POLYPHONIC PITCH DETECTION BY MATCHING SPECTRAL AND AUTOCORRELATION PEAKS

Sebastian Kraft, Udo Zölzer

Department of Signal Processing and Communications
Helmut-Schmidt-University, Hamburg, Germany
sebastian.kraft@hsu-hh.de

ABSTRACT

This paper describes a polyphonic multi-pitch detector which selects peaks as pitch candidates in both the spectrum and a multi-channel generalised autocorrelation. A final pitch is detected if a peak in the spectrum has a corresponding peak within the same semitone range in at least one of the autocorrelation channels. The autocorrelation is calculated in octave bands and all pre-processing steps like filtering, whitening and non-linear distortion are applied exclusively in the frequency domain for maximum flexibility in the parametrisation and high computational efficiency. An evaluation with common data sets yields good detection accuracies comparable to state of the art algorithms.

***Index Terms*—** polyphonic pitch detection, music information retrieval, autocorrelation, spectral processing

1. INTRODUCTION

The autocorrelation and its variants like the cepstrum are standard features in the area of monophonic pitch detection but are rarely used for the analysis of polyphonic music (e.g. [1]). Recent algorithms that reached good accuracy scores of up to about $70\,\%$ in the *MIREX Multiple F0 estimation task*[1] of the last few years are nearly exclusively based on short time Fourier transform (STFT) representations of the signal content. This mid-level representation is then for example further processed by spectrogram factorization [2] or spectral peak and partial selection [3,4] to extract the fundamental frequencies. A complete overview of the history and latest developments in this research field can be found in [5].

Most musical instruments produce harmonic tones consisting of a fundamental frequency ($F0$) and several associated overtone partials. This harmonicity causes a regular pattern in the spectrum which is the main cue being analysed by all the above mentioned spectral algorithms. However, a pitch is not only harmonic but also periodic and periodicity can be observed as regular repetitions at integer multiples of a base lag in the autocorrelation function (ACF). Therefore, the idea of the presented algorithm is to combine cues from both sources for a stable and accurate detection of pitches.

The standard ACF is not well suited for the analysis of polyphonic music and several pre-processing steps like whitening, non-linear distortion and octave-band filtering similar to [1, 6] have to be applied. In the resulting multi-channel generalised autocorrelation function (MCACF) all peaks are selected as pitch candidates together with all the peaks from the spectrum. Usually, for a set of spectral peaks it is not clear which one is caused by a fundamental frequency or a harmonic. Vice versa, in the MCACF the ambiguity is in the decision between the fundamental and its sub-harmonics. Thus, the potential errors in both domains are opposed and a simple criterion to filter the candidates can be derived. To be finally detected, a candidate from the spectrum needs to have a corresponding candidate in the same semitone range in at least one of the MCACF channels.

Although this procedure appears to be comparatively simple, it is capable to remove a lot of candidates which would otherwise be false positive detections. Together with a careful parametrisation of all processing stages the proposed pitch detector achieves good accuracy values in an evaluation with common polyphonic data sets.

2. ALGORITHM

The time domain input signal $x(n)$ is split into overlapping blocks of length $N_W = 4096$ with a hop size $N_H = {}^{N_W}\!/_4$ between consecutive blocks. Each block is weighted with a Hann-window $w(n)$, zero-padded to a length $N_{\text{DFT}} = 16384$ and transformed into the frequency domain to yield the magnitude spectrum in a time-frequency representation

$$X(k, b) = \left| \text{DFT} \left\{ x(n + b\, N_H) \cdot \frac{w(n)}{N_W} \right\} \right|, \qquad (1)$$

with the frequency index k and block index b. However, b will be omitted for an improved readability in the following.

The range of the considered fundamental frequencies is limited by $F0_{\text{min}}/F0_{\text{max}}$ with the corresponding spectral bins $k_{\text{min}}/k_{\text{max}}$ or MCACF time lags $m_{\text{min}}/m_{\text{max}}$. Most of the relevant signal energy is found below 12 kHz and the spectrum is

[1] Music Information Retrieval Evaluation eXchange
http://music-ir.org/mirexwiki/

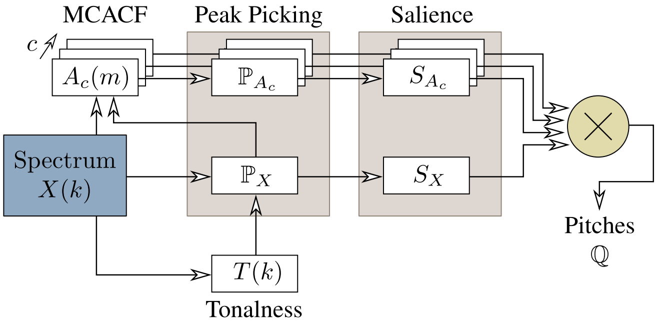


Fig. 1: Block diagram overview of the algorithm.

only evaluated up to a maximum bin $k_B = {}^{12\,\text{kHz}}/_{f_s} \cdot N_{\text{DFT}}$, where f_s denotes the sampling frequency. An overview of the different stages and the signal flow inside the algorithm is depicted as a block diagram in Fig. 1.

2.1. Tonalness estimation

A first step in the processing is the discrimination between noisy and tonal (sinusoidal) spectral components. Therefore, a tonalness measure

$$T(k) = \mathsf{t}_{\text{PK}}(k) \cdot \mathsf{t}_{\text{AT}}(k) \tag{2}$$

of each spectral bin is calculated as a multiplicative combination of the *peakiness* and *amplitude threshold* feature as described in [7].

2.2. Spectral peak picking

All K local maxima at the frequency indexes k_i, where the tonalness and magnitudes are above the thresholds

$$\begin{aligned} T(k_i) &> 0.7 \\ X(k_i) &> 0.001 \cdot \max\left[X(k)\right], \end{aligned} \tag{3}$$

are collected in the set of spectral peaks

$$\mathbb{P}_X = [k_1, \ldots, k_i, \ldots, k_K], \tag{4}$$

where k_i is limited to a range

$$k_{\text{min}} \leq k_i \leq k_B.$$

Every peak has a corresponding salience value

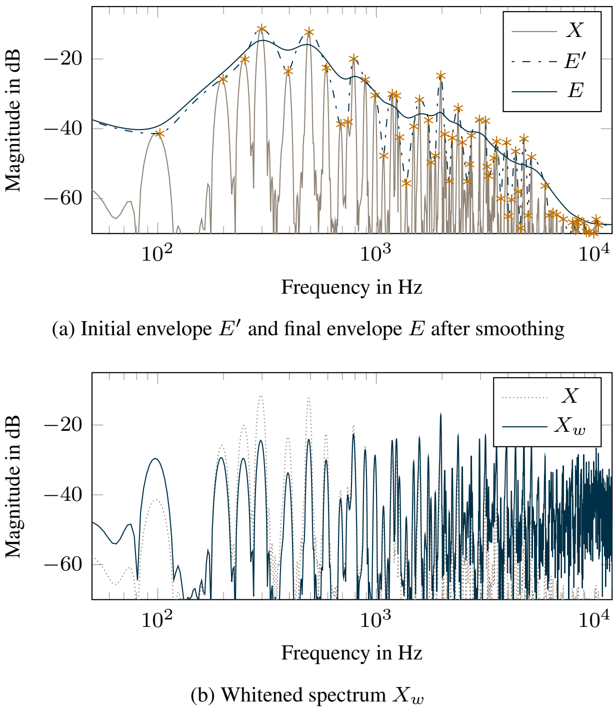
$$S_X(k_i) = \sum_{p=1}^{3} \Big(X\left(k_p\right)\Big)^{0.25} \tag{5}$$

which is the sum of the amplitudes of the first 3 harmonics at the positions $k_p = p \cdot k_i$. The spectrum is raised to a power of 0.25 before the summation to increase the influence of low energy regions in the salience calculation. To take a certain amount of inharmonicity into account, an improved salience calculation will search for a local maximum $\hat{k}_p$ in a surrounding Δ_k of the approximate position k_p and only fall back to k_p in the case that no local maximum was found.

For some instruments the fundamental frequency is considerably damped compared to the first harmonics and the threshold in (3) has to be as low as -60 dB to catch all possible $F0$ candidates. Naturally, these will then include a lot of false positives and after taking the harmonics into account with the salience calculation, all peaks which do not fulfil

$$S_X(k_i) > 0.1^{0.25} \cdot \max_{\forall k_i}\left[S_X(k_i)\right] \tag{6}$$

are removed again. However, this condition may become obsolete with an improved salience function or a more robust peak combination stage.



(a) Initial envelope E' and final envelope E after smoothing

(b) Whitened spectrum X_w

Fig. 2: Calculation of the spectral envelope (a) and final whitened spectrum with compensated envelope (b).

2.3. Multi-channel autocorrelation

Pre-whitening is performed to equalize the spectral envelope and to amplify low energy partials. An initial envelope E' is constructed as a curve through the spectral peaks $\mathbb{P}_X$ on a logarithmic frequency axis. It is recursively smoothed in both directions with a coefficient $\alpha = {}^{20}/_{N_W}$ and interpolated onto a linear frequency axis to yield the final envelope $E(k)$ (Fig. 2 a). The whitened spectrum

$$X'_w(k) = \frac{X(k)}{E(k)}, \tag{7}$$

$$X_w(k) = X'_w(k) \cdot \sqrt{\frac{\sum_{\kappa=0}^{k_B} X(\kappa)^2}{\sum_{\kappa=0}^{k_B} X'_w(\kappa)^2}} \tag{8}$$

is $X(k)$ divided by the envelope and additional normalization is applied to establish an equal power compared to the non-whitened spectrum in the important frequency region below k_B (Fig. 2 b).

The multi-channel autocorrelation (MCACF) is calculated in 5 bands with a width of one octave starting from the minimal pitched bin k_{min}. A set of filters

$$W'_c(k) = \begin{cases} \frac{4}{3\,k_c} \cdot k - \frac{1}{3}, & \frac{1}{4}\,k_c < k < k_c \\ 1, & k_c \le k \le 2k_c \\ -\frac{1}{18\,k_c} \cdot k + \frac{10}{9}, & 2\,k_c < k < 20\,k_c \\ 0, & \text{elsewhere} \end{cases} \tag{9}$$

with linear slopes is constructed where $k_c = 2^c \cdot k_{\text{min}}$ is the lower border of the current band and $c \in [0, 4]$ indexes the bands. The filters are additionally normalized

$$W_c(k) = \frac{W'_c(k)}{\sum_{\kappa=0}^{N_{\text{DFT}}} W'_c(\kappa)} \tag{10}$$

by the sum of their coefficients to compensate the increasing bandwidth and therefore higher energy in the upper octaves.

The slope of the bands appeared to have a huge impact on the quality of the resulting autocorrelation. On the one hand, it is necessary to remove high frequency components in order to avoid confusing their repetitions in the ACF with real pitches. On the other hand, a certain amount of partials will lead to much sharper located peaks in the ACF. The chosen parameters in (9) were found empirically and yield an ACF well suited for the following pitch detection step.

An efficient way to calculate the ACF is to take the inverse Fourier transform of the squared magnitude spectrum (Wiener-Khintchine theorem). By replacing the square in the exponent with an adjustable parameter the resulting ACF is non-linearly distorted. This results in the so-called generalised autocorrelation in channel c

$$A_c(m) = \text{IDFT}\left\{ \Big(X_w(k) \cdot N_W\Big)^{0.5} \cdot W_c(k) \right\} \tag{11}$$

where $X_w(k)$ is distorted by an exponent of 0.5 and weighted with the corresponding filter $W_c(k)$ prior to the IDFT. The variable m denotes the time lag and $X_w(k)$ is denormalized by N_W inverse to (1).

2.4. MCACF peak picking

All M_c local maxima at the time lag indexes m_j^c, where the MCACF is above the threshold

$$A_c(m_j^c) \quad > \quad 0.001 \cdot \textstyle\sum_c A_c(1), \tag{12}$$

are collected in the set of peaks

$$\mathbb{P}_{A_c} = [m_1^c, \ldots, m_j^c, \ldots, m_{M_c}^c],$$

where m_j^c is limited to a one octave range

$$2^{-(c+1)} \cdot m_{\max} \;\ge\; m_j^c \;\ge\; 2^{-c} \cdot m_{\max}.$$

Finally, the corresponding salience values

$$S_{A_c}(m_j^c) = \sum_{p=1}^{3} A_c(m_p) \tag{13}$$

are calculated for every peak and m_p is the approximate multiple $p \cdot m_j^c$. However, similar to Sec. 2.2, if there is a local maximum $\hat{m}_p$ in a range $\pm\Delta_m$ around m_p the amplitude at $\hat{m}_p$ will be taken instead. Negative values of the MCACF are not taken into account in the summation. In particular for short lags, associated with high pitches, the positions of the peaks are not accurate enough for a semitone resolution and it may be beneficial to calculate a refined base position $\hat{m}_j^c = \hat{m}_p/p$ from one of the multiples.

As there is a certain redundancy between the different bands due to the flat slopes of the filters, it is necessary to remove bands which do not carry enough information. Therefore, all bands c where

$$\max_{m_j^c \in \mathbb{P}_{A_c}} \left[A_c(m_j^c)\right] \;<\; 0.3 \cdot \max_{m > m_{\text{min}}} \left[A_c(m)\right] \tag{14}$$

are removed, which are bands where the maximum peak amplitude in $\mathbb{P}_{A_c}$ is significantly lower than the overall maximum in the MCACF apart from the zero lag. Like (6), this condition may be removed in case a more robust salience function or peak combination stage is found.

2.5. Peak combination

The frequency index and time lag values k_i and m_j^c of the peaks are translated to the corresponding frequencies in Hertz and quantised to the nearest semitones

$$Q_X(k_i) \quad = \quad \lfloor 69 + 2^{\frac{f_s}{k_i \cdot 440\,\text{Hz}}} \rceil, \tag{15}$$

$$Q_{A_c}(m_j^c) \quad = \quad \lfloor 69 + 2^{\frac{f_s \cdot m_j^c}{440\,\text{Hz}}} \rceil \tag{16}$$

in MIDI notation. Several pitch candidates from the spectrum or the MCACF may fall into the same semitone range. Hence, the salience vectors $S_{Q_X}(q)$ and $S_{Q_A}(q)$ for a semitone q

$$S_{Q_X}(q) \quad = \quad \underset{Q_X(k_i)=q}{\text{argmax}} \left[S_X(k_i)\right], \tag{17}$$

$$S_{Q_A}(q) \quad = \quad \underset{c}{\text{argmax}} \left[\underset{Q_{A_c}(m_j^c)=q}{\text{argmax}} \left[S_{A_c}(m_j^c)\right] \right] \tag{18}$$

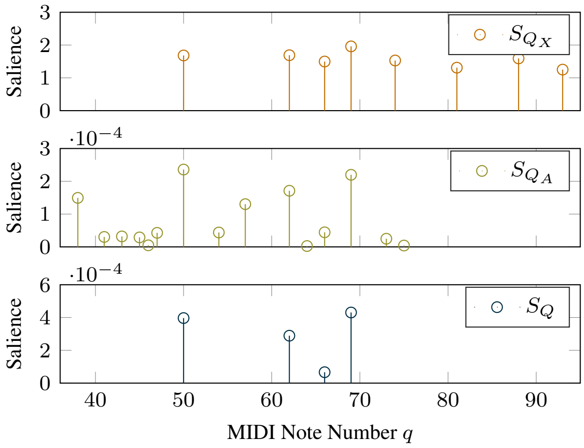


Fig. 3: Combination of spectral peaks (top) with MCACF peaks (middle) to yield the final detected pitches (bottom).

are unique mappings where only the maximum salience from the spectrum or MCACF in a semitone range remains and furthermore all channels c of the MCACF are summarized in a single vector. The final semitone salience

$$S_Q(q) = S_{Q_X}(q) \cdot S_{Q_A}(q) \qquad (19)$$

is the product of the individual saliencies. A last threshold is necessary to remove detections with very low and zero salience and all q where $S_Q(q) > 3 \cdot 10^{-5}$ are collected as the detected pitches in time frame b.

The process of combining pitch candidates is depicted as an example in Fig. 3 and in particular the candidates from the spectrum include a lot of false positives due to the harmonics. It would not be possible to set a threshold to reliably filter out these false positive candidates as the salience scores alone are not significant. However, by selecting candidates which are available in both sets, only true positive candidates remain in the bottom plot. It is obvious that this approach can just remove false positives and will not complete missing detections. Hence, it is important to assure that all pitches reliably evoke a peak in the MCACF as well as in the spectrum by selecting appropriate thresholds in (3) and (12). The proposed values were tweaked manually to achieve a balanced performance with various data sets.

3. EVALUATION

The presented algorithm was evaluated in two ways: First the influence of the polyphony level on the accuracy was investigated and afterwards three data sets were processed on the whole. In all evaluations the number of true positive, false positive and false negative detections were counted on a time grid of 10 ms throughout a single track. Based on these values the standard scores Precision, Recall, Accuracy and F-measure were retrieved [8]. The total score of a data set is the mean over the individual scores of the included tracks.

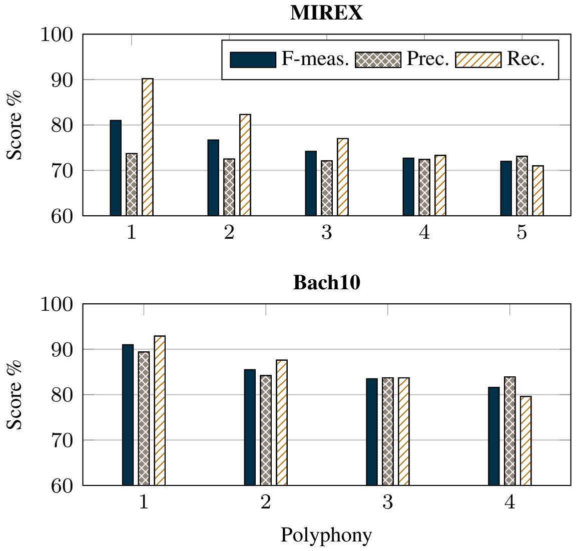


Fig. 4: Detection scores depending on the polyphony of the *MIREX Multi-F0 development* and *Bach10* data sets.

The input signals from the data sets have a sample rate $f_s = 44.1$ kHz and were normalized to a mean power of one to achieve a certain independence of the thresholds. The maximum search range for peaks in the spectrum and the MCACF is set to $\Delta_k = {}^{N_{\text{DFT}}}/_{3500}$ and $\Delta_m = {}^{f_s}/_{10000}$, respectively. The range of detectable pitches is limited to 5 octaves from $F0_{\text{min}} = 55$ Hz to $F0_{\text{max}} = 1750$ Hz.

3.1. Dependency on level of polyphony

The *Bach10* [9] and *MIREX Multi-F0 Woodwind Development* [8] data sets are available as single track recordings of monophonic instruments with separate ground truth information per track. This allows an easy recombination to achieve different levels of polyphony and results in 40 solo, 60 duet, 40 trio and 10 quartet tracks for the *Bach10* and 5 solo, 10 duet, 10 trio, 5 quartet and one quintet track for the *MIREX* data set. The detection results in dependency of the polyphony of the subsets are plotted in Fig. 4.

In both cases the F-measure and Recall values decrease with an increasing polyphony which is an expected behaviour. With the *Bach10* data set a good balance between Precision and Recall is kept independently of the polyphony level. However, the Precision values from the *MIREX* data set do not benefit from less polyphony.

3.2. Complete data sets

Additionally, the evaluation was performed with the *TRIOS* data set [10] and its results are compared with the *Bach10* and *MIREX* data sets in Table 1. For the latter ones these are identical to the respective values with the highest polyphony

Data set	F-meas.	Acc.	Prec.	Rec.
Bach10 [9]	81.6 %	69.0 %	83.9 %	79.6 %
MIREX [8]	72.0 %	56.3 %	73.1 %	71.0 %
TRIOS [10]	58.0 %	41.4 %	82.0 %	45.6 %

Table 1: Detection scores for full polyphony data sets.

in Fig. 4. Compared to the other sets, the *TRIOS* tracks are the most complex one. They consist of a polyphonic piano part mixed with one or two monophonic solo instrument voices. The solo voices are quite dominant and even for experienced listeners it is difficult to identify all voices of the piano apart from its main melody in the mixture.

The presented algorithm only reaches an F-measure of 58.0 % on the *TRIOS* data set which mainly suffers from a bad Recall of 45.6 %. Together with the high Precision score this indicates that most of the errors are missing detections and the algorithm simply cannot resolve the very dense arrangements. There are not a lot of reference results for the quite new *TRIOS* data set, yet, but Benetos [2] reported a 8 % higher F-measure (66.5 %). On the other hand, our achieved F-measure of 72.0 % with the *MIREX* data is 5 % better compared to the 67.2 % from [2] and also outperforms the 64.9 % from Cheng [11]. For the *Bach10* data set Duan [12] (without post processing) and Cheng [11] both report an F-measure of about 80 % which is similar to our 81.6 % in Table 1.

To summarize the evaluation, one can state that apart from the *TRIOS* results, the proposed approach reaches good scores which seem to reach into the range of state of the art algorithms. However, a more detailed evaluation as well as an analysis of the algorithm's parameters would be required for a final rating.

4. CONCLUSION

The autocorrelation was only rarely used for polyphonic pitch detection in the last years but in this paper it turned out to be a valuable mid-level signal representation. However, common modifications and subband processing are required to yield an autocorrelation that equally represents all necessary information. The simple matching of peaks in the spectrum and in the multi-channel autocorrelation as a basic criterion to detect pitches worked quite well and good F-measure values were achieved with the *MIREX* (72.0 %) and the *Bach10* (81.6 %) data sets. The results with the most complex *TRIOS* data set were not yet convincing, though. The main challenge for future developments would be to stabilize the Precision for low polyphony levels, e.g. by using a more complex scheme for the peak combination in order to remove false positives. In contrast, the bad Recall values require early optimisations in the spectrum and MCACF as these already seem to lack the necessary information and the combinational approach cannot reintroduce missing pitch candidates.

REFERENCES

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

[2] E. Benetos, S. Cherla, and T. Weyde, "An effcient shift-invariant model for polyphonic music transcription," in *Proc. 6th Int. Workshop on Machine Learning and Music*, 2013.

[3] K. Dressler, "Pitch Estimation by the Pair-Wise Evaluation of Spectral Peaks," in *Proc. 42th Int. AES Conference on Semantic Audio*, 2011.

[4] C. Yeh, A. Röbel, and X. Rodet, "Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1116–1126, Aug. 2010.

[5] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, pp. 407–434, 2013.

[6] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–20, Sept. 1997.

[7] S. Kraft, A. Lerch, and U. Zölzer, "The tonalness spectrum: feature-based estimation of tonal components," in *Proc. 16th Int. Conf. on Digital Audio Effects*, 2013.

[8] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *Proc. 10th Int. Society for Music Information Retrieval Conference*, 2009.

[9] Z. Duan, B. Pardo, and C. Zhang, "Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[10] J. Fritsch, *High Quality Musical Audio Source Separation*, Master, 2012.

[11] T. Cheng, S. Dixon, and M. Mauch, "A Deterministic Annealing EM Algorithm for Automatic Music Transcription.," in *Proc. 14th Int. Society for Music Information Retrieval Conference*, 2013.

[12] Z. Duan and D. Temperley, "Note-level music transcription by maximum likelihood sampling," in *Proc. 15th Int. Society for Music Information Retrieval Conference*, 2014.