A PERCEPTUALLY MOTIVATED MULTIPLE-F0 ESTIMATION METHOD

Anssi Klapuri

Institute of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
klap@cs.tut.fi

ABSTRACT

This paper describes a method for estimating the fundamental frequencies (F0) of several concurrent musical sounds. The method consists of a computational model of the human auditory periphery, followed by a novel periodicity analysis mechanism. Estimation of multiple fundamental frequencies is achieved by cancelling each detected sound from the mixture and by repeating the estimation for the residual. Computational load of the method remains reasonable since the peripheral hearing model (i.e., the hardest part) needs to be computed only once. The method is relatively straightforward to implement and outperformed two state-of-the-art reference methods in simulations.

1. INTRODUCTION

This paper proposes a method for the estimation of the fundamental frequencies (F0) of several concurrent sounds in polyphonic music signals. This is a central problem in many music signal processing applications, including music transcription (i.e., recovering the musical notation of a piece), structured audio coding, music information retrieval, and interactive music systems which respond to music signals in a meaningful way.

Several different approaches have been proposed to tackle the problem. Kashino extracted sinusoid tracks from a music signal and grouped them to sound sources based on the acoustic features of the tracks and on musical information [1]. de Cheveigné [2] and Tolonen and Karjalainen [3] proposed methods based on modeling human hearing. Davy and Godsill employed a parametric signal model and statistical methods to infer the F0s [4]. Abdallah investigated data-driven techniques for sound separation [5]. For a more extensive review, see [6, Ch.5].

The term 'perceptually motivated multiple-F0 estimator' is used here to refer to methods which build upon computational models of human pitch perception. The analytical ability of hearing is very good in complex mixture signals. Humans are able to hear the pitches of several co-occurring sounds and human musicians are the best music transcribers for the time being. It is thus quite natural to pursue multiple-F0 estimation by investigating what happens in a human listener. This has been done e.g. in [2], [3].

This paper describes a multiple-F0 estimation method which was originally proposed in [6, Ch.4] and is here considerably improved and simplified. Simulations were carried out to compare the accuracy of the proposed method with the perceptually motivated method of Tolonen and Karjalainen [3] and with an earlier multiple-F0 estimator by the present author [7]. The proposed method outperformed both the reference methods and, especially, the advantages of the perceptual approach as compared with the spectral techniques employed in [7] are discussed.

2. PITCH PERCEPTION MODELS

This section provides a brief overview of pitch perception models. The space limitations do not allow but scratch the surface and an interested reader is referred to [8, Ch.8] and [9] for better coverage.

Pitch as a perceptual term is defined as the frequency of a sine wave that is matched to a target sound by human listeners. Pitch can be heard for very diverse kinds of acoustic signals and computational models of pitch perception attempt to replicate this phenomenon. The first two processing steps of pitch models typically correspond to the peripheral hearing. In brief:

1. An input signal is passed though a bank of bandpass filters which models the frequency selectivity of the inner ear.
2. The signal at each band (or, auditory "channel") is processed to model the transform characteristics of the inner hair cells which produce firing activity in the auditory nerve. In signal processing terms, the main characteristics involved are compression and level adaptation, half-wave rectification, and lowpass filtering.

The processing mechanisms in the brain are not accurately known. However, in all modern pitch models the following processing steps can be distinguished:

3. Periodicity analysis of some form takes place for the signals within the auditory channels [10]. (Small) phase differences between the channels become meaningless.
4. Periodicity information is combined across the channels.

In the model of Meddis and Hewitt [11], for example, autocorrelation function (ACF) estimates were computed in Step 3 and these were summed across channels in Step 4 to get a summary ACF where prominent peaks were used to predict the perceived pitch. Recently, many alternative periodicity analysis mechanisms have been proposed, too (see [9] for a review).

3. PROPOSED METHOD

Although the modern pitch models have been quite successful in predicting the perceived pitch of various kinds of acoustic stimuli, they are not sufficient as such for reliable F0 estimation in music signals. The most obvious shortcoming is that they typically account for a single pitch only: pitch perception in sound mixtures is not addressed. Also, the range of pitch values that can be meaningfully processed is typically restricted to values clearly below 1kHz [11]. In the following, modifications and extensions are described to the remove these shortcomings. Previous approaches towards doing this can be found in [2], [3], [9].

Figure 1 shows the block diagram of the proposed method. It consists of a model of the peripheral auditory system, followed

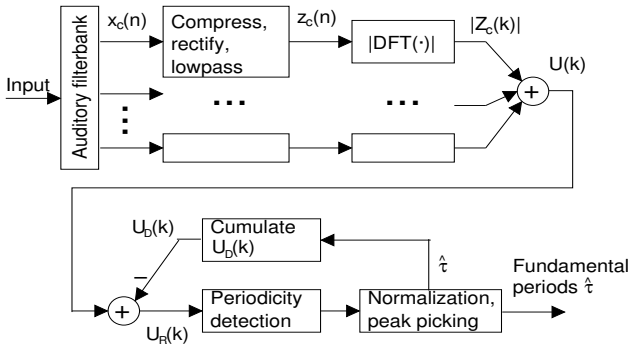

Figure 1: Block diagram of the proposed method.

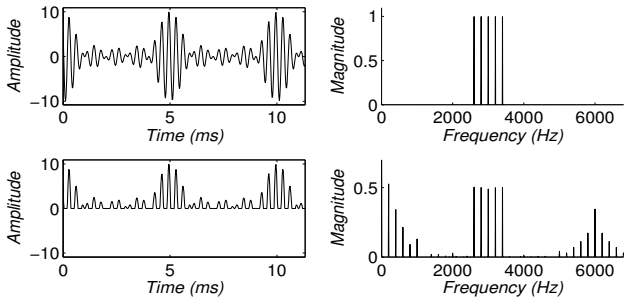

Figure 2: The upper panels show a signal consisting of the overtone partials 13–17 of a sound with F0 200Hz. The lower panels illustrate the signal after the half-wave rectification.

by a periodicity analysis mechanism were the F0s are iteratively estimated and cancelled from the mixture. The peripheral parts (Steps 1 and 2 in Sect. 2 above) are quite conventional and the proposed extensions are essentially directed at Steps 3 and 4.

3.1. Auditory filterbank and neural transduction

For the auditory filterbank, we employed the efficient implementation of "gammatone" filters as proposed by Slaney [12]. The equivalent-rectangular-bandwidths (ERB) of the filters we use are [8, p.175]

$$b_c = 0.108 f_c + 24.7\text{Hz}, \tag{1}$$

where f_c is the center frequency of the filter at channel c.

The center frequencies of the filters are uniformly distributed on the critical-band scale which is obtained by integrating the inverse of (1). This yields $\xi(f) = 21.4 \log_{10}(0.00437 f + 1)$, where f denotes frequency in Hertz and $\xi(f)$ gives the critical-band scale. A total of 72 filters between 60Hz and 5.2kHz are employed, but these parameters are not critical. The signal at the output of the auditory filter at channel c is denoted by $x_c(n)$.

Hair cell transduction is here modeled by a cascade of compression, half-wave rectification, and lowpass filtering for the subband signals $x_c(n)$. Compression was implemented by simulating the full-wave ν:th law compression (FWC) which is defined as

$$\text{FWC}(x) = \begin{cases} x^\nu, & x \geq 0 \\ -(-x)^\nu, & x < 0 \end{cases}. \tag{2}$$

For a narrowband signal—such as the output of an auditory filter, the effect of the FWC within the passband of the filter can be accurately modeled by simply scaling the signal with the factor $\gamma_c = \alpha(\sigma_c)^{\nu-1}$, where σ_c is the standard deviation of the signal at channel c and the factor α depends on ν but is common to all channes and can thus be omitted [6, p.37]. In addition to the mentioned scaling, the FWC generates small-amplitude distortion components at odd multiples of the channel center frequency. This can be avoided by using the model. The scaling factor γ_c normalizes the variances of the subband signal towards unity when $\nu < 1$. Here the value $\nu = 0.33$ is applied.

The compressed subband signals are subjected to half-wave rectification, defined as $\text{HWR}(x) = \max(x, 0)$. As shown in Fig. 2, HWR generates spectral components on zero frequency and on twice the channel center frequency. The former represents the spectrum of the amplitude envelope of the time-domain signal. It consists of beating components which correspond to the frequency intervals between the input partials. In the case of a harmonic sound, the interval corresponding to the F0 dominates.

The rectified signal is steeply lowpass-filtered in order to reject the distortion spectrum at twice the center frequency but to pass the subband signal along with its envelope spectrum.

The signal at channel c after the compression, rectification, and lowpass filtering is denoted by $z_c(n)$.

3.2. Periodicity analysis

To understand the proposed periodicity analysis method, let us first consider the ACF-based periodicity analysis, as used in [11] for example. Short-time ACF estimates within the channels can be efficiently computed as $r_{c,n}(\tau) = \text{IDFT}(|Z_{c,n}(k)|^2)$, where IDFT denotes the inverse Fourier transform and $Z_{c,n}(k)$ is the short-time Fourier transform of $z_c(n)$ computed in a time frame that is centered at n and zero-padded to twice its length before the transform.[1] The within-band ACFs are then summed to obtain the summary ACF, $s_n(\tau) = \sum_c r_{c,n}(\tau)$, where the F0 is estimated.

Since the IDFT and the summing are linear operations, their order can be reversed and we can write $s_n(\tau) = \text{IDFT}(S_n(k))$, where $S_n(k) = \sum_c |Z_{c,n}(k)|^2$. For real-valued (audio) signals, the IDFT can be written out as

$$s_n(\tau) = \text{IDFT}(S_n(k)) = \frac{2}{K} \sum_{k=0}^{K/2-1} \cos\left(\frac{2\pi\tau k}{K}\right) S_n(k), \tag{3}$$

where K is the size of the time frame after zero-padding.

In our method, three crucial modifications are made to (3). First, as can be seen in Fig. 1, magnitude spectra are summed across channels instead of power spectra. It was observed that raising the magnitude spectra to the second power accentuates timbral peculiarities that cannot be completely removed in polyphonic signals by the bandwise compression. Thus, the within-band magnitude spectra are summed to obtain a summary magnitude spectrum (SMS) denoted by $U(k) = \sum_c |Z_c(k)|$ (omitting time indexes for simplicity). This is analogous to the use of the "generalized" ACF in [3], where a real-valued exponent 0.67 was used instead of 2.

The second modification concerns the function $\cos(\cdot)$ in (3), which can be seen as a spectral template that matches overtone partials of F0 candidate K/τ. Here we replace $\cos(\cdot)$ by a response

[1] Zero padding is needed to compute the short-time ACFs in the frequency domain. In the proposed (modified) system, the same zero padding is used just to improve the frequency resolution.

that is more sharply tuned to the overtone frequencies and employs no negative weights between them (see (4) below). This alleviates the interference of co-occurring sounds and leads to a very efficient implementation computationally. The general approach is closely related to the "harmonic selection" techniques reviewed in [2].

The relative strength, or, *salience*, $\lambda(\tau)$ of a fundamental period candidate τ is calculated as

$$\lambda(\tau) = \frac{f_{\rm s}}{\tau} \sum_{j=1}^{\tau/2} \left(\max_{k \in \kappa_{j,\tau}} \left(H_{\rm LP}(k) U(k) \right) \right), \quad (4)$$

where $f_{\rm s}$ denotes the sampling rate and the factors $f_{\rm s}/\tau$ and $H_{\rm LP}(k)$ are related to the third modification to be explained later. The set $\kappa_{j,\tau}$ defines a range of frequency bins in the vicinity of the j:th overtone partial of the F0 candidate $f_{\rm s}/\tau$. More exactly, $\kappa_{j,\tau} = [k_{j,\tau}^{(0)}, \ldots, k_{j,\tau}^{(1)}]$, where

$$k_{j,\tau}^{(0)} = \lfloor jK/(\tau + \Delta\tau/2) \rfloor + 1, \quad (5)$$

$$k_{j,\tau}^{(1)} = \max(\lfloor jK/(\tau - \Delta\tau/2) \rfloor, k_{j,\tau}^{(0)}). \quad (6)$$

The scalar $\Delta\tau$ denotes spacing between successive period candidates τ. Here we use a constant sampling of lag values, $\Delta\tau = 1$, analogous to the ACF. In practice, the set $\kappa_{j,\tau}$ comprises exactly one frequency bin for all but the shortest periods τ.

The third modification in (4) compared to (3) is that individual partials in (4) are weighted by the factor $f_{\rm s}/\tau \times H_{\rm LP}(k)$, where

$$H_{\rm LP}(k) = \frac{1}{0.108 f_{\rm s} k/K + 24.7}. \quad (7)$$

By comparison with (1), we see that this is the reciprocal of the bandwidth of an auditory filter centered at frequency bin k. The factor $f_{\rm s}/\tau \times H_{\rm LP}(k)$ can therefore be written as $F(\tau)/b_c(jF(\tau))$ where $F(\tau) = f_{\rm s}/\tau$ is the F0 of the period candidate τ (= frequency interval between its overtones) and $b_c(jF(\tau))$ is the width of an auditory filter centered at its j:th overtone. The ratio of these two is interpreted as the *resolvability* of the partial j [6, p.45]: the lower-order overtones of a harmonic sound are resolved into separate auditory channels, whereas the higher-order overtones go to a same auditory channel with their neighbours and their frequencies cannot be perceived separately. Actually, $H_{\rm LP}(k)$ would belong to the within-band hair-cell modeling stage but, since the filter is the same for all channels, it is equivalent to apply it after the channels have been combined. The higher the center frequency of an auditory channel, the more the filter attenuates the spectrum at the passband of the auditory filter and thus gives it a smaller weight in relation to the envelope spectrum which is around zero frequency and not much affected. In psychoacoustic terms, this corresponds to the fact that, at higher auditory channels, the neural firing activity more and more follows the amplitude envelope of the subband signal and not its fine structure. As the resolvability is approximately inversely proportional to the harmonic index j when τ is fixed, the sum in (4) can be limited to $j \approx 20$.

Taken together, the computation of the salience function $\lambda(\tau)$ can be seen as a process where partials are picked from harmonic positions of the spectrum $U(k)$, their magnitudes are weighted by the modeled resolvability $f_{\rm s}/\tau \times H_{\rm LP}(k)$, and then summed. What makes all the difference is that the rectification within channels maps the contribution of higher-order partials to the position of the fundamental and its few multiples in the spectra $Z_c(k)$ (see Fig. 2). Moreover, the degree to which an individual overtone partial j is mapped to the position of the F0 increases as a function of j. This because the auditory filters become wider at higher frequencies and the partials thus have larger-magnitude neighbours with which to generate beating components in the envelope spectrum. As a consequence, the entire harmonic series of a sound contributes to its salience, despite the weighting with the resolvability. The proposed method is largely based on this basic observation. Organizing the higher-order partials to their respective sound sources in polyphonic music signals is a nightmare. The rectification operation in part accomplishes this "automatically" by mapping the support from higher-order harmonics to the position of F0 and its few multiples in $U(k)$.

Finally, the function $\lambda(\tau)$ is normalized in order not to favour either high or low F0s. The "balancing" operation weights $\lambda(\tau)$ linearly as a function of the logarithm of the corresponding F0:

$$\tilde{\lambda}(\tau) = (1 + b \ln(f_{\rm s}/\tau))\, \lambda(\tau), \quad (8)$$

where $b = -0.04$ gives a slight emphasis to low frequencies in a 93ms analysis frame and no balancing is needed in a 46ms frame.

3.3. Iterative estimation and cancellation

The global maximum of the function $\tilde{\lambda}(\tau)$ was found to be a robust indicator of one of the correct F0s in polyphonic signals. However, the next-highest weight was often assigned to half or twice of the firstly detected F0. Therefore, we employ an iterative technique where the F0 estimation is followed by the cancellation of the detected sound from the mixture and the estimation is then repeated for the residual signal. The algorithm is the following:

Step 1 A *residual SMS* $U_{\rm R}(k)$ is initialized to equal $U(k)$. A spectrum of detected sounds, $U_{\rm D}(k)$, is initialized to zero.

Step 2 A fundamental period $\hat{\tau}$ is estimated using $U_{\rm R}(k)$ and (4)–(8). The maximum of $\tilde{\lambda}(\tau)$ determines $\hat{\tau}$.

Step 3 Harmonic selection is carried out for the found period $\hat{\tau}$ according to (4)–(6). However, instead of summing up the magnitude values, the frequency and amplitude of each overtone partial is estimated and used to calculate its magnitude spectrum at the few surrounding frequency bins.

Step 4 The magnitude spectrum of the j:th partial is weighted by $f_{\rm s}/\tau \times H_{\rm LP}(k)$ and added to the corresponding position of the cumulative spectrum of all the detected sounds $U_{\rm D}(k)$.

Step 5 The residual SMS is recalculated as $U_{\rm R}(k) \leftarrow \max(0, U(k) - dU_{\rm D}(k))$, where $d = 0.5$ controls the amount of the subtraction and is a free parameter of the algorithm. Return to Step 2.

An important characteristic of the Step 4 is that, before adding the partials of a detected sound to $U_{\rm D}(k)$, they are weighted by their modeled resolvability in the same manner as at the F0 detection stage. As a consequence, the higher-order partials are not entirely removed from the mixture spectrum when the residual $U_{\rm R}(k)$ is calculated. This is essential in order not to corrupt the sounds that remain in the residual and have to be detected at the coming iterations. As explained above, the higher-order harmonics have been mapped to the position of the fundamental by the rectification and are thus effectively cancelled, too.

4. RESULTS

Simulation experiments were carried out to evaluate the performance of the proposed method and to compare it with two reference methods. The first of these was the perceptually motivated

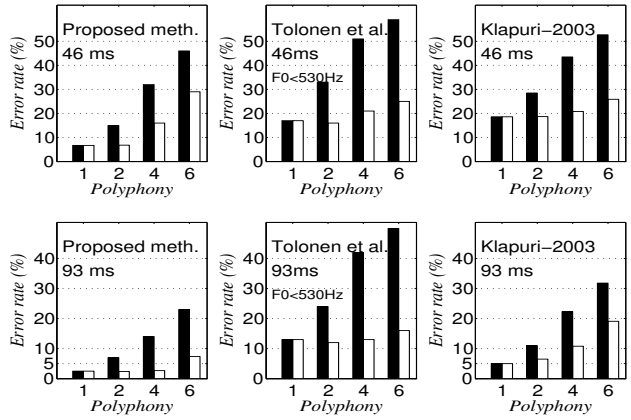

Figure 3: Error rates as a function of the number of concurrent sounds (polyphony) for the proposed method and for the two reference methods. The upper panels show the results for a 46-ms analysis frame and the lower panels for a 93-ms frame.

multiple-F0 estimator proposed by Tolonen and Karjalainen in [3]. The method was carefully implemented based on the reference, and the original code of the authors was used for the warped linear prediction part. The other reference method was proposed by Klapuri in [7] and is based on spectral techniques.

The acoustic material consisted of samples from the McGill University Master Samples collection, the University of Iowa website, IRCAM Studio Online, and of independent recordings for the acoustic guitar. There were altogether 2842 samples from 32 musical instruments, comprising brass and reed instruments, strings, flutes, the piano, the guitar, and mallet percussion instruments. Semirandom sound mixtures were generated by first allotting an instrument and then a random note from its playing range, restricting, however, the pitch between 65Hz and 2.1kHz. This was repeated to get the desired number of sounds which were mixed with equal mean-square levels. One thousand test cases were generated for mixtures of one, two, four, and six sounds.

One analysis frame immediately after the onset[2] of the sounds was given to the multiple-F0 estimators. The number of F0s to extract, i.e., the polyphony, was given along with the mixture signal. A correct F0 estimate was defined to deviate less than 3% from the reference F0, making it "round" to a correct musical note.

Figure 3 shows the results for the proposed method and for the two reference methods in 46-ms and 93-ms analysis frames. Two different error rates are shown. *Multiple-F0 estimation* rates (black bars) were computed as the percentage of all F0s that were not correctly detected in the input signals. In *predominant-F0 estimation* (white bars), only one F0 in the mixture was being estimated and it was defined to be correct if it matched the F0 of any of the component sounds. The test cases given to the proposed method and the second reference method were identical. For the method of Tolonen and Karjalainen, the pitch range was limited to three octaves between 65Hz and 520Hz. As reported by the original authors, accuracy of the method degrades rapidly above 600Hz. Computationally, the method was by far the most efficient.

The proposed method outperformed both the reference methods and good accuracy was observed especially in the 93-ms analysis frame. An important factor in the reported results is that the analysis frames were positioned immediately at the onsets of the sounds. Especially, the predominant-F0 error rates in the 46-ms frame shrink to about a third of the reported values (for all methods) if the analysis frame is positioned 80ms after the onset, where the noisy beginning transients have died off.

[2]The onset of the sounds was defined to be at the time where the waveform reached $1/3$ of its maximum value over the beginning 200ms.

5. CONCLUSIONS

A perceptually motivated multiple-F0 estimation method was described which can be used to process a large variety of pitched musical sounds. The author has been quite surprised at the relative ease at which the output of a peripheral hearing model can be used for accurate multiple-F0 estimation. A particularly important principle in an auditorily motivated analysis is that the higher-order overtones of a sound are processed collectively within each auditory channel; estimation and separation of individual higher-order partials is not attempted. For comparison, the reference method [7] is based on spectral techniques and requires quite long analysis frames to resolve and process the overtones of low-pitched sounds.

6. REFERENCES

[1] K. Kashino and H. Tanaka, "A sound source separation system with the ability of automatic tone modeling," in *Int. Computer Music Conf.*, Hong Kong, China, 1993.

[2] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model for auditory processing," *J. Acoust. Soc. of Am.*, vol. 93, no. 6, pp. 3271–3290, 1993.

[3] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[4] M. Davy and S. Godsill, "Bayesian harmonic models for musical signal analysis," in *Seventh Valencia International meeting Bayesian statistics 7*, Tenerife, Spain, June 2002.

[5] S. A. Abdallah, "Towards music perception by redundancy reduction and unsupervised learning in probabilistic models," Ph.D. dissertation, King's College, London, 2002.

[6] A. P. Klapuri, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, 2004.

[7] ——, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 804–815, 2003.

[8] B. C. J. Moore, Ed., *Hearing*, 2nd ed. San Diego, California: Academic Press, 1995.

[9] A. de Cheveigné, "Pitch perception models," in *Pitch*, C. Plack and A. Oxenham, Eds. New York: Springer, 2005.

[10] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. pitch and pitch salience," *Journal of Neurophysiology*, vol. 76, no. 3, pp. 1698–1716, 1996.

[11] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification." *J. Acoust. Soc. of Am.*, vol. 89, no. 6, 1991.

[12] M. Slaney, "An efficient implementation of the patterson holdsworth auditory filter bank," Perception Group, Apple Computer, Tech. Rep. 35, 1993.