Predominant-F0 estimation using Bayesian harmonic waveform models

Emmanuel Vincent, Mark D. Plumbley

PREDOMINANT-F0 ESTIMATION USING BAYESIAN HARMONIC WAVEFORM MODELS

Emmanuel Vincent and Mark D. Plumbley
Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom
emmanuel.vincent@elec.qmul.ac.uk

ABSTRACT

This paper describes a predominant pitch extraction system based on a family of Bayesian harmonic models. These models represent the short term waveform of the observed signal as a sum of harmonic partials and a residual noise. The amplitudes of the partials are modelled by a prior learnt on a training set, whereas the residual is modelled by a psycho-acoustically motivated prior. Efficient algorithms are provided to estimate the best fundamental frequency according to the MAP criterion. The performance of the method is evaluated in the framework of MIREX 2005 [1].

Keywords: Melody, Bayesian, sinusoidal model.

1 INTRODUCTION

Listeners tend to perceive music as a set of auditory objects (notes from instruments or singers, natural or electronic sounds, background noise) with various characteristics (instrument class, singer identity, playing style, pitch, loudness, onset/offset time, *vibrato*, *crescendo*, *etc*). Not all objects are equally important: melody notes, *i.e.* notes played by the lead instrument/singer, are more relevant than background noise for instance. The automatic detection of important objects allows to compress the information contained in music signals to a symbolic or a parametric description, that can be used for content-based retrieval or low bitrate encoding.

This paper focuses on predominant-F0 estimation, that is estimating the pitch of the most perceptually salient note at each instant within a music signal. When the lead instrument is playing this note belongs to the melody, otherwise it is part of the accompaniment. In the following, melody and accompaniment segments will be processed in the same way, but the performance of the method will be evaluated on melody segments only.

Many successful methods have been proposed to solve this problem in the simple monophonic case, *i.e.* when a single note is played at a time. Comb-filter methods subtract hypothesized pitches from the signal using a comb-filter and select the pitch which results in the smallest residual. Other similar methods use autocorrelation or difference functions [2]. These methods fail in polyphonic signals containing simultaneous notes with overlapping harmonics, because pitches which do not correspond to actual notes being played may still produce a small residual when their harmonics belong to different notes. Recently, one author has proposed to group peaks of the short-term spectrum of the signal using a prior model of their frequencies and amplitudes learnt on solo excerpts of the lead instrument [4]. Other authors have proposed to perform a full polyphonic transcription using note spectral templates and to select the most powerful note in a second step [5].

In this paper, we address the predominant-F0 estimation problem in a Bayesian framework using a family of probabilistic waveform models. These monophonic models are adapted from the polyphonic model proposed previously by the authors for object coding purposes [7]. They represent the short term waveform of the observed signal as a sum of harmonic partials and a residual noise. They are also similar to the method described in [4], except that they do not rely on a particular instrument and that they work in the waveform domain instead of the spectral domain. The expected advantage is that spectral peaks corresponding to background noise cannot be mistaken with spectral peaks corresponding to actual notes, and that the transcribed melody can readily be resynthesized from the estimated sinusoidal parameters. The results also give an indication whether our object coding method encodes melody prioritarily compared to other objects.

The rest of the paper is organized as follows. We describe the proposed waveform model in Section 2 and we present evaluation results in Section 3. We conclude by pointing out further research directions in Section 4.

2 BAYESIAN MELODY INFERENCE

2.1 Proposed waveform model

Existing Bayesian harmonic waveform models [1, 9] suffer some drawbacks for predominant-F0 detection. Firstly the number of partials per note follows a sparse prior which does not depend on the pitch of the note, which results in upper partials not being taken into account for the transcription of low pitch notes. Secondly, the distri-

[1] Second Music Information Retrieval Evaluation eXchange. URL: http://www.music-ir.org/mirexwiki/

bution of the residual (white or autoregressive Gaussian noise) does not correspond to the auditory significance of events. Also the parameters of these models are estimated by computationally intensive particle filtering methods.

The generative model we propose writes as follows. Let x_t be the t-th frame of the observed signal x defined by $x_t(u) = w(u)x(tS+u)$ where w is a window of length N and S is the stepsize. We express x_t as

$$x_t(u) = \sum_{h=1}^{H_t} a_{ht} w(u) \cos(2\pi f_{0t} h u + \phi_{ht}) + e_t(u), \quad (1)$$

where f_{0t} is the pitch of the predominant note on this frame, (a_{ht},ϕ_{ht}) are the amplitude and phase of its h-th partial and e_t is the residual signal. We compute the complex Fourier transform of this residual for positive frequencies $0 \leq f \leq N/2$ by $\tilde{e}_{tf} = \sum_{u=0}^{N-1} e_t(u) e^{-2i\pi fu/N}$.

2.2 Local inference

A first way to use this model is to estimate the predominant-F0 on each frame separately by setting local priors on the parameters. The pitch of the predominant note f_{0t} is associated with a fixed latent pitch F_{0t} belonging to the discrete MIDI semitone scale. We set a multinomial prior on F_{0t} and we define the number of partials H_t such that $H_t F_{0t}$ is just below the Nyquist frequency. The prior for f_{0t} is set to a log-Gaussian

$$P(\log f_{0t}) = \mathcal{N}(\log f_{0t}; \log F_{0t}, \sigma^f), \quad (2)$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ is the univariate Gaussian density of mean μ and standard deviation σ. The amplitudes of the partials are described as the product of a fixed normalized spectral envelope (m_h), a latent log-Gaussian amplitude factor r_t and a log-Gaussian residual, *i.e.*

$$P(\log a_{ht} | r_t) = \mathcal{N}(\log a_{ht}; \log(r_t m_h), \sigma^a), \quad (3)$$

$$P(\log r_t) = \mathcal{N}(\log r_t; \mu^r, \sigma^r). \quad (4)$$

The phases of the partials are assumed to be uniformly distributed

$$P(\phi_{ht}) = 1/2\pi. \quad (5)$$

Finally the prior on the residual is motivated by psychoacoustical properties. We define the excitation power of the signal in the auditory band centered at frequency f on frame t by $E_{tf} = \sum_{b=0}^{N/2} v_{fb} |\tilde{x}_{tb}|^2$ where $(\tilde{x}_{tb})$ is the complex Fourier transform of x_t and (v_{fb}) are coefficients modeling the frequency spread of the auditory band. We set

$$P(\tilde{e}_{tf}) = \mathcal{N}(\tilde{e}_{tf}; 0, \sigma^e E_{tf}^{(1-\alpha)/2} g_f^{-\alpha/2}), \quad (6)$$

where g_f is the frequency response of the outer and middle ear at frequency f. The meaning of this prior depends on the value of α. When $\alpha = 1$, it models the quantitative importance of the signal in each auditory band as proportional to its power. This is similar to the usual Gaussian white noise prior, except that the frequency response of the ear is taken into account to decrease the importance of components below 200 Hz that are less well perceived. Lower values of α give more importance to low power frequency bins. For instance, $\alpha = 0.5$ weights auditory bands proportionally to their amplitude, $\alpha = 0.25$ proportionally to their loudness and $\alpha = 0$ weights all auditory bands equally. Recent results in perceptual audio coding [3] suggested that $\alpha = 0.25$ was a promising value for object coding purposes.

For each possible value of the discrete pitch F_{0t}, maximum *A Posteriori* (MAP) estimates of all the parameters are obtained by an approximate second order Newton method. Then the posterior probability of (F_{0t}, f_{0t}, r_t) is computed by integrating over $(\log a_{oht})$ and (ϕ_{oht}) around their optimal values using the approximate Laplace integration method [6]. The value of f_{0t} for which this posterior probability is maximal is selected as the predominant-F0.

2.3 Temporal smoothing

The performance of the method may be improved by replacing the multinomial prior on F_{0t} by a first order Markov chain which imposes a temporal persistence prior. Decoding is then addressed by a standard Viterbi algorithm, using the local posterior probabilities estimated previously. The parameters of the Markov model are the initial probability of each discrete note, the mean duration of a note and the probability of each discrete interval between two successive notes.

3 EVALUATION

3.1 Parameter learning

We evaluate the performance of our model in the range between MIDI 45 and MIDI 84. Signals are downsampled to 22.05 KHz and frames are computed with half-overlapping Hanning windows of length 1024 (46 ms). The parameters of the waveform model (m_h), σ^a, μ^r and σ^r are learnt for each discrete pitch using the 20 sound excerpts and ground truth transcriptions of predominant-F0 on melody frames from MIREX 2004[2]. The parameters of the Markov chain are learnt using the ground truth symbolic melody transcriptions of the same files. Finally σ^f is set to 0.1 and σ^e is chosen (depending on α) in order to maximize the performance of the method on this training set.

3.2 Preprocessing with modified YIN

Because of the computationally intensive nature of our method, we select a few F0 candidates in a preprocessing step using a modified YIN method. The standard YIN [2] consists in computing the difference function $d_t(d) = \sum_{u=0}^{N-1} |w(u)(x(tS+u) - x(tS+u+d))|^2$ and in setting $f_{0t} = 1/d_{0t}$ where d_{0t} is the first local minimum of d_t below an adaptive threshold. This method is not reliable on highly polyphonic files: when the threshold is low most frames are classified as accompaniment, but when it is high the first local minima correspond to multiples

[2] URL: `http://ismir2004.ismir.net/ISMIR_Contest.html`

of the actual fundamental frequency. In the spirit of our model, we modify YIN by computing the difference function in the frequency domain as

$$d_t(d) = \sum_{f=0}^{N/2} g_f^{\alpha/2} E_{tf}^{(\alpha-1)/2} |(1 - e^{2i\pi fd/N})\tilde{x}_{tf}|^2 \quad (7)$$

and by setting $f_{0t} = 1/d_{0t}$ where d_{0t} is the global minimum of d_t. Instead of this single F0 value, several F0 candidates may be selected by finding the lowest local minima of d_t.

Taking into account the frequency response of the outer and middle ear when $\alpha = 1$ greatly improves the performance of the preprocessing. For instance with a single candidate a performance of 70.8% is obtained on the training set instead of 56.4% only. Using lower values of α further improves the performance with four candidates or more. The detailed performance of this preprocessing step is described in Table 1. In the following five candidates were selected on each frame with $\alpha = 0.5$ (or less if the difference function did not exhibit enough local minima).

Table 1: Percentage of melody frames in the training set for which the ground truth value of F0 is among the candidate values (1/4 tone tolerance).

Candidates	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.25$	$\alpha = 0$
1	70.8%	59.4%	51.1%	54.3%
5	86.4%	88.8%	88.2%	86.0%
15	89.9%	93.4%	94.9%	93.3%

3.3 Selection of the best α parameter

The performance of our method is first tested on the training set in order to select the best value of α. Table 2 shows that $\alpha = 0.25$ provides the best performance overall, whereas the usual setting of $\alpha = 1$ results in the worst performance by far. Moreover temporal smoothing always improves the quality of local estimates.

Table 2: Percentage of correctly transcribed melody frames on the training set (1/4 tone tolerance).

Method	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.25$	$\alpha = 0$
Local	61.3%	71.3%	75.8%	73.2%
Smoothed	63.8%	73.5%	77.6%	76.3%

3.4 Results

Two algorithms were evaluated on a separate testing set within the MIREX 2005 evaluation framework for Audio Melody Extraction[3]: the proposed method with $\alpha = 0.25$ and temporal smoothing, and the baseline method obtained by selecting the first candidate from the modified YIN method with $\alpha = 1$.

[3]Due to the nature of MIREX 2005, only a limited number of algorithms could be submitted for evaluation.

The testing set contained 25 excerpts of 10-40 s from the following genres: rock, r&b, pop, jazz and solo classical piano. The baseline method resulted in 59.6% of correctly transcribed melody frames and the proposed method in 59.8%[4]. The total running time was less than 5 minutes for the baseline method and about one day for the proposed method. The performance figures for other systems (submitted to MIREX 2005 by other participants) varied between 58.5% and 68.6%.

Several comments derive from these results. First the performance of both algorithms was noticeably lower on the testing set than on the training set. Some other participants experienced a similar performance decrease. This is probably because the testing set contains more difficult excerpts with a large degree of polyphony and thus the melody pitch is less often the loudest one. Moreover, the proposed method did not perform significantly better than the baseline method on the testing set. The most obvious reason for this is that the model suffered from overlearning and did not adapt well to the testing data. Other experiments should be conducted to validate this explanation.

4 CONCLUSION

This article discussed a method for predominant-F0 extraction based on a family of Bayesian harmonic waveform models. Compared with other methods modeling short-term magnitude spectra, the advantage of the proposed method is that it models better destructive interferences between partials from different notes and that it provides a straightforward way to distinguish between harmonic and noisy-like parts of the signal. Also, the transcribed melody can be directly resynthesized in the end. Its disadvantage is that it is slower, because it implies estimating the phase of the partials in addition to their amplitudes.

Experiments showed that the proposed psychoacoustically motivated priors for the residual provided better predominant-F0 estimates than the isotropic Gaussian prior used in harmonic waveform models proposed previously in the literature. The best results were obtained by weighting auditory bands of the residual proportionally to their loudness. This validates partially the choice of this model in an object coding context by ensuring that the melody will be encoded prioritarily on a large proportion of frames. Also the proposed estimation algorithm ran faster than existing algorithms.

Ideas to improve the performance of the proposed method for predominant-F0 extraction can be found in other algorithms. For example the model could be used to perform a full polyphonic transcription and to select the most powerful note in a second step as in [5], or the partials' amplitudes for each note could be modeled by a nonlinear subspace instead of a single template [8]. The running time could also be much reduced and the convergence of the Newton algorithm improved by using a more limited number of partials per note, or by adapting the model to represent short-term magnitude spectra

[4]See `http://www.music-ir.org/evaluation/mirex-results/audio-melody/` for a more complete set of results.

(but keeping the same psycho-acoustically motivated prior for the residual). New experiments involving these modified methods will help concluding whether full waveform modeling can improve the performance of predominant pitch extraction compared to magnitude spectrum modeling.

References

[1] M. Davy and S. Godsill. Bayesian harmonic models for musical pitch estimation and analysis. Technical Report CUED/F-INFENG/TR.431, Cambridge University, 2002.

[2] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111:1917–1930, 2002.

[3] R. Der, P. Kabal, and W.-Y. Chan. Towards a new perceptual coding paradigm for audio signals. In *Proc. ICASSP*, 2003.

[4] J. Eggink and G.J. Brown. Extracting melody lines from complex audio. In *Proc. ISMIR*, pages 84–91, 2004.

[5] M. Goto. A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models. In *Proc. ICASSP*, pages 3365–3368, 2001.

[6] D.J.C. McKay. Choice of basis for Laplace approximation. *Machine Learning*, 33(1):77–86, 1998.

[7] E. Vincent and M.D. Plumbley. A prototype system for object coding of musical audio. In *Proc. WASPAA*, 2005.

[8] E. Vincent and X. Rodet. Music transcription with ISA and HMM. In *Proc. ICA*, pages 1197–1204, 2004.

[9] P.J. Walmsley, S.J. Godsill, and P.J.W. Rayner. Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proc. WASPAA*, pages 119–122, 1999.