MULTI-PITCH AND PERIODICITY ANALYSIS MODEL FOR SOUND SEPARATION AND AUDITORY SCENE ANALYSIS

Matti Karjalainen and Tero Tolonen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Finland
matti.karjalainen@hut.fi, tero.tolonen@hut.fi

ABSTRACT

A model for multi-pitch and periodicity analysis of complex audio signals is presented that is more efficient and practical than the Meddis and O'Mard unitary pitch perception model, yet exhibits very similar behavior. In this paper we also demonstrate how to apply this model to source separation of complex audio signals such as polyphonic and multi-instrumental music and mixtures of simultaneous speakers. Such analysis techniques are important for automatic transcription of music and structural representation of audio signals. (See also: http://www.acoustics.hut.fi/~ttolonen/icassp99/pitchdet/)

1. INTRODUCTION AND MOTIVATION

Many principles have been proposed for the modeling of human pitch perception and for practical pitch determination of audio or speech signals [1, 2]. 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. The task becomes even more difficult if the signals have to be separated, i.e., each source signal or its spectral representation must be taken apart.

The concept of pitch [3] 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 which 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 [4, 5]. Such models are attractive since auditory processes may be simulated with relatively straightforward 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 [4] and its predecessors by Meddis and Hewitt [5] 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 propose a simplified model for pitch analysis that is computationally much more efficient than the Meddis and O'Mard model, yet has very similar behavior, as will be demonstrated below. Additional features will be proposed in order to allow for further analysis of multi-pitch signals, such as musical chords and speech mixtures. The application of the model to sound source separation, useful in computational auditory scene analysis (CASA) and structural representation of audio signals, is demonstrated.

2. REDUCED COMPLEXITY MODEL OF AUDITORY PITCH PERCEPTION

A block diagram of the Meddis-O'Mard unitary pitch perception model is depicted in Fig. 1. A band-pass filterbank, most often a gammatone filterbank [6], is used to simulate the frequency selectivity of the peripheral hearing. The signal is split into channels such as ERB (equivalent rectangular bandwidth) channels and each channel is half-wave rectified and lowpass filtered (about 1 kHz) in order to simulate the activity of the hair cells. Signal periodicity is next extracted in each channel by computing its autocorrelation function (ACF) or a similar periodicity measure. Finally, the ACFs are summed from each channel to yield a summary autocorrelation (SACF) that shows the overall periodicity properties of the incoming signal. For more details, see [4, 5].

Meddis and O'Mard have compared the behavior of the unitary pitch perception model with results from psychoacoustical experiments and shown that the model is capable of simulating several important or interesting special cases of perception, at least qualitatively [4]. A problem with the unitary model from a practical application point of view is that computing of filterbank, such as 32–120 channels and their autocorrelations, is a heavy although straightforward task. To reduce this computation has been one of our motivations in this study.

The proposed simplified pitch analysis model is illustrated in Fig. 2. The middle part of the model ($x_2 \rightarrow x_3$) corresponds to the functionality of the Meddis-O'Mard model and the lower part shows extensions that will be discussed below. The simplicity of our model is based on the division of the audio frequency range to only two subchannels. Low frequencies below 1 kHz are analyzed directly by autocorrelation while high frequencies above 1 kHz

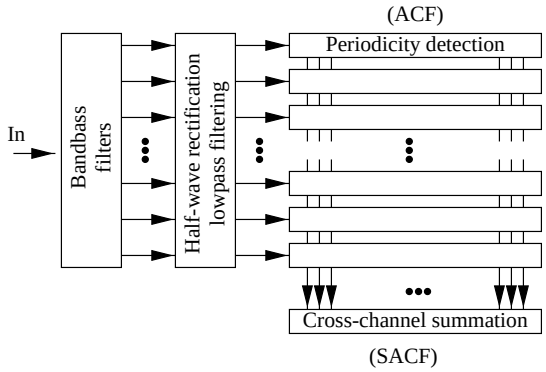

Figure 1: A block diagram of the unitary multi-channel pitch analysis model (after [4]).

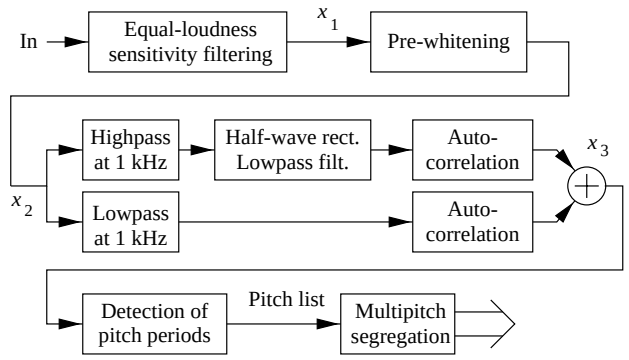

Figure 2: A block diagram of the proposed model.

are first (half-wave) rectified and low-pass filtered and then the autocorrelation is computed. Summary autocorrelation is now the sum from only two subchannels. This approach, in comparison to the model of Fig. 1, is based on assumption that at low frequencies for this particular task the auditory system acts as a simple linear channel before the periodicity detector as well as that above 1 kHz it does envelope following and then similar periodicity detection of the envelope. This two-channel analyzer is naturally much more efficient computationally than a multi-channel pitch analyzer.

Several details of the model in Fig. 2 are important for proper functioning. First, prefiltering to simulate the equal loudness curve sensitivity of the human ear may be included. The second box in our model is a frequency-warped version of linear prediction (WLP, order 12 for sample rate of f_s = 22 kHz), as described in [7], in order to pre-whiten the signal. This effect may be considered somewhat similar to the adaptation in hair cell models. The block 'autocorrelation' deviates also from its normal definition. In our model it is computed through the discrete Fourier transform (DFT) and its inverse (IDFT) as $corr(\tau) = IDFT\{|DFT\{x(\tau)\}|^k\}$ where exponent $k = 2/3$ instead of $k = 2$ for normal autocorrelation and τ is the time lag variable. In the frequency domain this operation shows resemblance to the loudness scaling of magnitude. One more detail of implementation is that actually the two (second order) 1 kHz low-pass filters in the middle part of Fig. 2 model include a high-pass property, e.g. with a cutoff at 80 Hz, in order to remove the down-ramp baseline of the correlation function up from zero time lag. These implementation features have been selected based on experimentation with various practical signals.

Another important issue in adapting the model to human auditory behavior is to consider the temporal resolution of pitch analysis. Again, based on experimentation, a Hamming window length of 46.4 ms (frame size 1024 samples for $f_s = 22$ kHz), yielding about 25 ms effective window length, was found practical. For shorter windows the SACF function shows too many spurious peaks and for longer windows the response to time-varying pitches is all too slow. The selected temporal resolution is also not too far from the JND threshold of pitch percept formation with short sine bursts. The hop size of SACF computation window in our model is 10 ms. It is also known that full resolution of pitch analysis in human perception is only achieved after about 100–150 ms of sine wave onset. This means that for steady state sounds we can smooth the SACF function over such a time span.

The two last blocks at the bottom of Fig. 2 are post-processing blocks where the left one takes the SACF function and produces a list of pitch objects and the right-hand side block does, if desired, separation of the signal components based on pitch estimate information. These blocks are discussed in more detail below.

3. COMPARISON OF THE MODELS

The performance and validity of the proposed two-channel SACF model (without pre-filtering and pre-whitening, using running autocorrelation similar to [5]) in pitch periodicity analysis is evaluated by a comparison with the multichannel SACF model of Meddis and Hewitt. AIM software [8] was used to compute the Meddis–Hewitt SACFs. The test signals were chosen according to [5].

In the "missing fundamental" experiment the test signal consisted of three equal-amplitude sinusoids with frequencies 600, 800, and 1000 Hz. In this case a listener would hear a tone with a fundamental frequency of 200 Hz. The SACFs computed using the proposed two-channel and the Meddis–Hewitt model are depicted in the top and the bottom plot of Fig. 3, respectively. The functions have been normalized so that the maximum of each function equals one. Both SACFs peak at lags of 5, 10, and 15 ms. The first peak corresponds to a fundamental frequency of 200 Hz. Both methods are clearly capable of resolving the missing fundamental. It is interesting to note that while the scales of the SACFs differ, the curves are almost identical, in this case.

The results of the "musical chord" experiment with the two-channel and the multichannel models are illustrated in the top and the bottom plots of Fig. 4, 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 four first harmonics, and the C^5, and E^5 tones contained three first harmonics each. All

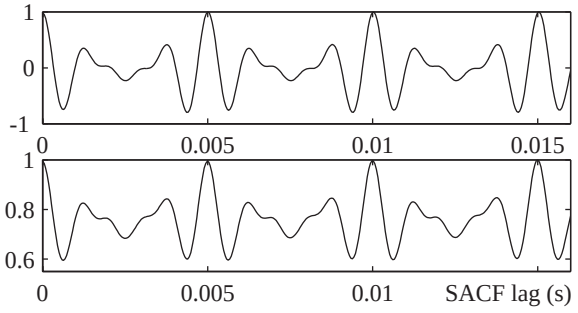

Figure 3: Comparison of the Meddis–Hewitt multichannel and the proposed two-channel SACF functions using the "missing fundamental" test signal. The two-channel SACF is plotted on the top and the Meddis–Hewitt SACF on the bottom.

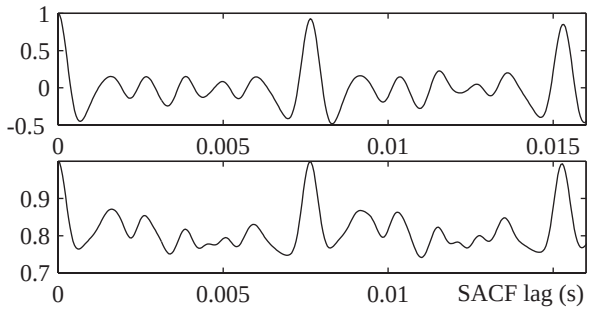

Figure 4: Comparison of the SACF functions of the two models using the "musical chord" test signal. The two-channel SACF is plotted on the top and the Meddis–Hewitt SACF on the bottom.

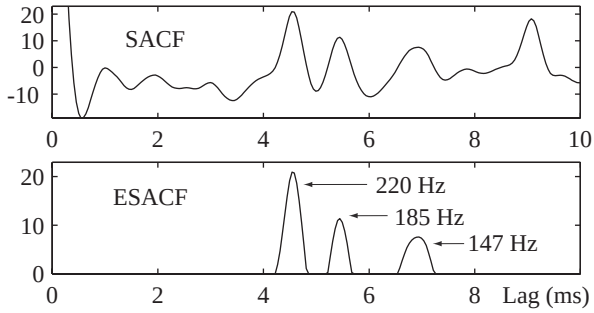

Figure 5: An example of multipitch detection. A test signal with three clarinet tones with fundamental frequencies 147, 185, and 220 Hz, and relative rms values of 0.4236, 0.7844, and 1, respectively, was analyzed. Top: two-channel SACF, bottom: two-channel ESACF after periodicity detection.

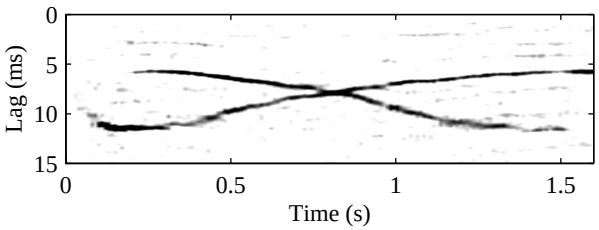

Figure 6: Example of pitch analysis (lag vs. time map) of a signal consisting of two similar vowels with time-varying pitches.

the harmonic components were of equal amplitude. Both models exhibit a SACF peak at a lag of 7.7 ms. This corresponds to a fundamental frequency of 130 Hz (tone C^3) which is the root tone of the chord. As before, the curveforms of the two summary autocorrelation functions are similar although the scales differ.

While it is only possible to report these two experiments in this context, the models behave similarly with a broader range of test signals. More examples of SACF analysis are available at WWW address: http://www.acoustics.hut.fi/~ttolonen/icassp99/pitchdet/.

4. PERIODICITY DETECTION

The peaks in the SACF curve produced as x_3 output of the model in Fig. 2 are relatively good indicators of potential pitch periods in the signal being analyzed, see Figs. 3 and 4. Such a summary periodicity function contains, however, much redundant and spurious information that makes it difficult to estimate which peaks are true pitch peaks. The autocorrelation function generates peaks at all integer multiples of the fundamental period. Furthermore, in case of musical chords the root tone, the common periodicity, often appears very strong though in most cases it should not be considered as a fundamental period of any source sound. To be more selective, a peak pruning technique similar to [9] 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 all repetitive peaks with double the time lag where the basic peak is higher than the duplicate. This 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 multiplies 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. 5. It presents the pitch detection results of one analysis frame that is computed from 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 all the sources are not resolved reliably. One further idea to improve the pitch resolution with complex mixtures, maybe with relatively different amplitudes, is to use an iterative algorithm whereby the most prominent sounds are detected and filtered out (see sound separation) or attenuated properly and the pitch analysis is repeated for the residual.

For further processing the enhanced SACF function can be represented as discrete pitch objects in the form of a pitch list. For each analysis frame the potential pitch periodicities are extracted as objects that code the time lag (fundamental period) of the pitch, the prominence (such as the peak value of the ESACF function), and possibly a measure of confidence of the pitch object if proper criteria are available. Such pitch lists from consecutive analysis frames constitute temporal trajectory information of the multi-pitch behavior of a multiple source signal that can be utilized in further sound separation or auditory scene analysis.

Two other examples of pitch detection are shown in Figs. 6 and 7. Figure 6 illustrates the temporal evolutution of the enhanced SACF function in the analysis of a mixture signal containing two simultaneus vowel pitch glides. Both vowels are Finnish /a/ sounds mixed from the same speaker, one gliding in pitch from low to high and the second one from high to low pitch. The pitch analysis information is shown as a spectrogram-like presentation which clearly indicates the fundamental periodicity trajectories. Some spurious peaks appear at low level but no systematic high level phantom trajectories are found in this relatively easy case where the two vowels have about the same amplitude.

The third example, an analysis of a two-vowel mixture, is illustrated in Fig. 7 and discussed in the context of sound separation.

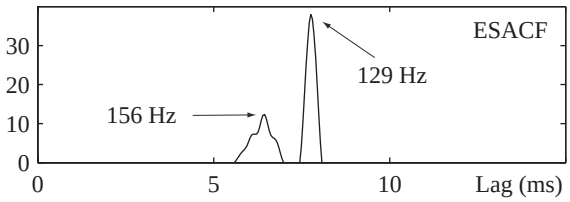

Figure 7: Result of pitch estimation on a signal consisting of vowels /a/ (129 Hz) and /i/ (156 Hz).

5. SOUND SOURCE SEPARATION

The discrete pitch objects may be used to derive separate representations for harmonic sound sources. The pitch-lag locations provide estimates of the fundamental period, and the prominence values may be used to decide in which order the harmonic signals are segregated if an iterative algorithm is used.

In this study, we applied a comb-notch filtering method to separate two Finnish vowels /a/ and /i/ from a mixed voice signal. The length of the analyzed segment was 93 ms (2048 samples at 22 kHz sampling rate). The ESACF illustrated in Fig. 7 provided the pitch objects of the two vowel sounds. The peak locations in the ESACF were used as fundamental period estimates, and the separated sound signals $s_{\mathrm{a}}(n)$ and $s_{\mathrm{i}}(n)$ were obtained by filtering the mixed signal with transfer functions $H_{\mathrm{i}}(z)$ and $H_{\mathrm{a}}(z)$, respectively. The two digital filters were designed to remove the pitch periodicities and they were implemented as $H(z) = (1 - z^{-P_{\mathrm{I}}} H_{\mathrm{f}}(z))^2$, where P_{I} is the integral part of the pitch period in samples and $H_{\mathrm{f}}(z)$ is a fractional delay filter [10] which implements the non-integral part of the pitch period. A second-order Lagrange filter was used for $H_{\mathrm{f}}(z)$.

Linear prediction (LP) spectra (order 24 at 22 kHz) of the separated signals $s_{\mathrm{a}}(n)$ and $s_{\mathrm{i}}(n)$ were computed to illustate the separation results. For comparison, the LP spectra of the original vowel signals were also computed before they were mixed. Fig. 8 shows the results. The original (solid line) and the estimated (dashed line) LPs are depicted on the top for the /a/ sound and on the bottom for the /i/ sound. The example demonstrates that although a simple digital filter was used in the separation, the estimated LPs resemble the originals quite well. The performance may be improved by elaborating the design of separation filters more carefully.

6. DISCUSSION

The experiments above with the proposed model represent realistic yet relatively simple cases where multi-pitch analysis and sound separation was found successful. It is easy, however, to find more complex mixtures of harmonic sounds where the analysis model as such does not resolve subsignals. The problem itself can be arbitrarily difficult since even the human auditory system has severe limitations in separating complex sound mixtures.

One limitation where the present model does not compete with human perception is when the relative amplitude levels of subsignals differ substantially. Iterative removal or attenuation of subsignals, starting from the most prominent one, is one possible strategy to improve the resolution. Another task for further study is to compare the model more carefully with the performance of human hearing in various aspects including temporal resolution of pitch detection. Sound separation and computational auditory scene analysis pose numerous open questions that have so far been studied very little.

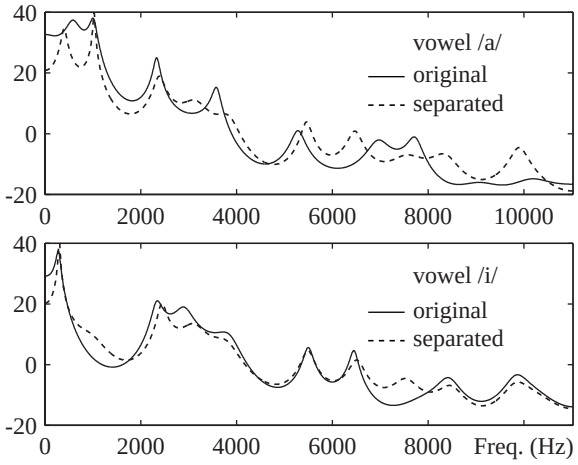

Figure 8: Separation of LP spectra from a mixture of two vowels. Top: original and reconstructed LP spectra of vowel /a/ with fundamental frequency of 129 Hz. Bottom: original and reconstructed LP spectra of vowel /i/ with fundamental frequency of 156 Hz.

7. ACKNOWLEDGMENT

This work has been financially supported by the GETA Graduate School at Helsinki University of Technology.

8. REFERENCES

[1] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *J. Acoust. Soc. Am.*, vol. 100, pp. 3491–3502, Dec. 1996.

[2] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[3] "USA Standard Acoustical Terminology." American National Standards Institute, S1.1-1960, 1960.

[4] R. Meddis and L. O'Mard, "A unitary model for pitch perception," *J. Acoust. Soc. Am.*, vol. 102, pp. 1811–1820, Sept. 1997.

[5] 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.

[6] R. D. Patterson, "The sound of the sinusoid: Spectral models," *J. Acoust. Soc. Am.*, vol. 96, pp. 1409–1418, Sept 1994.

[7] U. K. Laine, M. Karjalainen, and T. Altosaar, "Warped linear prediction (WLP) in speech and audio processing," *Proc. IEEE ICASSP'94*, pp. III.349 –III.352, 1994.

[8] 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. Am.*, vol. 98, pp. 1890–1894, Oct. 1995.

[9] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, Cambridge, Massachusetts, USA, June 1996.

[10] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay—tools for fractional delay filter design," *IEEE Signal Proc. Mag.*, vol. 13, pp. 30–60, Jan 1996.