COMPARISON OF FEATURES FOR MUSICAL INSTRUMENT RECOGNITION

Antti Eronen

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

ABSTRACT

Several features were compared with regard to recognition performance in a musical instrument recognition system. Both mel-frequency and linear prediction cepstral and delta cepstral coefficients were calculated. Linear prediction analysis was carried out both on a uniform and a warped frequency scale, and reflection coefficients were also used as features. The performance of earlier described features relating to the temporal development, modulation properties, brightness, and spectral synchronity of sounds was also analysed. The data base consisted of 5286 acoustic and synthetic solo tones from 29 different Western orchestral instruments, out of which 16 instruments were included in the test set. The best performance for solo tone recognition, 35% for individual instruments and 77% for families, was obtained with a feature set consisting of two sets of mel-frequency cepstral coefficients and a subset of the other analysed features. The confusions made by the system were analysed and compared to results reported in a human perception experiment.

1. INTRODUCTION

Automatic musical instrument recognition is a fascinating and essential subproblem in music indexing, retrieval, and automatic transcription. It is closely related to computational auditory scene analysis. However, musical instrument recognition has not received as much research interest as speaker recognition, for instance.

The implemented musical instrument recognition systems still have limited practical usability. Brown has reported a system that is able to recognize four woodwind instruments from monophonic recordings with a performance comparable to that of human's [1]. Martin's system recognized a wider set of instruments, although it did not perform as well as human subjects in a similar task [2].

This paper continues the work presented in [3] by using new cepstral features and introducing a significant extension to the evaluation data. The research focuses on comparing different features with regard of recognition accuracy in a solo tone recognition task. First, we analyse different cepstral features that are based either on linear prediction (LP) or filterbank analysis. Both conventional LP having uniform frequency resolution and more psychoacoustically motivated warped linear prediction (WLP) are used. WLP based features have not been used for musical instrument recognition before. Second, other features are analysed that are related to the temporal development, modulation properties, brightness, and spectral synchronity of sounds.

The evaluation database is extended to include several examples of a particular instrument. Both acoustic and synthetic isolated notes of 16 Western orchestral instruments are used for testing, whereas the training data includes examples of 29 instruments. The performance of the system and the confusions it makes are compared to the results reported in a human perception experiment, which used a subset of the same data as stimuli [2].

2. FEATURE EXTRACTION

2.1. Cepstral features

For isolated musical tones, the onset has been found to be important for recognition by human subjects [4]. Motivated by this, the cepstral analyses are made separately for the onset and steady state segments of a tone. Based on the root mean square (RMS) -energy level of the signal, each tone is segmented into onset and steady state segments. The steady state begins when the signal achieves its average RMS-energy level for the first time, and the onset segment is the 10 dB rise before this point.

For the onset portion of tones, both LP and filterbank analyses were performed in approximately 20 ms length hamming windowed frames with 25% overlap. In the steady state segment, frame length of 40 ms was used. If the onset was shorter than 80 ms, the beginning of steady state was moved forward so that at least 80 ms was analysed. Prior to the analyses, each acoustic signal was preemphasized with the high pass filter $1, -0.97z^{-1}$ to flatten the spectrum.

The LP coefficients were obtained from an all-pole approximation of the windowed waveform, and were computed using the autocorrelation method. In the calculation of the WLP coefficients, the frequency warping transformation was obtained by replacing the unit delays of the predicting filter with first-order all-pass elements. In the *z*-domain this can be interpreted by the mapping

$$z^{-1} \rightarrow \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} . \qquad (1)$$

In the implementation this means replacing the autocorrelation network with a warped autocorrelation network [5]. The parameter λ is selected in such a way that the resulting frequency mapping approximates the desired frequency scale. By selecting λ=0.7564 for 44.1 kHz samples, a Bark scale approximation was obtained [6]. Finally, the obtained linear prediction coefficients a_n are transformed into cepstral coefficients c_n with the recursion [7, pp. 163]

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} . \qquad (2)$$

The number of cepstral coefficients was equal to the analysis order after the zeroth coefficient, which is a function of the channel gain, was discarded.

For the mel-frequency cepstral coefficient (MFCC) calculations, a discrete Fourier transform was first calculated for the windowed waveform. The length of the transform was 1024 or 2048 point for 20 ms and 40 ms frames, respectively. 40 triangular bandpass filters having equal bandwith on the mel-frequency scale were simulated, and the MFCCs were calculated from the log filterbank amplitudes using a shifted discrete cosine transform [7, p.189].

In all cases, the median values of cepstral coefficients were stored for the onset and steady state segments. Delta cepstral coefficients were calculated by fitting a first order polynomial over the cepstral trajectories. For the delta-cepstral coefficients, the median of their absolute value was calculated. We also experimented with coefficient standard deviations in the case of the MFCCs.

2.2. Spectral and temporal features

Calculation of the other features analysed in this study has been described in [3] and will be only shortly summarized here.

Amplitude envelope contains information e.g. about the type of excitation; i.e. whether a violin has been bowed or plucked. Tight coupling between the excitation and the resonance structure is indicated by a short onset duration. To measure the slope of the amplitude decay after the onset, a line was fitted over the amplitude envelope on a dB scale. Also, the mean square error of the fit was used as a feature. Crest factor, i.e. maximum / RMS value was also used to characterize the shape of the amplitude envelope.

Strength and frequency of amplitude modulation (AM) was measured at two frequency ranges: from 4-8 Hz to measure tremolo, i.e. AM in conjunction with vibrato, and 10-40 Hz for graininess or roughness of tones.

Spectral centroid (SC) corresponds to perceived brightness and has been one of the interpretations for the dissimilarity ratings in many multidimensional scaling studies [4]. SC was calculated from a short time power spectrum of the signal using logarithmic frequency resolution. The normalized value of SC is the absolute value in Hz divided by the fundamental frequency. The mean, maximum and standard deviation values of SC were used as features.

Onset asynchrony refers to the differences in the rate of the energy development of different frequency components. A sinusoid envelope representation was used to calculate the intensity envelopes for different harmonics, and the standard deviation of onset durations for different harmonics was used as a one feature. Another feature measuring this property is obtained by fitting the intensity envelopes of individual harmonics into the overall intensity evelope during the onset period, and the average mean square error of those fits was used as a feature.

Fundamental frequency (f0) of tones is measured using the algorithm from [8], and used as a feature. Also, its standard deviation was used as measure for vibrato.

3. EXPERIMENTAL SETUP

Samples from five different sources were included in the validation database. First, the samples used in [3] consisted of the samples from the McGill University Master Samples Collection (MUMS) [9], as well as recordings of an acoustic guitar made at Tampere University of Technology. The other sources of samples were the University of Iowa website, IRCAM Studio Online (SOL), and a Roland XP-30 synthesizer. The MUMS and SOL samples are recorded in studios with different acoustic characteristics and recording equipment, and the samples from Iowa University are recorded in an anechoic chamber. The samples from the Roland synthesizer were played on the keyboard and recorded through analog lines into a Silicon Graphics Octane workstation. The synthesizer has a dynamic keyboard, thus these samples have varying dynamics. The samples from SOL include only the first 1.5 seconds of the played note.

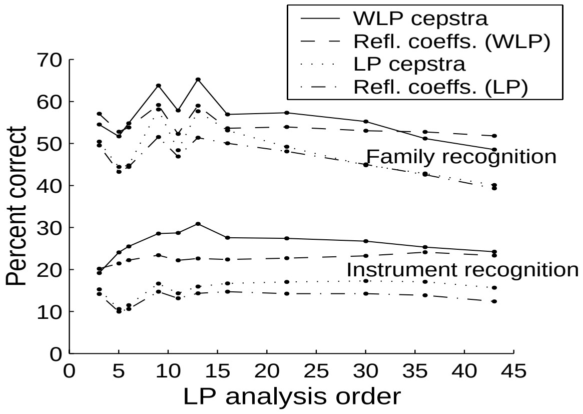

Figure 1. *Classification performance as a function of analysis order for different LP based features.*

Cross validation aimed at as realistic conditions as possible with this data set. On each trial, the training data consisted of all the samples except those of the particular performer and instrument being tested. In this way, the training data is maximally utilized, but the system has never heard the samples from that particular instrument in those circumstances before. There were 16 instruments that had at least three independent recordings, so these instruments were used for testing. The instruments can be seen in Figure 4. A total of 5286 samples of 29 Western orchestral instruments were included in the data set, out of which 3337 samples were used for testing. The classifier made its choice among the 29 instruments. In these tests, a random guesser would score 3.5% in the individual instrument recognition task, and 16.7% in family classification.

In each test, classifications were performed separately for the instrument family and individual instrument cases. A *k*-nearest neighbours (kNN) classifier was used, where the values of *k* were 11 for instrument family and for 5 individual instrument classification. The distance metric was Mahalanobis with equal covariance matrix for all classes, which was implemented by using the discrete form of the Karhunen-Loeve transform to uncorrelate the features and normalize the variances, and then by using the euclidean distance metric in the normalized space.

4. RESULTS

Different orders of the linear prediction filter were used to see the effect of that on the performance of several LP and WLP based features. The results for instrument family and individual instrument recognition are shown in Figure 1. The feature vector at all points consisted of two sets of coefficients: medians over the onset period and medians over the steady state. The optimal analysis order was between 9 and 14, above and below which per-

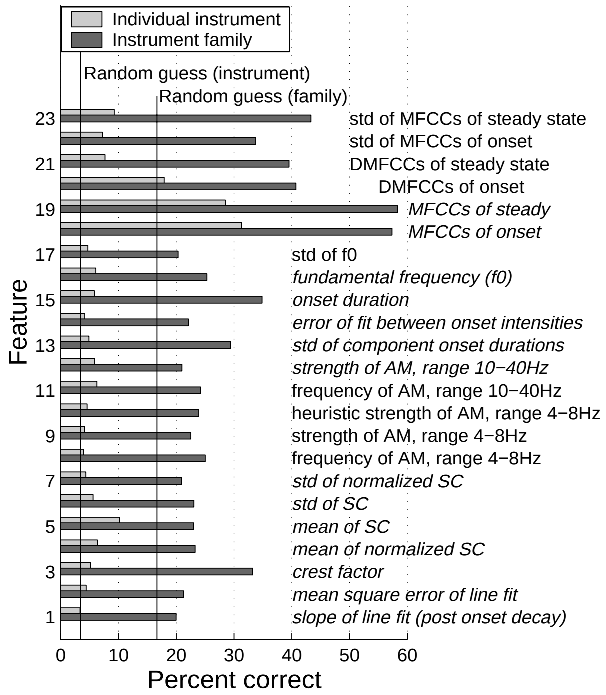

Figure 2. *Classification performance as a function of features. The features printed in italics were included in the best performing configuration.*

formance degrades. The number of cepstral coefficients was one less than the analysis order. WLP cepstral and reflection coefficients outperformed LP cepstral and reflection coefficients at all analysis orders calculated. The best accuracy with LP based features was 33% for individual instruments (66% for instrument families), and was obtained with WLP cepstral coefficients (WLPCC) of order 13.

In Figure 2, the classification accuracy is presented as a function of features. The cepstral parameters are mel-frequency cepstral coefficients or their derivatives. The optimal number of MFCCs was 12, above and below which the performance slowly degraded. However, optimization of the filter bank parameters should be done for the MFCCs, but was left for future research. By using the MFCCs both from the onset and steady state, the accuracies were 32% (69%). Because of computational cost considerations the MFCC were selected as the cepstrum features for the remaining experiments. Adding the mel-frequency delta cepstrum coefficients (DMFCC) slightly improved the performance, using the MFCCs and DMFCCs of the steady state resulted in 34% (72%) accuracy.

The other features did not alone prove out very successful. Onset duration was the most successful with 35% accuracy in instrument family classification. In individual instruments, spectral centroid gave the best accuracy, 10%. Both were clearly inferior to the MFCCs and DMFCCs. It should be noted, however, that the MFCC features are vectors of coefficients, and the other features consist of a single number each.

The best accuracy 35% (77%) was obtained by using a feature vector consisting of the features printed in italics in Figure 2. The feature set was found by using a subset of the data and a simple backward select algorithm. If the MFCCs were replaced with order 13 WLPCCs, the accuracy was 35% (72%).

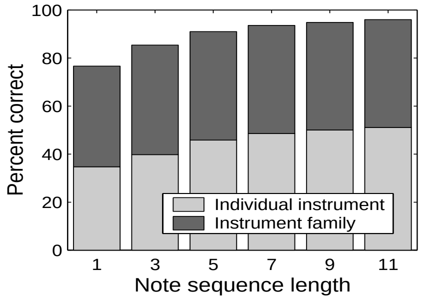

Figure 3. *Classification performance as a function of note sequence length.*

In practical situations, a recognition system is likely to have more than one note to use for classification. A simulation was made to test the system's behaviour in this situation. Random sequences of notes were generated and each note was classified individually. The final classification result was pooled across the sequence by using the majority rule. The recognition accuracies were averaged over 50 runs for each instrument and note sequence length. Figure 3 shows the average accuracies for individual instrument and family classification. With 11 random notes, the average accuracy increased to 51% (96%). In instrument family classification, the recognition accuracy for the tenor saxophone was the worst (55% with 11 notes), whereas the accuracy for the all other instruments was over 90%. In the case of individual instruments, the accuracy for the tenor trombone, tuba, cello, violin, viola and guitar was poorer than with one note, the accuracy for the other instruments was higher.

The recognition accuracy depends on the recording circumstances, as may be expected. The individual instrument recognition accuracies were 32%, 87%, 21% and 37% for the samples from MUMS, Iowa, Roland and SOL sources, respectively. The Iowa samples included only the woodwinds and the French horn, which were on the average recognized with 49% accuracy. Thus, the recognition accuracy is clearly better for the Iowa samples recorded in an anechoic chamber. The samples from the other three sources are comparable with the exception that the samples from SOL did not include tenor or soprano sax. With synthesized samples the performance is clearly worse, which is probably due to both the varying quality of the synthetic tones and the varying dynamics.

5. DISCUSSION

The confusion matrix for the feature set giving the best accuracy is presented in Figure 4. There are large differences in the recognition accuracies of different instruments. The soprano sax is recognized correctly in 72% of the cases, while the classification accuracies for the violin and guitar are only 4%. French horn is the most common target for misclassifications.

It is interesting to compare the behaviour of the system to human subjects. Martin [2] has reported a listening experiment where fourteen subjects recognized 137 samples from the McGill collection, a subset of the data used in our evaluations. The differences in the instrument sets are small, Martin's samples did not

Responded Presented	French horn	Trumpet	Bach trumpet	Bass trombone	Tenor trombone	Alto trombone	Tuba	Bass sax	Baritone sax	Tenor sax	Alto sax	Soprano sax	English horn	Oboe	Contrabass clar.	Bass clarinet	E-flat clarinet	B-flat clarinet	Contrabassoon	Bassoon	Bass flute	Alto flute	Flute	Piccolo	Double bass	Cello	Violin	Viola	Guitar
French horn	**50**	3		2	12		18					1								8			1		5	1			
Trumpet	8	**23**	7		24		2			11		2		2				3					5	1	3	1	4	4	1
Tenor tromb.	31	17		24	**10**	6	6													5			1						
Tuba	76			8	4		**7**													3					2				
Tenor sax	6	2	2	2	9				15	**22**	2	6							4	2						7	6	17	
Alto sax		8			1				1		**64**	5	2	1			3	1		1					2		1	12	
Soprano sax	4	3			4						2	**72**		2				5					10						
Oboe	3	7			1						1	6	3	**68**				3					3	2				3	
B-flat clar.	6	4			1		1			2	11	16		4		1	17	**30**		1			5			1	1	3	
Bassoon	16	1			3		1				1								1	**70**			3		1				
Flute	1	1	8		6	2			1		4	1	1	1				2		3	1	4	**59**	2	1	1		2	
Double bass	2	1									2														**56**	31	2	5	
Cello	1									1	4												1		31	**30**	5	28	
Violin	1	1	2								3	3	1					2					4	1	3	8	**4**	67	
Viola				1						2	4	1		1			1	1							6	25	45	**13**	
Guitar	2	8			1		1	1			2	1													43	38	1	1	**4**

Figure 4. *Confusion matrix for the best performing feature set. Entries are expressed as percentages and are rounded to the nearest integer. The boxes indicate instrument families.*

include any sax or guitar samples, but had the piccolo and the English horn, which were not present in our test data. In his test, the subjects recognized the individual instrument correctly in 45.9% of cases (91.7% for instrument families). Our system made more outside family confusions than the subjects in Martin's test. It was not able to generalize into more abstract instrument families as well as humans, which was also the case in Martin's computer simulations [2]. In individual instrument classification, the difference is perhaps smaller.

The within-family confusions made by the system are quite similar to the confusions made by humans. Examples include the French horn as tenor trombone and vice versa, tuba as French horn, or B-flat clarinet as E-flat clarinet. The confusions between the viola and the violin, and the cello and the double bass were also common to both humans and our system. In the confusions occurring outside the instrument family, confusions of the B-flat clarinet as soprano or alto sax were common to both our system and the subjects.

6. CONCLUSIONS

Warped linear prediction based features proved to be successful in the automatic recognition of musical instrument solo tones, and resulted in better accuracy than what was obtained with corresponding conventional LP based features. The mel-frequency cepstral coefficients gave the best accuracy in instrument family classification, and would be the selection also for the sake of computational complexity. The best overall accuracy was obtained by augmenting the mel-cepstral coefficients with features describing the type of excitation, brightness, modulations, synchronity and fundamental frequency of tones.

Care should be taken while interpreting the presented results on the accuracy obtained with different features. First, the best set of features for musical instrument recognition depends on the context [2,4]. Second, the extraction algorithms for features other than cepstral coefficients are still in their early stages of development. However, since the accuracy improved when cepstral features were added with other features, this approach should be further developed.

7. ACKNOWLEDGEMENT

The available samples in the web by the University of Iowa (http://theremin.music.uiowa.edu/~web/) and IRCAM (http://soleil.ircam.fr/) helped greatly in collecting our database. The warping toolbox by Härmä and Karjalainen (http://www.acoustics.hut.fi/software/warp/) was used for the calculation of WLP based features. Our MFCC analysis was based on Slaney's implementation (http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/).

8. REFERENCES

[1] Brown, J. C. "Feature dependence in the automatic identification of musical woodwind instruments." *J. Acoust. Soc. Am.*, Vol. 109, No. 3, pp. 1064-1072, 2001.

[2] Martin, K. D. *Sound-Source Recognition: A Theory and Computational Model.* Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999. Available at: http://sound.media.mit.edu/Papers/kdm-phdthesis.pdf.

[3] Eronen, A. & Klapuri, A. "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features". *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, June 5-9, 2000.

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

[5] Härmä, A. et.al. "Frequency-Warped Signal Processing for Audio Applications". *J. Audio Eng. Soc.*, Vol. 48, No. 11, pp. 1011-1031, 2000.

[6] Smith, J. O. & Abel, J. S. "Bark and ERB Bilinear Transforms". *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 6, pp. 697-708, 1999.

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

[8] 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.

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