AUTOMATIC TRANSCRIPTION OF POLYPHONIC PIANO MUSIC USING A NOTE MASKING TECHNIQUE

Ronan Kelly and Jacqueline Walker

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

ABSTRACT

This paper describes a polyphonic note detection system incorporating a simple masking technique that can accurately transcribe chords and polyphonic piano music. The system, developed in MATLAB, will take input files in .wav format. The music is segmented by using Note Average Energy (NAE) onset detection. Onsets are used to segment the music into note windows which are then analysed using the FFT. Following compilation of the frequency peaks in each note window, an iterative masking procedure is used to detect and successively extract the notes. The masking procedure uses a database of note masks which are compiled from multiple note examples using both monophonic and polyphonic examples.

1. INTRODUCTION

Music transcription is a complex cognitive task requiring a trained musician to listen to a piece of music and write down the notes played. While automatic transcription of monophonic music is largely considered a solved problem and many examples exist [1], [2], automatic transcription of polyphonic music, to a high level of accuracy, remains a considerable challenge [3]. Researchers into automatic music transcription of polyphonic music have considered a range of different approaches, including neural networks [4], [5], auditory-based approaches [6], [7], probabilistic inference [8], and heuristic signal-processing based techniques [9], [10], [11].

Transcription systems described to date have had limited success. In [4] and [5] it was found that with increasing polyphony the number of spurious notes also increased. Repeat notes where a note was played in succession several times, can also affect results [12], while in [9] it was found that detecting notes with a fundamental frequency less than 200Hz was difficult to achieve accurately. Another major factor in causing errors is the detection of notes that are an octave apart. As the note an octave above is, by definition, a doubling of frequency, it also corresponds to the second harmonic of the lower note. Octave errors are common throughout the automatic transcription field [4], [7], [12], [13].

Approaches to automatic music transcription can be divided into two broad types. One approach taken is to try and reproduce the characteristics of the human hearing system [14], [15]. In these approaches, a simplified model of the human auditory system is constructed and used for initial processing of the musical signal. The second approach, which is the focus of this paper, is to try and transcribe exactly the notes that were played using traditional signal processing techniques [9], [10], [11], [16].

In this work we use a masking approach to polyphonic note detection. Once a note has been identified, it is removed from the signal and the system attempts to identify the next note and so on. The procedure continues until no more notes can be found in that segment of the music. In the auditory field, *masking* occurs when a louder sound adjacent (in frequency) to a quieter sound prevents the quieter sound being heard [17]. In computer science, a *bit mask* is used to conceal bits in a word to allow the retrieval of information of interest [18]. In this research the term *masking* is used in both of these ways: it refers both to how the presence of overlapping harmonics from different notes in the same window leads to an apparent masking of notes by other more dominant notes and to the use of a mask for blocking out all but a particular note from a set of frequencies in an analysis window. In music signal identification and transcription, one of the most problematic issues is amplitude variability: notes can be played softly as well as loudly, the amplitudes varying along a continuum. To solve this problem, the amplitudes must be analysed in proportion to the values in the rest of the window. The key insight in the masking scheme is that frequencies present in a window that overlap from different sources (notes) are additive in amplitude. The main assumptions in the work reported here are that the music is polyphonic Western Tonal music consisting of either isolated piano chords or polyphonic piano music. As the music is polyphonic, it is not known a priori how many simultaneous notes exist in each analysis window. It is further assumed that all notes have the fundamental present. The range of notes considered is C2 to B6.

2. IMPLEMENTATION

An onset detection system is first used to segment the music. In this work, the Note Average Energy (NAE) onset detection system was used [19]. By determining the changing profile of the average energy within notes, the method is insensitive to both the dynamic range of the overall energy level of the music and to whether the song is monophonic or polyphonic. Once the position of a note is known, an FFT is performed on the note as a whole, rather than analysing the note frame by frame. The resulting information from the frequency domain has been found to be more robust at detecting the fundamental frequency of a note rather than using sliding, fixed size, overlapping windows. We refer to a ‘note window’ to emphasise that the whole duration of the note is under analysis. The note window approach has also been found to be less susceptible to noise, as the noise in a window of the entire note duration is small in comparison with the overall frequency amplitudes in the window. A sampling frequency of 12 kHz is used as it was found to give a balance between accuracy and computational efficiency. Figures 1a and 1b show two notes, sampled at 12 kHz, played in isolation. The note in Figure 1a lasts

for approximately 0.83 seconds, sampled at 12kHz (frequency

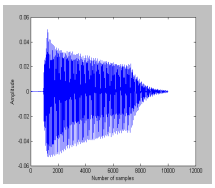

Figure 1a: Note with about 10000 samples.

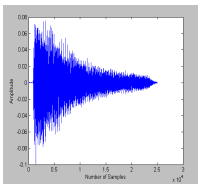

Figure 1b: Note with about 25000 samples.

resolution of 1.2Hz), and contains 10,000 samples and that in Figure 1b lasts for approximately 2.08 seconds, sampled at 12kHz (frequency resolution of 0.48Hz), and contains 25,000 samples. The number of samples in a window is rounded down to the nearest multiple of 1024 samples for use in the FFT calculation.

The first stage in the polyphonic note detection scheme is the compilation of a list of significant frequency peaks for a note window by a repeated procedure of peak-picking the highest frequency peak in the window and then removing it from the note window. Amplitudes in the note window are normalised to the amplitude of the maximum frequency peak found in the note window. The frequency value at this peak and its amplitude start the list of significant frequency peaks in the note window and the frequency peak itself is removed from the note window. The peak with the next highest amplitude in the note window is then identified and its frequency value and amplitude are added to the list and this frequency peak is removed from the note window. This process continues until the amplitude of the highest peak remaining in the modified note window falls below a threshold. The threshold was set heuristically at 20% of the amplitude of the maximum frequency peak in the note window as, following analysis of the notes in the note database (as described below), it was found that peaks below 40% of the fundamental are usually high (> 5^{th}) harmonics or noise (inharmonic). The threshold is set at half this empirical value to allow for the harmonics of other notes in the note window which have a small (relative to the largest peak) fundamental frequency amplitude.

In the second stage of the process, the list of significant frequency peaks found in the note window has to be linked to notes. As the fundamental frequency of a note typically has the largest amplitude of all the frequency components associated with a note, the first frequency peak in the list, which will also be the lowest frequency in the list, is identified with the closest corresponding fundamental frequency of a note. A mask, corresponding to the identified note, is applied to the significant frequency peaks list. The mask removes all of the amplitude of the first frequency peak, from the list of significant frequency peaks. As the mask also contains a proportional representation of the harmonics of the typical note to which it corresponds, the appropriate proportion of the amplitude of the harmonics of the note are removed from the frequencies which are present in the list of significant frequency peaks.

Once the first frequency has been removed from the list of significant frequency peaks, the system moves to the next peak in the list. Note that the amplitude of this peak may have been reduced by a previously run mask. If the amplitude of the frequency peak is still above the threshold, its frequency value is compared to note fundamental frequencies and the note corresponding to the closest fundamental frequency is selected as the next note identified in the note window. A mask corresponding to the newly identified note is run on the list of significant frequency peaks. If the amplitude of the frequency peak is not above the threshold, then that frequency is discarded, since it is likely that any remaining energy at that frequency is due to noise and does not indicate the presence of a note. This procedure continues until the list of significant frequency peaks is exhausted.

To create a note mask in the prototype system, 7 monophonic and 13 polyphonic examples of each note in the range C2 to B6 were recorded and analysed using an FFT on a note window. The amplitudes of the harmonics are converted to percentages of the amplitude of the fundamental frequency. Taking all the examples of each note, average percentage amplitudes of each of the harmonics of the note in relation to the fundamental frequency are calculated. For the polyphonic samples, each note in the range is recorded with another note played at the same time. Since the theoretical frequency values of the harmonics for each note in the combination are known, the amplitude of the harmonics of each note can be identified. If the harmonics of the 2 notes coincide, then this value is discarded as the amplitude value will be the sum of a harmonic from each note and not a harmonic from a single note and there is no guaranteed formula to apportion them. However, in future, analysis of such examples and their addition to an expanded note database could greatly increase the power of this system. Again, as with the monophonic samples, the average amplitudes of the harmonics in relation to the fundamental frequencies are calculated as percentages. An overall average is obtained by combining the data from the polyphonic samples with the monophonic samples.

Figure 2 shows an example of how a mask is used when two notes, D4 and A4 are played together. The first column shows the list of significant frequency peaks. The first frequency in the list is identified as the fundamental of D4 and added to start the list of notes found in the note window. Column 2 shows the values for the mask for frequency 294 Hz (D4) and column 3 gives the results when the 294 Hz (D4) mask is applied to the original list of significant frequency peaks from column 1. The first frequency, 440 Hz in the list after application of the mask is then identified with a note (A4) and added to the list of notes found and the mask in column 4, for the frequency of 440 Hz (A4) is run on the remaining values in the list of significant frequency peaks. Column 5 shows that all the information has been removed from the list of significant frequency peaks after the 440 Hz (A4) mask has been run.

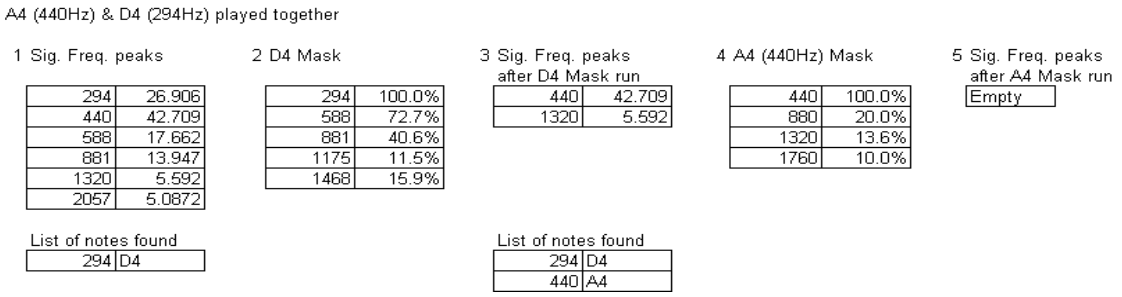

Figure 2: Masking example.

3. RESULTS

The polyphonic note detection system has been tested in two different settings. The first test involves chords played with approximately 0.5s separating each set of notes. The note onsets are straightforward to detect and so the masking system is mostly unaffected by any limitations of the onset detection system. While this type of transcription is not realistic with regards to transcribing music played on a piano, it does give a good indication of the potential of the masking system. Table 1 gives the results of the transcription system when it is applied to music with varying degrees of polyphony and where the chords or multiple notes are played approximately a half second apart. The percentage error is calculated as $\%E = ((m+x)/n)100\%$ where m is the number of missing notes, x is the number of extra notes and n is the total number of notes detected.

	Notes Detected	Total in Error	% Error
Chords	1906	146	7.66
High Polyphony Chords (5 – 8 notes)	225	18	8.0
Chords (Triads/4 notes)	638	20	3.13
Total	2769	184	6.64

Table 1: Results when note onsets are approximately 0.5s apart.

From the results in Table 1, the masking system can accurately detect what notes were played with an overall success rate of 93.36 %. When higher levels of polyphony are applied, the error rate increases. Our results also show that using a higher sampling rate is not necessarily beneficial as error rates were higher with some music files with sampling rates of 24 and 48 kHz, which gave an average error rate of 9.35%. A major cause of errors is the octave problem, which accounts for 39% of the errors in common with the findings of many others [4], [7], [16]. The second test for the system was transcription of piano music at normal tempo. In this case, the results are not as good and the error rate dramatically increases. The overall error rate for transcription was 42.6%. A major cause of error in this case is the onset detection system. As the errors caused by the onset detection system were so great, the onsets and offsets of the notes were extracted manually. When onset detection was manual, the results improved and total error rate became 20.96%.

Tune	Time (s)	Beats/s	Notes Detected	Total in Error	% Error
1Desperado	61	40	84	16	19.05
1Desperado	61	60	123	18	14.63
1Desperado	61	80	170	25	14.71
2Beethoven – Für Elise	61	40	142	31	21.83
2Beethoven – Für Elise	61	60	130	35	26.92
2Beethoven – Für Elise	52	80	122	25	20.49
3Danny Boy	61	40	94	21	22.34
3Danny Boy	45	60	98	23	23.47
3Danny Boy	41	80	99	25	25.25
Total			1062	219	20.62

Table 2: Transcribing piano tunes using manual onset detection.

4. DISCUSSION

A major cause of errors is the onset detection system which returns the correct onset times only about 80% of the time. It produces two types of errors: it may return an extra onset or it may miss an onset altogether. In the first case, where an extra onset is detected, this can sometimes be overcome by the polyphonic note detection system. Since the system considers an entire note window, if an extra onset is detected then it will analyse that as a note window. If that note window contains just noise where no note was played, then it will be eliminated from the list because the noise values will not be above the threshold and no note will be detected. If a spurious onset is detected in the middle of a note that is still being played, an actual note will be divided up into two different note windows. The polyphonic note detection system will likely detect the note in the first window and the same note in the second window as a continuation of the note in the first window.

Most problems occur when an onset is missed and only one note window will be used, when in fact there were two or more notes in succession. Clearly, several different errors could occur depending on the characteristics of the music being analysed. One possibility is that all the notes present will be detected together and it will seem that all the notes in the window were played at the same time. It is also possible, particularly if the fundamental frequencies of the successive notes are close together and the notes are close together in time, that there will not be sufficient frequency resolution to distinguish between the two frequencies.

The masking system is based on information gathered from building a model of the harmonic structure of notes played on the piano. The polyphonic note detection system may fail if a note deviates significantly from the assumed model. An example of this situation is if the second harmonic is abnormally large in comparison with the first harmonic. The polyphonic note detection system will still detect the first harmonic of the note but will not remove enough of the atypical second harmonic, as it is too large in amplitude. Because of this there is an extra peak left in the modified note window and this peak could then, depending on its size, be detected as the fundamental of an additional note despite its actually being a harmonic.

The inverse problem can also occur and lead to notes being missed. If, for example, two notes are present one octave apart, e.g. C4 and C5. If the amplitude of the fundamental frequency of C5 is smaller than typical then the C4 mask removes too much of the C5 value. As a result, note C5 will not be detected. Then, because C5 was not detected, the C5 mask will never be run. C6, the second harmonic of C5, may then be detected as a note because very little may have been removed by the C4 mask, as C6 is the fourth harmonic of C4. This problem mainly occurs in the third (C3, C#3, D3…B3), and fourth (C4, C#4, D4…B4), ranges of notes, and is more likely to occur in relation to C and D notes, including sharp, (C#), and flat, (D_b), notes.

The final limitation of the polyphonic masking system as it currently exists arises when two notes are played in quick succession. The peak of the first note causes interference with the detection of the peak of the second note. When applied to frequencies which have overlapping harmonics, the problem can get quite complex and lead to notes being detected in error similar to the problems caused by the onset detection system missing notes. This problem suggests that a more sophisticated masking system is required to deal with this problem and this will be the subject of further work. However, our work shows that encouraging results can be achieved with a simple system based on a note database and using high frequency resolution facilitated by today's increased computing power.

5. REFERENCES

1. Monti, G. and Sandler, M. "Monophonic transcription with autocorrelation," Proc. COST G-6 Conf. on Digital Audio Effects (DAFX-00), Verona, Italy, 2000.

2. Serman, M., Griffith, N. and Serman N. "MusicTracker: a system for modelling melodic dynamics in music performance," Proc. ICMC, Berlin, Germany, 2000.

3. Klapuri, A. "Automatic transcription of music," Proc. Stockholm Music Acoustics Conference (SMAC 03), Stockholm, Sweden, 2003.

4. Marolt, M. "Transcription of polyphonic piano music with neural networks". Proc. 10th Mediterranean Electrotechnical Conference, vol. 2, pp. 512-515, 2000.

5. Marolt, M. "A Connectionist Approach to Automatic Transcription Of Polyphonic Piano Music," IEEE Trans. Multimedia, vol. 6, no. 3, pp. 439-449, Jun. 2004.

6. Tolonen, T. and Karjalainen, M. "A computationally efficient multi-pitch analysis model," IEEE Trans. Speech Audio Process., vol. 8, no. 6, pp. 708-716, Nov. 2000.

7. Martin, K.D. "A Blackboard System for Automatic Transcription of Simple Polyphonic Music". MIT Media Lab, Technical Report #385, 1995.

8. Walmsley, P. J., Godsill, S. J. and Rayner, P. J. W. "Bayesian modelling of harmonic signals for polyphonic music tracking," Cambridge Music Processing Colloquium, 30 Sep. 1999.

9. Hamid Nawab, S., Ayyash, S.A. and Wotiz, R. "Identification of Musical Chords using Constant-Q Spectra". Proc. ICASSP, vol. 5, pp. 3373-3376, 2001.

10. Klapuri, A., Virtanen, T. and Holm. J.-M. "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," Proc. COST-G6 Conf. on Digital Audio Effects, DAFx-00, Verona, Italy, 2000.

11. Klapuri. A. "Number Theoretical Means of Resolving a Mixture of Several Harmonic Sounds," Proc. EUSIPCO, Rhodes, Greece, 1998.

12. Dunne, E.G. "This note's for you: A mathematical temperament,"http://www.research.att.com/%7Enjas/sequences/DUNNE/TEMPERAMENT.HTML,Sep. 15, 2000.

13. Lao, W., Tan, E.T. and Kam, A.H. "Computationally Inexpensive and Effective Scheme for Automatic Transcription of Polyphonic Music". Proc. IEEE Int. Conf. Multimedia and Expo, pp. 1775-1778, 2004.

14. Cooke, M. Modelling auditory processing and organisation, CUP, Cambridge, 1993.

15. Slaney, M. "Lyon's Cochlear Model," Apple Computer Technical Report #13, Apple Computer, Inc., Cupertino, CA, 1988.

16. Privošnik, M. and Marolt, M. "A System for Automatic Transcription of Music Based on Multiple-Agents Architecture," Proc. 9th Mediterranean Electrotechnical Conference, vol. 1, pp. 169-172, 1998.

17. Moore, B. C. J. An Introduction to the Psychology of Hearing, Academic Press, 2003.

18. Humpage, W. D. and Nguyen, T.T. Elements of System-Level Software Engineering, Dept. of Electrical and Electronic Engineering, UWA, Perth, 1987.

19. Liu, R., N. Griffith, J. Walker, P. Murphy. 2003. "Time domain note average energy based music onset detection". Proc. of the Stockholm Music Acoustics Conference (SMAC'03) Stockholm, pp. 553-556, Aug. 6-9, 2003.