MUSICAL INSTRUMENT RECOGNITION USING CEPSTRAL COEFFICIENTS AND TEMPORAL FEATURES

Antti Eronen and Anssi Klapuri

Signal Processing Laboratory, Tampere University of Technology
P.O.Box 553, FIN-33101 Tampere, FINLAND
eronen@cs.tut.fi, klap@cs.tut.fi

ABSTRACT

In this paper, a system for pitch independent musical instrument recognition is presented. A wide set of features covering both spectral and temporal properties of sounds was investigated, and their extraction algorithms were designed. The usefulness of the features was validated using test data that consisted of 1498 samples covering the full pitch ranges of 30 orchestral instruments from the string, brass and woodwind families, played with different techniques. The correct instrument family was recognized with 94% accuracy and individual instruments in 80% of cases. These results are compared to those reported in other work. Also, utilization of a hierarchical classification framework is considered.

1. INTRODUCTION

Music content analysis in general has many practical applications, including e.g. structured coding, database retrieval systems, automatic musical signal annotation, and musicians' tools. A subtask of this, automatic musical instrument identification, is of significant importance in solving these problems, and is likely to provide useful information also in other sound source identification applications, such as speaker recognition. However, musical signal analysis has has not been able to attain as much commercial interest as, for instance, speaker and speech recognition. This is because the topics around speech processing are more readily commercially applicable, although both areas are considered as being highly complicated.

First attempts in musical instrument recognition operated with a very limited number of instruments and note ranges. De Poli and Prandoni used mel-frequency cepstrum coefficients calculated from isolated tones as an inputs to a Kohonen self-organizing map, in order to construct timbre spaces [2]. Kaminsky and Materka used features derived from an rms-energy envelope and used a neural network or a k-nearest neighbour classifier to classify guitar, piano, marimba and accordion tones over a one-octave band [5].

The recent systems have already shown a considerable level of performance, but have still been able to cope with only a quite limited amount of test data. In [7], Martin reported a system that operates on single isolated tones played over the full pitch ranges of 15 orchestral instruments and uses a hierarchical classification framework. Brown [1] and Martin [8] have managed to build classifiers that are able to operate on test data that include samples played by several different instruments of a particular instrument class, and recorded in environments which are noisy and reverberant. However, even the recent systems are characterized either by a limited application context or by a rather unsatisfactory performance.

In this paper, we aim at utilizing a widest range of features characterizing the different properties of sounds. This is done in order to handle a certain defect in the earlier proposed systems: failure to make simultaneous and effective use of both spectral and temporal features, which is suggested by the work in psychoacoustics. Signal processing methods were implemented that attempt to extract cues about the temporal development, modulation properties, irregularities, formant structure, brightness, and spectral synchronicity of sounds. Although all the factors in sound source identification, and especially their interrelations are not known, a large number of them have been proposed. Thus it looked particularly attractive for us to utilize as much as possible of that information simultaneously in a recognition system, and to see if that would allow us to build a more robust instrument recognition system than described in experiments so far.

Our current implementation handles the isolated tone condition well, and we are hoping that it will generalize to still more realistic contexts. A practical goal of our research is to build an instrument recognition module that can be integrated to an automatic transcription system [6].

This paper is organized as follows. In Section 2, we shortly review the literature in sound source identification and perception. In Section 3, we first take a look at the features used in instrument recognition systems and discuss the approach taken in this paper. Then we describe our feature extraction algorithms. In Section 4, the selected features are validated with thorough simulations and the classification results are compared to those of earlier studies.

2. DIMENSIONS OF TIMBRE

A considerable amount of effort has been done in order to find the perceptual dimensions of *timbre*, the 'colour' of a sound. Often these studies have involved multidimensional scaling experiments, where a set of sound stimuli is presented to subjects, who then give a rating to their similarity or dissimilarity. On the basis of these judgements a low-dimensional space, which best accommodates the similarity ratings, is constructed and a perceptual or acoustic interpretation is searched for these dimensions.

Two of the main dimensions described in these experiments have usually been spectral centroid and rise time [3][9]. The first measures the spectral energy distribution in the steady state portion of a tone, which corresponds to perceived brightness. The second is the time between the onset and the instant of maximal amplitude.

Table 1: Feature descriptions	
1	Rise time, i.e., the duration of attack
2	Slope of line fitted into rms-energy curve after attack
3	Mean square error of line fit in 2
4	Decay time
5	Time between the end of attack and the maximum of rms-energy
6	Crest factor, i.e., *max / rms* of amplitude
7	Maximum of normalized spectral centroid
8	Mean of normalized spectral centroid
9	Mean of spectral centroid
10	Standard deviation of spectral centroid
11	Standard deviation of normalized spectral centroid
12	Frequency of amplitude modulation, range 4-8Hz
13	Strength of amplitude modulation, range 4-8Hz
14	Heuristic strength of the amplitude modulation in range 4-8Hz
15	Frequency of amplitude modulation, range 10-40Hz
16	Strength of amplitude modulation, range 10-40Hz
17	Mean error of the fit between each of steady state intensities and mean steady state intensity
18	Mean error of fit between each of onset intensities and mean onset intensity
19	Overall variation of intensities at each band
20	Fundamental frequency
21	Standard deviation of fundamental frequency
22-32	Average cepstral coefficients during onset
33-43	Average cepstral coefficients after onset

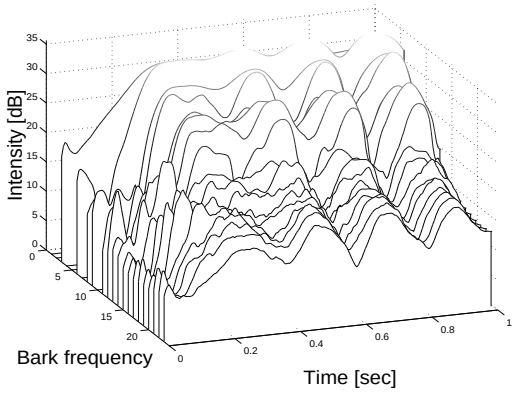

Figure 1. Flute tone: intensities as a function of Bark frequency. Especially amplitude modulation can be seen clearly.

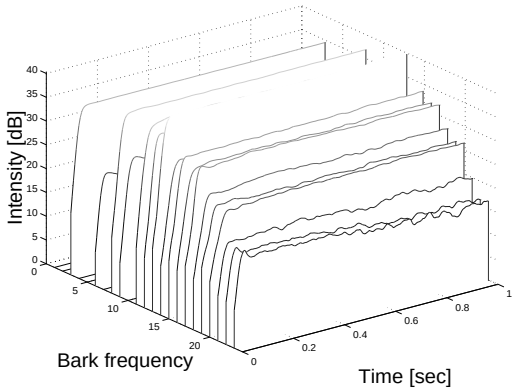

Figure 2. Clarinet tone: intensities as a function of Bark frequency plot. At the low end of clarinet playing range the odd partials are much stronger than the even partials.

The psychophysical meaning of the third dimension has varied, but it has often related to temporal variations or irregularity in the spectral envelope. A good review over the enormous body of timbre perception literature can be found in [4]. These available results provide a good starting point for the search of features to be used in musical instrument recognition systems.

3. CALCULATION OF FEATURES

Traditionally, the features provided by the timbre research can be divided into spectral and temporal ones. In instrument recognition systems reported so far, only features of either type have been used. For instance, Kaminsky and Materka used temporal features derived from a short time rms-energy envelope [5]. In the research of Martin [7][8], a selection of temporal features calculated from the outputs of a log-lag correllogram was used, but the spectral shape was not considered at all. Brown reports good results been achieved with cepstral coefficients calculated from oboe and saxophone samples [1]. Her analysis method did not take into account the temporal evolution of the spectral features, but rather resembled a method often used in speaker identification systems.

We wanted to test if combining the two types of features, cepstral coefficients and temporal features, would yield the necessary extra discriminating power needed for instrument recognition with a wider set of instruments. The feature set we used is presented in Table 1.

3.1 Feature extraction methods

The short-time rms-energy envelope contains information especially about the duration of excitation. We estimated rise-time, decay-time, strenght and frequency of amplitude modulation, crest factor and detected exponential decay from the rms-energy curve calculated in 50% overlapping 10ms frames.

The spectral centroid of the signal is calculated over time in 20ms windows. At each window, the rms-energy of the spectrum is estimated using logarithmic frequency resolution. After that, the spectral centroid is calculated. We use both the absolute value of spectral centroid and a normalized value, which is the absolute value divided by the fundamental frequency. The fundamental frequency estimation method used here is the one presented by Klapuri in [6].

Sinusoid track representation provides many useful temporal features. We first calculate the harmonic amplitude on each of Bark scale bands, which resemble the frequency resolution of the cochlea. Knowledge about the fundamental frequency is applied in

order to resolve whether any harmonics are found on each band. The amplitude envelopes of single harmonic frequencies can be calculated efficiently with an *O(n)* algorithm, where *n* is the sample length. If more than one harmonic frequencies are found, then amplitude envelopes are calculated separately and the resulting band-amplitude is the mean of these. The band-wise intensity is calculated by multiplying the amplitude by the center frequency of the band.

The intensities are decimated by a factor of about 5ms to ease the following computations and smoothed by convolving with a 40ms half-hanning (raised-cosine) window. This window preserves sudden changes, but masks rapid modulation. Figures 1 and 2 display intensity versus Bark frequency plots for 261Hz tones produced by flute and clarinet, respectively.

When the intensity matrix is calculated, a number of features can be easily extracted. The similarity of shape between intensity envelopes is measured by fitting the envelopes into a mean envelope and calculating the mean of mean squre errors. This is done separately for the onset period and the rest of the waveform. The error value of the onset period, accompanied with the standard deviation of banwise rise times, can be considered as a measure of onset asynchrony. Another measure that can be extracted from the intensity envelope curves is the overall variation of intensities at each band.

The spectral shape of tones is modelled with cepstral coefficients, which are calculated with a method adapted from an automatic speech recognition system described in [11]. Calculation procedure is done in 25% overlapping windowed frames of size approximately 20ms. Autocorrelation sequence is calculated first and then used for LPC coefficient calculation with Levinson-Durbin algorithm. LPC coefficients are then converted into cepstral coefficients, which have been found to be a robust feature set for use in speech and instrument recognition [1]. Instead of using averaged coefficients over the whole signal, which has been done in [1], we used two sets of 11 coefficients, averaged over the onset and the rest of the sample.

4. CLASSIFICATION

Musical instruments form a natural hierarchy, which includes different instrument families. In many applications, classification down to the level of instrument families is sufficient for practical needs. For example, searching a database to find string music would make sense. In addition to that, a classifier may utilize a hierarchical structure algorithmically while assigning a sound into a lowest level class, individual instrument. This has been proposed and used by Martin in [7][8]. In the following, we give a short review of his principles. At the top level of the taxonomy, instruments are divided into pizzicato and sustained. Second level comprises instrument families, and the bottom level individual instruments. Classification occurs at each node, applying knowledge of the best features to distinguish between possible subclasses. This way of processing is suggested to have some advantages over direct classification at the lowest end of the taxonomy, because the decision process may be simplified to take into account only the small number of possible subclasses.

Table 2: Classification results

	Hierarchy 1	**Hierarchy 2**	**No hierarchy**
Pizzicato / sustained	99.0%	99.0%	99.0%
Instrument families	93.0%	94.0%	94.7%
Individual instruments	74.9%	75.8%	80.6%

In our system, at each node a Gaussian or a k-NN classifier was used with a fixed set of features. The Gaussian classifier turned out to yield the best results at the highest level, where the number of classes is two. At the lower levels, k-NN classifier was used. Bad features are likely to decrease classifying performance, which makes evaluating the salience of each feature essential. The features used at a node were selected manually by monitoring feature values of possible subclasses. This was done one feature at a time, and only the features making clear distinction were included into the feature set of the node.

We implemented a classification hierarchy similar to that presented by Martin in [7], with the exception that his samples and taxonomy did not include piano. In our system the piano was assigned to an own family node because of having a unique set of some features, especially cepstral coefficients. According to Martin, classification performance was better if the reeds and the brass were first processed as one family and separated at the next stage. We wanted to test this with our own feature set and test data and tried the taxonomy with and without the Brass or Reeds node, which is marked with a '*' in Figure 3.

5. RESULTS

Our validation database consisted of 1498 solo tones covering the entire pitch ranges of 30 orchestral instruments with several articulation styles (e.g. pizzicato, martele, bowed, muted, flutter), as illustrated in Figure 3. All tones were from the McGill Master Samples collection [10], except the piano and quitar tones which were played by amateur musicians and recorded with a DAT recorder. In order to achieve comparable results to those described by Martin in [7], similar way of cross validation with 70% / 30% splits of train and test data was used. A difference to the method of Martin was to estimate the fundamental frequency of the test sample before classification, which was then compared to the pitch ranges of different instruments, taking only the possible ones into classification.

In Table 2, we present the classification results made in the three different ways. Hierarchy 1 is the taxonomy of Figure 6 without the Brass or Reeds node. In the No-hierarchy experiment classification was made separately for each classification level. The Hierarchy 2 proved out to yield slightly better results, like Martin reported in [7]. But interestingly, in our experiments, the direct classification at each level performed best at both tasks, which was not the case in Martin's experiments where the Hierarchy 2 yielded the best results. At the current implementation, classification result at the lower level of hierarchy is totally dependent on the results of the higher levels, and the error cumulates as the classification proceeds.

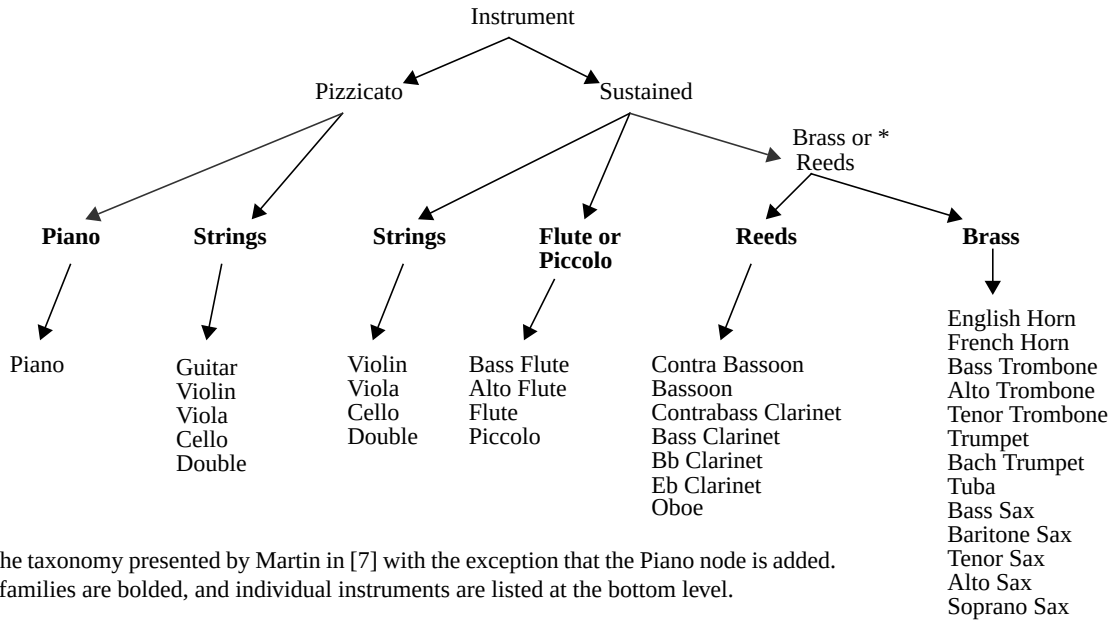

Figure 3. The taxonomy presented by Martin in [7] with the exception that the Piano node is added. Instrument families are bolded, and individual instruments are listed at the bottom level.

No significant advantage was achieved with hierarchical classification. Perhaps the biggest benefit of hierarchical approach would be got if more than one possible choises at each node were taken into account and the salience of the features was automatically evaluated. Classification of this kind has been used in [8].

The achieved performance both in instrument family and individual instrument classification was better than reported by Martin in [7]. His system's classification accuracies were approximately 90% in instrument family and 70% with individual instruments, while the data set consisted 1023 samples of 15 different instruments, being a subset of our data. Comparison to other systems is not reasonable because of the different amount of instruments or different method of performance evaluation used [1][5][8].

6. CONCLUSIONS

A system for musical instrument recognition was presented that uses a wide set of features to model the temporal and spectral charactreristics of sounds. Signal processing algorithms were designed to measure these features in acoustic signals. Using this input data, a classifier was constructed and the usefulness of the features was verified. Furthermore, experiments were carried out to investigate the potential advantage of a hierarchically structured classifier.

The achieved performance and comparison to earlier results demonstrates that combining the different types of features succeeded in capturing some extra knowlege about the instrument properties. Hierarchical structure could not bring further benefits, but its full potential should be reconsidered when a wider data set including more instruments, as well as different examples from a particular instrument class is available. Future work will concentrate on these areas, and on integrating the recognizer into a system that is able to process more complex sound mixtures.

7. REFERENCES

[1] Brown, J. C. "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features." J. Acoust. Soc. Am. 105(3) 1933-1941.

[2] De Poli, G. & Prandoni, P. "Sonological Models for Timbre Characterization". Journal of New Music Research, Vol 26 (1997), pp. 170-197, 1997.

[3] Grey, J. M. "Multidimensional perceptual scaling of musical timbres". J. Acoust. Soc. Am. 61(5), 1270-1277, 1977.

[4] Handel, S. "Timbre perception and auditory object identification". In Moore (ed.) Hearing. New York: Academic Press.

[5] Kaminskyj, I. & Materka, A. "Automatic source identification of monophonic musical instrument sounds". Proceedings of the 1995 IEEE International Conference of Neural Networks, pp. 189-194, 1995.

[6] Klapuri, A. "Pitch Estimation Using Multiple Independent Time-Frequency Windows". Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999.

[7] Martin, K. D. "Musical Instrument Identification: A Pattern Recognition Approach". Presented at the 136th meeting of the Acoustical Society of America, 1998.

[8] Martin, K. D. "Sound-Source Recognition: A Theory and Computational Model". Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[9] McAdams S., Winsberg S., Donnadieu S., De Soete G., Krimphoff, J. "Perceptual scaling of synthesized musical timbres: common dimensions, specifities, and latent subject classes". Psychological Research, 58, pp 177-192, 1995.

[10] Opolko, F. & Wapnick, J. "McGill University Master Samples" (compact disk). McGill University, 1987.

[11] Rabiner, L. R. & Juang, B. H. "Fundamentals of speech recognition". Prentice-Hall 1993.