Continuity preserving signal processing
Andringa, Tjeerd

Publication date:
2002

Citation for published version (APA):
The estimation of pitch-contours in noise is difficult. The main reason for this is the general inability to detect and estimate features reliably (conclusion 1.16). Even for clean situations it not trivial to implement a reliable pitch estimation algorithm, this situation exacerbates in noise when the signal-in-noise-paradox (conclusion 1.12) makes it difficult to decide on the focus of the search effort. In fact, the difficulties encountered during the formulation of a reasonably functioning pitch-extraction algorithm for noisy situations formed the basis for the formulation of the framework developed in chapter 1.

This chapter focuses on two methods to estimate pitch. The first method has been introduced in section 2.9 and aims to estimate the best period contour at any time in unknown noisy situations. The second method is a ridge-based variant of well-known correlogram-based approaches (Meddis 1997) that are based on correlogram summation. Both implementations are a proof-of-concept. They are not optimized and do not use the full potential of the techniques presented so far. In particular the CPC (as discussed in section 4.3) is not utilized. Yet both form a useful starting point for future developments.
5.1 Fundamental Period Contour Estimation in Noise

Section 2.9 outlined an algorithm for pitch estimation in noise. This section provides additional information and focuses on the design decisions that formed the basis for its formulation. Most pitch estimation techniques are developed for clean conditions or conditions with considerable constraints on the noise and are typically based on cues (O'Shaughnessy 2000) like:

- periodicity of the short term FFT-based energy spectrum, a cepstrum (Gold 2000), a harmonic sieve (Duifhuis 1982) or subharmonic summation (Parson 1976, Hermes 1988).
- the first large peak in the autocorrelation
- correlogram summation (Meddis 1997), similar to the technique in section 5.2.
- zero crossings

These techniques assume a single source, assume known noise or apply quasi-stationarity before the signal is split into individual components for which quasi-stationarity can be guaranteed (see conclusion 1.21). Consequently, these techniques will produce incorrect results whenever the quasi-stationarity assumption is violated. The set of situations in which the pitch estimation technique is guaranteed to function (or fail) is extremely difficult to quantify since it can depend on the pitch value that still has to be estimated. Consequently, these pitch estimation algorithms are not as robust as is required for a system that can deal with arbitrary signals (definition 1.3).

The basic design choices that led to the formulation of the robust period contour estimation algorithm are based on an informal study of a large number of very noisy (SNR < +3 dB) speech signals. The speaker characteristics, as well as the characteristics of the noise were varied and the resulting cochleograms, ridges and instantaneous frequency contours were inspected to determine which features where most robust and could be used to base a pitch estimation algorithm on. Good indicators for reliable sources of information were:

1. the strongest ridges at each moment
2. long ridges
3. smooth ridges

1. Note that these features result from (strong) local dominance.
4. ridges with frequencies that correspond to the local characteristic frequency

Ridges in which several of these features are combined are particularly reliable. Among the less reliable indicators are direct extensions of typical techniques that work well in noiseless environments. For example, a standard harmonic sieve applied to the local instantaneous frequency estimates (LIF, see section 4.5) along ridges results in successions of very good local estimates alternated with very poor or wrong estimates. Since the quality of the estimates could not be determined automatically, this method did not lead to reliable results.

To facilitate implementation, a multi-stage algorithm that operates on the complete utterance was chosen. Figure 5.1 provides an overview of the fundamental period contour estimation algorithm as introduced in section 2.9 on page 65. The input for the algorithm is the information as represented in figure 2.10 and encompasses the cochleogram, the ridges and the local instantaneous period. These values are sampled every 5 ms. The first stage of the algorithm is the selection and smoothing of the strongest ridges. This algorithm starts with the detection of instantaneous periods whose corresponding characteristic segment (see section 3.3) are more than one segment from the ridge’s characteristic segment. These period values are replaced by the segments characteristic period. Each ridge is followed and, as long as successive periods are within 5% of each other, the local periods are assigned to the same period contour. When successive periods are not within 5%, an additional check is performed to check if the next value is within 5%. If a valid next value can be found, the gap is filled with the average of its neighbors, otherwise a new contour is started.

All contours are augmented with a smoothed version of the contour. Smoothing is performed with a 5-point (25 ms) moving average. In the middle of the contour the smoothed local period is based on a local neighborhood of 2 frames on each side. In the two first or last points of a contour the smoothed period values are based on the corresponding values of a first order approximation. Finally, the average ordinality of each contour is computed. The ordinality is a measure of the relative importance in terms of energy. A segment of the strongest ridge at time \( t \) has ordinality 1, the second most energetic segment has ordinality 2, etc. A period contour is accepted

\[2\] This is a task that can be improved upon with the CPC as a measure of reliability.
whenever its length exceeds 50 ms and its average ordinality is smaller or equal than 2, or alternatively whenever its length exceeds 75 ms. So far, these practical values have not been optimized formally.

The second step assumes that all possible fundamental periods lie between 13.3 ms (75 Hz) and 2.5 ms (400 Hz), a range that spans most speakers (Furui 1989). It is assumed that each contour represents a single harmonic number from start to end. The range of possible harmonic numbers can be estimated using the limits of the pitch range and equation 2.11. As visualized in the upper panel of figure 2.12, the smoothed period contours are multiplied by each possible harmonic number and copied to all possible fundamental periods. This involves a change in the corresponding characteristic segments.
of the contours, and since each segment has its own group delay this implies a temporal shift according to:

\[ t \rightarrow t + d(s_{np}) - d(s_p) \quad n \in \{1, 2, \ldots\} \]  

(5.1)

\(d(s_p)\) and \(d(s_{np})\) are the group delays associated with the segments that are most sensitive to period \(p\) and period \(np\), respectively. Note that this time-shift implicitly defines the instantaneous fundamental period as the instantaneous period of the first harmonic as expressed on the BM.

The third step combines the copied (cloned) contours into smooth fundamental period contour hypotheses. This is a complicated process since contours can sometimes be combined in different ways. When the local periods of two cloned contours match, on average, within 3% the clones are combined into a single hypothesis. Clones that extend each other are combined when a second order fit can be estimated that matches both contours within 3% during 25 ms. Note that this criterion is more strict than the 5% that was used for contour estimation in the first step. When this criterion is loosened the algorithm may produce unreliable results due to the combination of contours of different sources. The time-shift of equation 5.1 is very important because it allows a reliable comparison between multiple contours. When this form of group delay compensation is absent, contours of the same source will not be combined during rapid changes of pitch.

Next, concurrent fundamental period contour hypotheses are compared. Concurrent hypotheses that assimilate more contours and lead to the longest hypotheses are favored over short hypotheses that consist of fewer contours. Based on this criterion some of the least supported hypotheses are discarded. This decision is applied with care so that only the least supported hypotheses are discarded. Hypotheses based on less than three concurrent ridges are always discarded. The remaining set of period contour hypotheses is likely to contain the correct period contours.

The fourth and last step is the forced choice between concurrent contour hypotheses. For ASR systems this is a very important stage since this choice decides which part of the signal will be interpreted according to the expectations and limitations of the recognition system. Any error at this stage leads to recognition errors. This warrants a very careful decision process that is based on all available information: i.e., all ridges and their corresponding instantaneous periods. The decision process chooses at most one period contour for each moment. The selected hypothesis maximizes the number of instantaneous period values that it can claim as a possible harmonic, in
Fundamental Period Contour Estimation

combination with a comparable number of odd and even harmonics. The number of claimed harmonics per fundamental period contour hypothesis \( p(t) \) is determined by counting the number of instantaneous period values that satisfy:

\[
\cos \left( \frac{p(t + d_s)}{p_{s,t}} \right) > 0.95
\]

(5.2)

\( p_{s,t} \) is the instantaneous period value derived from a ridge at time \( t \) in segment \( s \). And \( p(t + d_s) \) is the fundamental period hypothesis that is group delay corrected with a value \( d_s \) to denote the expected instantaneous fundamental period of segment \( s \). This is again a situation in which group delay correction is necessary, because instantaneous frequency information of different regions of the basilar membrane is compared. The cosine based criterion of equation 5.2 is equivalent to accepting a deviation of 5.1% around the expected value.

A variant of equation 5.2 can be used to count the number of odd and even harmonics that are within 5.1% of the expected value:

\[
N_{p(t)} = N^o_{p(t)} + N^e_{p(t)}
\]

\[
= \sum_i \left[ \cos \left( \frac{p(t + d_i)}{p_i} \pi \right) < -0.95 \right] + \sum_i \left[ \cos \left( \frac{p(t + d_i)}{p_i} \pi \right) > 0.95 \right]
\]

(5.3)

The square brackets denote a Boolean value: 1 if the statement is true, 0 if the statement is false. The index \( i \) refers to the period values \( p_{s,t} \) in the selected set of ridges, while \( p(t + d_i) \) is the required group delay corrected value for the local instantaneous fundamental period reflected at time \( t \) in segment \( s \). \( N_{p(t)} \) is the total number of accepted harmonics, \( N^o_{p(t)} \) and \( N^e_{p(t)} \) are the number of odd and even harmonics. The odd harmonics fall around the minimal values, while the even harmonics coincide with the maximal values of the cosine function.

The best hypothesis of two or more concurrent hypotheses is the one that maximizes:

\[
\text{Average # harmonics per frame} \cdot \text{Fraction odd harmonics}
\]

(5.4)

Both criteria are important. The average number of claimed harmonics is a measure of the quality of the hypothesis: efficient hypotheses that claim a large number of harmonics per frame are usually to be preferred over hypotheses that claim a lower number of harmonics per frame. If the fraction of odd harmonics is low, it is likely that the fundamental period contour is an octave too low. This happens quite often, because spurious contributions tend
to increase the average number of claimed harmonics. The combined criterion reduces to the average number of odd harmonics per frame. This simple criterion has a high probability to select the correct hypothesis. Similar criteria can be found in literature (e.g., Duifhuis 1982).

The selected information allows a reestimate and smoothing of the fundamental period contour based on all claimed harmonics. The local moving average for each frame is based on all data points within 12.5 ms of the center of the frame. Because equation 5.1 computes the fundamental period contour as the instantaneous period contour of the first harmonic as expressed on the BM it must be time-shifted to reflect the instantaneous period of the source. The final output of the algorithm is a sequence of parameters, defining the local instantaneous fundamental period and its temporal derivative. These are used to approximate the local fundamental period value $T_s(t)$ (equation 4.15) according to:

$$\begin{align*}
T_s(t) &= T(t + d) = T(t) + d \frac{dT(t)}{dt} \\
&= T_s(t) + d \frac{dT_s(t)}{dt} \\
&\approx T_s(t) + d \frac{dT_s(t)}{dt}
\end{align*}$$

The actual period may fluctuate around the estimated values. The local optimization technique described in section 4.6 decides on the final and optimal value.

The current implementation is, as mentioned before, a first proof-of-concept, it is not fully optimized for quality of estimation, stability or speed. Informal experiments suggest that the algorithm can correctly estimate 95% or more for most noisy situations with an SNR higher than 0 dB. In these cases it allows a very good TAC-estimation. Between 0 and -3 dB the probability of a correct estimation reduces to 70%; below -3 dB the algorithm breaks down rapidly. The algorithm fails when its basic assumptions are invalid, i.e., when the pitch of the target is out of range or when it is presented with a concurrent mixture of period contours. The algorithm seems to works slightly better with voices within the female pitch range, than in the typical male pitch range. This difference results from the higher harmonic density of male speech and can probably be reduced by minor adjustments.

Although the current implementation requires the whole signal, it is possible to reimplement the system in a way that estimates period contours with a delay of 100 ms or less. The lower limit of this delay is determined by a combination of group delay effects, the temporal scope required for the computation of local frequencies and, most importantly, the number of period hypotheses the system is allowed to produce. With a delay of 50 ms the system...
Fundamental Period Contour Estimation

has less information available to reduce the number of likely fundamental period candidates than when it is allowed to combine evidence over 100 ms. Optimally, the delay ought to depend on the signal itself: very redundant signals require a small delay, while less redundant signals require more and longer processing.

5.2 Fundamental Period Contour Estimation for Clean Speech

A fundamental period estimation algorithm that can be applied to clean signals has been developed for the speech recognition test. This algorithm ought to be stable, efficient and reliable, because every major error in the period estimation algorithm leads to recognition errors. The general algorithm of the previous section leads to very good estimates in clean circumstances, but it is a first version of a complex algorithm; its stability cannot be guaranteed for use on large database and it is not optimized for speed. Moreover, a simpler algorithm suffices for clean situations. The algorithm presented in this section is intended to become a reliable and relatively fast alternative for the more general period estimation technique for clean signals.

The demands for a fundamental period estimation algorithm to test the robustness of a speech recognition system are slightly different from a system that aims to select and track as much of the source as possible. The latter was optimized in the general fundamental period estimator. For an ASR-test it is necessary to produce a signal representation that resembles the stored templates as well as possible. This entails that spurious contributions should contaminate the selection as little as possible. During onset, but more often during offset, the signal energy might be relatively low and the probability of spurious contributions relatively high, while little linguistic information is conveyed. For example, the information after \( t=360 \) ms in the word /NUL/ in figure 2.3 on page 49 is of little consequence, while a rising pitch can be estimated for at least another 100 ms. During these last 100 ms, the signal-to-noise ratio decreases rapidly which results in a more contaminated TAC-selection. To reduce this contamination it is beneficial to be conservative when determining whether or not the start or end of a signal is voiced. This is implemented by restricting both absolute energy and the decay behavior of the ridges in the low-frequency half of the basilar membrane model. This part of the basilar membrane is hardly affected by unvoiced signal components.
When the energy loss corresponds to 50% or more in 10 ms, or when the energy does not exceed 1% of the average maximal energy of the utterance, the frames are considered unvoiced. This combined criterion works only for well calibrated databases and rather short utterances. It must be replaced by more sophisticated criteria when the data behaves more like spontaneous speech (Furui 1989). Speech recognition results suggest that less than 0.5% of the voiced sound events were incorrectly assigned as unvoiced.

The decay-criterion is a bit more restrictive than the decay of the leaky integration process in the absence of input. The decay in 10 ms associated with a leaky-integration time constant of 10 ms is $e^{-10/10} = e^{-1} = 0.37$ while the applied threshold is 0.5. For speech signals this threshold is very efficient. Coherent voiced speech is rarely incorrectly broken up in subparts due to the decay criterion. The combination of the thresholds for absolute energy and decay leads to fundamental period contours that tend to have earlier offsets and later onsets.

The fundamental period algorithm is based on a summation of the autocorrelation along ridges. As mentioned in the context of figure 2.8, the autocorrelations along ridges that stem from the same source agree on the fundamental period as the first common periodicity that all ridges share. Figure 5.2 shows a typical example of a set of autocorrelations and the corresponding summation. Note that all autocorrelations are simply added and no group delay correction has been performed. Consequently, the result is an approximation. The optimization in the selection algorithm (see section 4.6) determines the final instantaneous fundamental period.

In each frame the three highest peaks in the summed autocorrelation with values higher than 0.3 times the local energy along the ridge are selected and sorted according to the autocorrelation value (the highest peak first). When no peak satisfies the criterion, the frame is considered as unvoiced. It is assumed that one of these autocorrelation lags corresponds to the desired fundamental period value for this frame. The candidate autocorrelation lags are depicted in the upper panel of figure 5.3. A blue dot denotes the highest local

3. In the odd case that a period contour is broken up in subparts, the leaky-integration will smoothen small gaps in the selection.

4. This algorithm is similar to correlogram based algorithms that claim to model aspects of human pitch perception (e.g., Meddis 1997). The main difference is the use of the running autocorrelations under ridges, instead of computing and summing the complete correlogram.
autocorrelation value, a green dot the second value and the red dot the lowest selected value. The selected peaks are combined into temporal contours. Contours shorter than 25 ms are discarded. In each frame the remaining contours are compared with the corresponding characteristic frequency of the segment of the lowest ridge. This is depicted in the lower panel. The contour that falls 60% or more of the time within 10% of the characteristic frequency of the lowest ridge is chosen, the other contours are discarded. Finally, the selected period contours are smoothed with the same procedure as described in section 5.1. The final output of the algorithm is, conform the demands of the TAC-selection algorithm in section 4.6, a parameter set defining a first order approximation of the local instantaneous fundamental period per frame.

This technique combines two strategies that complement each other: periodicity information in the running autocorrelation provides an accurate

5. This implementation does not work properly with a missing fundamental.
Fundamental Period Contour Estimation for Clean Speech

To determine a lower bound of the quality of the period contour estimation a standard MFCC-based HMM system was trained on the voiced frames, as estimated by the algorithm, of the female voices of the TIDIGIT database. A recognition report based on a test with the training data is shown in figure 5.4. Less than 0.5% of the voiced sound events are missed or estimated with a major error that prevents correct recognition. Two thirds of the deletions occur in words with short vowels like /SIX/ and /EIGHT/. The overall performance of the speech recognition system shows that it is possible to reach 1% error-

6. Note that the two sources of knowledge correspond to time and place coding in the auditory nerve.
5.3 A Comparison of Fundamental Period Estimation Algorithms

This thesis presents several ways to estimate and reestimate fundamental period contours. Section 5.1 presented a general algorithm that is optimized to function in very noisy situations (SNR = 0). Section 5.2 presented an algorithm tuned to the demands of an ASR system that assumes a clean speech signal. Both algorithms provide a smoothed version of the actual period.
contour that is based on a selection of the available data. The TAC-selection can, in principle, provide a reestimation of the instantaneous local fundamental period estimate because it combines all available periodic information.

The quality of the reestimation is limited by the sample frequency and the integration time constant. When the time constant is relatively long (which is the case for most fundamental periods), temporal averaging reduces the temporal resolution. Frequency resolution is limited by the sample frequency. With a sample frequency of 20 kHz the difference of one sample corresponds to 0.5% (0.5 Hz) around periods corresponding to 100 Hz, while around 400 Hz each sample leads to a difference of 2% (8 Hz).

The natural fluctuations of the period-to-period variation in duration are called jitter and are typically between 0.5% and 1% (O'Shaughnessy). The combination of an integration time constant of 10 ms, and the sample frequency leads to a sensitivity that is insufficient to study the jitter in detail, but the sensitivity is not low enough to average out all effects of jitter. The current system settings are therefore unsuitable for a detailed comparison of the pitch estimation techniques. Yet it is useful to compare the results and consequences of the different algorithms.

Figure 5.5 shows some examples, in terms of pitch, of the two fundamental period contour algorithms of the previous sections. The solid lines give the output of the period estimation algorithms. The green line denotes the output of the 'general' algorithm, the black line that of the algorithm for clean speech that was optimized for ASR-purposes. In general, the agreement between both is very good. The main difference between the two estimates occurs at the end of the third contour, where the general algorithm interpreted some noise as a valid extension of the contour. This type of error is, due to the complexity of the general algorithm, extremely difficult to reduce, because the reduction of the number of one type of error leads usually to an increase in the number of other errors. In part, this is a consequence of conclusion 1.16 which formulated a limitation of the measurement process. It is, however, possible to evaluate multiple hypotheses to find the best interpretation of the local periodicity information.

The plus-signs and open circles in Figure 5.5 denote the frequencies that result from the reestimation process during TAC-selection. The plus-signs refer to a clean signal and the open circles refer to the signal with added babble noise at 0 dB SNR. These values fluctuate around the original estimates that were estimated from clean signals: the green contour for the red values and the
black contour for the blue values. In clean situations the fluctuations are minimal and usually due to natural pitch fluctuations and the discrete periodicity values. In noise, the fluctuations are only marginally larger. This increase is primarily the result of the opportunistic local optimization of the positive area under the compressed selection (section 4.6). At each time-step it is possible that spurious contributions are selected. The TAC selects the information that correlates with the fundamental period contour more efficiently than the uncorrelated contributions. This entails that spurious contributions usually have a small and local effect that shows itself by a small increase in the noise of the reestimated fundamental period value.

The close-up in the main figure shows that both fundamental period estimation techniques failed to estimate a small increase in frequency between 1380 and 1420 ms that shows up very clearly in both the clean reestimates and the noisy reestimates. This implies that the information under the ridges does not
not always provide sufficient and completely correct information. In this case the local frequency estimation is completely dominated by the first two harmonics: the higher harmonics do not give rise to consistent ridges. The reestimation based on the selection uses all available information and consequently allows an improved pitch estimation.

The other interesting effect in the third contour is the estimation error at the end of the red, ‘general’ contour. Here the signal had ended, but the ‘general’ period contour estimation was still able to find some supportive evidence for continuation. The corresponding reestimated period values in clean condition (marked with ‘+') vary considerably. In noise the reestimated values vary less because some spurious contribution provided a consistent target.

This period estimation error has only a minor effect on the final TAC-selection as depicted in figure 5.6. In clean conditions (the upper panels) hardly any difference is visible: the estimation error occurred when signal energy was
low and decreasing. In noisy conditions the effect of the estimation error is still small. The lower left hand panel shows the additional signal contribution after $t=1450$ ms, while the lower right hand panel only shows the effect of the leaky integration. This example shows that pitch estimation errors do not necessarily lead to large estimation errors.