14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

USE OF CONTINUOUS WAVELET-LIKE TRANSFORM IN AUTOMATED MUSIC TRANSCRIPTION

Aliaksandr Paradzinets, Hadi Harb, Liming Chen

Département Mathématiques Informatique, Ecole Centrale de Lyon
36 avenue Guy de Collongue, 36130 Ecully, France
{aliaksandr.paradzinets, hadi.harb, liming.chen}@ec-lyon.fr

ABSTRACT

This paper describes an approach to the problem of automated music transcription. The Continuous Wavelet-like Transform is used as a basic time-frequency analysis of a musical signal due to its flexibility in time-frequency resolutions. The signal is then sequentially modeled by a number of tone harmonic structures; on each iteration a dominant harmonic structure is considered to be a pitch candidate. The transcription performance is measured on test data generated by MIDI wavetable synthesis both from MIDI files and played on a keyboard. Three cases: monophonic, polyphonic and complicated polyphonic are examined.

1. INTRODUCTION

The problem of automated music transcription, i.e. detection of note events with correct pitch in natural polyphonic music still remains not completely resolved for the general case where the number and origin of instruments involved cannot be assumed prior to analysis. At the same time, it is a significant stage in various music analysis tasks as, for instance, automatic music indexing and intelligent navigation (query by humming, search by similarity), computer participation in live human performances, etc. The intelligent music navigation is one of the sharpest issues since there is an exponential growth of music databases and collections either in private or professional sectors and a poorness or sometimes inexistence of means to simplify the navigation and to make it intelligent.

The question of automated music transcription is the question of multiple F0 (pitch) estimation. Being a very difficult problem it is widely addressed in the literature. Lots of works are dedicated to the monophonic case of pitch detection in singing/speech [1][2]. The polyphonic case is usually considered with a number of limitations like the number of notes played simultaneously or an assumption of instruments involved [3][4]. The general case, for example, CD recordings [5] remains less explored.

In this paper we address the problem of automated transcription of polyphonic music with the use of Continuous Wavelet Transform. Paragraph 2 gives information about CWT. In paragraph 3 we describe our approach of F0 estimation (note detection). Basic tests and results on MIDI synthesized samples are provided in the section 4.

2. CONTINUOUS WAVELET TRANSFORM vs. FFT

The Fast Fourier Transform and the Short-Time Fourier Transform have been the traditional techniques in signal analysis for detecting pitches. However, the frequency and time resolution is linear and constant across the frequency scale (Figure 1) while the frequency scale of notes as well as human perception of a sound is logarithmic.

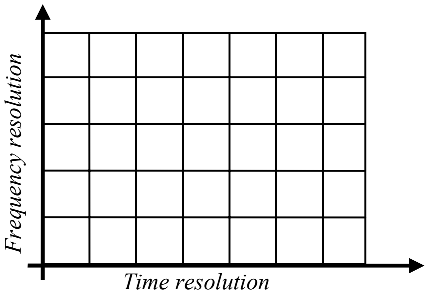

Figure 1. Time-frequency resolution of the Fourier Transform

The rule of calculating the frequencies of notes is well-known. So if we consider a frequency range for different octaves, it is growing as the number of octave is higher. Thus, to cover well with frequency grid the wide range of octaves large sized windows are necessary in the case of FFT; this affects the time resolution of the analysis. On the contrary, the use of small windows makes impossible to resolve frequencies of neighboring notes in low octaves.

The Continuous Wavelet Transformation (CWT) was introduced 15 years ago in order to overcome the limited time-frequency localization of the Fourier-Transform (FFT) for non-stationary signals and was found to be suitable in a lot of applications [6]. Unlike the FFT, the Continuous Wavelet Transformation has a variable time-frequency resolution grid with a high frequency resolution and a low time resolution in low-frequency area and a high temporal/low frequency resolution on the other frequency side. In that respect it is similar to the human ear which exhibits similar time-frequency resolution characteristics [7].

Also the scale of frequencies can be chosen as logarithmic which fits well for the note analysis. In that case the number of frequency bins is constant for each octave.

In our works an experimental "wavelet-like" function with logarithmic frequency scale was used to follow the musical note system:

$$\psi\left(x,a^{*}\right)=H_{x,m\left(a^{*}\right)}e^{jw\left(a^{*}\right)x} \qquad (1)$$

where a^{*} – relative scale of wavelet (0..1), *H(x,m)* – function of Hanning window of length *m*:

$$m\left(a^{*}\right)=L_{\max}k_{1}\cdot e^{-k_{2}a^{*}} \qquad (2)$$

$$w\left(a^{*}\right)=L_{\max}{}^{a^{*}}\Big/L_{\min}{}^{a^{*}+1} \qquad (3)$$

Here k_1, k_2 – time resolution factors (0.8 and 2.8), L_{max} and L_{min} – range of wavelet absolute scales (6000 and 14).

We have chosen this function because it has elements of windowed Fourier transform (with Hanning window) and classical wavelets. The frequency scale here is always logarithmic while the time resolution scale can be adjusted to be from liner to logarithmic. Time/frequency scale of the transform is shown on the Figure 2.

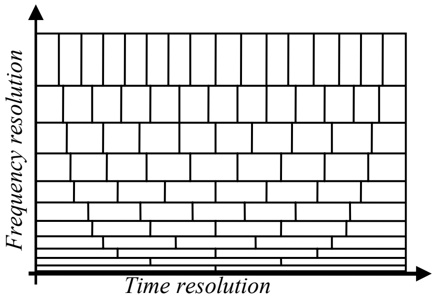

Figure 2. Time-frequency resolution of the Transform used in our work

The following example (Figure 3 a, b, c) shows two FFT representations (with different window size) and one CWT representation of a test signal containing two notes E1 and A1 playing constantly together with four notes B5 playing with a small interval of time (1/16 sec). Low octaves notes usually present bass lines which a slow changing and high octave notes playing main melody lines a much faster changing in time. The example shows either unresolved low frequency notes (a) or smoothed in time high notes (b) when the CWT spectrogram (c) is free from that problem.

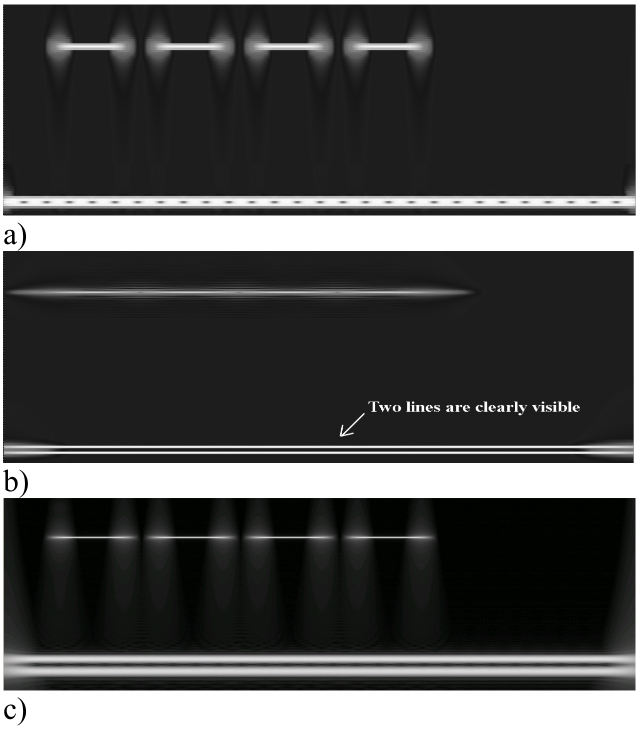

Figure 3. 3 representations of a test signal. a) FFT with small window, b) FFT with bigger window, c) CWT.

The use of CWT, however, has a negative point of costly computations.

3. F0 ESTIMATION

Numerous algorithms for F0 estimation (pitch detection) exist in the literature. [5] describes an advanced method PreFEst. Using EM algorithm, it basically estimates the F0 of the most predominant harmonic structure in the input sound mixture; it simultaneously takes into consideration all the possibilities of F0 and considers that the input mixture contains every possible harmonic structure with different weights (amplitude). Another pitch model based system is presented in [3].

In the paper [4], authors describe a computationally inexpensive scheme of transcribing monophonic and polyphonic music produced from a single instrument. The scheme is based on two steps (track creation and grouping) and uses discrete variable window-size comb-filter together with sharpening filter.

3.1. Our approach

We use a technique inspired by harmonic pitch models. The analysis procedure is divided into two parts (its diagram is shown on Figure 5). The first part consists of model generation. The model is simply a "fence" of peaks situated at the places where F0 and its harmonics 2*F0, 3*F0… etc. can be found on a CWT spectrogram. Recall that our CWT spectrogram has a logarithmic frequency scale, so the distances between corresponding harmonics on such spectrogram remain constant with the change of absolute value of F0. Only the forms of each peak are changing over the frequency scale due to the change of frequency and time resolution of wavelets. To obtain these forms we are using the CWT applied on sine waveforms with appropriate frequencies. A typical flat harmonic structure is depicted on Figure 4.

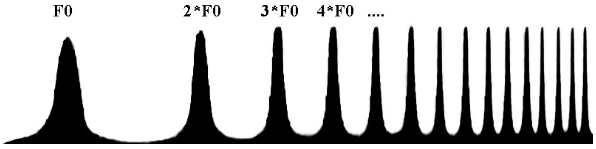

Figure 4. Harmonic structure in logarithmic scale.

The shape of the harmonic model-"fence" may be used either flat where all amplitudes are similar for all the harmonics or with raised low harmonics part (ratios 3 2 1.5 1 1 etc… for corresponding harmonics) which actually gives better results in general case. In particular, the shape of the harmonic model can be adjusted to the instrument assumed to be used in the play. In general case such assumption cannot be made.

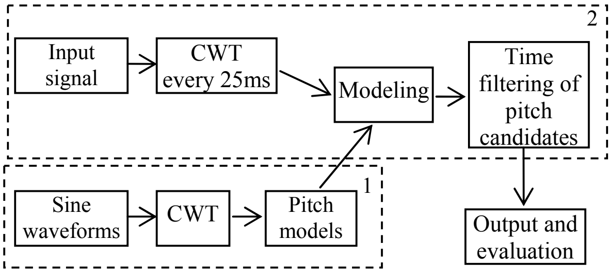

Figure 5. Transcription process

The second part of the analysis lies in analyzing of input wave signals for transcription. The input signal (16KHz, 16bit PCM) is processed by the CWT which has 1024 bins for frequencies lying in the range of 34-5500Hz every 25 ms. Obtained spectrum slices are analyzed in the following way. The above-mentioned harmonic structure is moved across the frequency scale of the CWT spectrogram slice and the correlation between the model and the spectrogram is computed. The place where the correlation has a maximum value on the spectrogram is assumed to be an F0 candidate. As it is found, the harmonic "fence" is subtracted from the currents slice of the spectrogram with the values on its peaks being taken from the actual values on the spectrogram. The procedure is repeated until no more harmonic-like structures are found in the spectrum (above the certain threshold) or the maximum number of harmonic structures to be searched defined in the algorithm is reached. We limit the maximum number of notes searched to four in our works.

The described algorithm has an advantage of its rapidity and it is working well in detection of multiple pitches with non-integer rates. However, there is a disadvantage of such algorithm. Notes with F0s being in integer rates whose partials intersect are rarely can be completely detected together. At the same time, two notes with a distance of an octave hardly can be separated, because the second note does not bring any new harmonics into the spectrum, but changes the amplitude of existent harmonics of the lower note, so some knowledge of instruments involved in the play might be necessary to resolve the problem.

Another possibility of the search for F0 candidates has been studied. Instead of successive subtractions of dominant harmonic structures found, one can use a local maximums search on the correlation curve. Every peak above a defined threshold is considered as an F0 candidate. Such approach can partly solve the problem of notes with overlapping partials while it generates "phantom" notes in one octave down to the note which actually present. With subtracting algorithm such notes never appear.

Finally, all pitch candidates are filtered in time in order to remove any noise notes with duration below a defined threshold.

4. EXPERIMENTS

The easiest way to make basic experiments in automated music transcription is to use MIDI files (plenty of them can be freely found in Internet) rendered into waves as input data and MIDI events themselves as ground truth. However, the real life results must be obtained from recorded music with true instruments and then transcribed by musical educated specialists.

In our work we use wave files synthesized from MIDI using hardware wavetable synthesis of Creative SB Audigy2 soundcard with high quality 140Mb SoundFont bank "Fluid-R3", where for example, acoustic grand piano is sampled every four notes from a real one. To make the conditions closer to reality in some tests we pass the signal over speakers and record it with a microphone.

The MIDI data itself is taken both from finished MIDI files and also from sequences played on a keyboard.

Recall and Precision measures were used to measure the performance of the note detection. Recall measure is defined as:

$$\text{Recall} = \frac{\textit{the number correct notes detected}}{\textit{the actual number of notes}} \qquad (4)$$

Precision is defined as following:

$$\text{Precision} = \frac{\textit{the number correct notes detected}}{\textit{the number of all notes deteced}} \qquad (5)$$

For the overall measure of the transcription performance, the *F1* measure was used

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \qquad (6)$$

All falsely detected notes including those with octave errors were considered as missdetected. For some tasks of music indexing as for instance tonality determination, the note basis, but not the octave number is important. For that reason, the performance of note detection without taking into account octave errors was estimated as well. In the result tables Perf.Oct is a performance of note detection without octave errors counted.

The following table gives information about the results obtained. The MONO case is a monophonic keyboard play. The POLY set is a polyphonic keyboard play. In the MIDI Classic section there are mostly piano/harpsichord polyphonic pieces (Back Fugue is played by Flute/Oboe/Clarinet/Basson, Vivaldi Mandolin used a Koto instrument). The MIDI multi-instrument case is a test set with polyphonic modern music titles played by numerous instruments including percussive.

Table 1. Precision of the note detection.

Name	Notes	Poly-phony	Performance			Perf Oct
		max / avg	Rec	Prec	F1	F1
MONO						
Piano Manual	150	1 / 1	100	100	**100**	100
Violin Manual	160	1 / 1	100	97	**98.5**	100
POLY						
Piano Manual	330	2 / 1.8	98.5	100	**99.5**	99.7
Piano Manual	214	5 / 2.2	95.8	100	**97.8**	99.1
Flute Manual	174	4 / 2	97.7	97.7	**97.7**	99.7
MIDI Classic						
Fur_Elize	924	6 / 1.6	91.1	88.7	**88.9**	95.6
Fur_Elize w/ microphone	924	6 / 1.6	88.1	86.9	**87.5**	95.4
Tchaikovsky 01	177	4 / 3.5	84.7	95.5	**89.8**	95.4
Tchaikovsky 16	186	4 / 2.6	86.5	100	**92.8**	97.2
Bach 01	687	5 / 1.7	91.1	88.7	**89.9**	98.2
Bach 03	549	5 / 2.1	98.9	91.9	**95.2**	96.8
Bach Fugue	252	5 / 2.4	83.7	76.1	**79.8**	93.2
Vivaldi Mandolin	1415	6 / 2.9	70.1	74.8	**72.4**	91.5
MIDI multi-instr POP w/ beats						
POP kminogue	2545	10 / 4.7	40.6	37.1	**38.8**	64.3
POP Madonna	2862	8 / 3.9	43.9	56.9	**49.5**	66.4
Soundtrk godfather	513	9 / 4.1	88.7	67.2	**76.5**	90.4

5. DISCUSSION

As it can be seen from the result table, the problem of *complicated* multi-instrument polyphonic music transcription remains still not very well resolved while the transcription of monophonic of polyphonic music with low number of voices and/or a single instrument has much better performance. At the same time one can use note information obtained from the partial transcription in such tasks as for instance tonality detection in polyphonic pieces or other issues not critical for missed notes and octave errors. To improve the precision results for the general case musical titles several techniques can be used as, for instance, beat detection and elimination.

6. CONCLUSION

In this paper we described a CWT-based approach of automated music transcription. Overall results show good precision rates for monophonic and polyphonic examples from classic pieces. However there are still unsatisfactory results on modern popular music with beats etc. We concentrate our future work on improving the performance of the note detection algorithm and on a direct application of the described approach to melody lines extraction and melodic similarity.

REFERENCES

[1] Abe T. et al., "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *in ICSLP 96*, pp. 1277–1280, 1996.

[2] Hu J., Sheng Xu., Chen J. "A Modified Pitch Detection Algorithm" *IEEE COMMUNICATIONS LETTERS*, VOL. 5, NO. 2, FEBRUARY 2001

[3] Klapuri A. "Pitch Estimation Using Multiple Independent Time-Frequency Windows" Proc. *1999 IEEE Workshop* on *Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 17-20, 1999

[4] Lao W., Tan E.T., Kam A.H."Computationally inexpensive and effective scheme for automatic transcription of polyphonic music" *ICME 2004*: 1775-1778

[5] Goto M. "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models" *ICASSP 2001* Proceedings, pp. V-3365-3368, May 2001.

[6] Kronland-Martinet R., Morlet J. and Grossman A. "Analysis of sound patterns through wavelet transform", I*nternational Journal of Pattern Recognition and Artificial Intelligence,*Vol. 1(2), 1987, 237-301.

[7] Tzanetakis G., Essl G., Cook P. "Audio Analysis using the Discrete Wavelet Transform" *Proc. WSES Int. Conf. Acoustics and Music: Theory* 2001 *and Applications (AMTA 2001) Skiathos, Greece*