A Computationally Efficient Multipitch Analysis Model

Tero Tolonen, *Student Member, IEEE,* and Matti Karjalainen, *Member, IEEE*

***Abstract*—A computationally efficient model for multipitch and periodicity analysis of complex audio signals is presented. The model essentially divides the signal into two channels, below and above 1000 Hz, computes a "generalized" autocorrelation of the low-channel signal and of the envelope of the high-channel signal, and sums the autocorrelation functions. The summary autocorrelation function (SACF) is further processed to obtain an enhanced SACF (ESACF). The SACF and ESACF representations are used in observing the periodicities of the signal.**

The model performance is demonstrated to be comparable to those of recent time-domain models that apply a multichannel analysis. In contrast to the multichannel models, the proposed pitch analysis model can be run in real time using typical personal computers. The parameters of the model are experimentally tuned for best multipitch discrimination with typical mixtures of complex tones.

The proposed pitch analysis model may be used in complex audio signal processing applications, such as sound source separation, computational auditory scene analysis, and structural representation of audio signals. The performance of the model is demonstrated by pitch analysis examples using sound mixtures which are available for download at http://www.acoustics.hut.fi/~ttolonen/pitchAnalysis/.

***Index Terms*—Auditory modeling, multipitch analysis, periodicity analysis, pitch perception.**

I. Introduction

MANY principles have been proposed for the modeling of human pitch perception and for practical pitch determination of simple audio or speech signals [1]–[3]. For regular signals with harmonic structure, such as clean speech of a single speaker, the problem is solved quite reliably. When the complexity increases further, e.g., when harmonic complexes of sounds or voices are mixed in a single signal channel, the determination of pitches is generally a difficult problem that has not been solved satisfactorily.

Computational algorithms for multipitch identification, for instance, in automatic transcription of polyphonic music, have been around for over 20 years. The first systems had typically substantial limitations on the content, and they were only able to detect up to two simultaneous harmonic tones [4]–[9]. The more recent systems have advanced in performance [10]–[14] allowing more simultaneous tones to be detected with greater accuracy.

The concept of pitch [15] refers to auditory perception and has a complex relationship to physical properties of a signal. Thus, it is natural to distinguish it from the estimation of fundamental frequency and to apply methods that simulate human perception. Many such approaches have been proposed and they generally follow one of two paradigms: place (or frequency) theory and timing (or periodicity) theory. Neither of these in pure form has been proven to show full compatibility with human pitch perception and it is probable that a combination of the two approaches is needed. Recently it has been demonstrated that a peripheral auditory model that uses time-domain processing of periodicity properties shows ability to simulate many known features of pitch perception which are often considered to be more central [16], [17]. Such models are attractive since auditory processes may be simulated with relatively straightforward digital signal processing (DSP) algorithms. Additional features may be readily included using, e.g., frequency domain algorithms if desired.

The unitary pitch analysis model of Meddis and O'Mard [16] and its predecessors by Meddis and Hewitt [17] are among the best known recent models of *time-domain* pitch analysis. The unitary model is shown to exhibit qualitatively good correspondence to human perception in many listening tasks such as missing fundamental, musical chords, etc. A practical problem with the model is that, despite of its quite straightforward principle, the overall algorithm is computationally expensive since the analysis is carried out using a multichannel auditory filterbank.

In this paper we present a multipitch analysis model that is computationally efficient. While it does not attempt to simulate the human auditory system in detail, it is still intuitive from the auditory modeling viewpoint. Our pitch[1] analysis model finds applications in complex audio signal processing tasks, such as sound source separation, computational auditory scene analysis [18]–[21], structural representation of audio signals, and content analysis techniques. The proposed model is, to a certain extent, a computationally superior simplification of the Meddis and O'Mard model which has very similar behavior, as will be demonstrated below. Additional features will be proposed in order to allow for further analysis of multipitch signals, such as musical chords and speech mixtures. The performance of the model is demonstrated by periodicity analysis examples using sound mixtures available at http://www.acoustics.hut.fi/~ttolonen/pitchAnalysis/.

[1]Following a common practice, we use term pitch even when fundamental period or pitch period could be more precise concepts.

Manuscript received January 18, 1999; revised December 27, 1999. This work was supported by the GETA Graduate School, Helsinki University of Technology, the Foundation of Jenny and Antti Wihuri (Jenny ja Antti Wihurin rahasto), Tekniikan edistämissäätiö, and Nokia Research Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dennis R. Morgan.

The authors are with Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, FIN-02015 Espoo, Finland.

Publisher Item Identifier S 1063-6676(00)09259-2.

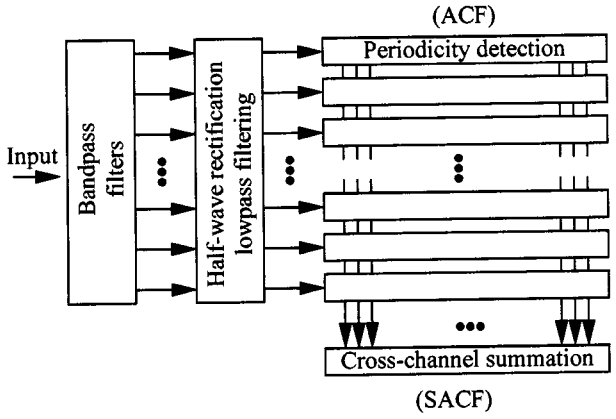

Fig. 1. Block diagram of the Meddis–O'Mard model [17].

The paper is organized as follows. In Section II, the proposed pitch analysis model is introduced and compared to pitch perception models reported in the literature. Section III describes how the periodicity representation can be enhanced so that periodicities may be more easily investigated. Section IV discusses the model parameters and shows with examples how they affect the behavior of the model, and Section V demonstrates the model performance in multipitch determination. Finally, Section VI concludes the paper with a summary and discussion.

II. Pitch Analysis Model

A. Multichannel Pitch Analysis

In many recent models of human perception, the key component is a filterbank that simulates the behavior of the cochlea. The filterbank separates a sound signal into subband channels that have bandwidths corresponding to the frequency resolution of the cochlea. A common choice is to use a gammatone filterbank [22] with channels corresponding to the equivalent rectangular bandwidth (ERB) channels of human audition [23].

Fig. 1 depicts the pitch perception model of Meddis and O'Mard [17] that uses the filterbank approach. The input signal is first divided into 40–128 channels depending on the implementation [16], [17], [21]. The signal in each channel is half-wave rectified and lowpass filtered. Essentially, this step corresponds to the detection of the envelope of the signal in each channel. From the envelope signals, a periodicity measure, such as the autocorrelation function (ACF), is computed within each channel. Finally, the ACFs are summed across the channels to yield a summary autocorrelation function (SAFC) that is used in pitch analysis.

In studies that have applied the pitch analysis paradigm of Fig. 1, several implementations are reported. In some systems, pre-processing of the signal is performed before the signal enters the filterbank. For instance, in [17] a bandpass filter is used for simulating the middle ear transfer function. In [17] the half-wave rectification and lowpass filtering block is replaced with a block that estimates the probability of neural activation in each channel. In [24], an automatic gain control block is added after the half-wave rectification and the lowpass filtering is removed.

There are several approaches for computation of the autocorrelation or a similar periodicity measure within each of the channels. The time domain approach is a common choice [16], [17], [21]. In these systems, an exponential window is applied with a window time constant that varies from 2.5 ms [17] to 25 ms [21]. Our experiments have determined that the effective length of the window should be approximately 10–30 ms so that the window spans more than one period of the pitched tone with all fundamental periods that are in the range of interest. Ellis [21] applies a logarithmic scale of the autocorrelation lag with approximately 48 samples for each octave. He motivates the use of such a scale by better resemblance with the pitch detection resolution of the human auditory system. This requires interpolation of the signals in the channels. Ellis notes that half-wave rectification is preferred over full-wave rectification in order to suppress octave errors.

Some of the pitch analysis systems prefer to use a discrete Fourier transform (DFT) based autocorrelation computation for computational efficiency [24]. This approach also allows for processing of the signal in the frequency-domain. As discussed below, nonlinear compression of the DFT magnitude may be used to enhance the performance of the pitch analysis. Such a compression is not readily implementable in a time-domain system.

Although the unitary pitch perception model of Meddis and O'Mard has been widely adopted, some studies question the general validity of the unitary pitch perception paradigm. Particularly, it has been suggested that two mechanisms for pitch perception are required: one for resolved harmonics and one for unresolved harmonics [25], [26]. The term resolved harmonics refers to the case when only one or no components fall within the 10-dB-down bandwidth of an auditory filter [27]. In the other case, the components are said to be unresolved. The present study does not attempt to answer the question on the pitch detection mechanism of the human auditory system. In fact, the proposed model has only two channels and does not attempt directly to follow human resolvability. Interestingly enough, as shown in the following subsection, the model still qualitatively produces similar and comparable results to those of the more elaborate multichannel pitch analysis systems.

The computational demands of multichannel pitch analysis systems have prohibited their use in practical applications where typically real-time performance is required. The computational requirements are mostly determined by the number of channels used in the filterbank. This motivates the development of a simplified model of pitch perception presented below that is more suitable in practical applications and still qualitatively retains the performance of multichannel systems.

B. Two-Channel Pitch Analysis

A block diagram of the proposed two-channel pitch analysis model is illustrated in Fig. 2. The first block is a pre-whitening filter that is used to remove short-time correlation of the signal. The whitening filter is implemented using warped linear prediction (WLP) as described in [28]. The WLP technique works as ordinary linear prediction except that it implements critical-band auditory resolution of spectral modeling instead of uniform frequency resolution, and can be used to reduce the

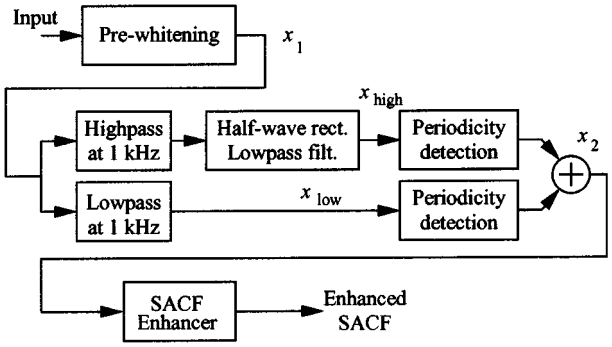

Fig. 2. Block diagram of the proposed pitch analysis model.

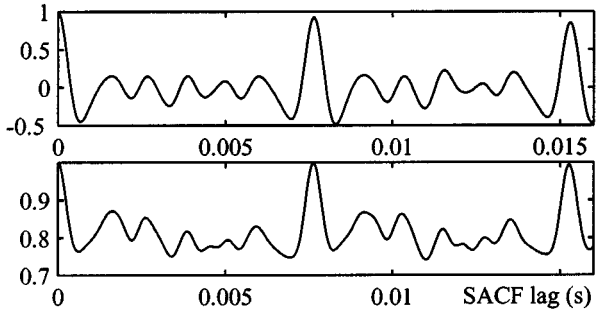

Fig. 3. Comparison of the SACF functions of the two models using the "musical chord" test signal. The two-channel SACF is plotted on the top and the Meddis–Hewitt SACF on the bottom.

filter order considerably. A WLP filter of 12th order is used here with sampling rate of 22 kHz, Hamming windowing, frame size of 23.2 ms, and hop size of 10.0 ms. Inverse filtering with the WLP model yields the pre-whitened signal. To a certain extent, the whitening filter may be interpreted as functionally similar to the normalization of the hair cell activity level toward spectral flattening due to the adaptation and saturation effects [29], [30].

The functionality of the middle part of Fig. 2 corresponds to that of the unitary multichannel pitch analysis model of Fig. 1. The signal is separated into two channels, below and above 1 kHz. The channel separation is done with filters that have 12 dB/octave attenuation in the stop-band. The lowpass block also includes a highpass rolloff with 12 dB/octave below 70 Hz. The high-channel signal is half-wave rectified and lowpass filtered with a similar filter (including the highpass characteristic at 70 Hz) to that used for separating the low channel.

The periodicity detection is based on "generalized autocorrelation," i.e., the computation consists of a discrete Fourier transform (DFT), magnitude compression of the spectral representation, and an inverse transform (IDFT). The signal x_2 in Fig. 2 corresponds to the SAFC of Fig. 1 and is obtained as

$$\begin{aligned} x_2 &= \mathrm{IDFT}(|\mathrm{DFT}(x_{\mathrm{low}})|^k) + \mathrm{IDFT}(|\mathrm{DFT}(x_{\mathrm{high}})|^k) \\ &= \mathrm{IDFT}(|\mathrm{DFT}(x_{\mathrm{low}})|^k + |\mathrm{DFT}(x_{\mathrm{high}})|^k) \end{aligned} \quad (1)$$

where x_{low} and x_{high} are the low and high channel signals before the periodicity detection blocks in Fig. 2. The parameter k determines the frequency domain compression [31]. For normal autocorrelation $k = 2$ but, as detailed in Section IV, it is advantageous to use a value smaller than 2. Note that periodicity computation using the DFT allows the control of the parameter k or the use of some other nonlinear processing of the frequency transform, e.g, application of natural logarithm resulting in the cepstrum. This is not directly possible with time-domain periodicity detection algorithms. The fast Fourier transform (FFT) and its inverse (IFFT) are typically used to speed the computation of the transforms. The last block of Fig. 2 presents the processing of the SACF (denoted x_2). This part of the algorithm is detailed in Section III.

Before comparing the performance of the models of Figs. 1 and 2, it is instructive to study the sensitivity to the phase properties of the signal in pitch analysis when using a multichannel model or a two-channel model. In the two-channel case, the low channel is phase-insensitive (except for the windowing effects) due to the autocorrelation [notice the modulus in (1)]. However, the high channel is phase-sensitive since it follows the amplitude envelope of a signal in the frequency band above 1000 Hz. Thus, all phase-sensitivity in our model is inherently caused by the high channel. This is different from the Meddis–Hewitt model where all channels are phase-sensitive since they follow the envelope of the signal in the corresponding frequency band. However, when lower channels resolve the harmonics, the difference is relatively small since in that case the autocorrelation computation removes the phase-sensitivity.

C. Comparison of Multichannel and Two-Channel Models

The performance and validity of the proposed two-channel SACF model (without pre-filtering and pre-whitening, using running autocorrelation similar to [17]) in pitch periodicity analysis is evaluated here by a comparison with the multichannel SACF model of Meddis and Hewitt. The AIM software [32] was used to compute the Meddis–Hewitt SACFs. The test signals were chosen according to [17].

The results of the "musical chord" experiment [17] with the two-channel and the multichannel models are illustrated on the top and the bottom plots of Fig. 3, respectively. In this case, the test signal consisted of three harmonic signals with fundamental frequencies 392.0, 523.2, and 659.2 Hz corresponding to tones G^4, C^5, and E^5, respectively. The G^4 tone consisted of first four harmonics, and the C^5 and E^5 tones contained the first three harmonics each. All the harmonic components were of equal amplitude. Both models exhibit an SACF peak at a lag of 7.7 ms. This corresponds to a frequency of 130 Hz (tone C^3), which is the root tone of the chord. The waveforms of the two summary autocorrelation functions are similar although the scales differ.

While it is only possible to report this experiment here, the models behave similarly with a broader range of test signals. More examples of SACF analysis are available at http://www.acoustics.hut.fi/~ttolonen/pitchAnalysis/.

III. Enhancing the Summary Autocorrelation Function

The peaks in the SACF curve produced as output x_2 of the model in Fig. 2 are relatively good indicators of potential pitch periods in the signal being analyzed as shown in Fig. 3. However, such a summary periodicity function contains much redundant and spurious information that makes it difficult to estimate

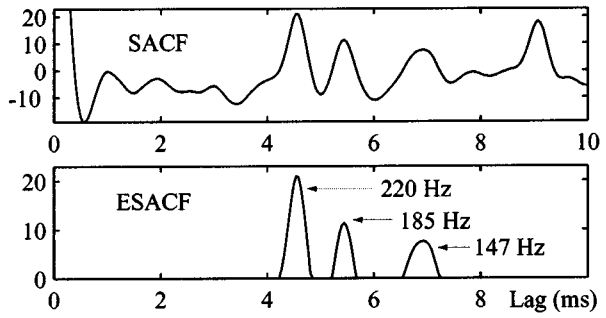

Fig. 4. An example of multipitch analysis. A test signal with three clarinet tones with fundamental frequencies 147, 185, and 220 Hz, and relative rms values of 0.4236, 0.7844, and 1, respectively, was analyzed. Top: two-channel SACF, bottom: two-channel ESACF.

which peaks are true pitch peaks. For instance, the autocorrelation function generates peaks at all integer multiples of the fundamental period. Furthermore, in case of musical chords the root tone often appears very strong though in most cases it should not be considered as the fundamental period of any source sound. To be more selective, a peak pruning technique similar to [21], but computationally more straightforward, is used in our model.

The technique is the following. The original SACF curve, as demonstrated above, is first clipped to positive values and then time-scaled (expanded in time) by a factor of two and subtracted from the original clipped SACF function, and again the result is clipped to have positive values only. This removes repetitive peaks with double the time lag where the basic peak is higher than the duplicate. It also removes the near-zero time lag part of the SACF curve. This operation can be repeated for time lag scaling with factors of three, four, five, etc., as far as desired, in order to remove higher multiples of each peak. The resulting function is called here the enhanced summary autocorrelation (ESACF).

An illustrative example of the enhanced SACF analysis is shown in Fig. 4 for a signal consisting of three clarinet tones. The fundamental frequencies of the tones are 147, 185, and 220 Hz. The SACF is depicted on the top and the enhanced SACF curve on the bottom, showing clear indication of the three fundamental periodicities and no other peaks. We have experimented with different musical chords and source instrument sounds. In most cases, sound combinations of two to three sources are resolved quite easily if the amplitude levels of the sources are not too different. For chords with four or more sources, the subsignals easily mask each other so that some of the sources are not resolved reliably. One further idea to improve the pitch resolution with complex mixtures, especially with relatively different amplitudes, is to use an iterative algorithm, whereby the most prominent sounds are first detected and filtered out (see Section VI) or attenuated properly, and then the pitch analysis is repeated for the residual.

IV. Model Parameters

The model of Fig. 2 has several parameters that affect the behavior of pitch analysis. In the following, we show with examples the effect of each parameter on the SACF and ESACF representations. In most cases, it is difficult to obtain the correct value for a parameter from the theory of human perception. Rather, we attempt to obtain model performance that is similar to that of the human perception or approximately optimal based on visual inspection of analysis results. The suggested parameter values are collected in Table I.

TABLE I
Suggested Parameter Values for the Proposed Two-Channel Model

k	window size	pre-whitening	filter order
0.67	46.4 ms	12-order WLP	2/trans. band

In the following section, there are two types of illustrations. In Figs. 5 and 7, the SACFs of one frame of the signal are plotted on the top, and one or two ESACF's that correspond to the parameter values that we found best suited are depicted on the bottom. In Figs. 6 and 8, consecutive ESACFs are illustrated as a function of time. This representation is somewhat similar to the spectrogram: time is shown on the horizontal axis, ESACF lag on the vertical axis, and the gray level of a point corresponds to the value of the ESACF. Test signals are mixtures of synthetic harmonic tones, noise, and speech signals. Each synthetic tone has the amplitude of the first harmonic equal to 1.0, and the amplitude of the nth harmonic equal to $1/n$. The initial phases of the harmonics are 0. The noise that is added in some examples is white Gaussian noise. In this work, we have used speech samples that have been recorded in an anechoic chamber and in normal office conditions. The sampling rate of all the examples is 22 050 Hz.

A. Compression of Magnitude in Transform Domain

In Section II we motivated the use of transform-domain computation of the "generalized autocorrelation" by two considerations: 1) it allows nonlinear compression of the spectral representation and 2) it is computationally more efficient. The following examples concentrate on magnitude compression and suggest that the normal autocorrelation function ($k = 2$) is suboptimal for our periodicity detection model. Exponential compression is easily available by adjusting parameter k in (1).

While we only consider exponential compression in this context, other nonlinear functions may be applied as well. A common choice in the speech processing community is to use the natural logarithm, which results in the cepstrum. Nonlinear compression in the transform domain has been studied in the context of pitch analysis of speech signals [31]. In that study, the periodicity measure was computed using the wideband signal directly without dividing into channels. It was reported that the compression parameter $k = 0.5$ gave the best results. It was also shown that pre-whitening of the signal spectrum showed a tendency of improving the pitch analysis accuracy.

Fig. 5 illustrates the effect of magnitude compression to the SACF. The upper four plots depict the SACF's that are obtained using $k = 0.2$, $k = 0.67$, $k = 1.0$, and $k = 2.0$, whereas the bottom plot shows the ESACF that is obtained from the SACF with $k = 0.67$. A Hamming window of 1024 samples (46.4 ms with a sampling frequency of 22 050 Hz) is used.

The test signal consists of two synthetic harmonic tones with fundamental frequencies 140.0 and 148.3 Hz and white Gaussian noise with signal-to-noise ratio (SNR) of 2.2 dB. The fundamental frequencies of the harmonic tones are separated

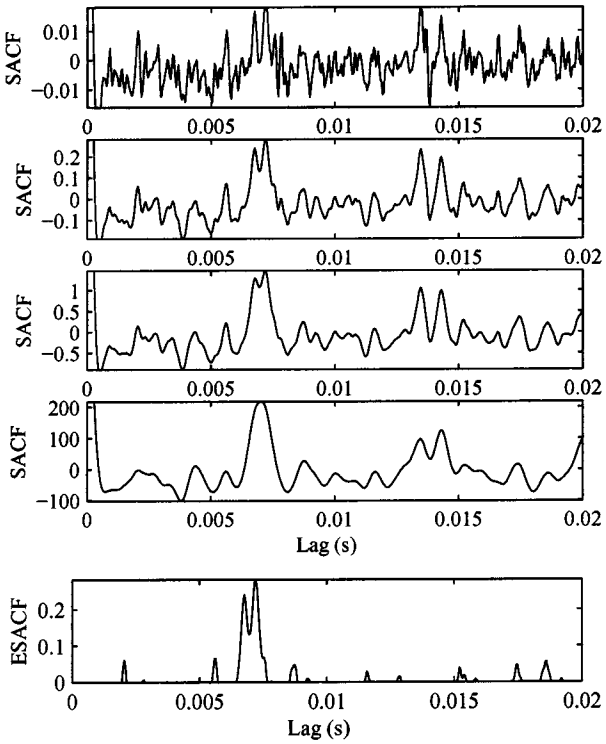

Fig. 5. An example of magnitude compression. Test signal consists of two harmonic tones with added Gaussian white noise (SNR is 2.2 dB). The first four plots from the top illustrate the SACF computed with $k = 0.2$, $k = 0.67$, $k = 1.0$, and $k = 2.0$, respectively. The bottom plot depicts the ESACF corresponding to the second plot with $k = 0.67$.

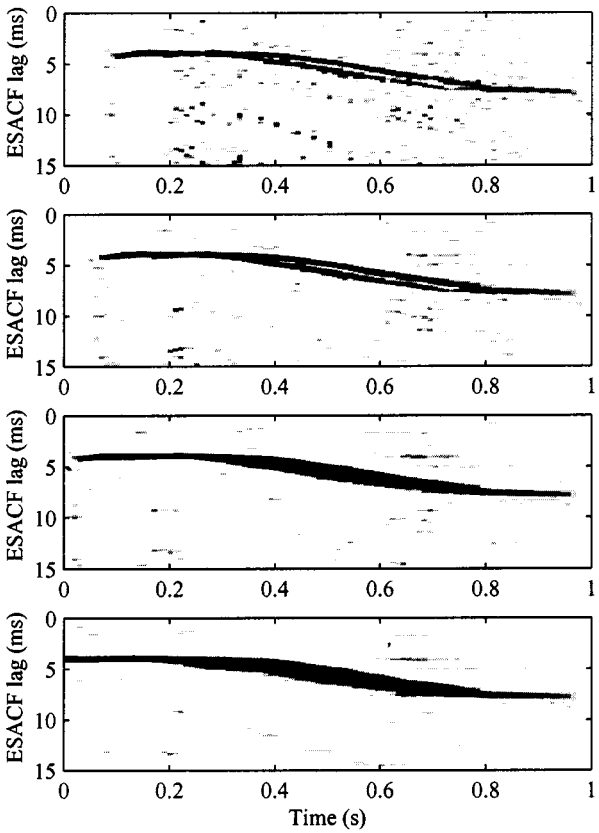

Fig. 6. An example of the effect of the window length. Test signal consists of the Finnish vowel /a/ spoken by a male. The pitch of the vowel is falling. The signal is added onto itself after a delay of 91 ms. Plots from top to bottom illustrate the ESACF computed with window length 23.2, 46.4, 92.9, and 185.8 ms, respectively.

by one semitone. The two tones are identifiable by listening to the signal, although the task is much more involved than with the signal without the additive noise (see the examples at the WWW). The example shows that the SACF peaks get broader as the value of k increases. However, the performance with low values of k is compromised by sensitivity to noise, as seen by the number and level of the spurious peaks in the top plot. According to this example, $k \simeq 0.67$ is a good compromise between lag-domain resolution and sensitivity to noise.

While we prefer the use of $k = 0.67$, in some computationally critical applications it may be useful to use $k = 0.5$ if optimized routines are available for the square-root operation but not for the cubic-root operation. It is interesting to compare exponential compression to the cepstrum where logarithmic compression is used. Cepstral analysis typically results in higher lag-domain resolution than exponential compression with $k = 0.2$, but it may be problematic with signals with low amplitude levels, since the natural logarithm approaches $-\infty$ as the signal amplitude approaches 0 [31].

B. Time-Domain Windowing

The choice of the time-domain window in correlation computation affects the temporal resolution of the method and also sets a lower bound to the lag range. The shape of the window function determines the leakage of energy in the spectral domain. It also determines the effective length of the window that corresponds to the width of the main lobe of the window in the spectral domain. We have applied a Hamming window in our pitch analysis model, although other tapered windows may be used as well. For stationary signals, increasing the window length reduces the variance of the spectral representation. However, the sound signals typically encountered in pitch analysis are far from stationary and a long window is bound to degrade the performance when the pitch is changing. Fig. 6 illustrates this trade-off.

Consecutive frames of ESACF's computed on the test signal are plotted in a spectrogram-like fashion in Fig. 6. The darker a point is in the figure, the higher is the value of the ESACF. The test signal consists of Finnish vowel /a/ spoken by a male. The pitch of the vowel is falling. The vowel signal is added onto itself after a delay of 91 ms. Listening to the signal reveals that the human auditory system is easily able to distinguish two vowels with falling pitches. The value of the hop size parameter is 10 ms, i.e., consecutive frames are separated in time by 10 ms. The Hamming window lengths of the analysis are, from the top to the bottom, 23.2, 46.4, 92.9, and 185.8 ms. The two plots on the top of Fig. 6 exhibit two distinct pitch trajectories, as desired. However, as the window length is increased, the two trajectories are merged into one broader trajectory, as shown in the two bottom plots. Clearly, the two cases on the top are preferred over the two on the bottom.

When the two top plots of Fig. 6 are compared, it is noticed that the one using a shorter window exhibits more undesired peaks in the ESACF. As expected, the use of a longer window reduces the artifacts that correspond to noise and other

un-pitched components. From Fig. 6 it is suggested that a Hamming window with a length of 46.4 ms is a good compromise.

When a tapered time-domain window, such as the Hamming window, is used, the hop size is typically chosen to be less than half of the window length. This ensures that the signal is evenly weighted in the computation. The hop size may be further reduced to obtain finer sampling of the ESACF frames in time, if desired. We have chosen a hop size equal to 10 ms which, from our experiments, seems a good compromise between the displacement of consecutive ESACF frames and computational demands. A similar hop size value is very often used in speech and audio processing applications.

C. Pre-Whitening

The pre-whitening filter that is used before the filterbank removes short-time correlation from the signal, e.g., due to formant resonance ringing in speech signals. In the spectral domain, this corresponds to flattening the spectral envelope. We thus expect the whitening to give better resolution of the peaks in the autocorrelation function.

Since whitening flattens the overall spectral envelope of a signal, it may degrade the signal-to-noise ratio of a narrowband signal since the noise outside the signal band is strengthened. However, as the following example illustrates, the whitening does not typically degrade the two-channel pitch estimator performance. Fig. 7 demonstrates the effect of pre-whitening with test signal of Fig. 5, i.e., two harmonic tones with fundamental frequencies separated by one semitone and additive white Gaussian noise. The first two plots illustrate the SACF without (top) and with (second) pre-whitening. The peaks of the whitened SACF are better separated than those without whitening. The spurious peaks that are caused by the noise still appear at a relatively low level. This is confirmed by investigation of the corresponding ESACFs in the third and the fourth plot of Fig. 7.

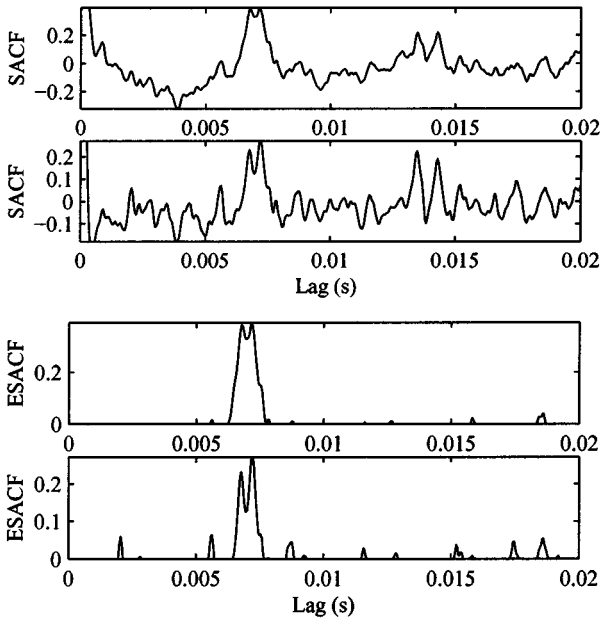

Fig. 7. An example of the effect of pre-whitening. The test signal is the same as in Fig. 5. The first two plots illustrate the SACF without (top) and with (second) pre-whitening. The third and the fourth plot depict the ESACF corresponding to the first and the second plot, respectively.

D. Two-Channel Filterbank

The choice of the filterbank parameters affects the performance of the periodicity detector quite significantly, as demonstrated by the following example. The most important parameters are the filter orders and cut-off frequencies. As discussed in Section II, the two filters are bandpass filters with passband from 70–1000 Hz for the lower channel and from 1000–10 000 Hz for the upper channel. The lowest cut-off frequency at 70 Hz is chosen so that DC and very-low-frequency disturbances are suppressed while the periodicity detection of low-pitched tones is not degraded. The crossover frequency at 1000 Hz is not a critical parameter; it may vary between 800–2000 Hz (see, e.g., [30], [33]). It is related to the upper limit of fundamental frequencies that may be estimated properly using the method and naturally also affects the lag domain temporal resolution of periodicity analysis. When a tone with a fundamental frequency higher than the crossover frequency is analyzed, the SACF is dominated by the contribution from the high channel. The high-channel compressed autocorrelation is computed after the low-pass filtering at the crossover frequency, thus, the high-channel contribution for fundamental frequencies above the crossover frequency is weak. The method is really a periodicity estimator: it is not capable of simulating properly the spectral pitch, i.e., a pitch that is based on resolved harmonics with fundamental frequency higher than 1000 Hz. Note that the other aforementioned filterbank methods are also periodicity detectors and have their fundamental frequency detection upper limit related to the lowpass filter after the half-wave rectification.

Both filters are of the Butterworth type for maximally flat passbands. The filter order that is used for every transition band is a parameter that has a significant effect on the temporal resolution of the periodicity detector. Fig. 8 shows an example of the effect of channel-separation filtering. The test signal is the same as that in the example of Fig. 6, i.e., a Finnish vowel /a/ with a falling pitch is added to itself after a delay of 91 ms. The plots from top to bottom illustrate the ESACF's when filter orders 1, 2, and 4 (corresponding to a rolloff of 6, 12, and 24 dB/octave) are used for each transition band, respectively. It is observed that the spurious peaks in the ESACF representations are reduced as the filter order is increased. By examination of Fig. 8 we conclude that filter order 2 for each transition band is the best compromise between resolution of the ESACF and the number of spurious peaks in the ESACF.

V. Model Performance

The performance of the pitch analysis model is demonstrated with three examples:

1) resolution of harmonic tones with different amplitudes;
2) musical chords that are played with real instruments;
3) ESACF representation of a mixture of two vowels with varying pitches.

The test signals can be downloaded at http://www.acoustics.hut.fi/~ttolonen/pitchAnalysis/.

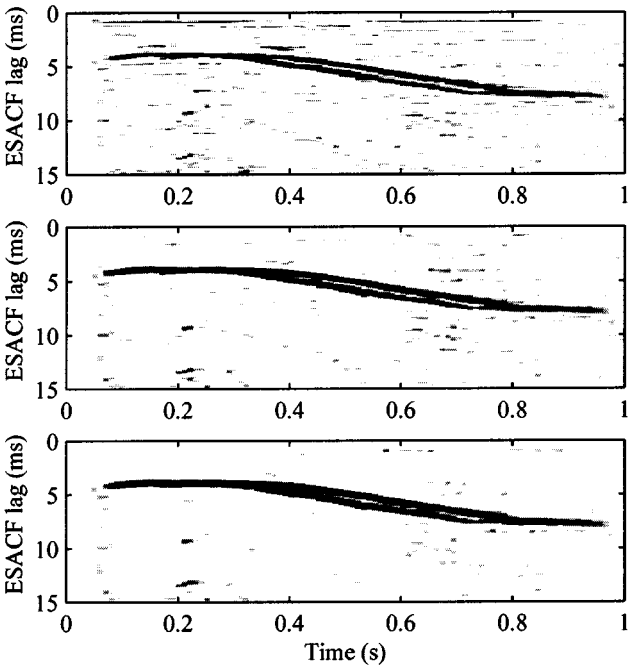

Fig. 8. An example of the effect of channel-separation filtering. Test signal is the same as in Fig. 6. The plots from top to bottom illustrate the ESACF's when the filter orders 1, 2, and 4 (corresponding to a roll-off of 6, 12, and 24 dB/octave) are used for each transition band, respectively.

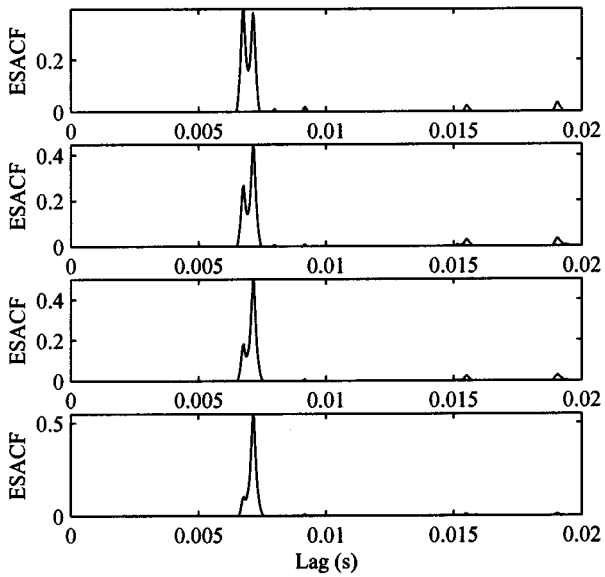

Fig. 9. Example of the resolution of harmonic tones with varying amplitudes. The amplitude ratios of the two tones are, from the top to the bottom, 0.0, 3.0, 6.0, and 10.0 dB.

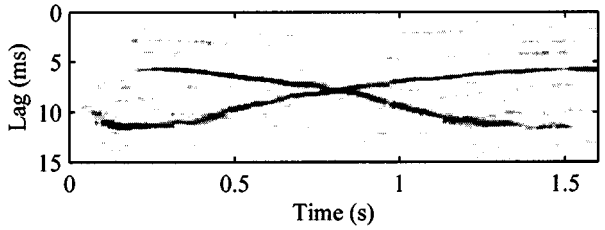

Fig. 10. An example of the pitch analysis of two vowels with crossing pitch trajectories. The test signal consists of two Finnish vowels /ae/, one with a raising pitch and the other one with a falling pitch.

Fig. 9 shows the first example where the test signal consists of two synthetic harmonic tones with different amplitude ratios. The fundamental frequencies of the tones are 140.0 and 148.3 Hz. The amplitude ratios in the plots of Fig. 9 are, from the top to the bottom, 0.0, 3.0, 6.0, and 10.0 dB. The tones are clearly separated in the top plot and the separation degrades with increasing amplitude ratio, as expected. This is in agreement with perception of the test signals; the test signal with 0.0 dB amplitude ratio is clearly perceived as two tones while at 10.0 dB the weaker signal is only barely audible.

Fig. 10 shows an ESACF representation of a mixture of two vowels with varying pitches. The two Finnish /ae/ vowels have been spoken by a male in an anechoic chamber and mixed. The figure shows two distinct trajectories with few spurious peaks.

The example of Fig. 11 illustrates pitch analysis on musical chords. The test signals consist of 2–4 clarinet tones that are mixed to get typical musical chords. The tones that are used for mixing are D_3 (146.8 Hz), $F\sharp_3$ (185.0 Hz), A_3 (220.0 Hz), and C_4 (261.6 Hz). The plots from the top to the bottom show the ESACF of one analysis frame. The test signals are, respectively, tones D_3 and A_3; D_3 and $F\sharp_3$; D_3, $F\sharp_3$, and A_3; and D_3 $F\sharp_3$, A_3, and C_4.

This example allows us to investigate the performance of the model when tones are added to the mixture. The little arrows above the plots indicate the tone peaks in the ESACF representations. The first two plots show the performance with only two tones present. In both cases, the two tones are shown with peaks of equal height. The maximum value of the peak corresponding to the tone D_3 at lag of $\sim$7 ms is almost 30 in the top plot and only a little more than 20 in the second plot. In the third plot, three tones are present. The tones are the same that were used for the first two plots, but now the D_3 peak is more pronounced than the peaks corresponding to tones $F\sharp_3$, and A_3. Finally, inclusion of tone C_4 again alters the peak heights of the other peaks. This dependence of the peak height is caused partly by the computation of the ESACF representation from the SACF representation, and partly since the tones have harmonics with colliding frequencies. In all the cases, however, the tones are clearly distinguishable from the ESACF representation.

VI. Summary and Discussion

The multipitch analysis model described above has been developed as a compromise between computational efficiency and auditory relevance. The first property is needed to facilitate advanced applications of audio and speech signal analysis. Computational auditory scene analysis (CASA) [18]–[21], structured and object-based coding of audio signals, audio content analysis, sound source separation, and separation of speech from severe background noise are among such applications. Auditory relevance is advantageous in order to enable comparison of the system performance with human auditory processing.

Computational efficiency was shown by testing the algorithm of Fig. 2 on a 300 MHz PowerPC processor (Apple Macintosh G3). Computation of the SACF using WLP pre-whitening, sample rate of 22 kHz, and frame size of 1024 samples (46 ms) took less than 7.0 ms per frame. With a 10 ms hop size, only a

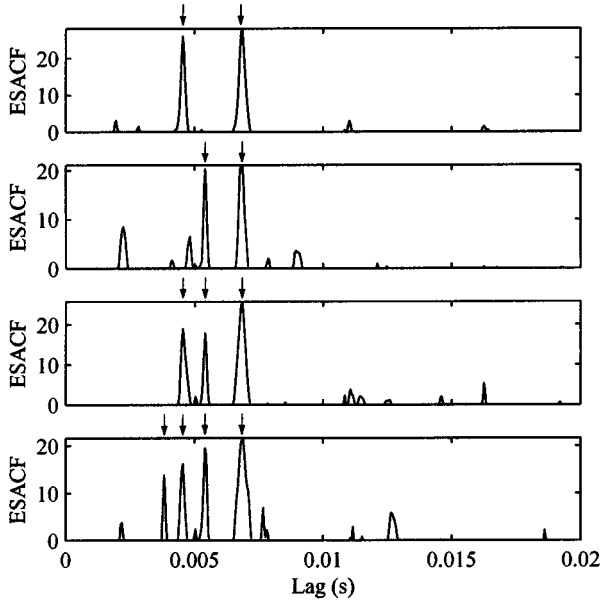

Fig. 11. Example of pitch analysis of musical chords with clarinet tones. The plots show ESACF's of signals consisting of tones D_3 and A_3 (top plot); D_3 and $F_3^\sharp$ (second plot); D_3, $F_3^\sharp$, and A_3 (third plot); and D_3 $F_3^\sharp$, A_3, and C_4 (bottom plot).

part of processor's capacity is used in real-time pitch analysis. A multichannel model with correlation in each channel could not be implemented as real-time analysis using current general purpose processors.

Although the auditory analogy of the model is not very strong, it shows some features that make it easier to interpret analysis results from the point of view of human pitch perception. The pre-whitening of the input signal, often considered useful in pitch analysis, may be seen to have minor resemblance to spectral compression in the auditory nerve.

The division of the audio frequency range into two subranges, below and above about 1 kHz, is a kind of minimum division to realize direct time synchrony of neural firings at low frequencies and envelope-based synchrony at high frequencies. Note that the model is a periodicity analyzer that does not implement spectral pitch analysis, which is needed especially if the fundamental frequency of the input signal exceeds the synchrony limit frequency of about 1 kHz. In this study we focused on the pitch analysis of low-to-mid fundamental frequencies and did not try to include spectral pitch analysis.

The computation of time-lag correlation (generalized autocorrelation) is difficult to interpret from an auditory point of view since it is carried out in the frequency domain. The only interesting auditory analogy is that the exponent for spectral compression is close to the one used in computation of the loudness density function [34]. It would be interesting to compare the result with neural interspike interval histograms which, however, are less efficient to compute.

The enhanced summary autocorrelation (ESACF) is an interesting and computationally simple means to prune the periodicity of autocorrelation function. In a typical case this representation helps in finding the fundamental periodicities of harmonic complex tones in a mixture of such tones. It removes the common periodicities such as the root tone of musical chords. This is useful in sound source separation of harmonic tones. In music signal analysis, the complement of such pruning, i.e., detection of chord periodicities and rejection of single tone pitches, might as well be a useful feature for chord analysis.

An interesting question of pitch analysis is the temporal integration that the human auditory system shows. As with a large set of other psychoacoustic features, the formation of pitch percept takes 100–200 ms to reach its full accuracy. Using a single frame length of 46 ms, corresponding to an effective Hamming window length of about 25 ms, is a compromise of pitch tracking and sensitivity to noise. Averaging of consecutive frames can be used to improve the stability of SACF with steady-pitch signals. Better temporal integration strategies may be needed when there is faster pitch variation.

This paper has dealt with multipitch analysis of an audio signal using the SACF and ESACF representations. The next step in a typical application would be to detect and describe pitch objects and their trajectories in time. Related to such pitch object detection is the resolution of the analysis when harmonic tones of different amplitudes are found in a mixture signal. As shown in Fig. 9, separation of pitch objects is easy only when the levels of tones are relatively similar. An effective and natural choice to improve the pitch analysis with varying amplitudes is to use an iterative technique, whereby the most prominent pitches are first detected and the corresponding harmonic complexes are filtered out by FIR comb filtering tuned to reject a given pitch. Tones with low amplitude level can then be analyzed from the residual signal.

The same approach is useful in more general sound (source) separation of harmonic tones [35]. The pitch lags can be used to generate sparse FIR's for rejecting (or enhancing) specific harmonic complexes in a given mixture signal. An example of vowel separation and spectral estimation of the constituent vowels is given in [35]. This can be considered as kind of multipitch prediction of harmonic complex mixtures.

REFERENCES

[1] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *J. Acoust. Soc. Amer.*, vol. 100, pp. 3491–3502, Dec. 1996.
[2] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.
[3] W. J. Hess, "Pitch and voicing determination," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, ch. 6, pp. 3–48.
[4] J. A. Moorer, "On the segmentation and analysis of continuous musical sound by digital computer," Ph.D. dissertation, Dept. Music, Stanford Univ., Stanford, CA, May 1975.
[5] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38, pp. 956–979, June 1990.
[6] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2254–2263, Apr. 1994.
[7] C. Chafe, D. Jaffe, K. Kashima, B. Mont-Reynard, and J. Smith, "Techniques for note identification in polyphonic music," Dept. Music, Stanford Univ., Stanford, CA, Tech. Rep. STAN-M-29, CCRMA, Oct. 1985.
[8] C. Chafe and D. Jaffe, "Source separation and note identification in polyphonic music," Dept. Music, Stanford University, Stanford, CA, Tech. Rep. STAN-M-36, CCRMA, Apr. 1986.
[9] M. Weintraub, "A computational model for separating two simultanious talkers," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 81–84, 1986.

[10] K. Kashino and H. Tanaka, "A sound source separation system with the ability of automatic tone modeling," *Int. Computer Music Conf.*, pp. 248–255, 1993.
[11] A. Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," in *Proc. Signal Processing IX: Theories Applications*, vol. 4, 1998, p. 2365.
[12] K. D. Martin, "Automatic transcription of simple polyphonic music: Robust front end processing," Mass. Inst. Technol., Media Lab Perceptual Computing, Cambridge, Tech. Rep. 399, 1996.
[13] ——, "A blackboard system for automatic transcription of simple polyphonic music," Mass. Inst. Technol. Media Lab. Perceptual Computing, Cambridge, Tech. Rep. 385, 1996.
[14] D. P. W. Ellis and B. L. Vercoe, "A wavelet-based sinusoid model of sound for auditory signal separation," in *Int. Computer Music Conf.*, 1991, pp. 86–89.
[15] *USA Standard Acoustical Terminology*, Amer. Nat. Stand. Inst., S1.1-1960, 1960.
[16] R. Meddis and L. O'Mard, "A unitary model for pitch perception," *J. Acoust. Soc. Amer.*, vol. 102, pp. 1811–1820, Sept. 1997.
[17] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery—I: Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, pp. 2866–2882, June 1991.
[18] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[19] M. P. Cooke, "Modeling auditory processing and organization," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 1991.
[20] G. J. Brown, "Computational auditory scene analysis: A representational approach," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 1992.
[21] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, June 1996.
[22] R. D. Patterson, "The sound of the sinusoid: Spectral models," *J. Acoust. Soc. Amer.*, vol. 96, pp. 1409–1418, Sept. 1994.
[23] B. C. J. Moore, R. W. Peters, and B. R. Glasberg, "Auditory filter shapes at low center frequencies," *J. Acoust. Soc. Amer.*, vol. 88, pp. 132–140, July 1990.
[24] M. Slaney and R. F. Lyon, "A perceptual pitch detector," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 357–360, 1990.
[25] R. Carlyon and T. M. Shackleton, "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?," *J. Acoust. Soc. Amer.*, vol. 95, pp. 3541–3554, June 1994.
[26] R. P. Carlyon, "Comments on 'a unitary model of pitch perception'," *J. Acoust. Soc. Amer.*, vol. 102, pp. 1118–1121, Aug. 1997.
[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.
[28] U. K. Laine, M. Karjalainen, and T. Altosaar, "Warped linear prediction (WLP) in speech and audio processing," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. III.349–III.352, 1994.
[29] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381–1403, Nov. 1979.
[30] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.
[31] 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 *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1985, pp. 11.11.1–11.11.4.
[32] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890–1894, Oct. 1995.
[33] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation—I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Amer.*, vol. 102, pp. 2898–2905, Nov. 1997.
[34] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.
[35] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, 1999, pp. 929–932.

Tero Tolonen (S'98) was born in Oulu, Finland, in 1972. He majored in acoustics and audio signal processing and received the M.Sc.(Tech.) and Lic.Sc.(Tech.) degrees in electrical engineering from the Helsinki University of Technology (HUT), Espoo, Finland, in January 1998 and December 1999, respectively. He is currently pursuing a postgraduate degree.

He has been with the HUT Laboratory of Acoustics and Audio Signal Processing since 1996. His research interests include model-based audio representation and coding, physical modeling of musical instruments, and digital audio signal processing.

Mr. Tolonen is a student member of the IEEE Signal Processing Society and the Audio Engineering Society.

Matti Karjalainen (M'84) was born in Hankasalmi, Finland, in 1946. He received the M.Sc. and Dr.Tech. degrees in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1970 and 1978, respectively. His doctoral dissertation dealt with speech synthesis by rule in Finnish.

From 1980 to 1986, he was Associate Professor and, since 1986, Full Professor of acoustics with the Faculty of Electrical Engineering, Helsinki University of Technology, Espoo, Finland. His research activities cover speech synthesis, analysis, and recognition, auditory modeling and spatial hearing, DSP hardware, software, and programming environments, as well as various branches of acoustics, including musical acoustics and modeling of musical instruments.

Dr. Karjalainen is a fellow of the AES and a member of ASA, EAA, ICMA, ESCA, and several Finnish scientific and engineering societies. He was the General Chair of the 1999 IEEE Workshop on Applications of Audio and Acoustics, New Paltz, NY.