Toward Automatic Transcription - Pitch Tracking in Polyphonic Environment

Keerthi C. Nagaraj

Submitted to: Dr. Brian L. Evans

Course: EE381K Multi dimensional Digital Signal Processing

05/08/2003

Abstract

With growing traffic of multimedia on the Internet the need for better coding techniques for music, importance of efficient music indexing and MIDI conversion cannot be over stressed. For any such application automatic music transcription is a key process. The development of robust automatic music transcription systems depends on the efficiency with which the polyphonic components of music can be resolved. An efficient pitch estimation technique that handles real world signals is presented. A Bayesian probability based model is used to weigh the pitch candidates in order to select the best pitch estimate. This method gives the flexibility to impose constraints on the time varying characters of the incoming data. In comparison to the conventional autocorrelation based method, we find that the method to be more consistent across pitch transitions although computationally expensive.

1 Introduction

In the field of music signal processing, one of the most intriguing problems has been of automatic music transcription. Its applications include Karaoke, Music-Minus-One systems, content based audio retrieval and music indexing. From a musical point of view, the most important attribute of harmonic sounds is their fundamental frequency. Thus the basic requirement for a music transcription system is to detect the fundamental frequencies of sounds, and attribute them to the corresponding sources of generation. A key component in any automatic music transcription system therefore is a robust pitch tracker.

Pitch tracking of monophonic signals is practically a solved problem. However, monophonic pitch trackers fail to segregate the sound sources when the signal is polyphonic, as is the case in any real world music recording. A reliable algorithm for multi-pitch contour tracking is critical for many auditory processing tasks. However, due to the difficulty of dealing with the interference from noise intrusions and mutual interference among multiple harmonic structures, the design of such an algorithm is very difficult and still an open arena for research.

Musical signals can be well represented as a sum of harmonically related sinusoids motivated by consideration of the sound generation mechanisms of many musical instruments. We consider the problem of estimating the harmonic model parameters in a Bayesian framework which has the potential to incorporate a priori knowledge about the structure of the data and of its parameters. Constraints can be imposed upon the time-domain variation of model parameters to reflect the high correlation of frequency and harmonic structures over time. A pitch estimation approach, which uses knowledge in this format, in order to resolve the simultaneously occurring partials is presented. Unlike the temporal correlation approach, this method integrates data across frames to verify the sanity of pitch estimate. Comparison is made with the conventional Summary AutoCorrelation Function method. Some observations and avenues for future work are discussed.

2 Background

Several pitch tracking algorithms have been proposed for tracking musical signals. Brown[1] proposed a method which uses Constant Q spectrum for recovering the spectral information from the signals . Several enhancements have been proposed to this algorithm[2, 3]. However, this algorithm is more effective in monophonic environment.

Auditory modelling of musical signals is another approach to handle polyphonic pitch tracking[4]. This method is popular since, it not only explains a wide range of psychoacoustic phenomena in hearing (such as 'missing pitch' phenomenon), but also tries to organize sounds to their sources of production. An automatic transcription system using this method and a 'blackboard' architecture to organize data has been proposed by Martin[5]. However, this approach does not use auditory cues which are necessary to resolve harmonics.Kashino and Tanaka [6] have proposed an approach based on automatic tone modeling. This approach extracts tone models from the signal which is being analyzed. There are several limitations on this method of pitch tracking. Since, the model is self learning, there are stricts constraints on the type of data that it can handle.

In the method proposed in this report, an auditory model is used as the front end for signal processing. It acts as a tool to extract data about fundamental frequency, its harmonics, and their corresponding amplitudes. A simple bayesian probability network is used to organize this information as hyper data. A probabilistic weightage is given to all the possible harmonics and the best pitch estimate is selected.

3 Implementation

3.1 Auditory Model Front End:

The model proposed by Meddis and O'Mard[7] is popular for accurately representing the actual behaviour of humar cochlea and is especially sensitive to "missing pitch" phenomenon. It consists of a filterbank, followed by a nonlinearity function and a periodicity detection

stage. Each bandpass filter is implemented by four cascaded second order filter sections, which realize an 8th order filter with a "gammatone" impulse response. The bandwidth of each filter is set to *Equivalent Rectangular Bandwidth* (ERB) of the cochlear tuning curve. The nonlinearity of half-wave rectification and smoothing are added to the filter bank output to simulate the unidirectional response and 'phase locking' behavior of the inner hair cells of the cochlea. This signal output is fed in to a short averaging window, and then autocorrelated. Defining a windowing function $w(t)$, we can summarize the operation as:

$x_w(t, t_0) = x(t)w(t - t_0);\ R_{x_w x_w}(\tau, t_0) = \int_{-\infty}^{\infty} x_w(t, t_o)x_w(t - \tau, t_0)dt$

At a fixed time, summing up the autocorrelation outputs across all the cochlear channels for each time lag, a peak is detected at a lag at which, the auditory system percieves a pitch. The inverse of this time lag is pitch frequency. A block diagram of the complete operation is given in **Fig. 2**.

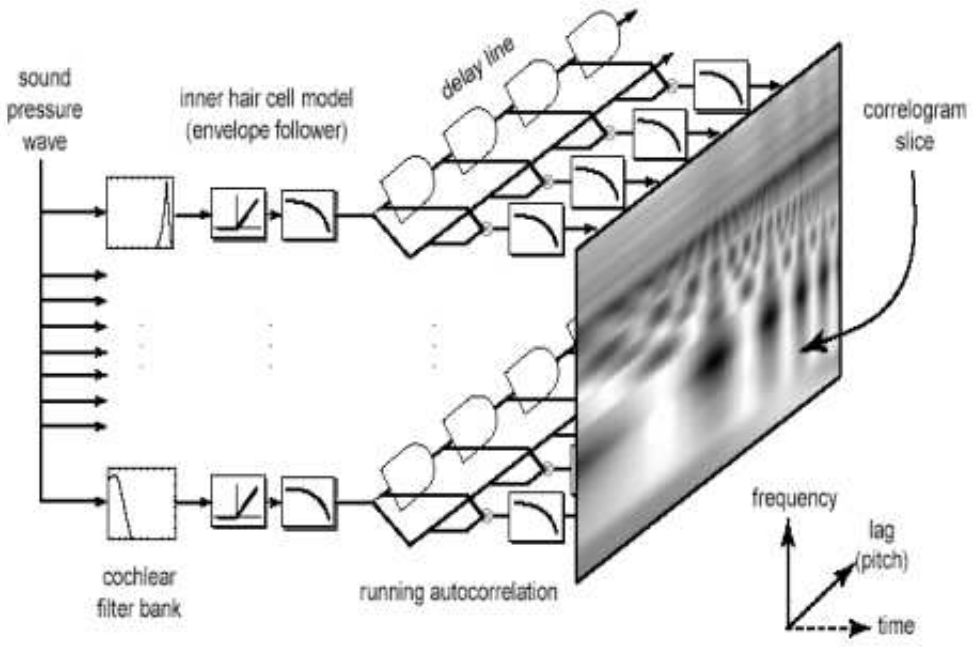

Fig. 2. Complete operation of auditory model front end [4] with gammatone filterbank of 54 channels followed by halfwave rectification and autocorrelation

3.2 Bayesian Modeling of Data

The data is segmented into frames d_i of length N, chosen to make the frame duration around 20 ms, during which time we assume the data is stationary. The model is constructed as a sum of unknown number of concurrently sounding *notes*, where the parameters of note q in frame i are: fundamental frequency ω_i^q, number of harmonics $H_i^q$and harmonic amplitudes b_i^q. A maximum limit of Q notes is imposed, but each note can be switched in to or out of the model via a binary indicator variable γ_i^q, which is estimated along with the other parameters. Hence the model order selection is implicitly carried out within the estimation process. Each note is expressed as a general linear model, in terms of a harmonic basis matrix G_i^q, the amplitudes b_i^q, and an error term e_i which is assumed gaussian, independant and identically distributed with variance $\sigma_{e_i}^2$.

$$d_i = \Sigma_{q=1}^{Q} \gamma_I^q G_i^q b_i^q + e_i$$

$$G_i^q = [s(\omega_i^q)...s(H_i^q\omega_i^q)c(\omega_i^q)...c(H_i^q\omega_i^q)]$$

$$s(\omega) = [sin(\omega t_1)sin(\omega t_2)...sin(\omega t_N)]$$

$$c(\omega) = [cos(\omega t_1)cos(\omega t_2)...cos(\omega t_N)]$$

Denoting the notes in a frame q as $\Theta_i^q = \{\gamma_i^q, \omega_i^q, H_i^q, b_i^q\}$, the likelihood for a frame i is given by

$$p(d_i|\{\Theta_i^q; q = 1...Q\}, \sigma_{e_i}^2) = \frac{1}{(2\pi\sigma_{e_i}^2)^{\frac{N}{2}}} exp\left[-\frac{||e_i||^2}{2\sigma_{e_i}^2}\right]$$

A least-squares or maximum likelihood method would seek to maximize the above equa-

tion, but we here we pose the model in a bayesian framework which enables us to impart prior information into the model via *a priori* probability densities on the parameters, and which also provides a basis for probabilistic model selection.

Fig. 3 shows a graphical model representing the linear model for a set of $N_f$frames of data (termed a *block*). The *hyperparameters* $\{\Delta_\Theta^q\}$are any prior knowledge we may use about the notes parameters. Block hyperparameters decide the prior distributions of the note parameters and assist in converging to a local best-fit model.

The hyperparameters in this implementation are the average pitch ν^q, variation $\sigma_{\omega^q}^2$across $N_f$frames and the indicator function $\Gamma^q$of the particular note in the block. We can observe that instantaneous frequency in a frame, ω_i^q depends on $\{\nu^q, \sigma_{\omega^q}^2\}$. The parameters are assigned prior distribution which are same as listed in [8]. Since, the level of data hierarchy used in the current implementation is not very deep, only a uniform prior of the harmonics H_i^q and a uniform prior for the probability of a note being played in the frame are used.

The joint posterior distribution for all parameters is obtained as follows:

$p(\{\Theta_i^q\}, \{\Delta_\Theta^q\}, \{\sigma_{e_i}^2\}|\{d_i\}) \propto p(\{\Theta_i^q\}, \{\Delta_\Theta^q\}, \{\sigma_{e_i}^2\}) \times \Pi_{i=1}^{N_f} p(d_i|\{\Theta_i^q\}, \sigma_{e_i}^2)$

After marginalization of amplitudes and error variances, the posterior distribution of a model given the observed data is given by:

$p(\mathcal{M}_\mathbb{Q}, \{\tilde{\theta}^q\}_\mathbb{Q}|d, \sigma^2) \propto p(\{\tilde{\theta}^q\}_\mathbb{Q}|\mathcal{M}_\mathbb{Q}) p(\mathcal{M}_\mathbb{Q}) \frac{p(\{\tilde{\theta}^q\}_\mathbb{Q}|\mathcal{M}_\mathbb{Q}) p(M_\mathbb{Q})}{(1+\sigma^2)^{\frac{M}{2}} [\mathbf{d^t P_C d} + 2\beta]^\varepsilon}$

$\mathbf{P}_c = \mathbf{I}_\mathcal{N} - \frac{\sigma^2}{(1+\sigma^2)} \mathbf{G_c} (\mathbf{G_c^t G_c})^{-1} \mathbf{G_c^t}$

Where $M = 2\sum_{q \in \mathbb{Q}} H^q$, $\tilde{\theta}^q = \{\omega^q, H^q\}$,$\varepsilon = N/2 + \alpha$ and $p(\mathcal{M}_\mathbb{Q}) = p(\{\Gamma^q\}_{1:Q})$. A composite basis matrix $\mathbf{G_c}$is formed from the concatenation of the included basis matrices $q \in \mathbb{Q}$. The error variance hyper parameters $\alpha$and $\beta$are chosen to be small. This probability is maximised numerically by repeated iterations. A stream of dependent samples from the posterior are generated and used for estimation.The implementation consists of iteratively sampling for each note and minimizing the error residual r_i^q in terms of note parameters:

$r_i^q = d_i^q - \sum_{q' \neq q} \gamma_i^{q'} G_i^{q'} b_i^{q'}$

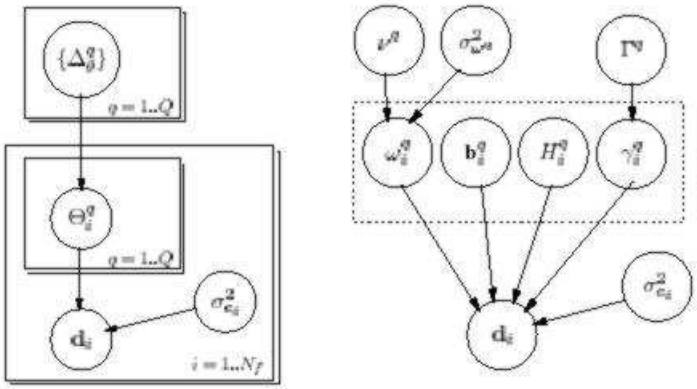

Fig 3. Bayesian modeling of data in pitch estimation process. The second diagram shows how the hyper parameters are related in hierarchy to the data.

4 Simulation and results

The estimation algorithm is a two stage process. In this first stage of the algorithm, the global kernels propose a state space move for note parameters for a particular note q but across all frames $i = 1...N_f$, whereas in the second stage, local kernels propose a state change locally within a frame. The algorigthm can be summarized as follows:

1. Assign prior distributions for ω_i^q, H_i^q.
2. For each frame:
   (a) Calculate the summary autocorrelation function
   (b) Perform peak picking and select possible fundamental frequency candidates above a threshold
   (c) Multiply the peaks with the reliability vector
   (d) pass the output through a weighted median filter to get the best pitch frequency corresponding to highest peak.

(e) Find error by comparing the evolving model and observed data

(f) Update the reliability vector enhancing the weight of the most recently chosen peak.

(g) Repeat the process to minimize the error r_i^q

A few samples of music with clearly distinguishable pitches and very less vibrato were chosen. The above discussed algorithm was compared with the conventional Summary AutoCorrelation Function (SACF) method. Given a good starting (*e.g.,* results from previous block), the convergence is rapid typically 100 -120 iterations. It is to be noticed that such iterations are not performed in the case of SACF method. Hence the SACF method is computationaly efficient. It was observed that since the SACF method handles data frame by frame, the pitch transitions are not correctly tracked by it. we can see erroneous pitch detections.The proposed algorithm however, applies salience information before declaring pitch estimates,hence eliminates the possibility of incorrect pitch estimates as can be seen in **Fig. 4**.

5 Conclusion

In this project, a probabilistic approach was applied to estimate pitch in polyphonic environment. Simple hyperparameters were used to wiegh the probability of pitch candidates. The algorithm is more efficient in tracking constant and slow variation of pitches. It is less error prone at pitch transitions in comparison to SACF method. It can give constant pitch estimates which come in more handy in the context of music transcription. However, the algorithm smears the strong attack transients due to integration across frames. It can be inferred that if the pitch detected by such an algorithm were used in music synthesis, then the music would be "mellowed" down. It was noted that addition of hyperparameters increases the computation exponentially. Hence, there is a need to wiegh the feasibility adding each new parameter. In this implementation, low level of hyperparameters (restricted to note

level) has been used. However, pitch estimation can be made more efficient by the use of higher level data such as chords, and permissible chord transitions.

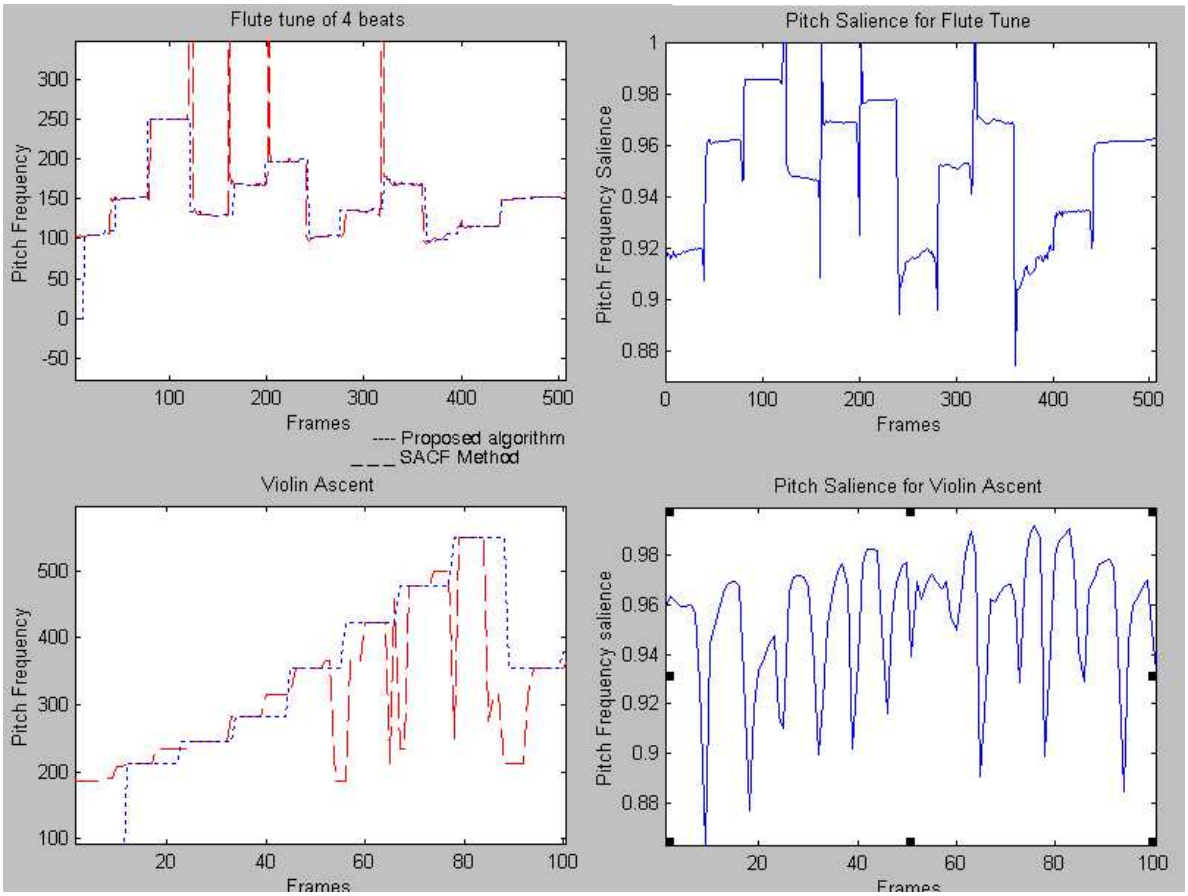

Fig. 4. Experimental runs of the proposed algortihm with samples of flute and violin. Pitch frequency and corresponding salience vectors are plotted

References

[1] J.C. Brown, "calculation of a Constant Q Spectral Transform," *J. Acoust. Soc. America.* Vol. 89, pp 425-434, 1992.

[2] J. C. Brown and M. S. Puckette, "An Efficient Algorithm for the Calculation of a Constant Q Tra nsform," *J. Acoust. Soc. Am.* Vol. 92 pp. 2698-2701, 1992.

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

[4] D. P. W. Ellis, "A Perceptual Representation of Sound for Source Seperation," *J. Acoust. Soc. Am.* Vol. 91, Issue4, pp 2334, April 1992.

[5] K.D Martin, "A Blackboard System for Automatic Transcription of Simple Polyphonic Music," Tech. Rep. No. 385, MIT Media Lab, Perceptual Computing Section, Massachussetts Institute of Technology, Massachussetts, CA July 1996

[6] K. Kashino and H. Tanaka, "A Sound Source Seperation System with the Abilityof Automatic Tone Modeling", *Proc. of International Computer Music Conference.* pp 248-255 Aug. 1993

[7] R.Meddis and L. O'Mard,"A Unitary Model for Pitch Perception," *J. Acoust. Soc. Amer.,*Vol. 102, pp 1810-1820, Sept. 1997

[8] P. J. Walmsley, S. J. Godsill and P. J. W. Rayner, "Polyphonic Pitch Tracking Using Joint Bayesian Estimation of Multiple Frame Parameters," *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 17-20,1999

[9] K. D. Martin, "Sound-Source recognition: A Theory and Computational Model," PhD Thesis, Massachusetts Institute of Technology, Massachusetts, CA, June 1999