A MISSING FEATURE APPROACH TO INSTRUMENT IDENTIFICATION IN POLYPHONIC MUSIC

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

Gaussian mixture model (GMM) classifiers have been shown to give good instrument recognition performance for monophonic music played by a single instrument. However, many applications (such as automatic music transcription) require instrument identification from polyphonic, multi-instrumental recordings. We address this problem by incorporating ideas from missing feature theory into a GMM classifier. Specifically, frequency regions that are dominated by energy from an interfering tone are marked as unreliable and excluded from the classification process. This approach has been evaluated on random two-tone chords and an excerpt from a commercially available compact disc, with promising results.

1. INTRODUCTION

Automatic transcription of music remains a challenging problem in digital audio processing. In order to derive a musical score from an acoustic waveform, the fundamental frequency (F0), onset time and offset time of each tone must be extracted, together with the identities of the instruments on which these tones were played. Although techniques exist for instrument recognition from monophonic audio recordings (e.g., see [1]), these cannot be directly applied to polyphonic music in which tones from different instruments overlap in time. Here, we describe an approach to this problem based on missing feature theory, which has been successfully applied in the fields of robust speech recognition [2] and speaker identification [3].

Only a few previous studies have attempted instrument recognition from multi-instrumental music. Kashino and Murase [4] describe an approach based on time-domain waveform templates and adaptive filtering. For three instruments (flute, violin and piano) they achieved classification performance of about 75% for specially-arranged ensemble recordings. However, the F0 and onset time of each note were supplied to their system. Work by Kinoshita et al. [5] proposed a frequency-domain approach, which used features related to the sharpness of onsets and the spectral distribution of partials. F0s were extracted prior to the recognition process. If it was estimated that partials from more than one tone contributed to a frequency component, the resulting feature value was either completely excluded from the recognition process, or was only used after an estimated mean value for the first identified instrument was subtracted. With this technique they achieved 66%-75% correct classification for random two-tone combinations (73%-81% correct if the true F0s were provided), again using three different instruments (clarinet, violin and piano).

Few workers have reported the performance of their algorithms on naturalistic sound recordings, such as commercially available compact discs (CDs). Brown et al. [1] have attempted instrument identification from such recordings, although their system is limited to monophonic music. They describe a classifier based on Gaussian mixture models (GMMs), which were trained with a range of features including cepstra. Training and test material were taken from different CD recordings of solo music. They achieved an average instrument classification performance of about 60% correct for four different woodwinds, which was comparable with human performance. Similar but slightly lower results have been reported by Martin [7], who used a wider range of instruments and a number of different features in a hierarchical classification scheme, and by Marques and Moreno [6], who compared techniques based on GMMs and support vector machines.

2. SYSTEM DESCRIPTION

The aim of the current study is to identify the instruments present in CD recordings of polyphonic, multi-instrumental music. Since tones from different instruments may overlap in time, a missing feature approach is employed in which frequency regions that are thought to be dominated by energy from a non-target tone are marked as unreliable and excluded from the recognition process. This idea is motivated by a model of auditory perception which proposes that listeners are able to recognise partially masked sounds from an incomplete acoustic representation [2]. In polyphonic music, harmonics of one tone often overlap with those of another tone. As a consequence, the energy values of these partials no longer correspond to those of either instrument, and conventional instrument recognition techniques will fail.

A schematic view of our system is shown in Figure 1. The first stage is a frequency analysis of the sampled audio signal. Subsequently, the F0s of all tones are extracted and frequency regions where partials of a non-target tone are found are marked as unreliable. Hence, a binary 'mask' is derived, that indicates the spectral features which should be employed by a GMM classifier. Note that our approach relies on the detection of F0s and approximately harmonic overtone spectra; it is therefore applicable to most musical instruments but excludes drums and other untuned percussion instruments.

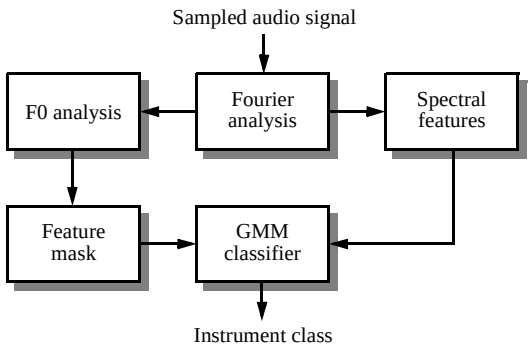

Figure 1: Schematic of the instrument classification system.

2.1. Acoustic features

Sampled audio recordings were divided into frames 40 ms in length with a 20 ms overlap. For each frame, a fast Fourier transform (FFT) was computed, the log magnitude was taken and the spectrum was normalised to a standard maximum value.

Spectral features were computed by summing the energy within 60 Hz frequency bands, with an overlap of 10 Hz between adjacent features. Broader frequency bands would have the advantage that features were more likely to represent formant regions, which are mostly independent from the F0. However, for the missing data approach local spectral features are required, since it is necessary to mask the partials of an interfering harmonic series without affecting the partials of the target tone. Spectral features are computed for bands between 50 Hz and 6 kHz, which includes the range of all possible F0s of the instruments used, and their formant regions.

Three issues are worthy of note here. First, only frame-based spectral features are used in our system, rather than features that encode the temporal evolution at the onset of tones (e.g., see [5], [7]). This has the advantage that recognition does not rely on exact onset detection, a task that is particularly hard in polyphonic music.

Secondly, although cepstral coefficients have been used successfully as features for instrument identification [1], [6], they are not used here because they do not fit naturally into the missing feature approach. A local frequency region which is corrupted by energy from a non-target tone has no clear correspondence in the cepstral domain. Hence, we employ local spectral features.

Finally, the spectral features used here are ordered on a linear scale and have a constant bandwidth (60 Hz). An alternative approach would be to use a scale based on human hearing, in which bandwidth increases quasi-logarithmically with increasing frequency. However, high-frequency regions of this scale will integrate energy from several partials into the same feature; again, this is not compatible with the missing feature approach, which requires local spectral features. Our initial studies confirmed that while logarithmic features performed well when all features were available for classification, linearly scaled features performed better in cases where some features were missing.

2.2. Fundamental frequency analysis

The F0 analysis used here is based on the notion of 'harmonic sieves' [9]. Each sieve consists of a pattern for the harmonic overtone series of each possible F0. The sieves are matched against the peaks in the spectrum, and the best fitting sieve is chosen as the corresponding F0. A score is derived for each sieve based on the power of the matching spectral peaks and their position within the sieve. Lower frequency slots in the sieve are given a higher weighting in the score; this reflects the power distribution of most natural instrument tones, and also helps to prevent sub-octave errors.

For a mixture of tones containing more than one F0, an iterative scheme is used in which the best matching sieve is identified, and the corresponding spectral peaks are removed. Further sieves are matched to the residual until all of the spectral peaks are accounted for.

To connect the frame based F0-hypotheses to longer tones, we use a simple birth-and-death algorithm. F0 estimates in adjacent frames are connected according to their proximity in frequency. Specifically, a running average of the F0 of the tone is computed, and further F0 estimates are recruited to the tone if they occur within a matching interval. This approach allows for small oscillatory changes in F0, such as vibrato, but breaks up chromatic scales. Candidate tones are rejected if their total power is too small or if their duration is too short.

2.3. Gaussian mixture model classifier with missing features

A GMM models the probability density function (pdf) of observed spectral features by a multivariate Gaussian mixture density:

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

Here, x is a D-dimensional feature vector and N is the number of Gaussian densities Φ_i, each of which has a mean vector μ_i, covariance matrix Σ_i and mixing coefficient p_i. Here, we assume a diagonal covariance matrix; although this embodies an assumption which is incorrect (independence of features) it is a widely used simplification (e.g., see [1]). Accordingly, (1) can be rewritten as

$$p(x) = \sum_{i=1}^{N} \prod_{j=1}^{D} \Phi_i(x_j, m_{ij}, \sigma_{ij}^2) \quad (2)$$

where m_{ij} and σ_{ij}^2 represent the mean and variance respectively of a univariate Gaussian pdf. Now, consider the case in which some components of x are missing or unreliable, as indicated by a binary mask M. In this case, it can be shown [3] that the pdf (2) can be computed from partial data only, and takes the form:

$$p(x) = \sum_{i=1}^{N} \prod_{j \in M'} \Phi_i(x_j, m_{ij}, \sigma_{ij}^2) \quad (3)$$

where M' is the subset of reliable features in M. Hence, missing features are effectively eliminated from the computation of the GMM pdf.

2.4. Training

Individual GMMs were trained for five different instruments from two different instrument families (flute, oboe and clarinet from the woodwind family, and violin and cello from the string family). To make the models as robust as possible they were trained with different recordings of each instrument, using both monophonic musical phrases and single tone recordings. Recordings from eight or nine different sources were used for every instrument, each approximately one minute in length. Silence was removed from the recordings prior to further processing by dropping all frames with an energy level below 0.5% of the peak energy. For some noisy recordings this approach was not completely reliable, and it was necessary to remove silent segments manually. After an initial clustering using a K-nearest neighbour algorithm, the parameters of the GMMs were trained by the expectation-maximisation (EM) algorithm. The number of Gaussian densities, N, was set to 120 after some experimentation; a further increase gave no improvement.

3. EVALUATION

The system was evaluated in four stages. First, a baseline performance was established for monophonic recordings, using

masks in which all spectral features were included. Secondly, random spectral deletions were introduced in order to assess the robustness of the classifier to missing features. Thirdly, mixtures of two simultaneous instrument tones were employed, with the true F0s known to the system. Finally, a duet recording from a commercially available CD was used, and instrument recognition was carried out using F0s extracted by the system.

3.1. Monophonic sounds

To establish an upper limit on identification performance with missing features, tests were carried out with monophonic recordings. Test material was taken from recordings which were not included in the training material, consisting of chromatic scales from the McGill Master Samples CD [8]. To avoid cues based solely on the different pitch range of the instruments, only tones from one octave (C4-C5) were used. The chromatic scales were manually cut into single tones and classification decisions were made for each tone. A confusion matrix for this test is shown in Table 1. Average instrument identification performance over all tones was 66%, with the worst performance for the violin (which was often confused with the cello). Discrimination between instrument families (woodwind or string) was 85% correct (i.e., when an instrument was confused, it tended to be confused with another instrument from the same family).

	Flute	Clarinet	Oboe	Violin	Cello
Flute	**77%**	15%	0%	0%	8%
Clarinet	15%	**62%**	0%	8%	15%
Oboe	0%	15%	**69%**	8%	8%
Violin	0%	0%	15%	**54%**	31%
Cello	0%	0%	15%	15%	**69%**

Table 1: Confusion matrix for instrument recognition of single notes from the McGill Master Samples CD [8].

Identification performance was also assessed on monophonic phrases from a number of classical music CDs, which were not used during training. Four recordings were used for the flute, clarinet and violin, and three recordings for the oboe and cello. All recordings were correctly identified except for two oboe recordings which were mistaken for flutes, and one cello for a clarinet. These results are comparable with those obtained by Brown [1] for monophonic woodwind phrases.

3.2. Monophonic sounds with random spectral deletions

To test the robustness of the models with missing features, we carried out preliminary studies with random spectral deletions in monophonic sound files. Different percentages of the features were marked as missing and recognition was performed only on the remaining features. Results were encouraging, as even with 80% deletions performance in terms of correctly identified frames dropped by no more than 10%-20%, both for single tones and monophonic phrases. These results are consistent with those of Cooke et al. [2] for random spectral deletions in an automatic speech recognition task.

3.3. Concurrent tones with masks based on given F0

As the next step towards realistic performance, we considered the task of identifying both instruments in a combination of two single notes, played concurrently by different instruments. Masks were based on the F0s of the tones, which in this case were supplied to the algorithm.

A test set was derived from the McGill samples by taking all possible combinations of two tones within one octave, excluding intervals in which both tones had the same F0 or were an octave apart. Before mixing, the tones were normalised to have equal peak amplitude. The length of each sound was determined by the shorter of the two tones to ensure that two instruments were present for the whole mixture.

	Flute	Clarinet	Oboe	Violin	Cello
Flute	**75%**	6%	0%	10%	9%
Clarinet	13%	**49%**	3%	22%	12%
Oboe	20%	13%	**25%**	28%	16%
Violin	3%	4%	10%	**57%**	25%
Cello	3%	9%	16%	36%	**37%**

Table 2: Confusion matrix for instrument recognition of two concurrent tones from different instruments, using masks derived from a given F0.

The F0s (according to the nominal frequency of the tones in equal temperament) were manually fed into the system. A perfect harmonic overtone series was assumed and all features into whose frequency range a harmonic from the non-target tone fell were marked as unreliable and excluded from the recognition process. Initial studies showed that the results improved when the exclusion was based on 'broadened' harmonics, where each harmonic from the non-target tone masked a small frequency range of ±5 Hz instead of an exact frequency. This can be explained by the fact that all of the tones exhibited some degree of vibrato, so that the frequencies of all harmonics changed slightly from frame to frame. Additionally, the overtone series for tones generated by real instruments is unlikely to be exactly harmonic. Since it was shown in Section 3.2. that identification is quite robust to missing features, it is preferable to exclude too many features than to risk including features which might be dominated by the non-target tone.

Using the F0-based masks, on average 49% of instruments were correctly identified; confusions are shown in Table 2. Discrimination between instrument families was 72% correct.

3.4. Duet recording with masks based on estimated F0

The final step was to test the system on a realistic recording. We chose 'Chôro n. 2', a duet for flute and clarinet from Villa-Lobos because a professional recording [10] and the score were locally available. The piece begins with a series of very fast notes, in which the F0 tracker was unable to extract the F0 of each instrument. We therefore used the slower beginning of the second part (bar 10, see Figure 2) for the recognition experiment. All tones were found except for one repeated onset of the A flat in the flute voice. Currently, the system does not transform the 'piano roll' output shown in Figure 3 into an actual score, so correct notes were identified manually from the estimated F0 values.

Two versions of the system were evaluated on this mixture. In the first, F0s were estimated and a mask was generated which assumed a perfectly harmonic overtone series. For this case, the instrument was correctly identified for 9 of the 12 tones.

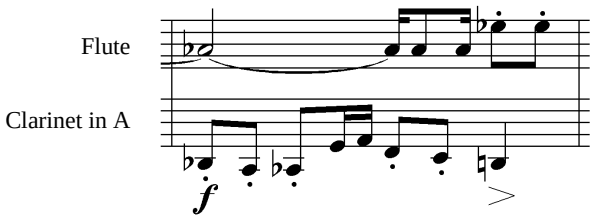

Figure 2: Original score of chôro n. 2 from Villa-Lobos.

A second version of the system constructed a mask from the estimated F0s without assuming a perfect overtone series. Specifically, frequencies were excluded from the mask at which there was a peak in the spectrum of the mixture that matched a harmonic in the sieve for the non-target tone. For this version of the system, the instrument was correctly identified for all 12 tones. Whilst strong conclusions cannot be drawn from a single example, our results appear to indicate that an accurate estimation of the overtone series of the interfering sound is important.

4. CONCLUSIONS AND FUTURE WORK

An instrument recognition scheme has been described based on missing feature theory and a GMM classifier. The system generalises well, giving good results on single note recordings and musical phrases. Importantly, the system is not limited to identification of instruments in monophonic music (although its performance on such tasks is comparable with previous approaches, e.g. [1], [6], [7]). Rather, by using a F0-based missing feature mask, the system is able to identify two different instruments playing concurrently.

The system was primarily evaluated on combinations of two isolated tones played by different instruments. However, we also obtained good recognition performance on a duet recording taken from a commercially available CD of classical music. The system is therefore promising as a tool for classification of instruments in naturalistic audio recordings, with possible applications in automatic transcription and information retrieval.

In future studies, we will include a wider range of test stimuli, including more instruments than the five used here, and complex musical mixtures which involve more than two concurrent instruments. Additional work needs to be done on the F0 extraction algorithm to make it more robust, especially with respect to short notes. An onset detection module could also help to identify repeated notes with the same F0, and to distinguish real notes from spurious ones.

Possibilities exist for improving instrument classification within the current system. For example, bounded marginalisation can be applied [2]. In this approach, an upper bound is placed on the value of a missing feature (typically, the observed energy) rather than discounting it completely.

Clearly, further work is required to transform the output shown in Figure 3 into a full musical transcription (such as that shown in Figure 2), and this will involve solving problems concerning quantisation in both time and frequency. However, our current results on instrument classification are promising, and suggest that our eventual goal – a system for transcribing classical chamber music from audio CDs – is achievable.

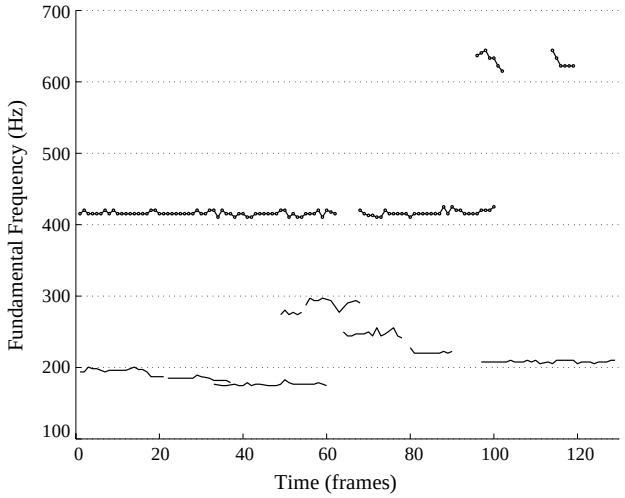

Figure 3: Output from the instrument classification algorithm in 'piano roll' format, corresponding to the score in Figure 2. Dotted lines indicate a flute tone, plain lines indicate a clarinet tone.

ACKNOWLEDGMENTS

JE is supported by the IHP HOARSE project. GJB is supported by EPSRC grant GR/R47400/01 and the MOSART IHP network.

REFERENCES

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

[2] Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (2001) Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Comm.* **34**, pp. 267-285.

[3] Drygajlo, A. & El-Maliki, M. (1998) Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. *Proc. ICASSP-98*, pp. 121-124.

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

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

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

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

[8] Opolko, F. & Wapnick, J. (1987) McGill University master samples (CD), Montreal, Quebec: McGill University.

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

[10] Villa-Lobos, H. (1994) *Chôro n. 2 for flute and clarinet.* Wind Music, Arts Music GmBH, CD 447200-2.