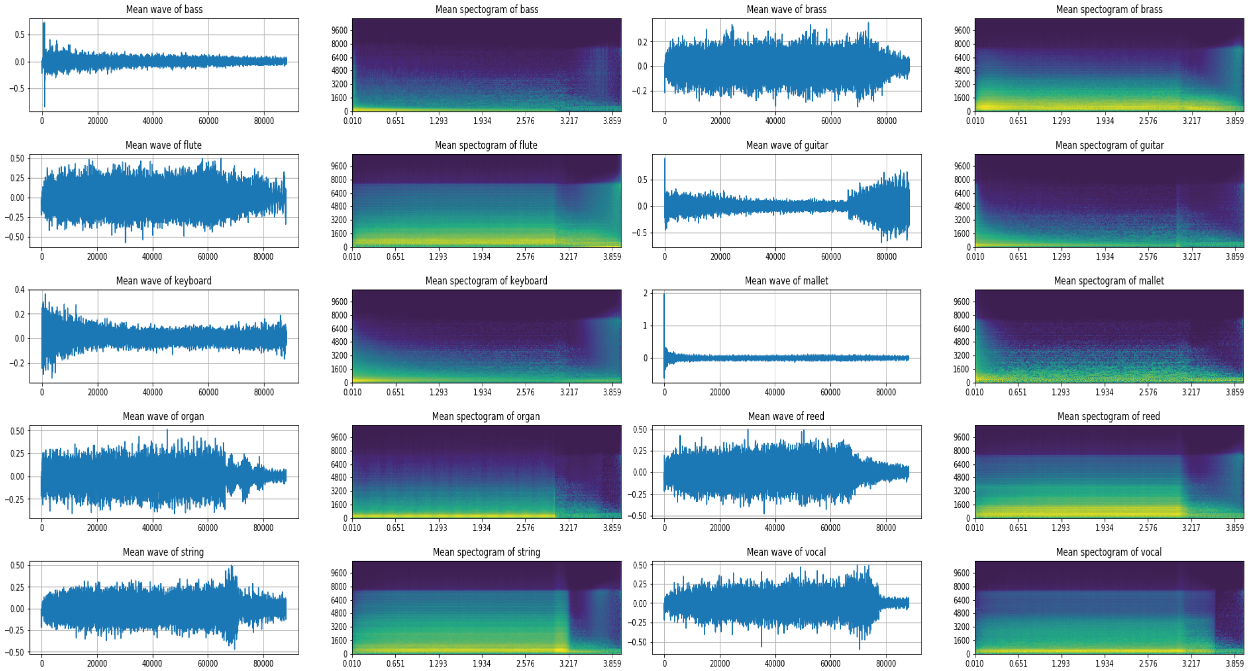

Figure 1: waveform, spectrogram of Instrument

3 Methods

3.1 Feature Extraction

Extraction of features is a very important part in analyzing and finding relations between different things. The data provided of audio cannot be understood by the models directly to convert them into an understandable format feature extraction is used. We use three feature group for representing timbre features, pitch/harmony features, rhythm features.

3.1.1 Timbre features

1) Zero Crossing Rate
The zero crossing rate(ZCR) indicates the number of times that a signal crosses the horizontal axis. ZCR is low for harmonic sounds and high for noisy sounds.

2) Spectral Centroid
The spectral centroid(SC) indicates at which frequency the energy of a spectrum is centered upon. This is like a weighted mean. Where $S(k)$ is the spectral magnitude at frequency bin k , $f(k)$ is the frequency at bin k. High centroid values mean 'brighter' textrue with more high frequency.

$$f_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)} \tag{1}$$

3) Spectral Bandwidth
Bandwidth is the difference between the upper and lower frequencies in a continuous band of frequencies. Where $S(k)$ is the spectral magnitude at frequency bin k,$f(k)$ is the frequency at bin k, and fc is the spectral centroid.

$$SB_p = \left(\sum_k S(k) \left(f(k) - f_c \right)^p \right)^{\frac{1}{p}} \tag{2}$$

4) Spectral Contrast
Spectral Contrast(SC) Feature considers spectral peak and valley intensity of each subband individually, so that the relative spectral characteristics of lower bands. spectral peaks and valleys are estimated by the average of a small

neighbourhood (given by α) around the maximum and minimum of the sub-band.k is frequency bin and $x_{k,i}$ is $k-th$ sub-band of the audio signal, sorted into descending order of magnitude.

$$Peak_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i}\right) \qquad Valley_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,N-i+1}\right) \tag{3}$$

$$SC_k = Peak_k - Valley_k \tag{4}$$

5) Spectral Rollof
The roll-off frequency is defined for each frame as the center frequency for a spectrogram bin such that at least roll percent (0.85 %) of the energy of the spectrum in this frame is contained in this bin and the bins below. Rolloff measure the spectral shape.

$$\sum_{k}^{R_t} S(k) = 0.85 \sum_{k} S(k) \tag{5}$$

6) Spectral Flux
Spectrum flux is a measure of how quickly the power spectrum of a signal calculated by comparing the power spectrum of one frame to the power spectrum of the previous frame. The result of spectral flux through $H(x)$ is mainly used to find the moment at which a musical instrument begins, since it contains only information about the increased energy compared to the previous frame.

$$SF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|) \qquad H(x) = \frac{x+|x|}{2} \tag{6}$$

7) Mel-Frequency Cepstral Coefficients
The mel-frequency cepstral coefficients(MFCCs) of a signal are a small set of featurey hich concisely describe the overall shape of a spectral envelope. MFCC take the STFT of (a windowed excerpt of) a signal. And map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows. Take the logs of the powers at each of the mel frequencies. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.The MFCCs are the amplitudes of the resulting spectrum.

8) Statistics of Mel scale
The MFCC removes the correlation between the Mel-spectogram information and overlapping windows and compresses the information. However, in order to reflect the interaction between information to the feature, Mel-Spectrum was changed in dB, and the mean value and standard deviation value of each bin were used as a feature.

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{7}$$

3.1.2 Pitch/Harmony features

1) Chroma STFT
Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

3.1.3 Rhythm features

1) Tempo
Tempo refers to the speed of a musical piece. More precisely, tempo refers to the rate of the musical beat and is given by the reciprocal of the beat period. Tempo is often defined in units of beats per minute (BPM).Tempo is expected to be a feature that distinguishes voices and instruments.

3.2 Machine Learning Algorithm

Classification is a problem of calculating the dependent variable category (class) which is most related to the independent variable value when the independent variable value is given. We solve the multi-class classification problem with 10 classes. The selection of the model was made by reflecting the characteristics of the classification model. Discriminate model(SVM, Softmax regression), generative model(Gaussian Naive bayes), non-parametric model(KNN, Decision Tree), ensemble model(Random Forest). Cross validation was performed by dividing 1200 train sets into 5 folds.

1) SVM
SVM assumes any decision boundaries. Among the data belonging to each class, the most front data nearest to the border is called a support vector. The sum of the distance between the support vector and the decision boundary is called the margin($||w||$). SVM is a technique for finding a classification boundary that maximizes margins.

$$obj(f) = max\frac{2}{\|w\|_2} \rightarrow \min \frac{1}{2}\|w\|_2^2 \qquad sub : y_i(w^T x_i + b) \geq 1 \tag{8}$$

2) Softmax regression
Softmax regression is a multi-label classification of logistic regression. It is a way to learn the most likely parameters of y-label within probability space. Where x(i) is a vector for all features $x(i)$ is a single column vector of shape [D,1]. is parameter of softmax regression and all label have each represented parameter.

$$\begin{bmatrix} P(y=1 \mid x) \\ \cdots \\ P(y=n \mid x) \end{bmatrix} = \begin{bmatrix} \frac{exp(\theta_1^T x)}{\sum_k exp(\theta_k^T x)} \\ \cdots \\ \frac{exp(\theta_n^T x)}{\sum_k exp(\theta_k^T x)} \end{bmatrix} \tag{9}$$

3) Gaussian Naive Bayes
If the feature vector x is multidimensional, that is, if $x = (x_1, ..., x_d)$, then the probability of all probability $x = (x_1, ..., x_D)$ $p(x_1, ..., x_D|y = k)$. However, the higher the dimension, the more likely it is that this multidimensional association probability is actually hard to obtain, so the assumption is that the individual independent variable elements of all dimensions are conditional independent from each other. The model applying this assumption to the Bayes classification model is the Naive Bayes classification model. For each independent variable x_d and for class k, the assumed Gaussian distribution is different.

$$P(y=k \mid x) \;\; \propto \;\; \prod_{d=1}^{D} P(x_d \mid y=k) \; P(y=k) \tag{10}$$

$$P(x_d \mid y=k) = \frac{1}{\sqrt{2\pi\sigma_{d,k}^2}} \exp\left(-\frac{(x_d - \mu_{d,k})^2}{2\sigma_{d,k}^2}\right) \tag{11}$$

4) KNN
KNN is a non-parametric methodology that predicts new data with the information of k closest neighbors of existing data when new data is given. KNN's hyper parameters are the number of neighbors to be searched k and the distance measurement method.

5) Decision Tree
Decision Trees (DTs) are a non-parametric methodology. The method of defining the classification rule is to find the best independent variable and reference value that makes the entropy between the parent node and the child node the lowest. Quantifying these criteria is information gain(IG). The key idea of information gain is that unlikely events are more informative than frequently occurring events.

$$IG[Y, X] = H[Y] - H[Y|X] \tag{12}$$

$$H\left(x\right) = -\sum_{x} P(x) \log P(x) \tag{13}$$

6) Random Forest
Random Forest is a model combining method that uses Decision Tree as an individual model. The random forest selects and uses only a part of the data feature dimension. However, we do not select the best independent variables by comparing all the independent variables at the time of node separation, but randomly reduce the independent variable dimension and then select the independent variable among them.

4 Experiments

In this report, the main goal was to increase the accuracy of the model, and to understand if the accuracy is going to increase when a certain feature is used. Therefore, we want to check 5fold cross validation results and test accuracy of 6 feature groups by combining the feature groups of 3 elements described above. Feature parameter are followed, sampling rate = 22050, Frequency size = 2048, Hop size = 512, MFCC dimension = 13, Number of Chorma Coefficients = 12, Mel-scale Coefficient = 10.

4.1 hypotheses set

We constructed six hypotheses through a combination of three feature groups. It is as follows.

- Timber
- Pitch/harmony features
- Rhythm features.
- Timber + Pitch/harmony
- Timber + Rhythm
- Pitch/harmony + Rhythm
- Timber + Rhythm + Pitch/harmony

4.2 Hyperparameter optimization

We use grid-search that is exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. With grid-search, also using 5fold cross validation.

5 Result

Classification accuracy varied depending on the nature of the machine learning algorithm and feature group. The final best result was the Timber + Rhythm feature with 92.5 accuracy and SVM model. The SVM model performed better than the other models overall. Timber features also showed better results than other features.

Table 1: Accuracy of Hypotheses set.ML-Algorithm

hypotheses set	SVM	Softmax	GaussianNB	KNN	DT	RF
Timber	91.0	88.0	77.5	84.5	70.5	88.0
Pitch/harmony	63.5	36.5	42	55.0	44.5	55
Rhythm	15.5	14.5	14.0	8.5	6.5	6.5
Timber + Pitch/harmony	90.5	88.0	30	71.0	63.5	89.0
Timber + Rhythm	92.5	88.0	73.0	85.5	61.5	88.5
Pitch/harmony + Rhythm	64	40.5	30.0	56.0	20.0	45.5
Timber + Rhythm + Pitch/harmony	91.5	87.0	77.5	73.0	62.5	90.5

6 Conclusion

6.1 Discussion

1) Important feature
Pitch / harmony + Rhythm features did not perform better than Timber feature. The reason for this is expected to be the difference between the absolute number of Timber features and the fact that it is an instrument voice file with a duration of 4 seconds. The Pitch / harmony + Rhythm feature is expected to have a good effect on more complex signal data, music, long sounds, and more structured data sets.

2) Interrupt feature
In the case of the Pitch / Harmony feature, it was found that the performance was lower when used with other features. It can be seen that this is a feature to insert disturbing information into the instrument classification model with duration of 4 seconds. It was confirmed that the increase in the number of features could not guarantee the increase in performance and that the number of computations and the number of parameters could be increased.

6.2 Future Work

1) Large-scale Data set
The data used at this time was 1200 train + validation sets of four seconds and 200 test sets. But it is not known whether it works well on larger scale train data. So we propose data argumentation. we must combine the background noise and white noise with the number of data samples, and then learn from a variety of larger data sets.

2) Pitch / harmony + Rhythm feature
The Pitch / harmony + Rhythm feature was not suited to short instrument classifications. Categorizing is necessary because of short sound. We need a method to create a cluster of sounds in a latent state, or a feature extraction technique to find rhythm in a shorter unit.

References

[1] G.Tzanetakis and P.Cook Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, pages 293-302. IEEE, 2002.

[2] Jiang, Dan-Ning and Lu, Lie and Zhang, Hong-Jiang and Tao, Jian-Hua and Cai, Lian-Hong Music type classification by spectral contrast feature In *Proceedings. IEEE International Conference on Multimedia and Expo*, pages 113–116. IEEE, 2002.

[3] McFee, Brian and Raffel, Colin and Liang, Dawen and Ellis, Daniel PW and McVicar, Matt and Battenberg, Eric and Nieto, Oriol librosa: Audio and music signal analysis in python *Proceedings of the 14th python in science conference*, 2015.