Selecting Proper Features and Classifiers for Accurate Identification of Musical Instruments

D. M. Chandwadkar and M. S. Sutaone

***Abstract*—Selection of effective feature set and proper classifier is a challenging task in problems where machine learning techniques are used. In automatic identification of musical instruments also it is very crucial to find the right set of features and accurate classifier. In this paper, the role of various features with different classifiers on automatic identification of musical instruments is discussed. Piano, acoustic guitar, xylophone and violin are identified using various features and classifiers. Spectral features like spectral centroid, spectral slope, spectral spread, spectral kurtosis, spectral skewness and spectral roll-off are used along with autocorrelation coefficients and Mel Frequency Cepstral Coefficients (MFCC) for this purpose. The dependence of instrument identification accuracy on these features is studied for different classifiers. Decision trees, k nearest neighbour classifier, multilayer perceptron, Sequential Minimal Optimization Algorithm (SMO) and multi class classifier (metaclassifier) are used. It is observed that accuracy can be improved by proper selection of these features and classifier.**

***Index Terms*—Feature extraction, classification, musical instrument identification.**

I. Introduction

Huge amount of digital audio material is available today as multimedia data on the World Wide Web. For automatic indexing of this data and for database retrieval applications, automatic identification of musical instruments is essential [1]. Instrument identification techniques can have many potential applications. Knowing various musical styles, audio editing, audio retrieval and transcription, play list generation, video scene analysis etc can be considered as some of the key applications.

Musical instruments can be identified by using the monophonic or polyphonic recordings. In this paper the monophonic signals (isolated notes played by various orchestral instruments) are used. The McGill University Master samples collection, a fabulous set of DVDs of instruments playing every note in their range, recorded in studio conditions are used as the database.

Instrument classification technique can generally be described as follows [1]:

- Lists of features are selected to describe the samples.
- Values for these features are computed.
- A learning algorithm that uses the selected features to discriminate between instruments is applied.
- The performance of the learning procedure is evaluated by classifying new sound samples (cross-validation).

One of the most crucial aspects in the above procedure of instrument classification is to find the right features [2] and classifiers. Most of the research on audio signal processing has been focusing on speech recognition and speaker identification. Few features used for these can be directly applied to solve the instrument classification problem.

In this paper, feature extraction and selection for instrument classification using machine learning techniques is considered. Four instruments: piano, acoustic guitar, xylophone and violin are identified using various features and classifiers. A number of spectral features, MFCCs, and autocorrelation coefficients are used. All features are first extracted. Instrument identification accuracy using these features for various classifiers is noted. Decision trees, k-nearest neighbour classifier, multilayer perceptron, Sequential Minimal Optimization Algorithm (SMO) and multi class classifier (meta classifier) are used. The performance of these features is assessed first individually, and then in combination with each other. The feature set is then reduced using principal component analysis and data set of reduced features is further tested with these classifiers using cross validation. A comparison of the classification accuracy is presented for further studies.

II. Features Used

The audio signal is described using various numerical values extracted from the signal. These are called as features of the signal. In this work the following features have been used:

Spectral shape features: features (instantaneous) computed from the Short Time Fourier transform (STFT) of the signal. These include centroid, spread, slope, skewness, kurtosis, roll-off, MFCC.

Temporal features: Autocorrelation coefficients.

Four musical instruments: piano, acoustic guitar, xylophone and violin are identified using the following features.

1) The spectral centroid (μ) is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound [3]. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights.

The centroid measures the spectral shape. Higher centroid values indicate higher frequencies.

For the time-domain signal x(t):

Manuscript received January 12, 2013; revised April 11, 2013.

D. M. Chandwadkar is with Department of Electronics and Telecommunication, K. K. Wagh Institute of Engineering Education & Research, Nashik, Maharashtra, India. (e-mail: dmc.eltx@kkwieer.org).

Dr. M. S. Sutaone is with the Department of Electronics and Telecommunication, Government College of Engineering, Pune, Maharashtra, India. (e-mail: mssutaone.extc@coep.ac.in).

$$A(f) = \left| F[x(t)] \right| \quad (1)$$

$$\text{Spectral centroid,} \quad \mu = \int f.p(f).df \quad (2)$$

$$\text{Where,} \quad p(f) = \frac{A(f)}{\sum_f A(f)} \quad (3)$$

2) The spectral spread (σ) is a measure of variance (or spread) of the spectrum around the mean value μ calculated in equation:

$$\sigma^2 = \int (f-\mu)^2.p(f)df \quad (4)$$

3) In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the *tail* on the left side of the probability density function is *longer* than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the *tail* on the right side is *longer* than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.

Thus, the spectral skewness is a measure of the asymmetry of the distribution around the mean value μ. The skewness (γ_1) is calculated from the 3rd order moment, m_3 as:

$$m_3 = \int (f-\mu)^3.p(f)df \quad (5)$$

$$\gamma_1 = \frac{m_3}{\sigma^3} \quad (6)$$

4) Spectral kurtosis (γ_2) indicates the flatness or peakedness of the energy distribution. Higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. It is calculated from the 4th order moment, m_4, using the value of μ as:

$$m_4 = \int (f-\mu)^4.p(f)df \quad (7)$$

$$\gamma_2 = \frac{m_4}{\sigma^4} \quad (8)$$

If kurtosis $\gamma 2 = 3$, then it indicates a normal (Gaussian) distribution. Spectra with $\gamma 2 < 3$ are flatter and conversely spectra with $\gamma 2 > 3$ have a more defined, sharper peak.

5) Spectral slope is a measure of how quickly the spectrum of an audio sound tails off towards the high frequencies, calculated using a linear regression. The spectral slope (m) gives an indication of the rate of decrease of the amplitude A(f). The slope is simply a linear regression of the spectral amplitude.

$$m = \frac{1}{\sum_f A(f)} \frac{N\sum_f f.A(f) - \sum_f f \times \sum_f A(f)}{N\sum_f f^2 - \left(\sum_f f\right)^2} \quad (9)$$

6) Spectral Roll-off is another measure of spectral shape. It is the point where frequency that is below some percentage (usually at 95%) of the power spectrum resides. It is often used as an indicator of the skew of the frequencies present in a window.

The spectral roll-off point (f_c) is the frequency for which 95% of the signal energy is below this frequency. Using the amplitude A(f):

$$\sum_0^{f_c} A^2(f) = 0.95 \sum_0^{f_{ny}} A^2(f), \quad (10)$$

where f_{ny} is the Nyquist frequency.

7) The autocorrelation of a signal is a measure of how well a signal matches with a time shifted version of itself. The autocorrelation of a frame represents the distribution of the signal spectrum but in the time domain. This feature was demonstrated to provide a good descriptor for classification by Brown [4].

Correlation is a mathematical tool used frequently in signal processing for analyzing functions or series of values, such as time domain signals. Correlation is the mutual relationship between two or more random variables. Autocorrelation is the correlation of a signal with itself. This is unlike cross-correlation, which is the correlation of two different signals.

Autocorrelation is useful for finding repeating patterns in a signal, such as determining the presence of a periodic signal which has been buried under noise, or identifying the fundamental frequency of a signal which doesn't actually contain that frequency component, but implies it with many harmonic frequencies.

Autocorrelation is implemented in the time domain as the convolution of a signal with itself reversed. Because convolution in the time domain becomes multiplication in the frequency domain, the autocorrelation can also be calculated as the Fourier transform of the power spectrum. The autocorrelation of a signal is zero phase and symmetric about t=0, so coefficients only need to be taken from one side. In this experiment 12 autocorrelation coefficients are used.

8) In sound processing, the MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

In short, MFCCs are cepstral coefficients used for representing audio in a way that mimics the physiological properties of the human auditory system. MFCCs were initially developed for speech, but they are also heavily used in other sound applications [5, 6]. MFCCs were successfully used to get the best accuracy in instrument family classification along with reduced computational complexity [7]. It has been proved that MFCCs are the better choice as compared to other features, both for musical instrument

modeling and for automatic instrument classification [8].

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

III. Feature Extraction

Following procedure has been used to compute features for Musical Instrument Identification using isolated solo notes played by the instruments with the help of MATLAB software.

1. From the waveform, the sampling frequency and other parameters are obtained.
2. The waveform is divided into small windows. To extract the features, music sound samples are segmented into 23 ms frames with 11.5 ms overlap. Hamming window is used. The music signals used from the McGill University master samples are sampled at 44.1 KHz. Hence in 23 ms, 1024 (210) samples are obtained. The FFT length is also taken as 1024.
3. For each window, the features are calculated. The 31 features are: spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral slope, spectral rolloff, twelve autocorrelation coefficients and thirteen MFCC coefficients. Feature vector of these 31 features for each window of 23 ms is obtained.
4. As the size of this feature vector is very large and depends on the length of the input wave file, instead of using these features directly for classification, their minimum value, maximum value, mean, standard deviation and variance are obtained using Matlab. In short, the statistical information of these features is used. Hence the number of features become 31*5=155, for one data sample.
5. These features are used for classification.

For piano, 86 data samples, for acoustic guitar 48, for xylophone 44, and for violin 21 samples were used. Thus, total 199 samples of data for various instruments have been used.

IV. Classifiers

Classification relies on the basic assumption that each observed pattern belongs to a category. Individual signals may be different from one another, but there is a set of features that are similar to patterns belonging in same class and different patterns for a different class. The feature sets are the base that can be used to determine class membership. Classification is domain-independent and provides many fast, elegant and well-understood solutions that can be adopted for use in musical instrument recognition.

A well known machine-learning scheme called WEKA (Waikato Environment for Knowledge Analysis) is used for identification of musical instruments using trained statistical pattern recognition classifiers [9]. It enables pre-processing, classifying, clustering, attributes selections and data visualizing. WEKA is employed when applying a learning method to a dataset and during analysis of its output to extract information about the data.

Classification results were tested using stratified ten-fold cross validation. Cross-validation (CV) is a standard evaluation technique in pattern classification, in which the dataset is split into *n* parts (folds) of equal size. *n*-1 folds are used to train the classifier. The *n*th fold that was held out is then used to test it.

The following classifiers found suitable for this application have been used.

1. Decision Trees: Class for generating a pruned or unpruned C4.5 decision tree.
2. K-nearest neighbours classifier: Can select appropriate value of K based on cross-validation. Can also do distance weighting.
3. Multilayer Perceptron: A Classifier that uses backpropagation to classify instances.
4. SMO (Sequential Minimal Optimization Algorithm): Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
5. Multiclass classifier (Meta classifier): A metaclassifier for handling multi-class datasets with 2-class classifiers. This classifier is also capable of applying error correcting output codes for increased accuracy.

Procedure for classification:

- Feature vectors for 31 features derived for 23 ms windows are obtained experimentally for four instruments (Piano, Acoustic Guitar, Xylophone and Violin). For this purpose the McGill University Database is used.
- Data for classification is prepared in required format.
- 10-fold cross-validation is performed.
- Classification accuracy is noted.
- Classification accuracies are recorded for various combinations of features.

V. Experimental Results

A. Classification Accuracy with One Feature

Initially only one type of feature is used at a time. Fig. 1 shows the confusion matrix for 13 MFCC coefficients used as a feature (their minimum value, maximum value, mean, standard deviation and variance as attributes) and multilayer perceptron (MLP) used as classifier.

a	b	c	d	← classified as
84	2	0	0	PIANO
0	46	1	1	ACOUSTIC_GUITAR
0	1	43	0	XYLOPHONE
0	0	0	21	VIOLIN

Fig. 1. Confusion Matrix for MLP used as classifier and MFCC as feature

Out of 199 instances 194 instances are classified correctly giving 97.48% classification accuracy.

Similar procedure is used for each feature with all classifiers mentioned above and the associated classification accuracy using only one feature at a time is as given in Table I.

TABLE I: CLASSIFICATION ACCURACY FOR INDIVIDUAL FEATURES

Feature	Classification accuracy (%) for classifier				
	Decision Tree J48	***kNN***	***MLP***	***SMO***	***Multiclass Classifier***
Spectral Centroid	82.41	81.40	84.00	63.00	74.37
Spectral Spread	67.33	74.37	79.90	46.73	71.35
Spectral Skewness	80.40	72.36	70.85	53.77	62.31
Spectral Kurtosis	79.00	69.00	71.86	45.73	58.79
Spectral Slope	43.21	80.90	83.00	63.32	73.37
Spectral Rolloff	83.00	80.40	80.40	54.77	78.39
All Spectral features (as above)	90.45	83.92	92.46	85.42	92.96
Autocorr coeff	82.41	77.38	94.00	88.44	85.00
MFCC	91.96	90.95	97.48	97.00	93.97

From the analysis it is observed that the overall classification accuracy with MFCC used alone is the best as compared to other features. It is also found that Multilayer Perceptron, k nearest neighbor classifier, Decision tree and Multi Class Classifier are giving good accuracy.

B. Classification Accuracy Using All Features

After testing the classification accuracy for each feature, all features were combined and the classification accuracy using these classifiers was found. Fig 2 shows confusion matrix for this combination with SMO used as classifier.

197 out of 199 instances are classified accurately (99%). Similar procedure is used for other classifiers mentioned above and the associated classification accuracy is noted.

a	b	c	d	← classified as
85	0	0	1	PIANO
0	48	0	0	ACOUSTIC_GUITAR
0	0	43	1	XYLOPHONE
0	0	0	21	VIOLIN

Fig. 2. Confusion Matrix for SMO used as classifier

Then principal components of the feature set are identified and the accuracy using all the classifiers is computed using these. The confusion matrix for Multiclass Classifier (Meta Classifier) used as classifier for these principal components is as shown in Fig 3. Here out of 199 instances, 195 are classified correctly giving 98% accuracy.

a	b	c	d	← classified as
86	0	0	0	PIANO
0	47	0	1	ACOUSTIC_GUITAR
0	1	42	1	XYLOPHONE
0	0	1	20	VIOLIN

Fig. 3. Confusion Matrix for Multiclass (Meta) classifier

The classification accuracy for these two cases for various classifiers is listed in Table II.

TABLE II: CLASSIFICATION ACCURACY WITH ALL FEATURES

Feature	Classification accuracy (%) for classifier				
	Decision Tree J48	***kNN***	***MLP***	***SMO***	***MultiClass Classifier***
All features (31*5=155)	94-97	90	98.49	99.00	98.49
Principal Components (33 PCs)	82.41	92.96	96	97.48	98

The classification accuracy using all 155 features is the best. Using 33 principal components also we get comparable accuracy. SMO (Sequential Minimal Optimization Algorithm), Multilayer perceptron and Multi Class Classifier are giving better results.

VI. CONCLUSION

It is observed that MFCC when used alone give better results as compared to the other features. Classification accuracy with all 155 features is better. We also get comparable accuracy using 33 principal components.

If the number of features is less (one feature used at a time), Multilayer Perceptron, k-nearest neighbor classifier, Decision tree and Multi Class Classifier are giving good accuracy. With large number of features, SMO (Sequential Minimal Optimization Algorithm), Multilayer perceptron and Multi Class Classifier are giving good accuracy.

We conclude that the set of features used (spectral centroid, spread, slope, skewness, kurtosis, roll-off, MFCC, and autocorrelation coefficients) can identify the four musical instruments (piano, acoustic guitar, xylophone and violin) with better accuracy (up to 99%). Multilayer perceptron, sequential minimal optimization algorithm and multi class classifier give better classification accuracy if the number of features is large.

REFERENCES

[1] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3, 2003.

[2] C. Simmermacher, D. Deng, and S. Cranefield, "Feature analysis and classification of classical musical instruments: An empirical study," in *Advances of Data Mining*, 2006, Springer LNAI 4065, pp. 444–458.

[3] G. Peeters, "A large set of audio features for sound description (similiarity and classification) in the CUIDADO project," in *Proc. 115th AES Convention*, 2004.

[4] J. Brown, "Cluster-based probability model for musical instrument identification" *Journal of the Acoustical Society of America*, vol. 101, pp. 3167, 1997.

[5] K. D. Martin, "Sound-Source Recognition: A Theory and computational model," PhD. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[6] J. H. Jensen, M. G. Christensen, M. Murthi, S. H. Jensen, "Evaluation of MFCC estimation techniques for music similarity", in *Proc. European Signal Processing Conference*, 2006, pp. 926–930.

[7] A. Eronen, "Comparison of features for musical instrument recognition," Workshop on Signal Processing for Audio and Acoustics(WASPAA), pp. 19-22, 2001.

[8] AB. Nielsen, S. Sigurdsson, L. Hansen, and J. Arenas-Garcia, "On the relevance of spectral features for instrument classification," in *Proc. Acoustics, Speech and Signal Processing (ICASSP-07)*, 2007.

[9] Ian H. Witten, Eibe Frank, *Data Mining- Practical Machine Learning Tools and Techniques*, Second Edition, Morgan Kaufmann Publishers: An Imprint Of Elsevier, 2005.

Dinesh M. Chandwadkar is a professor in E & TC Department at K. K. Wagh Institute of Engineering Education & Research, Nashik, India. He is pursuing Ph. D. under the guidance of Prof. Sutaone.

Dr. Mukul S. Sutaone is a professor in Electronics and Telecommunication department, and Dean, Alumni and International Affairs, at College of Engineering, Pune (COEP), India. He has Ph.D. in the field of Texture Analysis. His areas of research interests are Multimedia, Multidimensional Signal Processing and Multi-carrier Communication.