# Clustering Spectral Filters for Feature Extraction in Musical Instrument Identification

**Patrick Donnelly**
**Department of Computer Science, Montana State University**

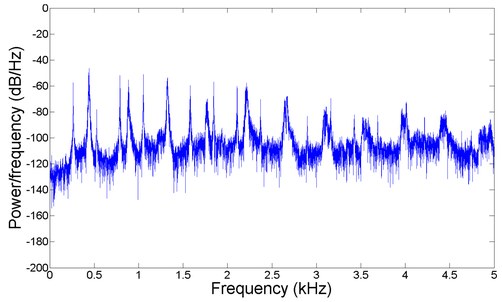
## Abstract

We propose a technique of training models for feature extraction using prior expectation of regions of importance in an instrument's timbre. Over a dataset of training examples, we extract significant spectral peaks, calculate their ratio to fundamental frequency, and use $k$-means clustering to identify a set of windows of spectral prominence for each instrument. These windows are used to extract amplitude values from training data to use as features in classification tasks. We test this approach on two databases of 17 instruments, cross evaluate between datasets, and compare with MFCC features.

## Motivation

Musical instrument identification is important to the tasks of music search, genre identification, and automatic score transcription. While there have been many approaches to recognize individual instruments, these are not extensible to the more complex case of identifying instruments present in polyphonic mixtures, which requires source separation.

Overlapping spectra of a Clarinet (261 Hz) and a Violin (440 Hz)




## Dataset & Signal Processing

We used the set of 17 instruments common to both the University of Iowa Musical Instrument Samples (MIS) and the Real-World Computing (RWC) datasets. These datasets contain several performers, articulations, and dynamic levels. Each example contains a single instrument playing a single note, at a 44.1 kHz sampling rate, 16-bit per sample, single channel waveform.

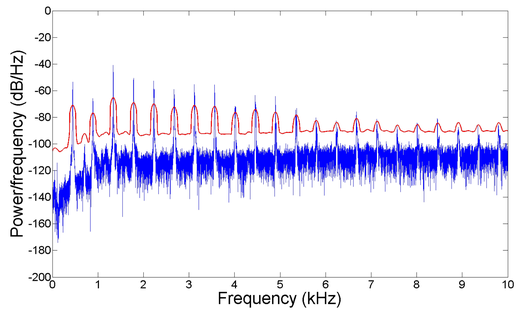
For each instrument example, we establish a smoothed frequency-dependent threshold calculated as the moving weighted average of amplitude values. We extract all peaks that exceed the threshold and for each peak we calculate the ratio of its location to the fundamental frequency.

## Signature Clustering & Feature Extraction

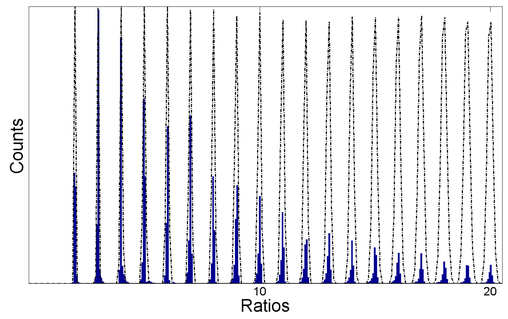
For each example the set of ratios for each example are extracted. The set of ratios are collected for each different instrument and are passed to a $k$-means cluster. For each learned cluster, we extract the mean and standard deviation and save as the spectral signature for that instrument. These clusters serve as a spectral filter to decide the specific locations from which to extract amplitude values as features.

For each example in the dataset, we extract the max amplitude value within each learned signature cluster. This set of amplitudes form the feature vector for that example.

Amplitude spectrum (blue) and threshold (red) of a Clarinet playing A4 (440 Hz)




Histogram of ratios (blue) overlayed with clusters learned for the Clarinet examples in the RWC dataset

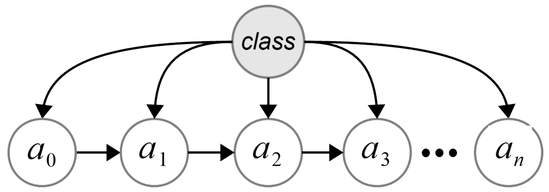



## Experiments

For each dataset, we construct a binary classifier for each instrument. Each classifier is trained with an equal number of positive and negative examples. We compare a Naïve Bayes (NB) classifier, a Bayesian network (BN), $k$-nearest neighbors algorithm ($k$-NN), and a support vector machine (SVM). We use 10-fold cross-validation to evaluate when training and testing on the same dataset. In the cross dataset experiments, we train models on one dataset and test on the other.

We evaluate across our two datasets RWC and MIS using three different feature sets. We train our signature models with 20 and 30 clusters, extract features, and test these features on four algorithms. For comparison, we test using the common feature set of the mean and standard deviation of the first 13 linear Mel-frequency cepstral coefficients (MFCCs), a total of 26 features. Our preset network structure for the BN classifier is not extensible to the MFCC feature set and is omitted.

Structure of Bayesian network (BN) with dependencies between frequency features




## Results

| Features | Algorithm | M M | M R | R R | R M |
|---|---|---|---|---|---|
| 20 Clusters | NB | 0.75 | 0.64 | 0.76 | 0.65 |
| | BN | 0.75 | 0.63 | 0.77 | 0.67 |
| | $k$-NN | 0.71 | 0.61 | 0.72 | 0.63 |
| | SVM | 0.71 | 0.64 | 0.72 | 0.67 |
| 30 Clusters | NB | 0.73 | 0.63 | 0.72 | 0.64 |
| | BN | 0.75 | 0.64 | 0.76 | 0.66 |
| | $k$-NN | 0.72 | 0.61 | 0.71 | 0.63 |
| | SVM | 0.70 | 0.66 | 0.72 | 0.68 |
| MFCC | NB | 0.86 | 0.58 | 0.80 | 0.71 |
| | $k$-NN | 0.91 | 0.71 | 0.91 | 0.80 |
| | SVM | 0.90 | 0.66 | 0.86 | 0.76 |

Results of cross dataset experiments showing the F-Measure averaged over the set of binary classifiers.

The top row shows the training set, M or R, and the bottom row indicates the test set.

| Instr. | NB | BN | $k$-NN | SVM |
|---|---|---|---|---|
| Violin | 0.74 | 0.77 | 0.69 | 0.75 |
| Viola | 0.72 | 0.72 | 0.65 | 0.62 |
| Violincello | 0.68 | 0.73 | 0.72 | 0.70 |
| Contrabass | 0.80 | 0.84 | 0.81 | 0.83 |
| Guitar | 0.80 | 0.85 | 0.90 | 0.85 |
| French Horn | 0.63 | 0.65 | 0.67 | 0.65 |
| Trumpet | 0.70 | 0.70 | 0.63 | 0.72 |
| Trombone | 0.65 | 0.70 | 0.63 | 0.72 |
| Tuba | 0.79 | 0.87 | 0.80 | 0.86 |

| Instr. | NB | BN | $k$-NN | SVM |
|---|---|---|---|---|
| Flute | 0.74 | 0.77 | 0.66 | 0.69 |
| Clarinet | 0.69 | 0.74 | 0.62 | 0.59 |
| Soprano Sax | 0.71 | 0.73 | 0.61 | 0.65 |
| Alto Sax | 0.68 | 0.70 | 0.61 | 0.60 |
| Oboe | 0.72 | 0.76 | 0.61 | 0.72 |
| Basson | 0.63 | 0.66 | 0.74 | 0.63 |
| Marimba | 0.83 | 0.84 | 0.87 | 0.84 |
| Piano | 0.76 | 0.81 | 0.87 | 0.79 |

Results of the RWC dataset with 30 features showing the F-Measure for each binary classifier

## Conclusion

Instead of using a single feature extraction scheme for all examples, we propose a strategy of training models for feature extraction using prior expectation of regions of importance in an instrument's timbre. We design this approach for extensibility to multilabel classification of polyphonic mixtures.

We demonstrate this approach over large datasets of single instrument tones containing multiple articulations, performers, and dynamic levels. Although the MFCC feature set outperformed our approach, MFCCs cannot be used as features in polyphonic mixtures without an prior attempt at source separation, which is a difficult problem in itself. Our approach, however, attempts spectral separation of the sources given known regions of spectral prominence for each instrument. In this work we demonstrated the validity of a feature extraction approach that relies on prior expectations of generalization of an instrument's unique timbre.

MONTANA STATE UNIVERSITY

Numerical Intelligent Systems Laboratory