Multipitch analysis of polyphonic music and speech signals using an auditory model

Anssi Klapuri, *Member, IEEE*
Institute of Signal Processing, Tampere University of Techonology,
Korkeakoulunkatu 1, FIN-33720 Tampere, Finland
tel. +358 3 3115 2124, fax +358 3 3115 4989, e-mail: anssi.klapuri@tut.fi

***Abstract*— A method is described for estimating the fundamental frequencies of several concurrent sounds in polyphonic music and multiple-speaker speech signals. The method consists of a computational model of the human auditory periphery, followed by a periodicity analysis mechanism where fundamental frequencies are iteratively detected and cancelled from the mixture signal. The auditory model needs to be computed only once, and a computationally efficient strategy is proposed for implementing it. Simulation experiments were made using mixtures of musical sounds and mixed speech utterances. The proposed method outperformed two reference methods in the evaluations and showed a high level of robustness in processing signals where important parts of the audible spectrum were deleted to simulate bandlimited interference. Different system configurations were studied to identify the conditions where pitch analysis using an auditory model is advantageous over conventional time or frequency domain approaches.**

***Index Terms*— Fundamental frequency estimation, pitch perception, music information retrieval, acoustic signal analysis.**

I. Introduction

Pitch analysis of polyphonic music and multiple-speaker speech signals is useful for many purposes. Applications include automatic music transcription, speech separation, structured audio coding, and music information retrieval. The task of estimating the fundamental frequencies (F0s) of several concurrent sounds – multiple-F0 estimation – is closely related to sound separation and auditory scene analysis, since an algorithm performing this task goes a long way towards organizing a complex signal into its constitutent sound sources [1]. This paper proposes a method for doing this in single-channel audio signals.

A number of different approaches have been proposed for multiple F0 estimation (see [2], [3] for a review). The first algorithms were developed to transcribe polyphonic music and were more or less heuristic in nature [4], [5], [6], [7]. Methods based on modeling the auditory scene analysis (ASA) function in humans were later proposed by Mellinger [8], Kashino and Tanaka [9], Ellis [10], Godsmark and Brown [11], and Sterian [12], presumably inspired by Bregman's work on human ASA [13]. Signal model based Bayesian inference methods were investigated by Goto [14], Davy et al. [15], Cemgil [16], and Kameoka et al. [17]. Most recently, unsupervised learning methods, such as independent component analysis, sparse coding, and non-negative matrix factorization, have been proposed by Casey and Westner [18], Lepain [19], Smaragdis and Brown [20], Abdallah and Plumbley [21], and Virtanen [22].

Auditory model based methods represent an important thread of work in this area since the early 1990s. By these we mean methods which employ a peripheral hearing model to calculate an intermediate data representation that is then used in further signal analysis. The rationale in doing this is that humans are very good at resolving sound mixtures and therefore it seems natural to employ the same data representation that is available to the human brain. Auditory model based methods have been proposed at least by Meddis and Hewitt [23], de Cheveigné [24], and Wu, Wang, and Brown [25] for speech signals, and by Martin [26], Tolonen and Karjalainen [27], and Marolt [28] for music signals. The mentioned methods are oriented towards practical pitch extraction in speech and music, as opposed to the work on pitch perception models themselves, which aim at reproducing psychophysical data and phenomena in an accurate manner. Excellent reviews on pitch perception models can be found in [29], [30].

This paper proposes a multiple F0 estimator which consists of a computational model of the human auditory periphery, followed by a periodicity analysis method where F0s are iteratively detected and cancelled from the mixture signal. For both parts, computationally efficient techniques are presented which make the overall method more than twice faster than real time on a PC with a 2.8 GHz Pentium 4 processor. In particular, a mechanism is described which speeds up the analysis at the subbands of an auditory model. In the periodicity analysis part, we replace the conventional autocorrelative analysis with a transform which is more robust in polyphonic signals and can be used for a wide pitch range between 40 Hz and 2.1 kHz. Some parts of this work have been previously published in two conferences papers [31] and [32].

One goal of this paper is to identify the conditions where an auditory model based pitch analysis has significant advantage over more conventional time or frequency domain approaches. It will be shown that these conditions include especially the processing of bandlimited signals or signals were parts of the spectrum are not usable due to bandlimited interference.

The method was evaluated using mixtures of musical sounds and mixed speech utterances. The results are compared with two reference methods [27] and [33]. Also, we compare alternative configurations of the proposed method where either the auditory model is disabled or the iterative estimation and cancellation mechanism is replaced with a joint estimator.

A. Klapuri is with Institute of Signal Processing, Tampere University of Technology, Tampere, Finland (e-mail: anssi.klapuri@tut.fi).

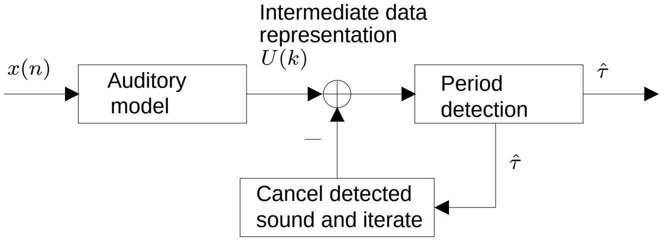

Fig. 1. An overview of the proposed method. An auditory model is followed by the iterative detection and cancellation of the most prominent period.

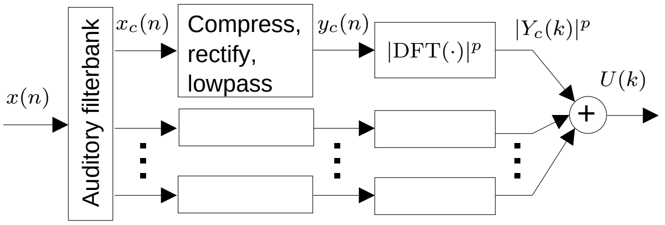

Fig. 2. Structure of the peripheral hearing model. An input signal is processed with a bandpass filterbank, after which the subband signals are compressed, rectified, and lowpass filtered. Short-time magnitude spectra are calculated within the bands, raised to power p, and then summed across bands.

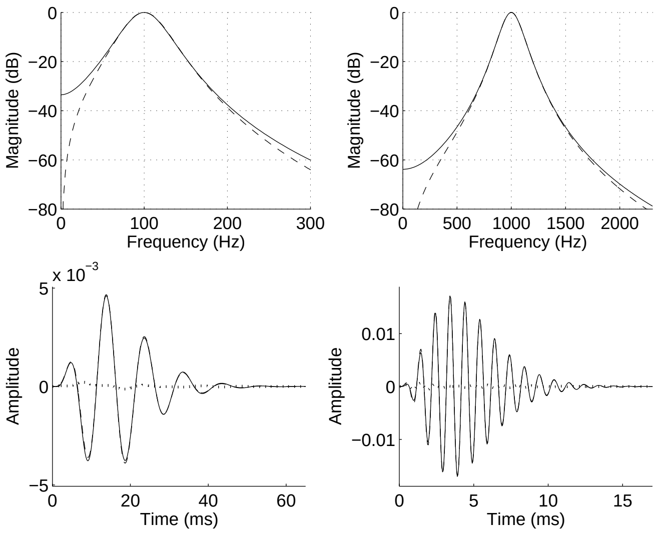

Fig. 3. The upper panels show the magnitude responses of two gammatone filters (solid line) and those of the proposed approximation (dashed line). The lower panels show the impulse responses of the two gammatone filters (solid line) and the difference between the impulse reponses of the gammatone filter and the proposed approximation (dotted line). The left and right panels correspond to center frequencies 100 Hz and 1 kHz, respectively.

II. Proposed method

Figure 1 shows an overview of the proposed method, where the two parts, the auditory model and the iterative F0 detection part, are clearly seen. The auditory model is detailed in Fig. 2. Except for the computational efficiency considerations which are described later, the auditory model follows the structure of modern pitch perception models. In these, a signal is generally processed as follows:

1) An input signal $x(n)$ is passed through a bank of linear bandpass filters which models the frequency selectivity of the inner ear [34], [35].
2) The signal $x_c(n)$ at band c (aka channel) is subjected to non-linear processing to obtain a signal $y_c(n)$ which models the level of neural activity in the auditory nerve fibers representing channel c [36], [37].
3) Periodicity analysis of some form takes place for the signals $y_c(n)$ within the channels [38], [39].
4) Periodicity information is combined across the bands.

As a concrete example of Steps 3–4, Meddis and Hewitt [38] computed short-time autocorrelation function (ACF) estimates $r_{c,t}(\tau)$ within the channels at successive times t, and then summed these to obtain a summary ACF, $\bar{r}_t(\tau) = \sum_c r_{c,t}(\tau)$, where prominent peaks were used to predict the perceived pitch.

Different parts of the system shown in Figs. 1–2 are now described in more detail.

A. Auditory filterbank

The most important parameter of the auditory filters is their bandwidth. The equivalent rectangular bandwidths (ERB)[1] of the filters we use are

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

where f_c is the filter's center frequency, b_c is the bandwidth, and $c = 0, 1, \ldots, C-1$ is the subband index. These bandwidths have been reported for humans in [40].

In order that the power responses of the auditory filters would sum approximately to a flat response, the center frequencies are distributed uniformly on a critical-band scale,

$$f_c = 229[10^{(\xi_1 c + \xi_0)/21.4} - 1], \tag{2}$$

where ξ_0 is the critical-band-number of the lowest band and $0 < \xi_1 < 1$ determines the band density. We use a total of 70 filters having center frequencies between 65 Hz and 5.2 kHz, corresponding to $\xi_0 = 2.3$ and $\xi_1 = 0.39$.

The power and impulse responses of the auditory filters have been studied in humans and other mammals and are quite accurately known [34], [35]. The *gammatone* filter provides an excellent fit to the experimental data, and is therefore widely used [41]. Figure 3 illustrates the frequency response and the impulse response of the gammatone filter.

Slaney has proposed a computationally efficient implementation of the gammatone filter by using a cascade of four second-order IIR filters [42]. We propose a different implementation for two reasons. Firstly, we wanted to attenuate the "tails" of the power response further away from the filter's center frequency, since the spectral variation of musical sounds is very large and we wanted to ensure that a certain filter is not dominated by frequency components too far from its center frequency. Secondly, the computational efficiency is improved by using filter sections which have only coefficient values ± 1 in the numerator of their z-transform, thus reducing the number of multiplication operations needed.

[1] The ERB of a filter is defined as the bandwidth of a perfectly rectangular filter which has an integral over its power response which is the same as for the specified filter.

The proposed filter structure uses two types of IIR resonators as building blocks, referred to as Resonator 1 and 2 in the following. An individual auditory filter consists of a cascade of these. The z-transform of Resonator 1 is of the form

$$\begin{aligned} H_1(z) &= \rho_1 \frac{(1-z^{-1})(1+z^{-1})}{(1-A\mathrm{e}^{\mathrm{i}\theta_1}z^{-1})(1-A\mathrm{e}^{-\mathrm{i}\theta_1}z^{-1})} \\ &= \rho_1 \frac{1-z^{-2}}{1-A\cos(\theta_1)z^{-1}+A^2z^{-2}}, \end{aligned} \tag{3}$$

where the parameters ρ_1, θ_1, and A are derived in Appendix A. Resonator 2 is of the same form but without the zeros, having a z-transform of the form

$$H_2(z) = \rho_2 \frac{1}{(1-A\mathrm{e}^{\mathrm{i}\theta_2}z^{-1})(1-A\mathrm{e}^{-\mathrm{i}\theta_2}z^{-1})}. \tag{4}$$

Appendix A describes the calculation of the parameters A, $\theta_{1,2}$, and $\rho_{1,2}$, and choosing the optimal configuration of second-order sections. It was found that a cascade of four resonators, two of each type, leads to the most accurate result. In a floating-point implementation, the factors $\rho_{1,2}$ can be combined into a single scalar to speed up the computation.

The upper panels of Figure 3 compare the frequency response of the gammatone filter with the proposed approximation at two different center frequencies, 100 Hz and 1 kHz. The biggest inaccuracies occur near zero frequency, where the proposed filter has a deeper notch than the gammatone filter. In practical applications, complete suppression of the dc component is merely a desirable feature. The lower panels illustrate the impulse responses of the two gammatone filters, with a dotted line showing the difference between the gammatone filter and the approximation.

B. Neural transduction

The signal $x_c(n)$ at each band is processed to model the transform characteristics of the inner hair cells (IHC) which produce firing activity in the auditory nerve. Several computational models of the IHCs have been proposed in the literature [36]. A problem with these is that a realistic IHC model depends critically on the absolute level of its input and has a dynamic range of about 25 dB only [43], [37]. As a consequence, most practical systems have replaced an accurate IHC model by a cascade of signal processing operations that model the main characteristics of the IHCs explicitly: (i) dynamic level compression, (ii) half-wave rectification, and (iii) lowpass filtering [10], [44], [25], [31]. This is also the approach followed here.

Compression was implemented with an automatic gain control, scaling the signal $x_c(n)$ within analysis frame t with the factor

$$\gamma_{c,t} = \sigma_{c,t}^{\nu-1}, \tag{5}$$

where $\sigma_{c,t}$ is the standard deviation of the signal $x_c(n)$ within the frame t. From the viewpoint of an individual analysis frame, the compression flattens ("whitens") the spectral energy distribution, since the scaling factors $\gamma_{c,t}$ normalize the auditory channel variances towards unity when $0 < \nu < 1$. Here the value $\nu = 0.33$ is applied. For comparison, Ellis [10] normalized the variances of the subband signals to unity, corresponding to $\nu = 0$. Tolonen and Karjalainen, in turn, applied inverse warped-linear-prediction filtering on the input wide-band signal which leads to a very similar result [27].

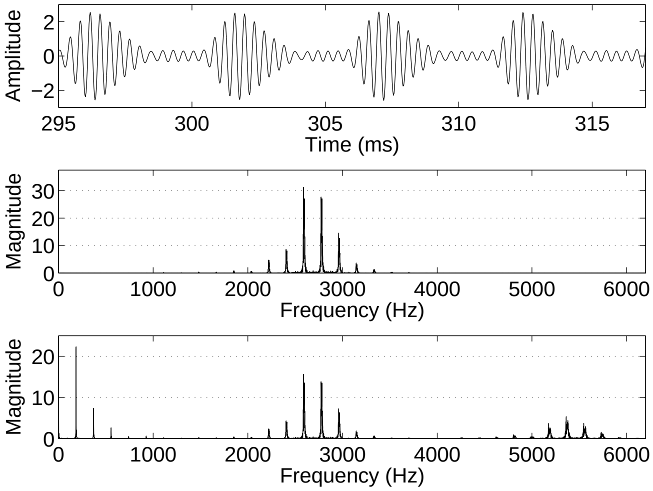

Fig. 4. The upper panel shows the subband signal $x_c(n)$ at a band with center frequency 2.7 kHz. The example signal is a trumpet sound with F0 185 Hz. The middle panels shows the magnitude spectrum of the subband signal and the lower panel shows the spectrum after half-wave rectification.

The compressed subband signals are subjected to half-wave rectification (HWR), defined as

$$\mathrm{HWR}(x) = \max(x,0) = \frac{1}{2}(x+|x|). \tag{6}$$

Figure 4 illustrates the HWR for a subband signal $x_c(n)$ of a trumpet sound at a band with center frequency 2.7 kHz. The upper two panels show the subband signal in time and frequency domains. The lowest panel shows the spectrum of the subband signal after rectification, that is, the spectrum of $\mathrm{HWR}(x_c(n))$. As can be seen, the rectification generates spectral components at the baseband and on twice the channel center frequency. The former represent the spectrum of the amplitude envelope of $x_c(n)$. 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 usually dominates.

Figure 5 illustrates the bandwise magnitude spectra of a trumpet sound after the within-band compression and rectification, $|\mathrm{DFT}\{\frac{1}{2}\gamma_c(x_c(n)+|x_c(n)|)\}|$. Note that here a logarithmic frequency scale is used. The rectified signal of Fig. 4 appears at the band with center frequency 2.7 kHz. As can be seen, the rectification maps the contribution of higher-order partials to the position of the F0 and its few multiples in the spectra. Moreover, the degree to which an individual overtone partial m is mapped to the position of the fundamental increases along with m. This is because the auditory filters become wider at the higher center frequencies and the partials therefore have more neighbours with which to generate the difference frequencies (beating) in the amplitude envelope. This is nice, since organizing the higher partials to their fundamental is very difficult in polyphonic music. The rectification does this "automatically", without the need to

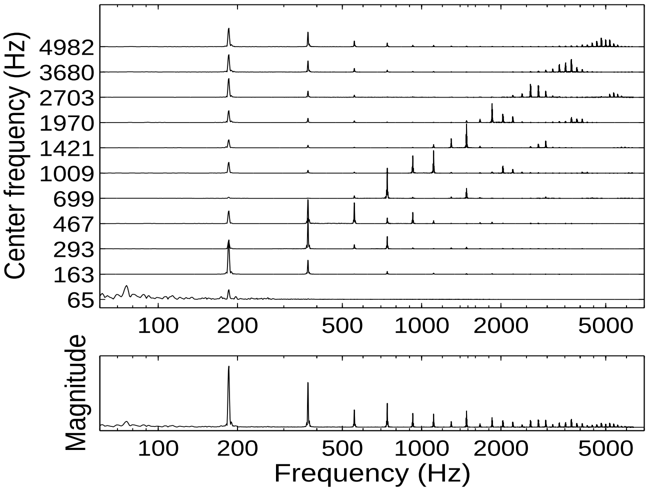

Fig. 5. The upper panel shows compressed and rectified spectra at a few auditory channels for a trumpet sound (F0 185 Hz). The lower panel shows a summary spectrum which was obtained by summing over the subbands.

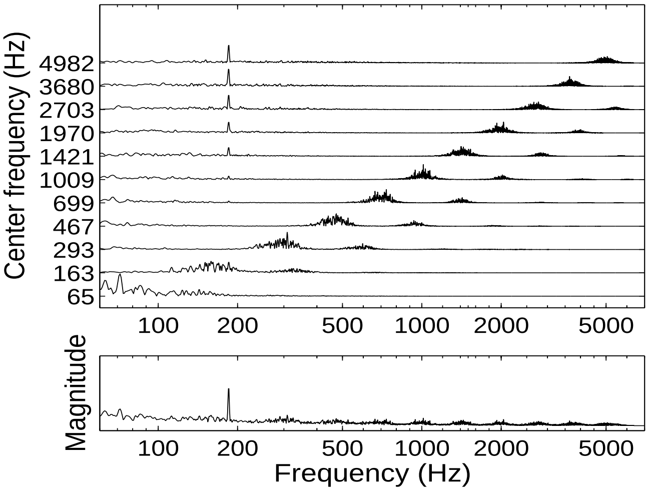

Fig. 6. The upper panel shows compressed and rectified spectra at a few auditory channels for an amplitude-modulated noise signal (modulation frequency 185 Hz). The lower panel shows a summary spectrum obtained by summing over subbands.

resolve individual higher-order partials.

An auditory model allows simulating the pitch perception for a wide range of signals. For example, let us consider amplitude-modulated white noise. It is known from psychoacoustics that such a signal will cause a pitch percept corresponding to the modulation frequency. Figure 6 shows the bandwise magnitude spectra of a white noise signal which was amplitude-modulated with the function $(1 + \cos(\omega n))$, where ω corresponded to 185 Hz. As can be seen, the spectrum is noisy at lower subbands, but at higher bands, the spectrum of the amplitude envelope (as generated by the HWR) shows a clear peak at 185 Hz, which is also visible in the summary spectrum in the lower panel. Although this particular case is trivial to reproduce with other methods too, auditory models by definition simulate hearing for a large variety of signals [38], [45].

The spectral components generated around $2f_c$ were not found useful, since these are not guaranteed to match the harmonic series of the sound, due to non-ideal harmonicity. On the contrary, lowpass filtering the rectified signal so as to reject the harmonic distortion around twice the center frequency (here called "distortion spectrum") was found to improve the F0 analysis. A difficulty in doing this, however, is that the passband of the auditory filter overlaps the distortion spectrum at the lowest channels. The problem can be solved by noting that HWR can be written as $\mathrm{HWR}(x) = \frac{1}{2}(x + |x|)$. In order to achieve a clean suppression of the distortion spectrum also at the lowest channels, the signal $x_c(n)$ is first full-wave rectified as $y'_c(n) = |x_c(n)|$, the resulting signal is lowpass filtered using a cutoff frequency f_c, summed with the original signal $x_c(n)$, and finally scaled down by two. In addition to improving the F0 analysis, this allows the overall system to be implemented very efficiently as will be explained in the next section. The signal at channel c after the compression, rectification, and lowpass filtering is denoted by $y_c(n)$.

C. Efficient computation of frequency-domain representation

The signals $y_c(n)$ are blocked into frames which are then Fourier transformed. In more detail, each frame is Hamming windowed, zero-padded to twice its length, and then the short-time Fourier transform is applied. The resulting transform at channel c and time frame t is denoted by $Y_{c,t}(k)$.

The bandwise spectra are raised to power p and then summed to obtain a "summary spectrum"

$$U_t(k) = \sum_c |Y_{c,t}(k)|^p. \tag{7}$$

This intermediate data representation is used in all subsequent processing.

To understand why a frequency-domain representation is computed, let us consider again as an example the summary ACF of Meddis's and Hewitt's model [38] which was mentioned in the beginning of Sect. II. The short-time ACF estimates within the subbands can be efficiently computed as $r_{c,t}(\tau) = \mathrm{IDFT}(|Y_{c,t}(k)|^2)$, where IDFT denotes the inverse Fourier transform and $Y_{c,t}(k)$ is the short-time Fourier transform of $y_c(n)$ in time frame t, zero-padded to twice its length before the transform. The summary ACF, in turn, can be computed as $\bar{r}_t(\tau) = \mathrm{IDFT}(\bar{R}_t(k))$, where $\bar{R}_t(k) = \sum_c |Y_{c,t}(k)|^2$. Note that the spectra $|Y_{c,t}(k)|^2$ can be summed *before* the IDFT because the IDFT and summing are linear operations and their order can therefore be reversed.

Based on the above discussion, we can see that the summary ACF representation of Meddis and Hewitt could be calculated simply using $p = 2$ in (7) and by replacing the period detection module in Fig. 1 with the inverse Fourier transform.[2] This is not what we will do, however, since the intermediate representation $U_t(k)$ allows a lot of flexibility in designing the periodicity analysis mechanism, and we will utilize that to make the estimator more robust in polyphonic signals.

[2] It should be noted, however, that the IHC model and some other details in [38] were different from those employed here.

It is clear that computation of the Fourier transforms $Y_{c,t}(k)$ at 70 subbands incurs a high computational load. In the following, we describe a technique which reduces this load roughly by factor 10.

Let us denote one windowed and zero-padded time frame of $x_c(n)$ by vector $\mathbf{x}_{c,t}$ and the corresponding compression scaling factor by $\gamma_{c,t}$ (see (5)). The compressed, rectified, and lowpass filtered signal $y_c(n)$ can then be written as

$$\mathbf{y}_{c,t} = \frac{1}{2}\gamma_{c,t}(\mathbf{x}_{c,t} + \mathbf{e}_c * |\mathbf{x}_{c,t}|), \tag{8}$$

where $\mathbf{e}_c$ is the impulse response of the distortion-suppression lowpass filter at band c, and $*$ denotes convolution. Using (8), the summary spectrum $U_t(k)$ can be written as

$$\begin{aligned} U_t(k) &= \sum_c \left| \text{DFT}\left[\frac{1}{2}\gamma_{c,t}(\mathbf{x}_{c,t} + \mathbf{e}_c * |\mathbf{x}_{c,t}|)\right]\right|^p \\ &= \frac{1}{2^p}\sum_c \gamma_{c,t}^p \left|\text{DFT}(\mathbf{x}_{c,t}) + \text{DFT}(\mathbf{e}_c * |\mathbf{x}_{c,t}|)\right|^p . \end{aligned} \tag{9}$$

In practice, the spectra of $\mathbf{x}_{c,t}$ and $\mathbf{e}_c * |\mathbf{x}_{c,t}|$ are non-overlapping in all except few lowest bands and (9) can be approximated by

$$\begin{aligned} U_t(k) \approx &\frac{1}{2^p}\sum_c \left|\gamma_{c,t}\text{DFT}\left(\mathbf{x}_{c,t}\right)\right|^p \\ &+ \frac{1}{2^p}\sum_c \left|\gamma_{c,t}\text{DFT}\left(\mathbf{e}_c * |\mathbf{x}_{c,t}|\right)\right|^p . \end{aligned} \tag{10}$$

The benefit of this form is that the first term on the right-hand side of (10) can be written as

$$\frac{1}{2^p}\sum_c \left|\gamma_{c,t}\text{DFT}\left(\mathbf{x}_{c,t}\right)\right|^p \approx \zeta \left|\Gamma_t(k) X_t(k)\right|^p , \tag{11}$$

where $X_t(k)$ is the spectrum of the *wideband* input signal $x(n)$ in frame t, ζ is a normalizing constant, and $\Gamma_t(k)$ is a frequency response obtained by linearly interpolating between the values $\gamma_{c,t}$ defined at the center frequencies f_c. This approximation is valid between the lowest and the highest subband center frequency, provided that the center frequencies and bandwidths obey (1)–(2).

It follows that bandwise Fourier transforms need to be computed only for the signals $\mathbf{e}_c * |\mathbf{x}_{c,t}|$ which represent the bandwise amplitude envelopes. As the bandwidth of these signals is narrow, the signals $\mathbf{e}_c * |\mathbf{x}_{c,t}|$ are decimated down to the sampling rate 5512.5 Hz before computing the DFTs. This means significant computational savings since an analysis frame of 2048 samples at 44100 Hz sampling rate, for example, shrinks to 256 samples at 5512.5 Hz sampling rate. After the decimation and the DFT, the calculated bandwise spectra are substituted to the second term on the right-hand side of (10). The first term is obtained from (11). This technique significantly improves the computational efficiently of the auditory model.

D. Periodicity analysis

As mentioned in the previous section, $U_t(k)$ can be used to compute the summary ACF by using $p = 2$ in (7) and simply inverse Fourier transforming $U_t(k)$ in each frame:

$$\bar{r}_t(\tau) = \text{IDFT}(U_t(k)). \tag{12}$$

Instead of (12), we will use a periodicity analysis method which improves the robustness in polyphonic signals and is able to handle the wide range of pitch values encountered in music.

In the proposed method, the *salience*, or strength, of a period candidate is calculated as a weighted sum of the amplitudes of the harmonic partials of the corresponding F0. More exactly, the salience $s_t(\tau)$ of a fundamental period candidate τ in frame t is calculated as

$$s_t(\tau) = \sum_{m=1}^{M} w(\tau, m) \max_{k \in \kappa_{\tau,m}} U_t(k), \tag{13}$$

where m is the partial index and the function $w(\tau, m)$ determines the weight of partial m of period τ in the sum (the weights will be explained later). The set $\kappa_{\tau,m}$ consists of a range of frequency bins in the vicinity of the m:th overtone partial of F0 candidate f_s/τ, where f_s denotes the sampling rate. More exactly,

$$\kappa_{\tau,m} = [\langle mK/(\tau + \Delta\tau/2)\rangle, \ldots, \langle mK/(\tau - \Delta\tau/2)\rangle], \tag{14}$$

where $\langle\cdot\rangle$ denotes rounding to the nearest integer and $\Delta\tau$ denotes spacing between successive period candidates τ. In the conventional ACF, $\Delta\tau = 1$, that is, the spacing between fundamental period candidates τ equals the sampling interval. Later in this section we will describe an algorithm which allows a very dense sampling of τ (small $\Delta\tau$). This has the consequence that all the sets $\kappa_{\tau,m}$ in (13) contain exactly one frequency bin, in which case the non-linear maximization operation vanishes and $s(\tau)$ becomes a linear function of $U_t(k)$, making it analytically more tractable.

The basic idea of (13) is intuitively appealing since the Fourier theorem states that a periodic signal can be represented with spectral components at integer multiples of the inverse of the period. Indeed, formulas and principles resembling (13) have been used for F0 estimation by a number of authors, under different names and in different variants – although these have used the DFT spectrum instead of an auditorily motivated representation. Already in 1960s and 70s, Schroeder introduced the *frequency histogram* and Noll the *harmonic sum spectrum* (see [46, p.414]). Parsons [47] and de Cheveigné [24] discuss *harmonic selection* methods, and more recently, Walmsley [48] uses the name *harmonic transform* for a similar technique. In the time domain, these techniques can be implemented using a bank of comb filters, where each filter has its characteristic feedback delay τ and the energy at the output of the filter defines the salience. In the auditory modeling literature, Cariani [49] proposed to use comb filters to separate concurrent vowels with different F0s. Also, the strobed temporal integration mechanism of Patterson [50, p.186] is closely related.

Equation (13) and other comb-filter like solutions have two advantages compared to the ACF. First, it is clear that (13) computes the salience of the period τ using only spectral components that are related to the period in question. This improves the robustness in polyphonic signals, since the spectral components between the partials have no effect on $s_t(\tau)$, which improves the signal-to-noise ratio (SNR) of the estimation. Secondly, it is very difficult to achieve a wide pitch range using the ACF. This is because any signal containing significant low frequency components shows high correlation for short lags (high frequencies). In polyphonic signals, the ACF is not robust above about 600 Hz: it is not able to handle the so-called "spectral pitch".[3] The proposed salience function (13) behaves robustly for a pitch range of at least 40 Hz – 2.1 kHz and has no theoretical upper limit.

The weights $w(\tau, m)$ determine the mapping from $U_t(k)$ to $s_t(\tau)$. These have been studied in [32], where the following parametric form was found:

$$w(\tau, m) = \frac{f_{\mathrm{s}}/\tau + \epsilon_1}{m f_{\mathrm{s}}/\tau + \epsilon_2}. \qquad (15)$$

Note that f_{s}/τ is the F0 value corresponding to τ and that (15) reduces to $1/m$ if the moderation terms ϵ_1 and ϵ_2 are omitted. The terms $\epsilon_1 \approx 20$ Hz and $\epsilon_2 \approx 320$ Hz are important for low-frequency partials and for low F0s. The sum in (13) can be limited to $M = 20$ terms, since weights beyond that are relatively small. As explained in Fig. 5, the higher partials are mapped to the position of the fundamental and its few multiples due to the rectification at subbands, and as a consequence, the entire harmonic series of a sound contributes to the salience.

The form of (15) allows fast computation of the saliences $s_t(\tau)$ as follows. First, $U_t(k)$ is filtered using only the denominator of (15), replacing the numerator with unity. This can be done since the denominator depends only on the frequency of the partial and not on the period. Then $s'_t(\tau)$ is computed using (13), but omitting the weights $w(\tau, m)$. Finally, each period $s'_t(\tau)$ is weighted by the numerator of (15).

It remains to choose the value of p in (7). We tested two values, $p = 1$ (magnitude spectrum) and $p = 2$ (power spectrum), optimizing the parameters ϵ_1 and ϵ_2 in (15) in both cases and monitoring the resulting salience functions. The value $p = 1$ led consistently to more reliable analysis results and was therefore chosen.

Varying p is closely related to the *generalized ACF* [52], defined as

$$r_{\mathrm{gen}}(\tau) = \mathrm{IDFT}(|\mathrm{DFT}(\mathbf{x})|^p), \qquad (16)$$

where $\mathbf{x}$ denotes the signal under analysis. The conventional ACF is obtained with $p = 2$. As discussed by Tolonen and Karjalainen in [27], choosing a proper value for p improves the reliability and noise robustness of the F0 analysis. They suggest using the value 0.67.

E. Iterative estimation and cancellation

The global maximum of the function $s_t(\tau)$ in frame t is a robust indicator of one of the correct F0s in polyphonic signals. However, the second or third-highest peak is often due to the same sound and located at τ that is half or twice the position of the highest peak. Therefore we employ an iterative technique where each detected sound is cancelled from the mixture before deciding the next F0. A similar idea has been previously utilized for example in [24], [33], and [15].

Algorithm 1: Fast search of the maximum of $s(\tau)$

```
1  $Q \leftarrow 1$
2  $\tau_{\mathrm{low}}(1) \leftarrow \tau_{\min}$
3  $\tau_{\mathrm{up}}(1) \leftarrow \tau_{\max}$
4  $q_{\mathrm{best}} \leftarrow 1$
5  while $\tau_{\mathrm{up}}(q_{\mathrm{best}}) - \tau_{\mathrm{low}}(q_{\mathrm{best}}) > \tau_{\mathrm{prec}}$ do
6      # Split the best block and compute new limits
7      $Q \leftarrow Q + 1$
8      $\tau_{\mathrm{low}}(Q) \leftarrow (\tau_{\mathrm{low}}(q_{\mathrm{best}}) + \tau_{\mathrm{up}}(q_{\mathrm{best}}))/2$
9      $\tau_{\mathrm{up}}(Q) \leftarrow \tau_{\mathrm{up}}(q_{\mathrm{best}})$
10     $\tau_{\mathrm{up}}(q_{\mathrm{best}}) \leftarrow \tau_{\mathrm{low}}(Q)$
11     # Compute new saliences for the two block-halves
12     for $q \in \{q_{\mathrm{best}}, Q\}$ do
13         Calculate $s_{\max}(q)$ using Equations (13)-(14)
           with $w(\tau, m) = \frac{f_{\mathrm{s}}/\tau_{\mathrm{low}}(q)+\epsilon_1}{m f_{\mathrm{s}}/\tau_{\mathrm{up}}(q)+\epsilon_2}$
               $\tau = (\tau_{\mathrm{low}}(q) + \tau_{\mathrm{up}}(q))/2$
14             $\Delta\tau = \tau_{\mathrm{up}}(q) - \tau_{\mathrm{low}}(q)$
15     end
16     # Search again the best block
17     $q_{\mathrm{best}} \leftarrow \arg\max_{q \in [1,Q]} s_{\max}(q)$
18 end
   Return $\hat{\tau} = (\tau_{\mathrm{low}}(q_{\mathrm{best}}) + \tau_{\mathrm{up}}(q_{\mathrm{best}}))/2$
19        $s(\hat{\tau}) = s_{\max}(q_{\mathrm{best}})$
```

Let us first look at an efficient way of finding the maximum of $s(\tau)$. Here we omit time indices for simplicity. Somewhat surprisingly, the global maximum of $s(\tau)$ and the corresponding value of τ can be found with a fast algorithm that does not require evaluating $s(\tau)$ for all τ. This is another motivation for the iterative estimation and cancellation approach where only the maximum of $s(\tau)$ is needed at each iteration.

Let us denote the minimum and maximum fundamental period of interest by $\tau_{\min}$ and $\tau_{\max}$, respectively, and the required precision of sampling τ by τ_{prec}. A fast search of the maximum of $s(\tau)$ can be implemented by repeatedly splitting the range $[\tau_{\min}, \tau_{\max}]$ into smaller "blocks", computing an upper bound for the salience within each block q, $s_{\max}(q)$, and continuing by splitting the block with the highest $s_{\max}(q)$. Let us denote the number of blocks by Q and the upper and lower limits of block q by $\tau_{\mathrm{low}}(q)$ and $\tau_{\mathrm{up}}(q)$, respectively. Index of the highest-salience block is denoted by q_{best}. The algorithm starts with only one block with upper and lower limits at $\tau_{\min}$ and $\tau_{\max}$, and then repeatedly splits the best block into two halves, as detailed in Algorithm 1.[4] As a result, it gives the maximum of $s(\tau)$ and the corresponding value of τ.

On lines 13–14 of the algorithm, in order to obtain an upper bound for the salience $s(\tau)$ within range $[\tau_{\mathrm{low}}(q), \tau_{\mathrm{up}}(q)]$, Eq. (13) is evaluated using the given values for $w(\tau, m)$, τ, and $\Delta\tau$. Splitting a block later on can only decrease the value of $s_{\max}(q)$ when computed for the new block-halves. Note that the best block has to be re-sought after each splitting in order to quarantee convergence to the global maximum.

In addition to being fast to compute, Algorithm 1 allows

[3] To the author's knowledge, the best solution so far for normalizing out this problem is the YIN algorithm by de Cheveigné and Kawahara [51].

[4] In practice, it is even more efficient to start with $[(\tau_{\max} - \tau_{\min})/\tau_{\mathrm{prec}}]^{1/2}$ blocks because this narrows the ranges $\kappa_{\tau,m}$ in (14).

TABLE I

Summary of the parameters of the proposed method

Auditory filterbank	70 subbands between 65Hz and 5200Hz
Level compression in (5)	$\nu = 0.33$
Moderation terms in (15)	$\epsilon_1 = 5$Hz/20Hz (46ms/93ms frame), $\epsilon_2 = 320$Hz
Spectrum power in (7)	$p = 1$
Cancellation weight	$d = 1$
Polyphony estim. in (17)	$\gamma = 0.66$

searching the maximum of $s(\tau)$ with a very high accuracy, that is, with a high precision of the found period $\hat{\tau}$.

The iterative estimation and cancellation goes as follows:

1) A residual spectrum $U_{\mathrm{R}}(k)$ is initialized to equal $U_t(k)$, and a spectrum of detected sounds $U_{\mathrm{D}}(k)$ to zero.
2) A fundamental period $\hat{\tau}$ is estimated using $U_{\mathrm{R}}(k)$ and Algorithm 1. The maximum of $s(\tau)$ determines $\hat{\tau}$.
3) Harmonic partials of $\hat{\tau}$ are located in $U_{\mathrm{R}}(k)$ at bins $\langle mK/\tau \rangle$. The magnitude spectrum of the Hamming window is translated to these frequencies, weighted by $w(\hat{\tau}, m)U_{\mathrm{R}}(\langle mK/\tau \rangle)$, and added to $U_{\mathrm{D}}(k)$.
4) The residual spectrum is recalculated as
   $U_{\mathrm{R}}(k) \leftarrow \max(0, U(k) - dU_{\mathrm{D}}(k))$,
   where $d \approx 1$ controls the amount of the subtraction.
5) If there are sounds remaining in $U_{\mathrm{R}}(k)$, return to Step 2.

Note that the purpose of the cancellation is ultimately to suppress harmonics and subharmonics of $\hat{\tau}$ in $s(\tau)$. This should be done in such a way that the residual is not corrupted too much to detect the remaining sounds at the coming iterations. These conflicting requirements are effectively met by weighting the partials of a detected sound by $w(\tau, m)$ in Step 3 before adding them to $U_{\mathrm{D}}(k)$. In practice this means that the higher partials are not entirely cancelled from the mixture since $w(\tau, m) \approx 1/m$.

When the number of sounds in the mixture is not given, it has to be estimated. This task, *polyphony estimation*, is accomplished by stopping the iteration when a newly-detected sound $\hat{\tau}_j$ at iteration j no longer increases the quantity

$$S(j) = \frac{\sum_{i=1}^{j} s(\hat{\tau}_i)}{j^{\gamma}}, \tag{17}$$

where $\gamma = 0.66$ was found empirically. Note that $S(j)$ would be monotonically decreasing for $\gamma = 1$ (average of $s(\hat{\tau}_i)$:s) and monotonically increasing for $\gamma = 0$ (sum). The value of j maximizing (17) is taken as the estimated polyphony $\hat{P}$.

Table I summarizes the parameters of the proposed method.

III. Results

Simulation experiments were carried out to evaluate the accuracy of the proposed method in analyzing polyphonic music and multiple-speaker speech signals. The results are compared with two reference methods [27] and [33], which have been shown to be quite accurate and for which reliable implementations were available. Also, we discuss alternative configurations of the proposed system where either (a) the auditory model is replaced with a DFT-based analysis front-end or (b) the iterative estimation and cancellation mechanism is replaced with a joint estimator.

A. Reference methods

The first reference method, denoted by "TK", has been proposed by Tolonen and Karjalainen in [27]. The authors used it to analyze mixtures of music and speech sounds. The method is motivated by an auditory model but divides an input signal into two channels only, below and above 1 kHz. An implementation was carefully prepared based on the reference, and the original code by the authors was used in the warped linear prediction part of the algorithm.

The second reference method, denoted by "AK", was proposed by the present author in [33] and is based on spectral techniques. The method was originally designed for polyphonic music transcription.

Two alternative configurations of the proposed method are used in the evaluations in order to investigate the importance and possible drawbacks of the described techniques. The first configuration, denoted by "alt-DFT", allows us to study the role of the auditory model. It is otherwise identical to the proposed method but does not apply half-wave rectification at the subbands (see Sect. II-B). As a result, the auditory filterbank does not need to be calculated at all, but $U_t(k)$ is obtained from (11), where the compression coefficients $\gamma_{c,t}$ were computed from the Fourier spectrum. All parameters of the system were separately optimized for this configuration.

Another configuration, denoted by "alt-JOINT", replaces the iterative estimation and cancellation with an algorithm where all F0s are estimated jointly. This allows us to investigate how the iterative search strategy affects the results. The joint estimator has been described in [32] and is not detailed here.

B. Results for music signals

Test cases for musical signal analysis were obtained by mixing recorded samples from musical instruments. 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 percussions.

Semirandom sound mixtures were generated by first allotting an instrument and then a random note from its playing range, restricting the pitch between 40 Hz and 2.1 kHz when 93 ms analysis frame was used and between 65 Hz and 2.1 kHz when 46 ms frame was used. This was repeated to get the desired number of sounds which were mixed with equal mean-square levels. Varying the relative levels would make the task even harder, but this was not tested. One thousand test cases were generated for mixtures of one, two, four, and six sounds. One analysis frame immediately after the onset of the sounds was given to the multiple-F0 estimators. The onset of a sound was defined to be at the time where the waveform reached $1/3$ of its maximum value over the beginning 200ms.

For the reference method TK, the test samples were limited below 530 Hz in pitch (2.1 kHz for the other methods), because the accuracy of the method degrades rapidly beyong that. This seems to be due to the limitations of ACF for high F0s as discussed in Sect. II-D.

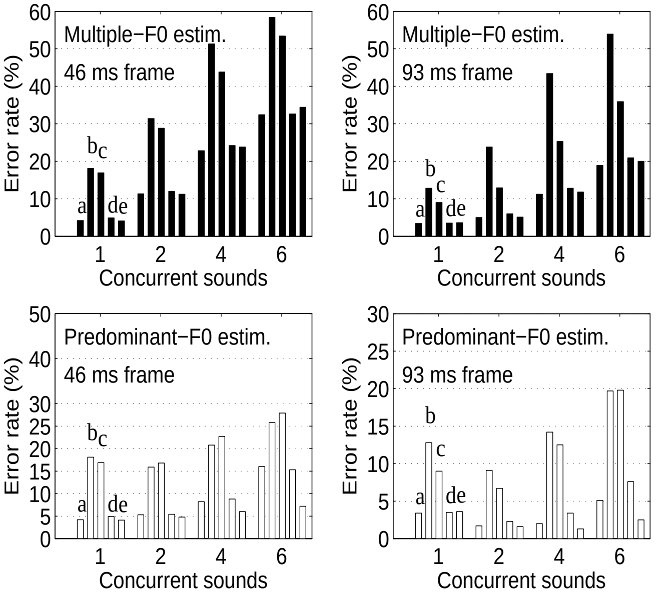

Fig. 7. Multiple-F0 estimation (top) and predominant-F0 estimation (bottom) results in 46 ms and 93 ms analysis frames. The number of concurrent sounds varied from 1 to 6. Reading left to right, each stack of six thin bars corresponds to the error rates of (a) proposed method, (b) reference TK, (c) reference AK, (d) configuration alt-DFT, (e) configuration alt-JOINT.

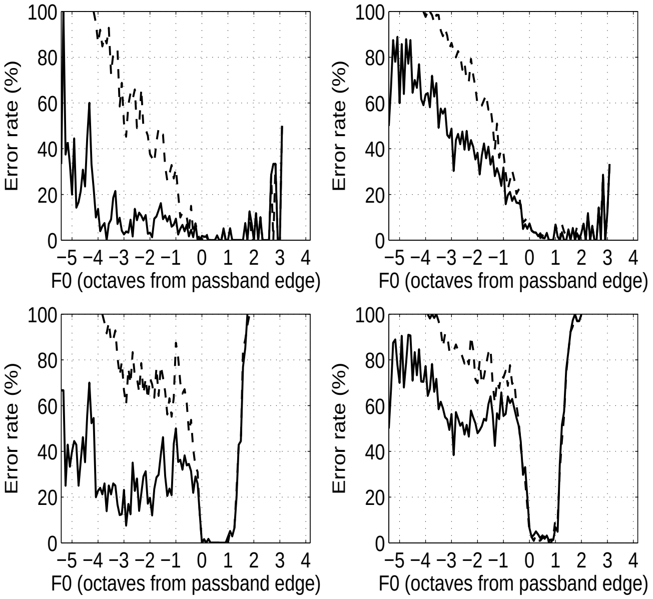

Fig. 8. Error rates for highpass filtered (top) and bandpass filtered signals (bottom). The left panels show results for isolated sounds and right panels for two-sound combinations. The errors are shown as a function of the F0, expressed in relation to the passband's lower edge. The solid line shows results for the proposed method and dashed line for the configuration alt-DFT.

Figure 7 shows F0 estimation results of the proposed and the reference methods in 46-ms and 93-ms analysis frames. Here the number of F0s to extract, the polyphony, was given as a side-information to the estimators: we will evaluate the polyphony estimation separately. 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. A correct F0 estimate was defined to deviate less than 3% from the reference F0, making it "round" to a correct musical note.

As can be seen, the proposed method outperforms the reference methods TK and AK clearly in all polyphonies. Interestingly, the configuration alt-DFT performs almost equally well in these clean, wideband signals. This would indicate that music signals that contain no drums can be processed quite well without resorting to the use of an auditory model. This is because most of the energy of the musical sounds is at their low harmonics, for which the bandwise non-linearity (rectification) is less important from the F0 analysis viewpoint. Concerning the iterative search procedure, in turn, the configuration alt-JOINT does not perform better in multiple-F0 estimation despite of being considerably more complex computationally (see [32]). In predominant-F0 estimation, the joint estimator is better in choosing the most reliable among the estimates that it has.

Figure 8 compares the robustness of the proposed method and the configuration alt-DFT, when only a part of the entire spectrum can be used for F0 estimation. This can be the situation for example when a noise source (such as drums) occupies the other bands. The upper panels show error rates for a highpass-filtered signals. Four cutoff frequencies, 250 Hz, 500 Hz, 1 kHz, and 2 kHz, were applied, and the results are averaged over these. The error rates are shown as a function of F0s, which vary from 5.5 octaves below the cutoff to 2.5 octaves above the cutoff frequency. The upper left panel shows results for isolated sounds and the upper right panel for two-sound combinations. The proposed method is significantly more robust than the alt-DFT configuration: for monophonic sounds, F0 estimation can be performed in about 90% of cases even when only partials 4 octaves above the fundamental are present. In brief, the auditory model based method is clearly better in utilizing the higher-order overtones of a harmonic sound. This is due to the rectification applied at subbands as explained around Fig. 5. In two sound combinations, the robustness difference between the methods is still clear, although often the estimation is confused by the other sound, especially if it has many strong partials at the passband.

The lower panels of Fig. 8 show F0 estimation results when only one-octave band of the signal is used. The lower boundary of the band was located at the above-mentioned four positions, and the results are averaged over these. F0 values at the x-axis are expressed in relation to the lower edge of the band. As expected, when the fundamental partial of the sound is within the passband (F0 between octaves 0 and 1 in the figure), errors are seldom made. On the other hand, F0 estimation beyond the band is hopeless since all the partials are filtered out. Again, the auditory model based method is significantly more robust than the alt-DFT configuration.

Figure 9 shows F0 estimation results in varying levels of wideband (50Hz – 10kHz) pink noise. As discussed in [33], this noise type is the most disturbing for F0 estimation, as

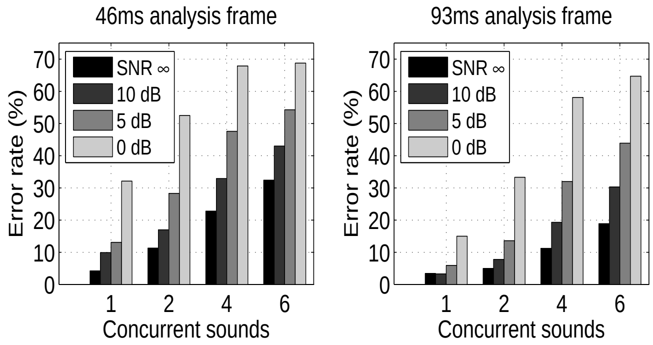

Fig. 9. F0 estimation results in varying levels of wideband pink noise. The left and right panels shows error rates in 46ms and 93ms analysis frames, respectively.

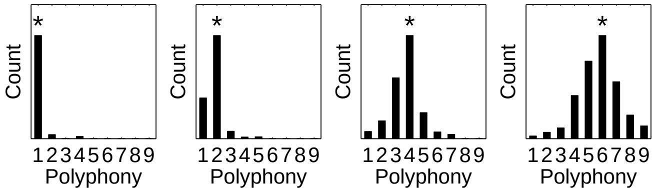

Fig. 10. The bars show histograms of polyphony estimates for the proposed method and a 93 ms analysis frame. The asterisks indicate the true polyphony (1, 2, 4, and 6, from left to right).

compared with same levels of white noise or drum sounds. The signal-to-noise ratio (SNR) is here defined as the ratio between the noise and the *sum* of the musical sounds in the analysis frame. Thus, the SNR from the viewpoint of an individual sound is much worse in higher polyphonies.

Figure 10 illustrates the results of estimating the number of concurrent sounds in a 93 ms analysis frame. The asterisk indicates true polyphony in each panel, and the bars show a histogram of the estimates. The estimation can be done only approximately, and it seems that more than one analysis frame would be needed to do it more accurately.

C. Results for speech signals

Speech signals were obtained from the CMU ARCTIC database of Carnegie Mellon University [53]. We used a total of 4×1132 recorded utterances from two male and two female US English speakers. Multiple-speaker speech signals were simulated by mixing signals from the database. The mixed signals were allotted independently from the database, however ensuring that the same speaker did not occur twice in a mixture. The root mean square levels of the signals were normalized over the entire utterance before mixing, and the mixture signals were truncated according to the shortest utterance. Two hundred independently randomized test cases were generated for one, two, and four-speaker mixtures.

Reference F0 curves were obtained by analyzing each utterance in isolation using the Praat program [54]. The CMU ARCTIC database includes pitch-marks extracted from the electroglottogram (EGG) signals using the CMU Sphinx program, but no hand corrections had been made on these, and especially the voicing information was found very unreliable. Therefore the Praat output is used as the "ground truth".

TABLE II

RESULTS FOR SINGLE-SPEAKER SIGNALS IN 32MS AND 64MS FRAMES

	Gross errors (%)		Fine errors (cent)	
Method	32ms	64ms	32ms	64ms
Proposed method	2.9	1.1	46	43
Reference TK [27]	6.3	3	40	49
alt-DFT configuration	1.2	0.9	34	43

To ensure that the Praat estimates were reasonably reliable, we compared them with the pitch-marks provided in the database, considering only segments where both sources claimed the signal to be voiced. As a result, gross discrepancies (>20% difference in F0) were found in 1.9% of the frames, and the standard deviation of the remaining fine errors was 27 cents (there are 1200 cents in an octave).

In all the results to be presented, the F0s were estimated independently in each analysis frame, without attempting to track a continuous pitch curve over the utterances.

Table II shows results for single-speaker signals (isolated utterances) using the proposed method, reference TK, and the alt-DFT configuration. The reference method AK performs poorly in short analysis frames and is therefore not used here. *Gross error* rates were computed as the percentage of time frames where the estimate differed more than 20% from the Praat reference. *Fine errors* were computed for the remaining frames as the standard deviation of the difference between the estimate and the Praat references, measured in cents. Only voiced frames were processed: voicing detection was not implemented. Both the auditory model based and the alt-DFT configuration perform well, within the limits of Praat's reliability.

Figure 11 shows the gross error rates for multiple-speaker speech signals. Estimating the number of speakers was not attempted, but the estimators were informed about the number of voiced speakers in each frame, and only this amount of F0s were extracted.[5] Here the reference method TK performs much better than for musical sounds, although still being inferior to the proposed method. The auditory model based method and the alt-DFT configuration perform approximately equally.

Figure 12 shows results for highpass-filtered speech signals, simulating the case that the lower portions of the spectrum are missing (defective audio reproduction) or corrupted by noise. The left panel shows results for individual utterances and the right panel for two-speaker mixtures. The proposed auditory model based method degrades gracefully as a function of the cutoff frequency, whereas the alt-DFT configuration gets confused (presumably by formants) as soon as the lowest and strongest partials are dropped.

D. Discussion

For practical reasons, only two reference methods (TK and AK) could be used above. Direct comparison with other methods is difficult since the experimental conditions vary

[5] In principle, the salience values could be used for voicing estimation (cf. (17)), but further optimization would be required for speech signals.

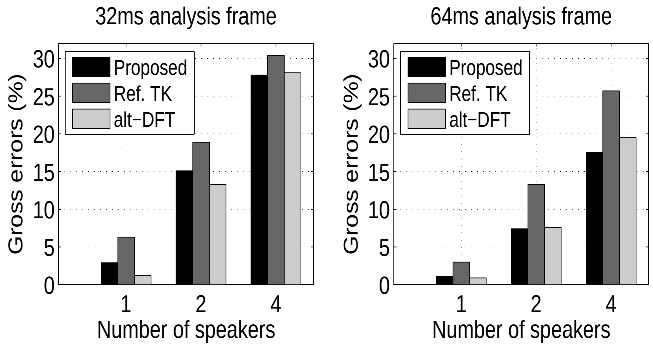

Fig. 11. Gross error rates for one, two, and four-speaker signals using the proposed method, the reference TK, and the alt-DFT configuration. The left and right panels correspond 32-ms and 64-ms analysis frames, respectively.

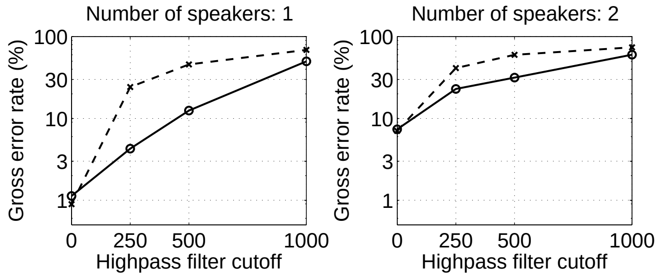

Fig. 12. Error rates for highpass-filtered speech signals. The solid line represents the proposed auditory model based method and the dashed line the alt-DFT configuration.

greatly. However, the method AK has been compared to human performance in [33], and was found to perform very similarly with trained human musicians in musical interval and chord identification tests (using 190 ms frame for the method and 1 s for humans). The proposed method achieves similar or better accuracy in twice shorter analysis frames.

In multiple-speaker pitch estimation, Wu, Wang, and Brown used the method TK as a reference in their recent work [25]. They too report clear improvement over the method TK in simulations. The experimental setup was quite different from here and prevents direct comparison of error rates. Single-speaker pitch estimation was recently studied by Cheveigné and Kawahara [51]. They report approximately 1% error rate for the best method. Here Praat was used to produce the ground truth pitch tracks, which does not allow accurate error measurement in the single-speaker case. However, the inaccuracies in the ground truth are quite harmless in multiple-speaker pitch estimation where the error rates are still relatively large.

IV. Conclusions

An auditory model based F0 estimator was proposed for polyphonic music and speech signals. A series of techniques was described to make the auditory model and the subsequent periodicity analysis computationally efficient and therefore practically appealing.

In the simulations, the method outperformed clearly the two reference methods TK and AK. Especially, both theoretical and experimental evidence was presented which shows that an auditory model based F0 estimator is good at utilizing the higher-order overtones of a harmonic sound. This is due to the nonlinearity applied at the subbands of a peripheral hearing model, which is difficult to reproduce in pure time or frequency domain methods. The improved processing of higher harmonics is particularly important in situations where the entire wideband signal is not available due to bad audio reproduction, or noise sources occupying parts of the audible spectrum. When analyzing clean, wideband signals, the auditory model did not have a clear advantage over the alt-DFT configuration, although it still performed very well. This seems to be due to the fact that most of the energy of music and speech sound is at the lowest partials, for which the bandwise nonlinearity is less important.

Robustness of the proposed pitch estimator for narrowband signals can be utilized for example in processing noise-contaminated speech signals. Time-frequency regions representing the clean speech can be located by processing the signal within subbands and by using the bandwise pitch values and their saliences as cues for speech separation. The properties of the proposed estimator were not fully exploited in the present work, which focused on processing individual time frames only. For example, the SNRs of different bands could be estimated and the subbands weighted accordingly in (7). In the future work, more effort is put on using the longer-term context in associating certain time-frequency regions to certain instruments/speakers or noise sources.

The proposed method has also been used for feature extraction in transcribing realistic musical recordings. The work has been reported in [55] and audio examples can be found at http://www.cs.tut.fi/sgn/arg/matti/demos/melofrompoly/.

V. Appendix A

This appendix describes the design of the two resonators (3) and (4) that are used to approximate the gammatone filter. The gammatone filter is defined by its impulse response as

$$g(t) = at^{\eta-1}\mathrm{e}^{-2\pi Bt}\cos(2\pi f_c t + \varphi), \tag{18}$$

where f_c is the center frequency of the filter, $a = (2\pi B)^\eta/\Gamma(\eta)$ ensures unity response at the center frequency, and $\Gamma(\cdot)$ denotes the gamma function. Choosing $\eta = 4$ leads to a shape of the power response that matches best with that found in humans. The parameter $B = 1.019b_c$ controls the ERB bandwidth b_c of the filter [56, p.256]. The phase parameter φ has no importance and a zero value can be used. We did not try to simulate any particular value.

In the following, we describe the calculation of parameters A, $\theta_{1,2}$, and $\rho_{1,2}$ in (3)–(4) assuming that the number J of resonator sections to be applied in a cascase is given. Choosing J and the optimal combination of the two resonator types can be done by trial and error since the number of alternatives is small. It was found that a cascade of four resonators, two of each type, leads to the best result.

Let us first consider the center frequency of Resonator 1. Power response of (3), after some straightforward algebra, can be written as

$$|H_1(\mathrm{e}^{\mathrm{i}\omega})|^2 = \frac{\rho_1^2(1-\cos^2(\omega))}{a_1 + a_2\cos(\omega) + A^2\cos^2(\omega)}, \tag{19}$$

where

$$a_1 = \frac{1}{4}(1-A^2)^2 + A^2\cos^2(\theta_1), \tag{20}$$

$$a_2 = -A(1+A^2)\cos(\theta_1). \tag{21}$$

Here we use angular frequencies $\omega = 2\pi f/f_{\text{s}}$ for simplicity, f_{s} denoting the sampling rate. The center frequency ω_c of the filter can be determined by differentiating with respect to $\cos(\omega)$ and setting the result to zero. This yields

$$\cos(\omega_c) = \frac{2A}{1+A^2}\cos(\theta_1). \tag{22}$$

Therefore the desired center frequency is obtained by substituting $\cos(\theta_1) = \frac{1+A^2}{2A}\cos(\omega_c)$ in (3), where $\omega_c = 2\pi f_c/f_{\text{s}}$.

The power response of Resonator 2 is obtained by replacing the numerator of (19) by $\rho_2^2/4$. Interestingly, the center frequency of this resonator obeys

$$\cos(\omega_c) = \frac{1+A^2}{2A}\cos(\theta_2). \tag{23}$$

Next, let us consider the resonator bandwidths. Provided that J resonators are applied in a cascade, we let each resonator reach $-3/J$ dB level at a point where the gammatone filter reaches -3 dB level, in order that cascade of J filters would have the desired bandwidth.

The 3-dB bandwidth of the gammatone filter (18) can be calculated as [56]

$$b_c^{\text{3dB}} = 2B\sqrt{2^{1/\eta}-1}. \tag{24}$$

A sufficiently accurate approximation of the bandwidth of the proposed resonators is obtained by assuming that only the closest pole affects their power response in the vicinity of the center frequency (see [57, p.88]). As a result, the value of A which leads to $-3/J$ dB bandwidth of b_c^{3dB} is obtained for both resonators as

$$A \approx \exp\left(-\frac{b_c^{\text{3dB}}\pi}{f_{\text{s}}\sqrt{2^{1/J}-1}}\right). \tag{25}$$

Finally, the scaling factors ρ_1 and ρ_2 in (3)–(4) are

$$\rho_1 = \frac{1}{2}(1-A^2), \tag{26}$$

$$\rho_2 = (1-A^2)\sqrt{1-\cos^2\theta_2}. \tag{27}$$

They were obtained by evaluating the power response at the filter's center frequency and requiring that to equal unity.

VI. Acknowledgements

Thanks to Tuomas Virtanen from Tampere University of Technology for constructive comments on an early version of the manuscript. This work was supported by the Academy of Finland, project No. 5213462 (Finnish Center of Excellence program 2006 - 2011).

References

[1] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley–IEEE Press, 2006.

[2] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[3] A. de Cheveigné, "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. J. Brown, Eds. Wiley–IEEE Press, 2006.

[4] J. A. Moorer, "On the transcription of musical sound by computer," *Computer Music Journal*, vol. 1, no. 4, pp. 32–38, 1977.

[5] C. Chafe, J. Kashima, B. Mont-Reynaud, and J. Smith, "Techniques for note identification in polyphonic music," in *International Computer Music Conference*, Vancouver, Canada, 1985, pp. 399–405.

[6] M. Piszczalski, "A computational model of music transcription," Ph.D. dissertation, Univ. of Michigan, Ann Arbor, 1986.

[7] R. Maher and J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, Apr. 1994.

[8] D. K. Mellinger, "Event formation and separation of musical sound," Ph.D. dissertation, Stanford University, Stanford, USA, 1991.

[9] K. Kashino and H. Tanaka, "A sound source separation system with the ability of automatic tone modeling," in *International Computer Music Conference*, Tokyo, Japan, 1993, pp. 248–255.

[10] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[11] D. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," *Speech Communication*, vol. 27, no. 3, pp. 351–366, 1999.

[12] A. Sterian, "Model-based segmentation of time-frequency images for musical transcription," Ph.D. dissertation, MusEn Project, University of Michigan, Ann Arbor, 1999.

[13] A. Bregman, *Auditory scene analysis*. Cambridge, USA: MIT Press, 1990.

[14] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[15] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic Western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, 2006.

[16] A. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 2, pp. 679–694, 2006.

[17] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004, pp. 297–300.

[18] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference*, Berlin, Germany, 2000.

[19] P. Lepain, "Polyphonic pitch extraction from musical signals," *Journal of New Music Research*, vol. 28, no. 4, pp. 296–309, 1999.

[20] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003.

[21] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *International Conference on Music Information Retrieval*, Barcelona, Spain, Oct. 2004, pp. 318–325.

[22] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, pp. 267–296.

[23] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 233–245, 1992.

[24] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model for auditory processing," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3271–3290, 1993.

[25] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[26] K. D. Martin, "Automatic transcription of simple polyphonic music: robust front end processing," MIT Media Laboratory Perceptual Computing Section, Tech. Rep. 399, 1996.

[27] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[28] M. Marolt, “A connectionist approach to transcription of polyphonic piano music,” *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[29] A. J. M. Houtsma, “Pitch perception,” in *Hearing—Handbook of Perception and Cognition*, 2nd ed., B. C. J. Moore, Ed. San Diego, California: Academic Press, 1995, pp. 267–295.

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

[31] A. P. Klapuri, “A perceptually motivated multiple-F0 estimation method for polyphonic music signals,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2005, pp. 291–294.

[32] ——, “Multiple fundamental frequency estimation by summing harmonic amplitudes,” in *International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

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

[34] R. D. Patterson, “Auditory filter shapes derived with noise stimuli,” *Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 640–654, 1976.

[35] E. de Boer and de Jongh H. R., “On cochlear encoding: Potentials and limitations of the reverse-correlation technique,” *Journal of the Acoustical Society of America*, vol. 63, no. 1, pp. 115–135, 1978.

[36] M. J. Hewitt and R. Meddis, “An evaluation of eight computer models of mammalian inner hair-cell function,” *Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 904–917, 1991.

[37] C. J. Plack and R. P. Carlyon, “Loudness perception and intensity coding,” in *Hearing—Handbook of Perception and Cognition*, 2nd ed., B. C. J. Moore, Ed. San Diego, California: Academic Press, 1995, pp. 123–160.

[38] R. Meddis and M. J. Hewitt, “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. II: Phase sensitivity,” *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2894, 1991.

[39] P. A. Cariani and B. Delgutte, “Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch,” *Journal of Neurophysiology*, vol. 76, no. 3, pp. 1698–1734, 1996.

[40] B. C. J. Moore, “Frequency analysis and masking,” in *Hearing—Handbook of Perception and Cognition*, 2nd ed., B. C. J. Moore, Ed. San Diego, California: Academic Press, 1995, pp. 161–205.

[41] R. D. Patterson and J. Holdsworth, “A functional model of neural activity patterns and auditory images,” in *Advances in speech, hearing and language processing*, W. A. Ainsworth, Ed. Greenwich, Connecticut: JAI Press, 1996, pp. 551–567.

[42] M. Slaney, “An efficient implementation of the Patterson Holdsworth auditory filter bank,” Perception Group, Advanced Technology Group, Apple Computer, Tech. Rep. 35, 1993.

[43] R. Meddis, “Simulation of mechanical to neural transduction in the auditory receptor,” *Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.

[44] M. Karjalainen and T. Tolonen, “Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, USA, 1999.

[45] B. C. J. Moore, Ed., *Hearing—Handbook of Perception and Cognition*, 2nd ed. San Diego, California: Academic Press, 1995.

[46] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin Heidelberg: Springer, 1983.

[47] T. Parsons, “Separation of speech from interfering speech by means of harmonic selection,” *Journal of the Acoustical Society of America*, vol. 60, no. 4, 1976.

[48] P. Walmsley, “Signal separation of musical instruments. Simulation-based methods for musical signal decomposition and transcription,” Ph.D. dissertation, Department of Engineering, University of Cambridge, Sept. 2000.

[49] P. Cariani, “Recurrent timing nets for auditory scene analysis,” in *International Joint Conference on Neural Networks*, Portland, Oregon, July 2003, pp. 1575–1580.

[50] R. D. Patterson, “Auditory images: How complex sounds are represented in the auditory system,” *Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 4, pp. 183–190, 2000.

[51] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[52] H. Indefrey, W. Hess, and G. Seeser, “Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain—preliminary results,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, Florida, 1985, pp. 415–418.

[53] J. Kominek and A. W. Black, “CMU ARCTIC databases for speech synthesis,” School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.

[54] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 4.3.14) [computer program],” 2005. [Online]. Available: http://www.praat.org/

[55] M. Ryynänen and A. Klapuri, “Transcription of the singing melody in polyphonic music,” in *International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[56] W. M. Hartmann, *Signals, sound, and sensation*. New York: Springer, 1998.

[57] K. Steiglitz, *A digital signal processing primer, with applications to digital audio and computer music*. Menlo Park, California: Addison Wesley, 1996.