INSTRUMENT RECOGNITION IN ACCOMPANIED SONATAS AND CONCERTOS

Jana Eggink and Guy J. Brown

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
Email: {j.eggink, g.brown}@dcs.shef.ac.uk

ABSTRACT

A system for musical instrument recognition is introduced. In contrast to most existing systems, it can identify a solo instrument even in the presence of an accompanying keyboard instrument or orchestra. To enable recognition in the presence of a highly polyphonic background, we use features based solely on the partials of the target tone. The approach is based on the assumption that it is possible to extract the most prominent fundamental frequency and the corresponding harmonic overtone series, and that these will most often belong to the solo instrument. Classification is carried out using a Gaussian classifier trained on examples of monophonic music. Testing our system on accompanied sonatas and concertos we achieved a recognition rate of 86% for 5 different instruments, an accuracy comparable to that of systems limited to monophonic music only.

1. INTRODUCTION

Interest in automatic music processing, and especially in information extraction from audio files, has grown significantly in recent years. In this paper we focus on the problem of instrument recognition from audio data. To know what instruments play in a musical recording can be useful for tasks related to automatic music transcription, automatic indexing and music analysis. Similarly in the growing field of musical information retrieval, knowledge about the instruments playing could lead to better results when searching for similar pieces of music (for example, different pieces played by a flute might be perceived as more similar to each other than a piece played by a cello). In a 'query-by-humming' context it would allow the user to specify a query which not only searches for a specific tune, but also specifies the instrument on which the tune should be played.

Previous work in automatic identification of musical instruments has mainly focused on monophonic recordings (e.g. [1], [10], [11]). While good results have been achieved for both isolated tones and recordings from commercially available compact discs (CDs), these studies assume that only one instrument is present at any moment in time. Only very few researchers have attempted instrument recognition in polyphonic music (e.g. [3], [7], [8]). These systems were only tested with a restricted number of simultaneous notes, typically 2 or 3, and relied on identifying the fundamental frequency (F0) of every tone. Duets for two melody instruments do exist in a classical repertoire, but most music is highly polyphonic; even sonatas for solo instrument are in the vast majority of cases accompanied by a keyboard instrument. Identifying all F0s in piano music is still an ongoing challenge, and while some good results have been achieved (e.g. [13]), it is not realistic to build a system relying on the identification of all F0s in a typical accompaniment.

In the present study we focus on the problem of instrument identification for solo instruments accompanied by a keyboard instrument (piano or cembalo) or a full orchestra.The advantage in trying to identify only the solo instrument is that it is normally played louder than the accompaniment, and the corresponding harmonic series is likely to stand out. The identification of the F0 of the most prominent tone is an easier task than identifying all other F0s, and some good results have been achieved when extracting melody lines from complex music [4].

Since the harmonics of the accompanying instrument(s) can span all frequency regions, cepstral features or features related to the energy in bandpass filters are unlikely to be able to distinguish between the target sound, i.e. the solo instrument, and the background accompaniment. We therefore use acoustic features based solely on the F0 and harmonics of an instrument sound.

2. SYSTEM DESCRIPTION

2.1. System Overview

The aim of the present study is to identify the solo instrument in an accompanied sonata or concerto. No attempt is made to decide whether the solo instrument is actually present at a particular moment, or to identify the type of accompaniment. Classification decisions are therefore made for entire sound files, without trying to identify instruments on a note by note level.

All processing is based on short, overlapping time frames of fixed length. For each frame, spectral peaks are extracted and the most prominent F0 is determined. The peaks belonging to the harmonic series of the estimated F0 form the basic features to be used by a Gaussian classifier, which identifies the instrument (Figure 1).

2.2. F0 Estimation

If necessary, the audio file is first converted to mono by mixing both channels of the stereo signal. Since most of the examples used are from commercially available CDs, they have a sampling rate of 44100 Hz, which is retained. The audio file is then divided into short frames of fixed length, with 50% overlap between successive frames. Every frame is multiplied with a Hanning window and a fast Fourier transform (FFT) is computed. To obtain a better frequency resolution, a highly zero-padded FFT (FFT size 16384) is used. Further processing is based on spectral peaks only. To locate peaks, the spectrum is convolved with a differentiated, 50 samples wide Gaussian. As a result the spectrum is smoothed and

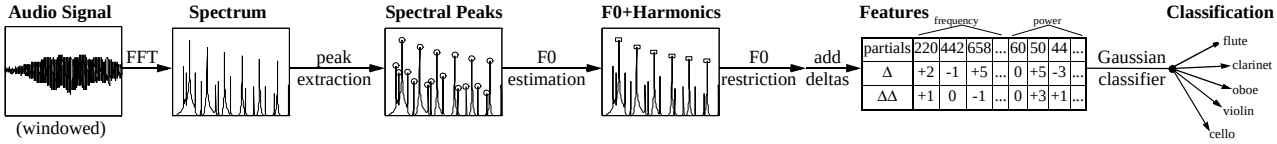

Figure 1: Schematic of the instrument classification system.

peaks are indicated by zero crossings in the convolution, which are easy to detect. Once the position of a peak is found, its frequency is defined by the frequency of the corresponding FFT bin.

The F0 estimation uses an approach based on a frequency-domain pattern matching technique, related to the so-called harmonic sieve [14]. A sieve consists of slots representing an ideal harmonic series, with one sieve for every possible F0. Each sieve is matched against the pattern of spectral peaks, and the more peaks coincide with the slots in a sieve, the more likely is the F0 represented by that particular sieve.

A common problem with this approach is octave confusion. A sieve half the frequency of the true F0 still captures all spectral peaks, with odd frequency slots left unmatched. A sieve double the true F0 has all its slots filled, but misses every other spectral peak. To solve this problem, a final score based on the power of the matched spectral peaks and their position within the sieve is computed. Generally, a F0 is more likely the more peaks it can account for. Additionally, lower harmonics are more powerful in most musical instrument tones, therefore a sieve with the strongest peak in the position of the first or second harmonic should be favoured over one with the highest peak in the position of e.g. the 10th harmonic. Hence, we use a weighting function based on a slowly decaying exponential, a heuristically determined measure that worked well in a number of preliminary studies. The power of the peaks allocated to a sieve is multiplied with the weighting function and the results are summed. Specifically, a peak corresponding to harmonic n (n=1 is the F0, n=2 the first overtone etc.) of a sieve has weighting $w(n)$ given by:

$$w(n) = \exp(-0.2 \times n) \quad (1)$$

The sieve that maximizes the sum of the weighted power of all its allocated peaks is taken as the most prominent F0 in that time frame.

The finer the spacing of the sieves on the frequency axis, the more exactly the F0 can be estimated. For musical purposes, a spacing corresponding to halftones is often enough, and saves significant computing time compared to a finer resolution. Our sieves are based on the F0s of all notes between C2 and C7, resulting in 61 sieves equally spaced on a logarithmic scale between 65 Hz and 2093 Hz. Since mistunings resulting in tones not coinciding with the F0 of the sieves are always possible, the frequency interval in which a spectral peak is considered to be a match to the slot of a sieve is relatively broad (a quartertone), so that all spectral peaks can be matched to a slot of at least one sieve.

2.3. F0 Restriction

Preliminary experiments using solo instruments with accompaniment revealed a common problem with the F0 estimation algorithm. Especially for woodwinds, F0 estimates were often below the range of the solo instrument. These F0 were either erroneous or corresponded to an accompanying instrument, the latter being inevitable in sections where the solo instrument is silent. Since the classifier was only trained to recognize solo instruments, but not different forms of accompaniment, these low F0s could not be used. A fixed frequency threshold, below which all F0 estimates are ignored, does not work because of the different pitch range of solo instruments; a piece for cello might only contain very low F0s. We therefore retained only the highest 50% of all estimated F0s, a heuristic rule that worked well with all instruments.

2.4. Acoustic Features

To allow recognition in the presence of a highly polyphonic background, we use features based solely on harmonics. The energy of instrument sounds is concentrated in their harmonics, which are evident as spectral peaks. If the solo instrument is louder than the accompaniment, as is common for classical sonatas and concerts, these peaks are likely to stand out in a spectral representation. While some information will be lost, e.g. inharmonic noise caused by the excitation method (air blowing against a hole, friction of a bow against strings), we expect peak-based features to be a more robust encoding of the solo instrument in the presence of background accompaniment.

Specifically, we use features based on the first 15 partials of the most dominant F0 at a given moment in time. Each feature vector contains 90 elements, including the frequency location and the normalised, log-compressed power of the spectral peaks corresponding to the lowest 15 partials. Added to these basic features are frame to frame differences (deltas) and differences of differences (delta-deltas) of both frequency and power within individual tones of a continuous F0. Including the exact frequency location of the partials allows instrument specific deviations from an ideal harmonic series to be coded. The relative power of the partials is likely to be the main source of information to discriminate among the instruments, being closely related to the perception of timbre. Delta and delta-deltas code spectral and amplitudes changes within a tone, most prominent in expressive gestures such as vibrato.

2.5. Classifier

Gaussian mixture models (GMMs) have been successfully employed in various classification tasks, including instrument recognition (e.g. [1], [3], [10]). A GMM models the probability density function (*pdf*) of an observed feature vector x by a multivariate Gaussian mixture density:

$$pdf(x) = \sum_{i=1}^{N} p_i \Phi_i(x, \mu_i, \Sigma_i) \quad (2)$$

Each of the N individual Gaussian densities Φ_i (centres) is characterized by its mean μ_i, covariance matrix Σ_i and mixing

coefficient p_i. Means, covariances and mixing coefficients are estimated during training. The data points of the training material are initially clustered using a k-means algorithm with random initialisation; final parameter values are then estimated using the expectation-maximisation (EM) algorithm [2].

To make the models as robust as possible, they were trained on different isolated note sample collections and approximately one minute each of 4 to 5 different monophonic recordings per instrument, taken from commercially available CDs. The F0s for isolated notes were known beforehand; for the monophonic recordings F0s were estimated by the system.

One model was trained for every possible F0 of every instrument considered (flute, clarinet, oboe, violin, cello), resulting in 210 different models. The alternative to training F0-dependent models would be to train one model for every instrument, which would result in only 5 models. However, incorporating F0 dependency can be advantageous for instrument classification, as the distribution of some features changes with the F0 [9]. Even a large number of F0-dependent models can be trained efficiently, since each model converges within very few iterations; in contrast, models trained over the whole pitch range of an instrument converge slowly. The recognition phase was also efficient, since models were restricted to those trained on the F0 detected at that point in time.

3. EVALUATION

The system was assessed on isolated monophonic samples (taken from the Iowa Musical Instrument Samples [5], the Ircam Studio Online [6], and the McGill University Master Samples [12]), on realistic phrases played by a single instrument, and most importantly on music played by a solo instrument accompanied by a keyboard instrument or a full orchestra. Both training and test examples spanned a wide stylistic range from baroque to 20th century music; no style-dependent differences could be detected in the results. Training and test material were always taken from different recordings. Classification decisions were made for each frame independently and the instrument which accumulated the most 'wins' over the duration of an audio file was taken as the overall classification for that example.

3.1. Parameter Estimation

A number of free parameters can influence the performance of the system. The first choice to be made concerns the window length for the initial signal segmentation. Longer windows allow a more accurate frequency resolution, but are more likely to include changing F0s. We tested window lengths of 1024 (23 ms) and 2048 samples (46 ms). All other parameters being the same, a shorter window length improved results by approximately 5% for isolated tones and 10% for realistic examples of accompanied solo instruments. For subsequent experiments, we therefore chose a window length of 23 ms.

Another choice concerns the covariance matrices of the GMMs. The matrices can be diagonal, assuming feature independence, or contain full covariances modelling the dependencies between the individual entries in the feature vector. Diagonal covariance matrices are more commonly used, as they are computationally less expensive (e.g. [1], [3], [10]). Testing our system using low numbers of centres (1 to 4), GMM-classifiers with full covariance matrices outperformed those with diagonal ones by approximately 10% to 20%. Recognition accuracy for classifiers using diagonal covariance matrices improved with the use of more centres, and might equal those using full covariance matrices with a sufficiently high number of centres. But with an increasing number of centres, more training iterations were needed for convergence. We therefore decided to use models with full covariance matrices. Only 1 centre per model was used, as a larger number of centres did not improve recognition accuracy. This indicates a very uniform distribution of features for each model, which is probably due to the fact that one model was trained for every F0. The classifiers used for further evaluation were therefore simple Gaussian classifiers, which are a limiting case of equation (2) for $N = 1$ and $p_1 = 1$.

3.2. Isolated Samples with Known F0

To asses the performance of the system without the influence of the F0 detector, we used isolated samples with a known F0. In a leave-one-out cross-validation scheme [2], models were trained using the same realistic examples but on only 2 sample collections, leaving the third one for testing. With classification decisions made for each tone independently, average recognition accuracy was 71% for both the McGill and the Ircam samples and 59% for the Iowa sample collection. A confusion matrix averaging across the 3 conditions is shown in Table 1.

stimulus \ response	Flute	Clarinet	Oboe	Violin	Cello
Flute	**76%**	9%	3%	12%	1%
Clarinet	16%	**64%**	9%	8%	2%
Oboe	6%	16%	**57%**	13%	7%
Violin	5%	1%	5%	**71%**	18%
Cello	3%	2%	7%	21%	**68%**

Table 1: Confusion matrix for instrument recognition of isolated notes.

3.3. Realistic Monophonic Phrases

As a next step towards a more realistic performance we tested the system on solo music without accompaniment. For every instrument, 5 short examples (2-10 sec) were taken from different recordings which were not used during training. The F0s were estimated by the system, introducing an additional possibility for errors. Overall recognition accuracy was 84%, both when all F0s were used and when they were restricted to the highest 50% (see section 2.3.).

3.4. Solo Instruments with Accompaniment

The main focus of the current application is to identify the solo instrument in accompanied sonatas and concertos. Overall 90 different examples were used for testing. They were taken from 8 different recordings per instrument, which were not used during training. Approximately half of the pieces were sonatas for solo instrument with piano or cembalo accompaniment, the other half consisted of concertos for solo instrument and orchestra. Since longer passages without the presence of the solo instrument are

quite common, mainly complete movements (semi-independent parts of a longer composition) were taken as examples to ensure a sufficient presence of the solo instrument. To limit computation time, only the first 3 minutes were taken from very long movements.

Overall classification accuracy in this task was 86%; a confusion matrix is shown in Table 2. Accuracy was slightly higher for sonatas (91%) than for concertos (80%), where a full orchestra plays in the background. This trend was true for 4 out of 5 instruments, with the oboe being the exception, but it is premature to say if the difference in accuracy is truly caused by the form of accompaniment. An orchestra accompaniment could be more difficult, because all instruments are in fact present and short solo passages for other than the nominal solo instrument are quite common.

response / stimulus	Flute	Clarinet	Oboe	Violin	Cello
Flute	**75%**	0%	0%	25%	0%
Clarinet	6%	**88%**	0%	6%	0%
Oboe	0%	0%	**82%**	18%	0%
Violin	0%	0%	0%	**88%**	12%
Cello	0%	0%	0%	6%	**94%**

Table 2: Confusion matrix for recognition of the solo instrument in accompanied sonatas and concertos.

4. CONCLUSIONS AND FUTURE WORK

The recognition results we obtained are very encouraging: there is no drop in performance between monophonic examples and music with accompaniment. Our initial assumption that a peak-based representation is robust against a highly polyphonic background is strongly supported by the results. The chosen representation based solely on the F0 and harmonics of an instrument tone also holds sufficient information to distinguish reliably among the instruments. Results for isolated samples are lower than for longer, realistic examples. This seems to be a common phenomenon ([3], [11]) and is probably due to the fact that phrases are longer and more varied, so that isolated, random errors are evened out.

The recognition accuracy of 86% for 5 different instruments achieved by our system is comparable with other approaches tested on realistic monophonic phrases, e.g. 82% for 6 instruments in [11], 70% for 8 instruments in [10], and up to 80% for 4 woodwind instruments in [1]. The main advantage of our system is that it does not assume a monophonic signal, but achieves equally high recognition accuracies even when the instrument is accompanied by a keyboard instrument or a complete orchestra.

Future work will concentrate on formally evaluating and improving the system for F0 estimation, which is crucial for the performance of the system. A first estimate of the solo instrument might be used in an iterative cycle to improve the F0 estimation, as specific assumptions about the distribution of energy between the partials can be made. This could result in a more reliable melody extraction, which would not only allow for a more accurate instrument identification, but would in itself be a useful achievement for various automatic music processing tasks.

For a realistic application more instruments have to be included, as for example both brass instruments and pianos are common solo instruments in concertos. The latter poses additional problems, as the piano is itself highly polyphonic, so that the assumption about a single dominant F0 might not hold true. In addition it would be desirable to include the singing voice as an 'instrument', and perhaps even to identify the singer. This would be especially useful when processing popular music, which is often highly dominated by a vocalist.

ACKNOWLEDGMENTS

Jana Eggink acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00276, HOARSE. Guy J. Brown is supported by EPSRC grant GR/R47400/01 and the MOSART IHP network.

REFERENCES

[1] Brown, J.C., Houix, O. & McAdams, S., "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Am.* 109, pp. 1064-1072, 2001

[2] Bishop, C.M., *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1995

[3] Eggink, J. & Brown, G.J., "A missing feature approach to instrument identification in polyphonic music," *Proc. ICASSP'03*, pp. 553-556, 2003

[4] Goto, M., "A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models," *Proc. ICASSP'01*, pp. 3365-3368, 2001

[5] *Iowa Musical Instrument Samples*, `http://theremin.music.uiowa.edu`

[6] *Ircam Studio Online (SOL)*, `http://www.ircam.fr`

[7] Kashino, K. & Murase, H., "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Comm.* 27, pp. 337-349, 1999

[8] Kinoshita, T., Sakai, S. & Tanaka, H., "Musical sound source identification based on frequency component adaptation," *Proc. IJCAI-99 Workshop on Computational Auditory Scene Analysis*, Stockholm, Sweden, 1999

[9] Kitahara, T., Goto, M. & Okuno, H.G., "Musical instrument identification based on F0-dependent multivariate normal distribution," *Proc. ICASSP'03*, pp. 421-424, 2003

[10] Marques, J. & Moreno, P., "A study of musical instrument classification using Gaussian mixture models and support vector machines," *Cambridge Research Laboratory Technical Report Series CRL/4*, 1999

[11] Martin, K., *Sound-source recognition: A theory and computational model*. PhD Thesis, MIT, 1999

[12] Opolko, F. & Wapnick, J., *McGill University Master Samples* (CD), Montreal, Quebec: McGill University, 1987

[13] Raphael, C., "Automatic transcription of piano music," *Proc. ISMIR'02*, 2002

[14] Scheffers, M.T.M., *Sifting vowels - auditory pitch analysis and sound segregation*. PhD Thesis, University of Groningen, 1983