TOWARDS AUTOMATIC TRANSCRIPTION OF AUSTRALIAN ABORIGINAL MUSIC

Andrew Nesbit [1]
School of Physics
University of Melbourne
Australia

Lloyd Hollenberg
School of Physics
University of Melbourne
Australia

Anthony Senyard
Dept of Computer Science
and Software Engineering
University of Melbourne
Australia

ABSTRACT

We describe a system designed for automatic extraction and segmentation of didjeridu and clapsticks from certain styles of traditional Aboriginal Australian music. For didjeridu, we locate the start of notes using a complex-domain note onset detection algorithm, and use the detected onsets as cues for determining the harmonic series of sinusoids belonging to the didjeridu. The harmonic series is hypothesised, based on prior knowledge of the fundamental frequency of the didjeridu, and the most likely hypothesis is assumed. For clapsticks, we use independent subspace analysis to split the signal into harmonic and percussive components, followed by classification of the independent components.

Finally, we identify areas in which the system can be enhanced to improve accuracy and also to extract a wider range of musically-relevant features. These include algorithms such as high frequency content techniques, and also computing the morphology of the didjeridu.

1. INTRODUCTION

The traditional music of Indigenous Australians is firmly entrenched in oral tradition. The songs are passed down through generations within a group, without written notation, and typically describe the history and culture of the group.

We have designed and implemented our transcription system with two styles of Australian music in mind. The first is Lirrga, a genre of music from northwest Australia. The performances we study are composed and performed by Pius Luckan (voice, clapsticks) and Clement Tchinburrur (didjeridu) [11]. The recordings were made in Wadeye (Port Keats, northern Australia). In the Marri Ngarr language (one of seven languages spoken at Wadeye), clapsticks are called *titir* and didjeridu is *karnbi*.

The second set of recordings is a collection of traditional songs from the Galpu clan of northeast Arnhem Land. The songs are arranged and performed by Gurritjiri Gurruwiwi (voice), with Djalu Gurruwiwi (didjeridu) [5]. The Yolngu people who reside here call the didjeridu *yiḏaki* and the clapsticks *bilma*.

The rhythmic structures of these musics are complex and highly expressive. Polyrhythms, changes of tempo and changes of metre are common and integral to the music.

Our initial motivation for creating this system was to construct a useful tool to aid ethnomusicologists studying Australian Aboriginal music. We had access to manually-created transcriptions of the Lirrga songs, and these served as a good model as to the level of detail our system should aim towards. Although our system has been designed with more than one style of music in mind, for this reason, and also for the fact that the two styles of music are very different, we have executed most of our evaluations on the Lirrga.

The system is designed to determine onsets of the clapsticks, and onsets and fundamental frequencies of the didjeridu parts. We assume that there is only one of each instrument playing at any given time, and that the fundamental frequency of the didjeridu is below 100 Hz, which works well for most of our samples. We do not attempt to transcribe vocals in this system.

As far as we aware, no published research on automatic transcription of Australian Aboriginal music exists. However, work has been done in studying the musical acoustics of the didjeridu, and recent studies may be found in [3], [4], [7].

Seminal studies into the history and cultural significance of the didjeridu include [14]. Also, research into didjeridu notation [13] provides guidelines as to the types of features we may wish to extract.

[1] The first author is currently with the Department of Electronic Engineering, Queen Mary, University of London.

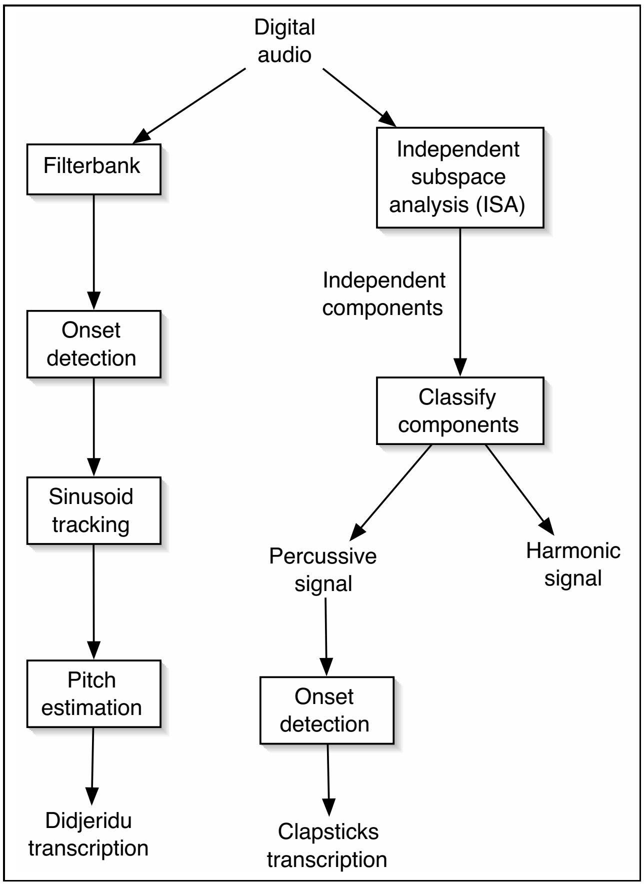

Figure 1. High-level overview of the system.

2. THE SYSTEM

As the techniques used for transcription of didjeridu (deterministic) are very different to those used for transcription of clapsticks (statistical), the system is essentially split into two disjoint "halves", as indicated in Figure 1. The left half indicates the data transformations that occur for didjeridu processing, and the right half describes the operations used for extracting clapsticks.

2.1. Extraction of didjeridu

The overall scheme used for this phase was based on a system for automatic transcription of bass lines [6], but with a different onset detection scheme. The original signal was passed through a bank of bandpass filters, emitting signals in the ranges 0–100 Hz, 100–200 Hz and 200–300 Hz. These ranges were chosen to capture the fundamental frequency of the didjeridu (typically below 100 Hz), and its first two upper harmonics into each frequency band. The other instruments carried very little energy in these frequency ranges, and so, within the context of this project, we assume that all musical information carried in these ranges belongs to the didjeridu.

2.1.1. Onset detection

Our informal experiments revealed that complex-domain onset detection [1] works well for low-frequency signals.

For each frequency band, complex-domain onset detection was applied. After the three frequency bands had been analysed in this way, the onsets from each of the bands were combined into one sequence. Each onset was considered in turn: if it was closer than 50 ms to another onset and its amplitude was less than the neighbouring onset, it was removed. Thus, the onsets for the didjeridu were given by the resulting sequence.

2.1.2. Frequency estimation

The next stage was to estimate the fundamental frequency at each onset. A frame as long as possible was considered, starting just after an onset and ending just before the next onset. For each such frame, a sinusoid extraction algorithm was applied. Rather than use the method detailed in [12] as suggested by [6], we opted for the triangle window method [9]. Every note under 100 Hz was considered to be a candidate fundamental frequency, and we make the assumption on our data that only one note is playing at a time.

For each fundamental frequency candidate F_0, we predict its harmonic series as $\{nF_0\}_{1\leq n\leq N}$. The actual harmonic series associated with each F_0 is determined from this by considering each nF_0 for $2 \leq n \leq N$ in turn, and searching for the extracted sinusoid whose frequency lies within 3% of its predicted value, and whose amplitude is a maximum. We found that a value of $N = 9$ gave good results, although this probably could have been made much smaller without noticable loss of accuracy in our results.

At this stage, we have one or more harmonic series corresponding to each onset. To determine the most probable harmonic series for each offset, we assign each series a confidence measure as described in [6]. The series with the highest confidence is deemed to be the correct one, and hence, the fundamental frequency is determined.

Note that our algorithm is a simplified version of the one it is based on. In particular, for each onset, the algorithm described in [6] tracks the harmonic series over time in order to determine the note offset for that series, and to determine the correct series using a more sophisticated measure. We chose this simpler technique because it was not practical to achieve the necessary frequency resolution for accurate determination of sinusoids for such low frequencies: the short frames required for accurate time resolution prohibited this.

2.2. Extraction of clapsticks

To extract the clapsticks, we used the method of independent subspace analysis (ISA) described in [15]. The following discussion is essentially a summary of that paper.

This technique is based on independent component analysis (ICA), and we use it to split the original signal into

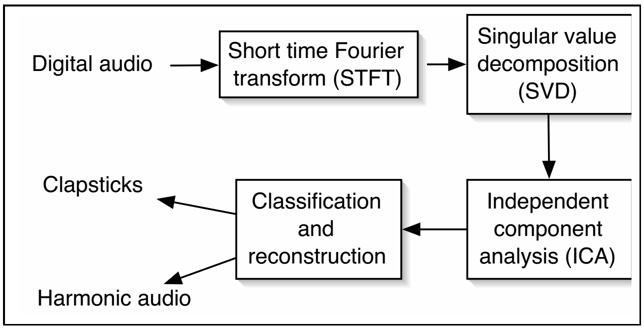
Digital audio
Short time Fourier transform (STFT)
Singular value decomposition (SVD)
Independent component analysis (ICA)
Classification and reconstruction
Clapsticks
Harmonic audio

Figure 2. Independent subspace analysis [15]

harmonic and percussive components. The classic formulation of blind source separation by ICA requires at least as many observed signals as there exist sources. In our case, we have one observation (the recording itself) and three sources (didjeridu, clapsticks and vocals). Figure 2 briefly indicate the steps that ISA performs to overcome this limitation. The original (single observation) time-series signal is transformed to the frequency domain by short-time Fourier transform (STFT). To get reliable separation, we used long frames (100 ms) with a half-frame overlap. Singular value decomposition (SVD) is performed on the resultant magnitude spectrogram, and a maximum-variance subspace of the original spectrogram, reduced to d dimensions, is then computed. Finally, amplitude envelopes and frequency weights of each of the d independent components are computed using ICA. (These d amplitude envelopes and frequency weights may be used to determine the independent spectrograms. We do not make use of these, however, so this is not done in our system.)

Through experimentation, we found that setting $15 \leq d \leq 20$ provides excellent separation with an acceptable computational cost.

The remaining task in this stage was to classify each of the d separated components into either harmonic or percussive categories. [15] describes five measurable features of the independent components, each of which gives an indication of the percussion-likeness of each of the independent components. Our system makes use of two of these features.

The first, percussiveness, is determined by computing a train of unit impulses, where each impulse is located at a maximum of the amplitude envelope (determined during the ICA), and convolving this impulse train with a model percussion template. This percussive impulse is modelled by an instantaneous onset and linear decay towards zero within 200 ms. The measure of percussiveness is given by the correlation coefficient between the output of the convolution and the amplitude envelope.

The second feature, noise-likeness, uses the component's vector of frequency weights (determined during the ICA). Similarly to the method described above, an impulse train corresponding to the maxima of the frequency vector is convolved with a Gaussian window. The noise-likeness is given by the correlation coefficient of the original frequency vector and the output of the convolution.

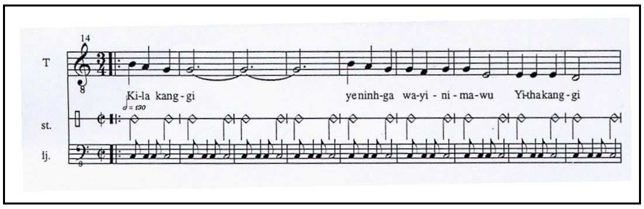

Figure 3. *Kila kanggi: Our mother* excerpt reference transcription (lirrga_1). Track 1 from CD [11]. Song text © composed by Clement Tchinburrur, sung by Pius Luckan, recorded by Chester Street, Pt Keats, 1985. Musical transcription © Linda Barwick, University of Sydney, 2002. Marri Ngarr morphemic analysis and translation © Lysbeth Ford, BIITE, 2002. Reproduced with permission.

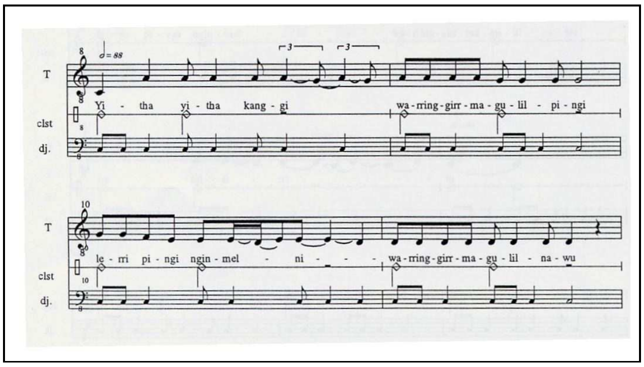
8
♩= 88
T
3
3
Yi - tha yi - tha kang - gi
wa-rring-girr - ma - gu - lil - pi - ngi
clst
dj.
10
T
le - rri pi - ngi ngin - mel - ni - - - wa-rring-girr - ma - gu - lil - na - wu
clst
dj.

Figure 4. *Yitha yitha kangki: Father, our Father* excerpt reference transcription (lirrga_2). Track 2 from CD [11]. Song text © composed by Clement Tchinburrur, sung by Pius Luckan, recorded by Chester Street, Pt Keats, 1985. Musical transcription © Linda Barwick, University of Sydney, 2002. Marri Ngarr morphemic analysis and translation © Lysbeth Ford, BIITE, 2002. Reproduced with permission.

The decision as to whether a component was percussive or harmonic was made by comparing the percussiveness and noise-likeness measures to predetermined thresholds.

To determine the onsets of the clapsticks, a note onset detection algorithm, as described in Section 2.1.1 was applied to the sum of the amplitude envelopes determined during the ICA process and corresponding to the percussively classified component.

3. RESULTS

The accuracy of the system, and also its sensitivity to inflections, will improve with the incorporation of more sophisticated techniques and algorithms. We discuss this in section 4.

We evaluated the performance of our system by running it on short (approximately 10 seconds) excerpts and comparing the results against the transcriptions Figures 3–6, which were prepared manually by ethnomusicologists.

3.1. Didjeridu transcription

To measure the transcription accuracy, we employ a metric described in [8]; this paper gives more than one metric, and we choose the more stringent alternative. It is defined

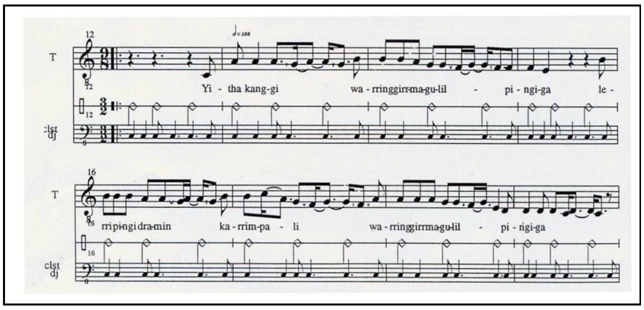
12
T
Yi - tha kang-gi wa - rringgirrmagu-lil - pi - ngi-ga le -
clst
dj
16
T
rripingidra-min ka - rrim-pa - li wa - rringgirrmagulil - pi - ngi-ga
clst
dj

Figure 5. *Yitha kanggi warringgirrmagulil: Our Father enter into us* excerpt reference transcription (lirrga_3). Track 4 from CD [11]. Song text © composed by Clement Tchinburrur, sung by Pius Luckan, recorded by Chester Street, Pt Keats, 1985. Musical transcription © Linda Barwick, University of Sydney, 2002. Marri Ngarr morphemic analysis and translation © Lysbeth Ford, BIITE, 2002. Reproduced with permission.

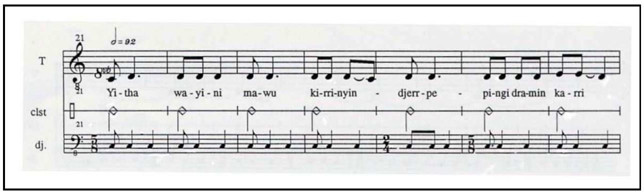
21
T
Yi - tha wa - yi - ni ma - wu ki-rri-nyin - djerr - pe - pi-ngi dra-min ka - rri
clst
dj.

Figure 6. *Father Deakin* excerpt reference transcription (lirrga_4). Track 5 from CD [11]. Song text © composed by Clement Tchinburrur, sung by Pius Luckan, recorded by Chester Street, Pt Keats, 1985. Musical transcription © Linda Barwick, University of Sydney, 2002. Marri Ngarr morphemic analysis and translation © Lysbeth Ford, BIITE, 2002. Reproduced with permission.

by the following relation:

$$R = \frac{\text{no. of notes found correctly}}{\text{no. of notes found in total} + \text{no. of notes missed}}$$

which, when applied to our results, gives these results (Table 1):

Excerpt	notes	found correct	found total	missed	R
lirrga_1	36	33	39	3	0.79
lirrga_2	30	23	37	7	0.53
lirrga_3	48	44	46	4	0.88
lirrga_4	21	14	17	7	0.58

Table 1. Results for didjeridu.

The overall accuracy (average R) was 70%. Almost all of the errors were a result of the note onset detection, rather than the frequency estimation. One reason for this is that some many note onsets are difficult to detect with an amplitude envelope type method, because a note onset does not necessarily correspond to a large increase in amplitude. Another reason is due to human subjectivity in formulating the reference transcriptions, and also in matching generated transcriptions to their references.

3.2. Clapsticks transcription

With respect to the separation of clapsticks from the harmonic components, errors did indeed occur in the classification stage. We used the metric outlined in [15] and found that overall, 71% of percussive components were found correctly. Spurious percussive classifications were at 31%.

The accuracy of note onset detection for clapsticks extraction was measured similarly to that of the clapsticks, by the preceeding formula. For all correctly classified clapstick tracks we obtain the following results (Table 2):

Excerpt	notes	found correct	found total	missed	R
lirrga_1	20	20	22	0	0.91
lirrga_2	8	8	8	0	1.0
lirrga_3	24	24	30	0	0.80
lirrga_4	7	7	7	0	1.0

Table 2. Results for clapsticks

This gives an overall accuracy of 93% for correctly classified clapstick tracks.

4. CONCLUSIONS AND FUTURE WORK

There are many ways in which we intend to increase the accuracy and scope of our system. We identify several measures for this.

First on our list is tracking inflection changes during a single note played by the didjeridu. Whilst the current didjeridu extraction method compares favourably with the model exemplified by our Westernised transcriptions, we wish to track the harmonic series associated with each onset. This poses a new problem: how does one map changes in harmonic series to changes in inflection? As described in Section 3.1, it is also the case that changes in inflection, rather than large changes in amplitude, correspond to new note onsets. This would therefore increase the accuracy ratings of our transcriptions. The complex-domain onset detection algorithm we have used [1] picks changes in inflection well, and so the problem remains: how to classify onsets determined this way, and how to use extra information to determine when inflection changes correspond to definite onsets.

Futhermore, clapsticks have considerable energy in the high frequency subbands, and so we would investigate tracking of transient energy in high frequency ranges [2] in order to augment note onset detection for clapsticks.

The morphology of the didjeridu often varies depending on the geographical location in Australia from which the music originates. This has an effect on the resonant frequencies of the instrument [3]. Therefore, by identifying notes played at higher resonant frequencies, we could compute the morphology of the didjeridu.

5. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of the following people: Linda Barwick for discussions leading to a greater understanding of the ethnomusicological and cultural aspects of the project, and also for providing the source music recordings and their transcripts; Aaron Corn for careful proofreading of our paper and insightful suggestions; Christian Dittmar for discussions leading us to a better understanding of independent subspace analysis.

This project was funded primarily by the School of Physics, University of Melbourne, and also partly by the Department of Electronic Engineering, Queen Mary, University of London.

6. REFERENCES

[1] Chris Duxbury, Juan-Pablo Bello, Mike Davies, and Mark Sandler. Complex-domain onset detection for musical signals. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 2003.

[2] Chris Duxbury, Mark Sandler, and Mike Davies. A hybrid approach to musical note onset detection. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002.

[3] Neville Fletcher. The didjeridu (didgeridoo). *Acoustics Australia*, 24:11–15, 1996.

[4] Neville Fletcher, Lloyd Hollenberg, John Smith, and Joe Wolfe. The didjeridu and the vocal tract. In *Proceedings of the International Symposium on Musical Acoustics*, pages 87–90, Perugia, Italy, 2001.

[5] Gurritjiri Gurruwiwi. *Waluka*. Yothu Yindi Foundation, 2001. Compact disc music recording, featuring Djalu Gurruwiwi on yiḏaki.

[6] Stephen W. Hainsworth and Malcolm D. Macleod. Automatic bass line transcription from polyphonic music. In *Proceedings of the International Computer Music Conference*, Havana, Cuba, September 2001.

[7] Lloyd Hollenberg. The didjeridu: Lip motion and low frequency harmonic generation. *Australian Journal of Physics*, 53:835–850, 2000.

[8] Kunio Kashino and Hiroshi Murase. Music recognition using note transition context. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, volume 6, Seattle, Washington, USA, May 1998.

[9] Florian Keiler and Udo Zölzer. Extracting sinusoids from harmonic signals. *Journal of New Music Research*, 30(3):243–258, 2001.

[10] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, USA, March 1999.

[11] Pius Luckan and Clement Tchinburrur. *Marri Ngarr church Lirrga Songs*, Unpublished. Musical performance recorded on 10 March, 1985, by Chester Street at Wadeye, NT, Australia.

[12] Malcolm D. Macleod. Nearly fast ML estimation of the parameters of real and complex single tones or resolved multiple tones. *IEEE Transactions on Signal Processing*, 46(1):141–148, January 1998.

[13] Alice Moyle. *Aboriginal Sound Instruments*. Australian Institute of Aboriginal Studies, Canberra, Australia, 1978.

[14] Alice Moyle. The didjeridu: A late musical intrusion. *World Archaeology*, 12(3):321–331, 1981.

[15] Christian Uhle, Christian Dittmar, and Thomas Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.