Monophonic Inustrument Identification

Aniket Shenoy

Abstract
This report talks about monophonic musical instrument recognition using cepstral coefficients as the features.The data set comprised of 347 samples of nine musical instruments belonging to the woodwinds, brass and string families. Classification was carried out using a multinomial logistic regression model and 10-fold cross-validation was used to validate the model. Using about 0.4 seconds of audio from the samples, the accuracy achieved in recognizing the family of instruments correctly was 77% and the accuracy in individual instrument identification was 50%.

Keywords
timbre — cepstrum — cepstral coefficients — FFT — spectrum — features — LR

Contents

Introduction

Instrument identification is a significant sub task of many complex music information processing and retrieval applications such as source separation, automatic transcription, etc.[3] Each instrument has a different pitch range and a unique timbre which is difficult to model or quantify. The human ability to distinguish between musical instruments has been a subject of investigation for a number of years.[2] Although humans can distinguish between various timbres and identify familiar instruments, it is a difficult task even for trained musicians to differentiate between timbres of instruments belonging to the same family when played in certain registers.

In varoius applications, classification down to the level of instrument families is sufficient for practical needs. One such example is searching a music database for brass sounds.[5] This report tries to distinguish between instrument families using cepstral coefficients. Cepstral coefficients have been used extensively in speech analysis and have more recently received attention in music analysis[2].

The outline of this report is as follows: Section 1 describes the data set, Section 2 talks about cepstral coefficients as features, Section 3 discusses the classification algorithm used, Section 4 showcases the results obtained, Section 5 gives a summary and conclusion of the report.

1. Data

The data was collected from University of Iowa's Electronic Music Studio database.[12] It consisted of 347 monophonic audio samples of nine instruments belonging to woodwind, brass and strings family. The samples were in aiff format and were converted to wav format while preserving the sample rate. Each sample was a single note being played on the instrument and was about 3 seconds long. The samples consisted of notes A-G played on all instruments. The instruments used are as shown in Table 1 and the families they belong to are shown in Table 2.

Table 1. Table of Instruments

Sr. No.	Instrument	Number of samples
1	Flute	39
2	Oboe	35
3	Bassoon	40
4	Horn	44
5	Tuba	37
6	Trumpet	36
7	Violin	40
8	Cello	45
9	Double Bass	29

Table 2. Table of Families

Woodwind	Brass	Strings
Flute	Horn	Violin
Oboe	Tuba	Cello
Bassoon	Trumpet	Double Bass

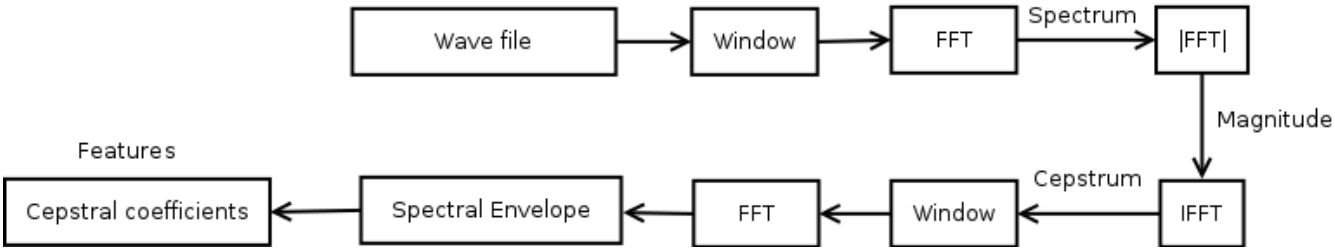

Figure 1. Flow of feature extraction

2. Feature Extraction

The cepstral coefficients were extracted from all the audio samples. FFT length of $N = 2^{14}$ was used. This amounted to about 0.4 seconds from when the instrument starts playing. Thus the feature vector consisted of a 347x8193 matrix where the 347 rows correspond to the 347 audio files, columns 1-8192 correspond to the cepstral coefficients and column 8193 is the class. The subsections below describe cepstral coefficients and how they are calculated.

2.1 Cepstral Coefficients

Cepstral coefficients model spectral energy distribution and characterize the steady state timbre of a signal. They make up the spectral envelope i.e. they are the coefficients of the spectral envelope. Below is a description of spectral envelope and how it is estimated.

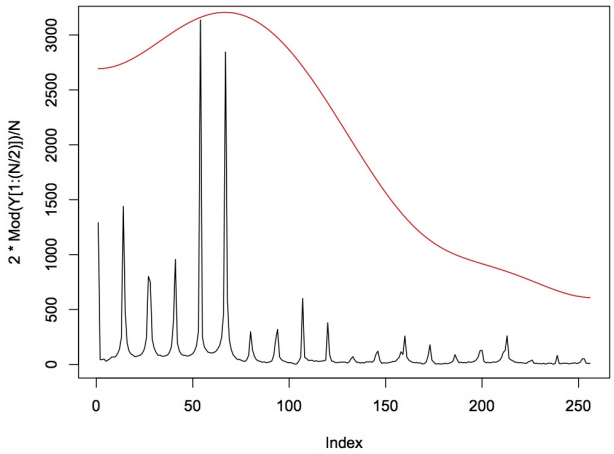

Figure 2. Spectral envelope

2.1.1 Spectral envelope

Spectral envelope is the curve in the frequency-amplitude domain that is derived from the magnitude spectrum of a signal. It is a defining factor for distinguishing timbre as it is distinctive to an instrument.[2] Spectral envelope is based on a smoothing function that passes through the prominent peaks of the magnitude spectrum of a signal. Figure 2 shows the spectral envelope of an oboe audio sample. Spectral envelope is extracted using cepstral analysis.

2.1.2 Cepstrum

Cepstrum is derived by taking the inverse Fourier transform (IFFT) of the magnitude spectrum of a signal.

$$c(j) = FFT^{-1}(|Y(j)|)$$

where,

$$Y(n) = FFT(y) = \sum_{j=0}^{N-1} y(j) e^{-\frac{2\Pi i jn}{N}}$$

The spectral envelope is estimated from the low frequencies of the cepstrum. The first components of the cepstrum correspond to the general shape of the spectrum.[4]

$$E = FFT(w.c(j))$$

i.e.

$$E = FFT(w.FFT^{-1}(|Y(j)|))$$

The curve is made smooth by applying a window function to the magnitude spectrum of the signal. Figure 1 shows the flow of how cepstral coefficients are calculated.

3. Classification

Classification was carried out using Multinomial Logistic Regression (LR). Multinomial LR generalizes LR to multiclass classification problems. Logistic regression can handle non-linear relationships between the response variable and the features. Also it does not assume normally distributed conditional attributes.[8][13] Like linear regression, multinomial LR also has a linear predictor function of the form:

$$f(k,i) = \beta_{0,k} + \beta_{1,k} x_{1,i} + \beta_{2,k} x_{2,i} + \cdots + \beta_{n,k} x_{n,i},$$

where, $\beta_{n,k}$ is the regression coefficient associated with the nth independant variable and the kth response; and x corresponds to the explanatory variables.[13]
This can also be written in vector form as:

$$f(k,i) = \beta_k \cdot \mathbf{x}_i,$$

where, $x_{0,i} = 1$

The independant variables were the $N/2$ cepstral coefficients and the dependant variable was the family to which the instrument belonged. Data was fit using a model of the form:

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{X}_i}}, i = 1, 2, ...k$$

where, $X_1, X_2, ...X_i$ are the independant variables describing the i^{th} observation, $\beta_1, \beta_2, ...\beta_k$ are the weights (regression coefficients) corresponding to the k^{th} outcome and Y is the response.[13]

In multinomial logistic regression, one category of the response variable is chosen as the reference category in order to estimate the intercept. In this case, the woodwind family was chosen as the reference category.

Stratified 10-fold cross-validation was used to validate the model with with a split of 90% train and 10% test data. This was repeated 10 times so that all data samples appear in both the train and test sets. Stratified cross-validation helps to prevent over-fitting by making sure that all the folds contain the same proportion of classes.[14] The accuracy of each hold-out set was then averaged to find the model's overall accuracy.[1]

4. Results

The accuracies attained in the iterations of the stratified 10-fold cross-validation were in the range of 60%-91%. The overall accuracy of the cross-validated model was 77%. Table 3 shows the confusion matrix of the model's predictions. These results were for classifying the instruments into their respective families. For individual instrument identification, the model's accuracy was 50%.

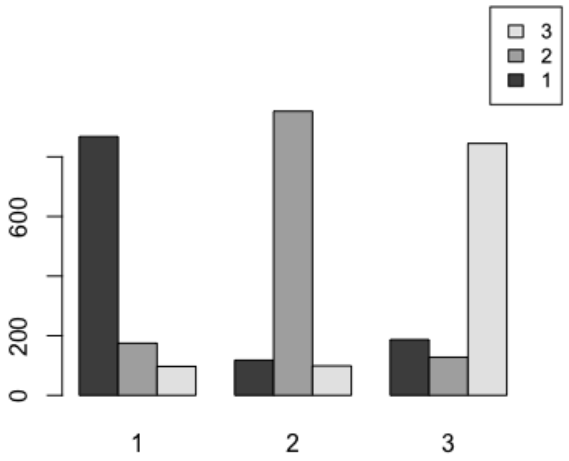

Figure 3. Visualization of results

Table 3. Confusion Matrix

	Class 1	Class 2	Class 3
Class 1	868	118	187
Class 2	175	953	128
Class 3	97	99	845

5. Summary and Conclusions

In this report, an attempt was made to classify monophonic instruments into their respective families by training a multinomial logistic regression model. The features used were the cepstral coefficients, which help to characterize the timbre of an instrument. The cepstral coefficients were extracted by calculating the cepstrum of a short window of the audio samples and then estimating the spectral envelope.

In a classification problem, feature selection is an extremely important step. In this case, cepstral coefficients were the only features used for identification and the model provided an accuracy of 77%. This implies that cepstral coefficients itself are very rich features when it comes to problems such as instrument identification.

Although the multinimial LR model was able to classify instruments into into their families with an overall accuracy of 77%, this is not enough for practical applications. In order to increase the accuracy, the potential improvements could be: using more features of different types that are effective in characterizing the unique traits of an instrument ; and using a more sophisticated machine learning algorithm. Future work will be to implement a more robust classifier and include identification of polyphonic instruments.

References

[1] Lewis Guignard and Greg Kehoe. *"Learning Instrument Identification"*. (2015)

[2] Róisín Loughran, Jacqueline Walker, Michael O'Neill and Marion O'Farrell. *"The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification"*. (2008).

[3] Greg Sell, Gautham J. Mysore, Song Hui Chon. *"Muscal Instrument Detection"* (2006)

[4] Janet Marques and Pedro J. Moreno. *"A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines"*. (1999)

[5] Antti Eronen and Anssi Klapuri. *"Musical Instrument Recognition using Cepstral Coefficients and Temporal Features"*

[6] Martin, K. D. *"Musical Instrument Identification: A Pattern Recognition Approach"*. (1998)

[7] Judith C. Brown. *"Computer identification of musical instruments using pattern recognition with cepstral coefficients as features"* (1998)

[8] Xin Zhang and Zbigniew W. Ras . *"Sound Isolation by Harmonic Peak Partition For Music Instrument Recognition"*.

[9] José Henrique Padovani. *"Spectral envelope extraction by means of cepstrum analysis and filtering in Pure Data"*.

[10] Sang Hyun Park. *"Musical Instrument Extraction through Timbre Classification"*.

[11] Mizuki Ihara, Shin-ichi Maeda, Shin Ishii *"Instrument Identification in Monophonic Music Using Spectral Information"*. (2007)

[12] University of Iowa Electronic Music Studios. *http://theremin.music.uiowa.edu/MISPost2012Intro.html*

[13] Multinomial Logistic Regression. *https://en.wikipedia.org/wiki/Multinomial*$_{l}$*ogistic*$_{r}$*egression*

[14] Cross-Validation. *https://en.wikipedia.org/wiki/Cross-validation*$_{(}$*statistics*$)$