Toward Automatic Transcription - Pitch Tracking In Polyphonic Environment

Literature survey by:

Keerthi C. Nagaraj

Date: 19th March 2003

Submitted to:

Professor Brian L Evans

EE381K- Multidimensional Digital Signal Processing

ABSTRACT

Accessing the perceptually relevant information contained in music signals is a classical multidimensional signal processing problem. Applications like automatic transcription, content based audio classification, music indexing etc. require identification and tracking of notes played by polyphonic instruments under noisy conditions. An important step in the direction of generic automatic transcription is the process of pitch tracking. The ease of pitch tracking depends on the clarity with which the instrument can be recognized in "ecological" music signals. Intuitively, identification of the perceptually significant parameters of an instrument must facilitate the process of pitch tracking of polyphonic instruments. Popular approaches for pitch tracking are use of sinusoidal models and auditory models. Previous methods of pitch tracking are discussed and analyzed for performance under polyphonic conditions. Possibility of integration of the knowledge about non-sinusoidal features music, and the process of pitch tracking is explored.

Introduction:

Transcription of music is defined to be the act of listening to a piece of music and of writing down musical notation for the notes that constitute the piece. According to musical practice, this requires extraction of notes, their pitches, timings, and classification of the instruments used. The corresponding sub-problems of automatic transcription are pitch tracking, rhythm detection and analysis of musical instruments. In this literature survey, I shall primarily concentrate on the first of these problems, i.e. tracking of pitch in polyphonic environment.

Firstly, pitch detection methods used for speech and monophonic signals are not appropriate as such to the detection of multiple pitches in music, since very often the frequency relations of simultaneous sounds in music either make several sounds appear as a single coherent sound, or a non-existent 'ghost' sound strongly arises just because of the joint effect of the others. Consider, as an example, two sounds that are an octave apart in pitch from each other: the harmonic partials of the higher sound match perfectly the positions of even partials of the lower one, making it appear as a single sound, and turning the separation of the sounds into an even theoretically ambiguous problem.

Secondly, as observed by Bregman [1], the human auditory system has a tendency to segregate a sound mixture to the physical sources, but orchestration is often called upon to oppose these tendencies and force the auditory system to create a single chimerical sound, which would be irreducible into perceptually smaller units. This is a problem in music transcription, as will be seen when an attempt is made to resolve a polyphonic musical.

Previous work:

Polyphonic pitch tracking algorithms have been around for more than 20 years. However, flexibility in the number of simultaneous harmonic tones has increased only recently. There are several approaches to this problem widely varying in principle and implementation [2, 3]. Such methods have been successfully employed in transcription systems [4,5]. A few popular methods are discussed here.

Sinusoidal modeling – Constant-Q Transform (CQT)

The transform is similar to Short Term Fourier Transform (STFT) but, the window length varies as function of frequency so that a constant number of periods are within the window at each frequency. The CQT has a geometrically distributed frequency resolution. It is better suited for music signal analysis, since we see a mirror of such a distribution in frequency resolution in direct comparison to the music scale. The STFT is tailored to match the contraction of the complex exponential as:

$$S^{CQ}(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s(\xi)h\left((t-\xi)\frac{\omega}{\omega_0}\right)\exp(-j\omega\xi)\,d\xi$$

Here the *h(.)* is the sliding window and ω_0 is the reference frequency at which the window is unaltered. If the fundamental frequency is estimated by some peak picking technique, then this transform gives the peaks indicating the pitch of the note. Further improvements to this basic algorithm have been proposed in [6] like faster calculation [7] and making the transform sensitive to phase changes [8]. This is primarily to accommodate the effects of tremolo and vibrato features of musical signals and to eliminate the limitation that the signal should stick to equally tempered scales.

Auditory modeling – Pitch perception model

In the pitch perception model proposed by Ellis [9], which is an enhancement over [10], the audio signal is first decomposed into frequency bands by a model of bassilar membrane mechanics (implemented by a gammatone filter bank). Each filter channel is further processed by a model of inner hair cell (IHC) dynamics approximated by half-wave rectification followed by smoothing (to eliminate carrier frequencies above the phase-locking limit of the hair cells) and onset enhancement. The output of each IHC is analyzed by short-time autocorrelation, yielding an estimate of periodic energy in each filter channel as a function of lag, or inverse pitch. Finally, the autocorrelations are summed across the filter bank channels, and the lag with the resultant largest peak is chosen as the "pitch percept". The advantage of this model is that it accounts not only for phenomenon and several of the "weak pitch" phenomena.

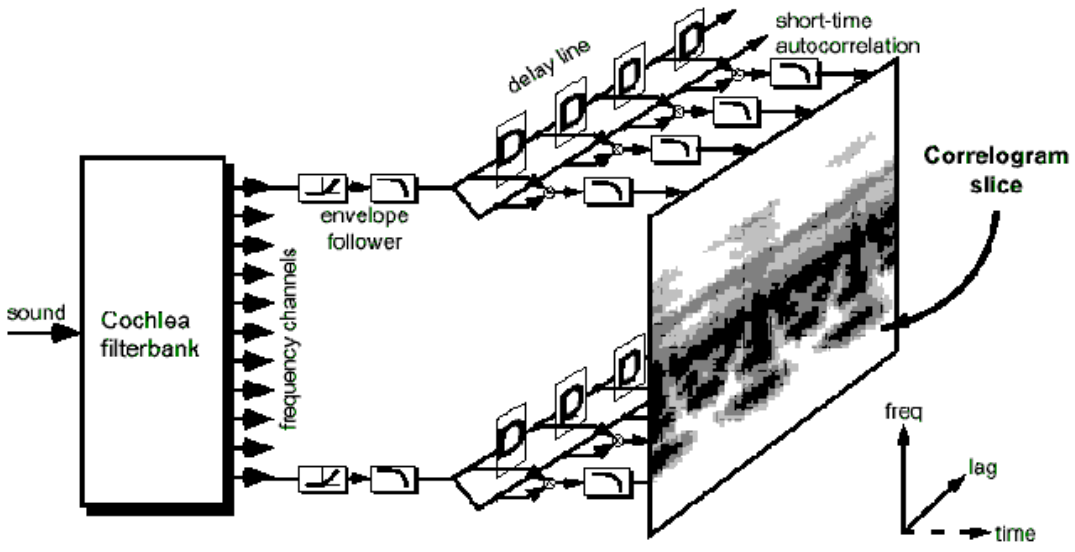

Figure (1) Pitch perception model which approximates the behavior of human cochlea

Ellis computes a "log-lag" correlogram, where the three axes of the correlogram volume are: filter channel frequency, lag (or inverse pitch) on a logarithmic scale and time. The output of each frequency/lag "cell" is computed by a simple filter structure, as shown in Figure (1). To compute the "pitch percept", Ellis normalizes the output of each frequency/lag cell by the energy in that filter bank channel (given by the output for that channel at zero lag), and averages across the filter bank channels, yielding what he calls the *summary autocorrelatio*n, or *periodogra*m. The log-lag (log-pitch) axis is an improvement over standard correlograms in that it more closely relates to the variation in pitch resolution ability of humans as a function of pitch.

Computationally efficient variant of the pitch perception model [11]

The method proposed by Karjalainen et al, [11] is a simplification of the perception model explained above. They claim that the perceptual difference between a signal processed by the filter bank of [10] and a highly simplified version of the same, i.e. [11], is not much. They claim that instead of using a filter bank, of more than 54 channels (one per half-tone), one could get similar results by using a simplified version of this filter bank constituted by only 2 channels.

The significant differences from the previous model are that some preprocessing of signal is necessary before it is fed into the channels and some post-processing is employed to remove ambiguities between very close prospective pitch candidates. The original Summary Autocorrelation Function (SACF) curve is first clipped to positive values and then time-scaled (expanded) by a factor of two and subtracted from the original clipped SACF function, and again the result is clipped to have positive values

only. This removes repetitive peaks with double the time lag where the basic peak is higher than the duplicate. It also removes the near-zero time lag part of the SACF curve. This operation can be repeated for time lag scaling with factors of three, four, five, etc., to remove higher multiples of each peak. This function is called SACF Enhancer. The proposed model is diagrammatically represented in figure (2).

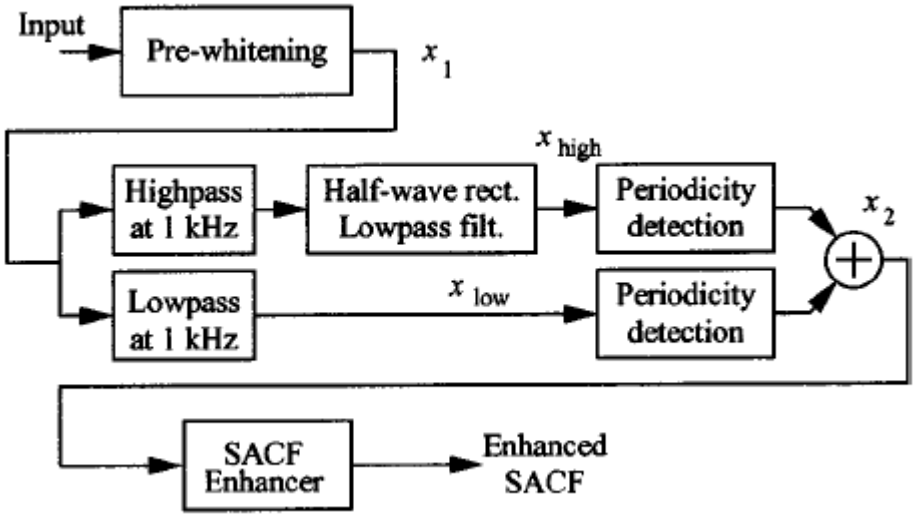

Figure (2). A computationally efficient variant of the pitch perception model of [10]

Analysis & Conclusion:

A general discussion of different approaches to pitch tracking in polyphonic environment was done. The approaches are observed to be tradeoffs between computational efficiency and auditory relevance. The efficient implementation of CQT [7] relies on FFT in such a way that it practically combines several bins of FFT to produce a logarithmic frequency resolution. It therefore fails to interchange the weaker frequency resolution at the high end to a better time resolution, which would be desirable to model human hearing. On the other hand, straightforward calculation of CQT is too

laborious to be practically useful. A variant of Q transform has also been predominantly used [12], which claims to overcome a few restrictions posed by CQT.

The chief drawback of using perceptual model is the computational complexity. Some authors contend that there are several assumptions made in modeling the human auditory system, the effect of which has not been analyzed in depth. So whether the use of such a complicated model for the purpose of pitch tracking is justified, is debatable. It is observed that pitch perception models apply a short-time autocorrelation to the outputs of a certain auditory filter bank. The filter bank is not intended to provide a sufficiently precise frequency resolution, but it is the subsequent autocorrelation that is used for tracking periodicity in the signal. Utilization of autocorrelation is a problem here, since autocorrelation fuses information on perceptual grounds in such a way that it prevents a separate treatment of each harmonic partial that is considered necessary in order to resolve musical polyphonies. This turns out to be a disadvantage in music transcription.

The actual analogy of the system [11] with auditory model is not strong enough for rigorous comparisons. It only facilitates interpretations of the results of analysis in terms of human pitch perception. In depth analysis in terms of pitch trajectories and resolution is still awaited.

Future work:

Comparative analysis of the pitch tracking mechanisms with the main aim of application to music transcription will be undertaken. A working re-synthesis model will be designed to compare the performances of different models. Possibility of enhancements in the pitch tracking method with the addition of prior knowledge about the musical instruments being played will be explored.

References:

[1] A. S. Bregman, *Auditory Scene Analysi*s. Cambridge, MA: MIT Press, 1990.

[2] K. Kashino and H. Tanaka, “A sound source separation system with the ability of automatic tone modeling,” *Int. Computer Music Conf*., pp.248–255, 1993.

[3] T. Virtanen, A.Klapuri, "Separation of Harmonic Sounds Using Linear Models for the Overtone Series", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2002

[4] K. D Martin, “Sound-Source Recognition: A Theory and Computational Model”, PhD Dissertation, Mass. Inst. Technol., Media Lab Perceptual Computing, Cambridge, 1999.

[5] K. Kashino, K. Nakadai, T Kinoshita, H. Tanaka, “Application of Bayesian Probability Network to Music scene Analysis”, Proceedings of the International Joint Conference on AI, CASA workshop, 1995.

[6] J C.Brown, ``Calculation of a Constant Q Spectral Transform" J. Acoust. Soc. Am. 89 425-434, 1991.

[7] J C. Brown, M S. Puckette, "An Efficient Algorithm for the Calculation of a Constant Q Transform", J. Acoust. Soc. Am. 92, 2698-2701, 1992.

[8] J C. Brown, M S. Puckette, "A High Resolution Fundamental Frequency Determination Based on Phase Changes of the Fourier Transform" , J. Acoust. Soc. Am. 94, 662 – 667, 1993

[9] D P W. Ellis, “A perceptual representation of sound for source separation”, J. Acoust. Soc. Am., Volume 91, Issue 4, p. 2334, April 1992

[10] R. Meddis and L. O’Mard, “A unitary model for pitch perception,” *J. Acoust. Soc. Amer*., vol. 102, pp. 1811–1820, Sept. 1997.

[11] T. Tolonen*, M* Karjalainen, “A Computationally efficient Multipitch Analysis Model”, IEEE Transactions On Speech And Audio Processing, Vol. 8, No. 6, November 2000

[12] A. Klapuri, “Automatic Transcription Of Music”, Department Of Information Technology, Tampere University Of Technology, MSc Thesis, April 1998