QUALITATIVE AND QUANTITATIVE ASPECTS IN THE DESIGN OF PERIODICITY ESTIMATION ALGORITHMS

Anssi Klapuri
Signal Processing Laboratory, Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, FINLAND
Tel: +358 3 3652124; fax: +358 3 3653857
e-mail: klap@cs.tut.fi

ABSTRACT
Several approaches have been taken to estimate the fundamental frequency of periodic or pseudoperiodic acoustic signals. However, since the algorithms operate in different transform domains, it has proven difficult to combine the desirable features of different algorithms, and to utilize simulation results across domains. In this paper, we set out to define a common ground on which to investigate the underlying models of different periodicity estimation algorithms. Implications of the analysis are applied in a new algorithm which aims at combining several advantages. Validation experiments were made using a database of musical instrument sounds, both in noise and in harmonic interference.

1 INTRODUCTION
Perception of periodic sounds plays an important role in human hearing and in understanding acoustic complexes. Psychoacoustic studies have attempted to model how pitch, percept of the fundamental frequency \( f_0 \), is formed [1,2,3,5]. Also, several audio applications, such as parametric speech coding systems, have pitch estimators as an integral part of the systems. We use the term periodicity analysis algorithm to refer to the core of these computations, where the periodicity of the signal is analyzed. A problem is that most often the background assumptions of different algorithms have not been indicated, which makes it hard to compare algorithms that operate in different transform domains. This, in turn, prevents from combining their advantages, and from utilizing simulation results and parameters across domains.

In the first part of this paper, we investigate the underlying models of different algorithms. In order to see the fundamental similarities and differences of the algorithms, they are all analyzed in the frequency domain. Comparison is done from two aspects. Qualitative characteristics refer to the type of information that algorithms utilize in pitch calculations. Quantitative properties determine the functions and weighting that relate the atomic pieces of information to pitch likelihoods. Model level causes to practical performance strengths or weaknesses are indicated.

In the second part of the paper, we apply the implications of the analysis to suggest a new pitch algorithm that aims at combining the strengths of several classic models. A collection of musical instrument sounds and singing voice is used to validate the performance of the algorithm. Simulations were run for noisy cases, and for signals that comprise several simultaneous harmonic sounds. Particular robustness was observed in the latter case.

2 COMPARISON OF THE UNDERLYING MODELS IN PITCH ALGORITHMS

2.1 Spectral place models
Autocorrelation function (ACF) based algorithms are among the most frequently used pitch estimators [6,9]. Usually maximum value in ACF is taken as the true pitch period, and if there are several maxima, the shortest period is selected. We denote a discrete time domain signal by \( x(k) \), and its discrete Fourier transform by \( X(k) \). Short-time autocorrelation \( r(n) \) of a \( K \)-length sequence \( x(k) \) is defined here as

\[
r(n) = \frac{1}{K} \sum_{k=0}^{K-n-1} x(k)x(k+n). \tag{1}
\]

This can be calculated in frequency domain as follows. First, \( x(k) \) is zero-padded to twice its length and transformed to frequency domain. Then square of the magnitude spectrum, \( |X(k)|^2 \), is taken and inverse transformed back to time domain, which yields \( r(n) \). Now, for real signals, \( r(n) \) can be expressed in terms of \( X(k) \) as

\[
r(n) = \frac{1}{K} \sum_{k=0}^{K-1} \left| X(k) \right|^2 \cos\left(\frac{2\pi nk}{K}\right) \tag{2}
\]

The calculations are illustrated in Figure 1 for the case when \( n \) corresponds to the true pitch period. Speaking in frequency domain terms, an important property of ACF is to weight spectral components according to their spectral locations. Thus we call it a spectral place type pitch estimator.

It is quite easy to see that \( r(n) \) has a local maximum in the position of the true pitch period \( n_0 \). However, “twice too low pitch” octave errors are likely to occur, since integer multiples of period \( n_0 \) also place positive weights at the harmonics frequencies. “Too high” octave errors are not probable, since in that case odd harmonics get a negative weight.

Cepstrum pitch detection has close model level similarities with ACF [4]. Cepstrum \( c(n) \) of a discrete signal \( x(k) \) is defined as the inverse discrete Fourier transform of the logarithm of the short-time magnitude spectrum:

\[
c(n) = \text{IDFT} \{ \log \text{DFT} \{ x(k) \} \} \tag{3}
\]

The only difference to frequency domain ACF calculations is that logarithm of the magnitude spectrum is used instead of the second power. Again, it can be shown that for real signals, \( c(n) \) can be written in frequency domain terms simply by writing \( \log |X(k)| \) in the place of \( |X(k)|^2 \) in Equation (2). Thus cepstrum is a spectral place type algorithm, too, and causes a tendency towards “too low” octave errors.

Although ACF and cepstrum pitch estimators are qualita-
“Too high” pitch errors are likely to occur, since doubling the position corresponding to the period of the spectrum. But, the model increase to $R_n$ two frequency components with a spectral interval. Thus the partials cannot be assumed to be found at harmonic spectrum positions, but are gradually shifted upwards in the spectrum [10]. This makes Equations (2) to fail in matching the harmonic series at higher frequencies. This is not a big concern in speech processing, but readily met when analyzing musical sounds at a wide band.

2.2 Spectral interval models

Spectrum autocorrelation method and its variants have been successfully used in several pitch estimators [7,8]. The idea is derived from the observation that a periodic but non-sinusoidal signal has a periodic magnitude spectrum, the period of which is the fundamental frequency. Autocorrelation function $R(n)$ over the positive frequencies of a $K$-length magnitude spectrum $|X(k)|$ is calculated as

$$R(n) = \frac{2}{K} \sum_{k=0}^{K/2-n-1} |X(k)||X(k+n)|. \tag{5}$$

The information on which the above equation bases pitch calculations is fundamentally different from that of an ACF or cepstrum pitch detector. Equation (5) does not pick spectral components according to their place, but instead, any two frequency components with a spectral interval $n$ give an increase to $R(n)$. This means that the spectrum can be arbitrarily shifted without affecting the output value. The model is here called a spectral interval type pitch estimator.

It is easy to see that $R(n)$ has a local maximum in the position corresponding to the period of the spectrum. But, contrariwise to the spectral place type algorithms, “twice too high pitch” errors are likely to occur, since doubling the true spectral period will pick every second harmonic of the sound. “Too low” octave errors are not probable since there is no spectral periodicity at half the $f_0$ rate.

A spectral interval type pitch estimator works relatively well for sounds that exhibit inharmonicities. Although the intervals between harmonic partials do not remain constant, they are more stable than the places of the partials, which shift cumulatively, as can be seen from Equation (4).

Quantitative properties of spectral interval type pitch estimators depend on the preprocessing that is done prior to the spectral autocorrelation calculation. Often highpass liftering is used to remove spectral envelope [8]. Basic phenomena are similar to those observed with ACF and cepstrum algorithms, but frequency domain gives some more flexibility.

2.3 Envelope periodicity models

There is at least one more fundamentally different algorithm that has been part of several pitch models, especially of those developed in the field of psychoacoustic research [2,1,3,5]. The idea of the model is derived from the observation that any signal $x(t)$ with more than one frequency component exhibits periodic fluctuations, beating, in its time domain amplitude envelope $\xi(t)$. The rate of the beating depends on the frequency difference of each two frequency components. In the case of a harmonic sound, interval $f_d$ will dominate and the fundamental beat is clearly visible in the amplitude envelope of the signal, as illustrated in Figure 2. We call these envelope periodicity (EP) models [5].

Time domain beating can be utilized in pitch estimation as follows. First, a discrete time domain signal is half-wave rectified and lowpass filtered to get $\xi(t)$. After this, autocorrelation or some other method is used to find the most prominent periodicity in $\xi(t)$. These operations are often done for several different bandlimited versions of the signal [2].

Because of the non-linear rectification operation, it is difficult to represent the algorithm in the frequency domain. However, certain properties of the underlying model can be revealed by analyzing how the algorithm processes an ideal harmonic sound, which consists of $H$ harmonic partials. Each harmonic has three parameters: frequency $f_i$, amplitude $a_i$, and phase $\theta_i$. Generally, it can be shown that any two partials $i$ and $j$ where $f_i > f_j$ cause amplitude enve-
lope beating with frequency \( f_b = f_i - f_j \), amplitude \( a_b = (1/\pi) \cdot \min(a_i, a_j) \), and phase \( \theta_b = \theta_i - \theta_j \). Thus each harmonic \( h \) and its companion \( h+i \) generate beating at rate \( i \cdot f_0 \). Let us use \( \Xi(k) \) to denote the spectrum of amplitude envelope \( \xi(k) \). The cumulative beating at rate \( i : f_0 \) can be roughly approximated as
\[
\Xi(ik_0) = \frac{1}{\pi} \cdot \sum_{h=1}^{H-i} \left[ \min(a_h, a_{h+i}) \cdot e^{i(\theta_h - \theta_{h+i})} \right]
\]
where \( k_0 \) is the sample spectrum corresponding to \( f_0 \), \( i=1,2,...,H-1 \), and \( \Xi(k)=0 \) for other frequencies.
If autocorrelation is chosen to estimate the fundamental period of amplitude envelope, \( \text{ACF}[\Xi(k)] \) can be written as
\[
p(n) = \frac{1}{K-n} \cdot \sum_{k=0}^{K-1} \left| \Xi(k) \right|^2 \cdot \cos\left(\frac{2\pi nk}{K}\right).
\]
However, since any other method could be used as well to estimate the periodicity of \( \xi(k) \), we are more interested in the properties of \( \xi(k) \) itself, and its spectrum \( \Xi(k) \).

Envelope periodicity models have three special qualitative properties, all of which can be seen from Equation (6). First, they are spectral interval oriented, not spectral place oriented. Moreover, if ACF is used in the last phase, also integer multiples of each spectral interval are summed in Equation (7). Second, EP models make implicit spectral smoothing, since \( \min(a_h, a_{h+i}) \) operation in Equation (6) filters out single clearly higher amplitude harmonic partials. We can only mention here that smoothing the spectrum of harmonically related partials before using them to pitch calculations was found to make a significant improvement in a multipitch estimation system that has been developed in our laboratory. Such smoothing has also been observed in the human pitch perception. A third property of EP models is that they are phase sensitive in summing up the harmonic partials. This may cause some advantage in multipitch estimation, since frequency components arising from a same physical source are often phase locked to each other. However, this speculation was not validated by experiments.

3 PROPOSED ALGORITHM

Analysis presented in this paper has implications that resulted in a new algorithm for the periodicity analysis of acoustic signals. An overview of the system is to calculate independent pitch estimates at separate frequency bands, and then combine the results to yield a global estimate. This solves several problems, one of which is inharmonicity. According to Equation (4), the higher harmonics may deviate from their expected spectral positions, and even the intervals between them are not constant. However, we can assume the spectral intervals to be piece-wise constant at narrow enough bands. Thus we utilize spectral intervals to calculate pitch likelihoods at separate frequency bands, and then combine the results in a manner that takes the inharmonicity into account. Another advantage of bandwise processing is that it provides robustness in the case of badly corrupted signals, where only a fragment of the whole frequency range is good enough to be used.

3.1 Quantitative decisions

For the sake of computational efficiency, bandwise processing is done in the frequency domain. A single fast Fourier transform is needed, after which local regions of the spectrum are separately processed. An obvious goal in quantitative sense is to combine the advantages of the algorithms that are either robust to noise or to the anomalies in the spectra of the analyzed sounds.

Applying logarithm to the magnitude spectrum causes a drastic dynamic compression, which equalizes the spectra of exotic sounds. After this, noise is removed by subtracting a moving average over a Bark-scale spectrum. Around each Bark’s critical band, we calculate the average of the logarithmic magnitude spectrum values in a window that is four times wider than the critical band, i.e., \( 4 \cdot \text{CB}_\text{Bark} \), where
\[
\text{CB}_\text{Bark} = 25 + 75 \cdot \left[ 1 + 1.4 \cdot \left( \frac{f_c}{1000} \right)^{-0.69} \right].
\]

Hamming window is applied to the values before averaging. An overall smooth average spectrum \( X_{av}(k) \) is then calculated by interpolating between the points at each Bark band. An important feature of \( X_{av}(k) \) is to pick more details at low frequencies, but to get smoother at high frequencies. Finally, we get an equalized spectrum, subject to pitch calculations
\[
X_e(k) = \log[X(k)] - X_{av}(k).
\]

The subtraction operation has certain analogy with inverse LPC filtering, which is clear by noticing that subtraction in a logarithmic scale is equivalent to multiplicative weighting in a linear scale. Also, highpass liftering in the system of Kunieda et al. is quite similar [8]. A fundamental difference in our method is to use a varying bandwidth in spectral flattening. Usage of a bank of bandpass lifters in the system of Lahat et al. may implicitly result in the same [7].

3.2 Qualitative decisions

The equalized spectrum \( X_e(k) \) is processed in 18 logarithmically distributed bands that extend from 50 Hz to 6000 Hz. Each band comprises a 2/3-octave wide region of the spectrum that is subject to weighting with a triangular window. Overlap between adjacent bands is 50 %, which makes them sum to unity when the windowing is taken into account.

The mechanism by which pitch likelihoods are calculated at each band is a combination of place and interval models. Samples at a band are denoted as \( X_k(k), k \in [b_1, b_2] \), where \( b_0=b_2-1 \) is the number of samples at the band. We use \( k_0 \) to refer to the spectrum value corresponding to a fundamental frequency \( f_0 \). The likelihood \( L_g(k_0) \) of \( f_0 \) candidates that are determined to have exactly one harmonic partial \( h \geq 1 \) at the band \( B \) is got by copying the value of \( X_e(k) \) that corresponds to that partial. Mathematically we can write this as
\[
L_g(k_0) = X_e(hk_0) \quad \forall k_0, \quad [3h \ (hk_0 \geq b_1, hk_0 \leq b_2)].
\]

Obviously, the place of a partial \( h \) is used to associate it to \( k_0 \). This is not a problem from inharmonicity point of view, since Equation (10) applies only to \( k_0 \) that have at maximum one partial at the band, i.e., \( k_0 > b_n \). In these cases \( h_0 < 4 \), and inharmonicity does not affect the lowest harmonics.
An interval model is used for fundamental frequency candidates \( k_0 < b_0 \), that may have more than one partial at the band. This is necessary since the series of higher harmonics may have shifted due to inharmonicity. \( L(k_0) \) is calculated as

\[
L_B(k_0) = \max_{m \in M} \left\{ \frac{W(N) \cdot \sum_{n=0}^{N} X_e(b_0 + nk_0 + m)}{N} \right\},
\]

where \( M=\{0,1,\ldots,k_0-1\} \) is the shift, and the number of harmonics in the sum \( N = \lfloor (b_0 - m - 1)/k_0 + 1 \rfloor \). This finds such a series of every \( k_0 \)th spectrum samples that maximizes the above quantity. \( W(N) = 0.7/N + 0.3 \) is used as a normalization factor, because \( N \) varies for different \( k_0 \) and \( m \).

A straightforward way to combine pitch likelihoods \( L_B(k_0) \) from different bands would be to sum over the vectors as \( L(k_0) = \sum L_B(k_0) \). However, this does not take the inharmonicity into account. It was observed in simulations that e.g. for string instruments, pitch increases clearly as a function of the band center frequency. Thus we wrote an algorithm which starts from \( L_B(k_0) \) at the band where the fundamental is located, and matches local maxima at higher bands in a way that allows a slight increase in pitch. The bandwise likelihoods were squared before summing to provide robustness in interference, where pitch is observable only at a limited band. In signals containing a mixture of several harmonic sounds, a clear improvement resulted from explicit smoothing of the amplitudes of a detected harmonic series before the final pitch decision (see Sec. 2.3).

### 3.3 Discussion

The presented algorithm bases pitch calculations only on the harmonically related frequency partials of each \( k_0 \), not on the entire spectrum. This is crucial when analyzing mixtures of several harmonic sounds. The presented algorithm is a core part of an automatic music transcription system, where usage of the overall spectrum turned out to be problematic, because the pitch models get confused by the partials of the other sounds, and the interrelations of sounds cause unpredictable effects. Picking only the appropriate partials provides robustness in sound mixtures as far as we do not rely on the detection of single partials, as is the case. Secondly, the algorithm is a combination of place and interval models. This provides some immunity against octave errors, since place models are more robust to "too high" octave errors, while interval models are robust to "too low" octave errors.

### 4 VALIDATION EXPERIMENTS AND DISCUSSION

A large amount of simulations was run to monitor the behaviour of the proposed algorithm. Here we summarize some main observations. Test material consisted of a database of sung vowels plus 26 different musical instruments comprising plucked and bowed string instruments, flutes, and brass and reed instruments. These introduce several different sound production mechanisms, and variety of spectra.

In each simulation, about half a thousand random sounds were selected by first allotting an instrument, and then a random note from its whole playing range, however, restricting the pitch over five octaves between 65 Hz and 2100 Hz. Samples were fed to the pitch algorithm that estimated pitch in a single time frame. A correct pitch was defined to deviate less than half a semitone (±3%), from the correct pitch.

In the first experiment, we tried to find the shortest time frame in which the algorithm still works without problems. Good results were achieved with down to lengths of 23ms, but 46ms was required for the lowest notes. Error rates were 4%, 2%, and 0.2% for time frames 46ms, 93ms, and 186ms.

In the second experiment, different levels of pink noise were added in the range 50Hz and 10kHz. The resulted error rates for signal-to-noise ratios (SNR) 10dB, 5dB, and 0dB were 4%, 10%, and 17%. In the third experiment, an octave-wide SNR-0dB noise band was added at a varying random position between 100Hz and 6000Hz. Resulting 4% error rate shows that even high levels of band-limited noise make hardly noticeable decrease to results, due to the bandwise processing approach. In these two noise experiments, 46ms frame was used and ±5% pitch deviations were allowed.

In the last experiment, we generated random mixtures of harmonic sounds, each with same RMS-level. Pitch estimate was defined to be correct if it was the correct pitch of one of the component sounds. A long 186ms window was used to cope with mixtures of low-pitched sounds. Error rate was 0.2% for monophonic signals, and remained 0.5% for all combinations from two to six sounds, showing significant robustness for polyphonic signals, as explained in Sec. 3.3.

### REFERENCES


