IMPROVING MONOPHONIC PITCH DETECTION USING THE ACF AND SIMPLE HEURISTICS

Carlos de Obaldía, Udo Zölzer

Department of Signal Processing and Communications
Helmut Schmidt Univeristy
Hamburg, Germany
deobaldia@hsu-hh.de

ABSTRACT

In this paper a study on the performance of the short time autocorrelation function for the determination of correct pitch candidates for non-stationary sounds is presented. Input segments of a music or speech signal are analyzed by extracting the autocorrelation function and a weighting function is used to weight candidates for assessing their harmonic strength. Furthermore, a decision is devised which alerts if there are possible non-related jumps on the fundamental frequency track. A technique to modify the spectral content of the signal is presented to compensate for these jumps, and a heuristic to return a steady fundamental frequency track for monophonic recordings is presented. The system is evaluated with several databases and with other algorithms. Using the compensation algorithm increases the performance of the ACF and outperforms current detection algorithms.

1. INTRODUCTION

Intonation in human perception corresponds to the perceivable tone that is registered by the human brain. This is perceived as pitch, which is in turn related to the fundamental frequency f_0 of a particular sound, that is, the main frequency component of the Fourier series expansion of a signal. In order to extract such information from a signal, several methods have been introduced which help find the perceivable intonation, or pitch, of a particular sound which is usually estimated by its fundamental frequency.

Among the methods used for determining f_0, the autocorrelation function (ACF) has always been of particular interest since it represents the periodicities encountered in a waveform. This is specially suited for determining the fundamental frequency in audio recordings. In this manner, prominent peaks of the ACF will give information about the perceived *pitch* or tone of a sound, particularly when the latter is of a stationary nature.

There are two major study problems at the time of finding pitch tracks in monophonic recordings: The first one is the problem of finding reliable pitch candidates. This can be approached, for example, by analyzing the spectral peaks in the frequency domain, or by determining the autocorrelation lags of a signal segment in a particular way. Several solutions have been proposed in the literature to mitigate ambiguities and perform a correct estimation of the pitch; by analysis of the response to a filter bank, by the use of linear predictive coding to extract pitch candidates [1], by analyzing the spectra with the use of correlation-based functions [2], by the use of spectral weighting functions [3], by applying correlation in the spectral domain [4], with the help of compound weighting functions [5], or by the use of the normalized cross correlation [6], among other methods.

The other issue is the tracking of the fundamental frequency f_0 when several disturbances in the pitch trajectory occur in order to select the correct pitch candidate. Examples of this problem can be when harmonics acquire a higher energy than the fundamental, where subharmonics surpass a decision threshold, or when a particular note is retained after the current note is played or sung. In these cases, candidate pitch lags are taken into consideration and a heuristic rule is created to compensate for shifts and ambiguities. The mitigation of these ambiguities have been approached by different monophonic pitch detection algorithms, but it has also been a matter of study for polyphonic recordings, such by taking different candidates of the autocorrelation-spectrum pairs [7], by statistical analysis of the pitch candidates [8], or by finding other domains where the ambiguity of the estimation can be diminished [9]. Another solution, for example, was considered by [10] in treating the common frequency trajectories based on a graph-solving problem in the spectrogram using continuous Hidden Markov Models (HMMs). Several algorithms also smooth the magnitude function of consecutive spectra to find the most prominent peak. In this work it is shown that the variability of the pitch candidates can be modeled in a simple way, whilst not having to reuse different heuristics for the detection, and a method is presented which provides a way to account for possible ambiguities.

2. BACKGROUND

While cross-correlating two different signals, similarities which exist between the signals at a moment in the present are found based on the history of the signal. Mathematically, the autocorrelation of a stationary discrete signal $x(n)$ is defined as

$$r_{xx}(m) = \sum_{n=m}^{N-1} x(n)x(n-m), \tag{1}$$

where $x(n)$ is a windowed signal of length N_w, and m is the index of the delayed sample, called the lag. This effectively compares an isolated signal segment with a time shifted version of itself. It can be inferred from Eq. 1 that the global maximum for the function appears at $m = 0$. If there exist any maxima in $r_{xx}(m)$ for $m > 0$, the signal is said to be periodic with $T_0 = m$ and will contain local maxima at multiples of the lag m. Since the autocorrelation of the signal at $m = 0$ equals the power in the signal, the height of the local maxima in $r'_{xx}(m)$ for $m > 0$ represents the relative harmonic power of the signal. Thus the ratio of the height

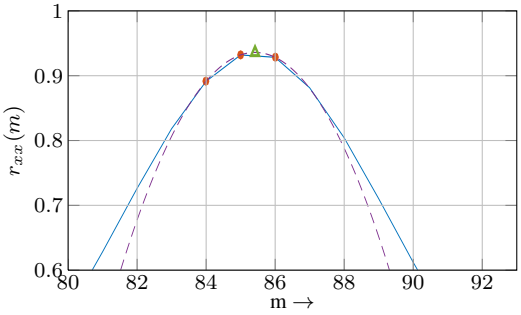

Figure 1: Interpolation of the candidates. The three red points are the anchors of the resulting interpolation shown with the dashed line. The corrected lag is the green triangle

of $r_{xx}(m)$ at specific lags m to the power of the signal,

$$r'_{xx}(m) \equiv \frac{r_{xx}(m)}{r_{xx}(0)}, \tag{2}$$

is considered to be the harmonic strength of the signal and indicates possible periodicities at the lags m whose harmonic strength approaches unity [11].

Since most of real-world signals are in the wide sense non-stationary, Eq.1 can be then calculated over a window $x_b(n)$ of length N_w centered at continuous places in time b spaced by N_h hop samples. This can give estimates of the fundamental frequency $f_0 = 1/T_0$ based on the harmonic strength at each frame b.

Another way to find lag candidates is with the use of the average magnitude difference function (AMDF)[5, 12], where a running sum is performed on the difference of the signal with itself, such that

$$g(m) = \sum_{n=0}^{N-1} (x(n) - x(n+m)). \tag{3}$$

This difference function can alternatively be used as a weighting function to help improve the accuracy of the ACF as a pitch estimator.

The local maxima of $r'_{xx}(m)$ of a particular segment provides a number of candidates which help determine the fundamental frequency in that segment along time. In this paper, several steps for preprocessing at the time of calculating the short time autocorrelation signal are assessed and evaluated, so that the most relevant steps for the calculation of the pitch candidates are taken into account. Furthermore, a method for weighting the harmonic strength of each candidate is presented. Methods to determine voiced and unvoiced parts in the signal are also presented, and a heuristic to determine and follow pitch trajectories is introduced as well. A spectral correction algorithm additionally improves the retrieval of candidates when the harmonic strength of multiples of the fundamental may mask the correct detection. Evaluation results and a comparison with several algorithms is presented and evaluated at the end.

3. PITCH EXTRACTION

Several algorithms which work on the spectral domain smooth the magnitude function of consecutive isolated spectra to find the most prominent frequency peak belonging to a particular frame [13]. In this work it is shown that the variability of the pitch candidates can also be modeled in a straightforward manner.

As it was introduced in Sec.2, the autocorrelation function has a global maximum at $m = 0$, where it represents the overall energy content of the particular isolated segment in which the function is calculated. Periodicities in $x_b(n)$ can then be found by analyzing the places where the pitch lag m is at other local maxima, or $r_{xx}(m_{peak})$ after $m = 0$. The relationship of these peaks to the maximum peak of the ACF would give a good cue of the fundamental period within a predefined interval $[m_{min}, m_{max}]$.

3.1. Pre-processing

For finding the fundamental frequency candidates, an incoming signal is hard - center clipped with a threshold of $\Gamma_c = 1 \times 10^{-3}$ to reduce the influence of other harmonic ratios on the signal [14]. The signal is then high-pass filtered with a butterworth filter of degree six at a cut off frequency of $f_c = 50$ Hz to reduce influences of low frequency noise.

At each windowed frame b of a signal $x(n)$ taken at a hop size $N_h = 0.01s \cdot f_s$, an isolated segment is extracted from $x(n)$ and multiplied with a hanning window

$$w(n) = \sin^2(\frac{\pi n}{N_w + 1}), \tag{4}$$

giving a signal segment $x_b(n)$ for $n = 0, ..., N_w + 1$ where N_w is the length of the window in samples. The autocorrelation of the segment is then calculated using Eq. 1.

The frame is then weighted using Eq. 3 such that

$$\hat{r}_{xx}(m) = \frac{r'_{xx}(m)}{g(m) + \alpha}, \tag{5}$$

where $\alpha = 1$ is taken. The output is then normalized to the maximum of the resulting signal.

3.2. Pitch Candidate Vector

The local maxima are found in the resulting weighted and normalized ACF $\hat{r}_{xx}(m)$ with a peak picking technique to generate a vector of pitch candidates

$$\mathbb{P} = [m_\kappa, ..., m_K], \;\; m_{min} < m_\kappa < m_{max}, \tag{6}$$

where m_{min} and m_{max} are the allowed minimum and maximum lags respectively, and $\kappa = \{1, \; ..., K\}$, where K is the number of detected pitch candidates above the threshold $g_r(m)$. After the previous step, just the positions of $\tilde{r}_{xx}(m)$ which surpass the threshold

$$q_r(m) = \frac{(\ln(f_s) - \ln(m))}{\ln(f_s)} \tag{7}$$

are taken in consideration and sorted in descending order with respect to the distance $\hat{r}_{xx}(m) - q_r(m)$. The resulting candidate vector $\mathring{\mathbb{P}}$ contains the current pitch estimation candidates for that particular frame.

3.3. Spectral Modification

Sometimes and due to the nature of particular sounds, the characteristics of the vocal tract, or the timbre characteristics of some instruments, will give away a formant structure which can detriment the detection performance of the most prominent lag [15].

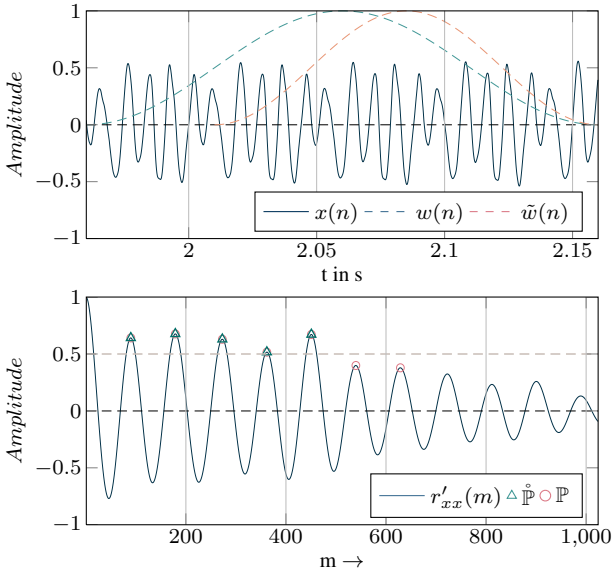

Figure 2: A case for finding the correct candidate with spectral modification. The top frame show the two windows used for calculating the lag. The lower picture shows the ACF of $w(n)$, and the peaks of $\mathring{\mathbb{P}}$ above a threshold.

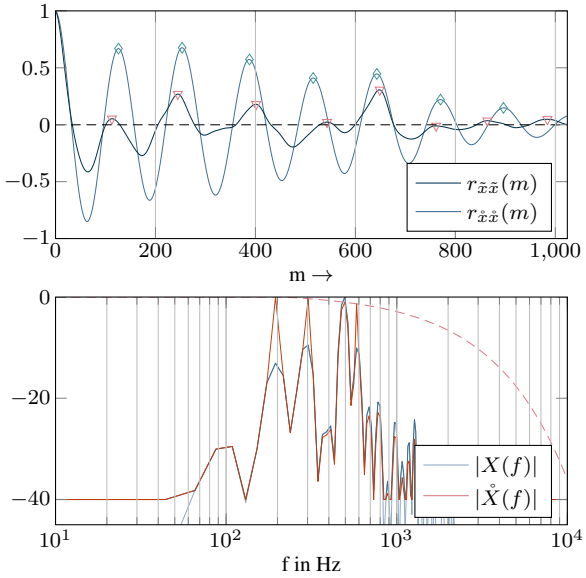

Figure 3: Spectral modification of the second window $\tilde{x}_b(n)$. The upper plot shows the resulting ACFs of $\tilde{x}_b(n)$ and the modified $\mathring{x}_b(n)$. The lower plot shows the spectrum $|\mathring{X}(f)|$ after modification.

To account for this, we introduce a candidate confirmation algorithm which follows the tracked pitch in the ACF, and determines its most plausible position. However some of the fragments may be accompanied by other harmonics which should not be taken in consideration, and the difference of the harmonic content between each frame should be taken into account.

To aid in this problem, the following technique makes use of the fact that if the harmonics of a particular fundamental frequency are equal in amplitude and since their harmonics are equally spaced in frequency, the distance between the resulting peaks of the autocorrelation from this modified waveform will match the fundamental period $T_0 = 1/f_0$.

The local maxima of $r_{xx}(m)$ above $q_r(m)$ are arranged in a vector $\mathring{\mathbb{P}}$, which contains the positions m of the pitch candidates sorted in descending order relative to $q_r(m)$. If the ratio of the harmonic strength of the first two lags $m_1,\ m_2$ in $\mathring{\mathbb{P}}$,

$$\frac{\min\left[\hat{r}_{xx}(m_1), \hat{r}_{xx}(m_2)\right]}{\max\left[\hat{r}_{xx}(m_1), \hat{r}_{xx}(m_2)\right]} \geq \gamma \tag{8}$$

taking γ=0.7, it will indicate that there is a possible mismatch with relation to the harmonic ratio of the frequency components of that particular signal segment. If there are more candidates in the vicinity of the harmonic strength for m_κ, a search is conducted to find a better estimate of the lag. A further window is thus applied on the frame to determine the correct pitch position, so that a second window in

$$\tilde{x}_b(n) = x(n - \frac{N_w}{4})\tilde{w}(n), \tag{9}$$

can be set, where $\tilde{w}(n)$ is a hanning window of length $\frac{N_w}{4}$ according to Eq.4 like in the upper plot of Fig.2. The spectrum of the signal snippet $\tilde{x}_b(n)$ is calculated taking

$$\begin{aligned} X_b(k) &= \mathcal{F}\{\tilde{x}_b(n)\} \\ &= \frac{1}{N_{\text{FFT}}} \sum_{n=1}^{N_{\text{FFT}}} x(n) \exp^{(\frac{-2\pi jkn}{N_{\text{FFT}}})} \forall k \in \{1, ..., N_{\text{FFT}}\}, \end{aligned} \tag{10}$$

where N_{FFT} is the size of the Fast Fourier Transform (FFT) and corresponds to the next power of two of N_w, and where $\tilde{x}_b(n)$ is zero padded accordingly. The spectrum is then normalized to its highest energy seen in any particular bin k, and the normalized power spectral density (PSD) of $\tilde{x}_b(n)$,

$$\tilde{X}_b(k) = \frac{|X_b(k)|^2}{\max[|X_b(k)|]}, \tag{11}$$

is calculated. The peaks of the spectral envelope which are above $-20\,dB$ are set to $0\,dB$ and a noise floor is in turn established at $-20\,dB$. We can then determine the harmonic content that remains in the signal from the frame before, by performing

$$\mathring{X}_b(k) = 2\tilde{X}_b(k) - \tilde{X}_{b-1}(k) \tag{12}$$

to reduce errors in transient regions. Furthermore, $\mathring{X}_b(k)$ is weighted with a roll-off factor of $40\,dB$ per octave starting at the first frequency bin k which is above $-20\,dB$ after Eq. 11, as it is depicted with a dashed line in Fig.3. The corresponding ACF of the delayed frame $\tilde{x}_b(n)$ is then calculated by using

$$r_{\mathring{x}\mathring{x}}(m) = \mathcal{F}^{-1}\{\mathring{X}_b(k)\}, \tag{13}$$

and the maximum peak $\mathring{m}$ in $r_{\mathring{x}\mathring{x}}(m)$ after $r_{\mathring{x}\mathring{x}}(0)$ is extracted. The resulting position in $\mathbb{P}$ which corresponds to the estimated fundamental frequency is found by calculating the closest m_κ in $\mathbb{P}$ with relation to $\mathring{m}$. Fig.3 shows the PSD of the second window in Fig. 2 which is used to determine the candidate m_κ for that frame b.

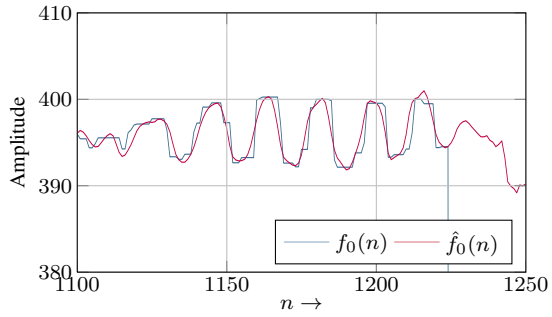

Figure 4: Result of applying the fractional lag calculation using the interpolation procedure for a vibrato around G_4 of a violin stem from Bach's 'Ach Gott und Herr'. The blue line represents the ground truth, $f_0(n)$.

3.4. Tracking Algorithm

Another problem at the time of finding the correct fundamental frequency is when the decay of a particular note is retained after another note starts. This has been shown to produce jumps in the fundamental frequency track, as the harmonic energy of two different fundamentals can be contained in the same frame.

To account for this, a method to provide a continuous tracking of f_0 is used. If there is a difference of more than $\delta_m = 10$ lags between consecutive estimates a decision is performed: If there exists a candidate in the current set $\mathbb{P}_b$ whose difference from $\mathbb{P}_{b-1}$ is at least 5 lags from the previous one, their positions are switched. This will prevent sporadic jumps in the track. Additionally, a guard interval is set such that this decision happens just when $\mathbb{P}_b$ of the ACF of 4 consecutive frames (b, ..., $b-3$) has contained at least one peak above the decision threshold.

3.5. Parabolic Interpolation

The ACF can be seen as a discrete and quantized signal where each pitch lag $m = 1, ..., M$ is an integer, so fractional tones can not be determined. A solution to this is to approximate the natural occurring tone with a parabolic interpolation taking as anchors the lag samples at positions between the local maxima [16]. If we take three points $y_1 = y(x_1)$, $y_2 = y(x_2)$, $y_3 = y(x_3)$ of a parabolic function of the form $y = ax^2 + bx + c$ around its local maximum $y(x_{max})$, and considering that $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ and $y_1 = r_{xx}(m_\kappa - 1)$, $y_2 = r_{xx}(m_\kappa)$, $y_3 = r_{xx}(m_\kappa + 1)$, then

$$\Delta x_{max} = \frac{1}{2} + \frac{1}{2}\frac{(y_1 - y_2)(y_2 - y_3)(y_3 - y_1)}{2y_2 - y_1 - y_3}, \tag{14}$$

will give the relative position of the maximum of the parabola with respect to $y_1 = r_{xx}(m_\kappa - 1)$, so that the lag is now estimated to be at $m_\kappa + \Delta x_{max}$. Although the approximation of the shape of the ACF to a parabola will introduce a bias to the estimation of the frequency, it represents a better approximation of the pitch. Fig. 1 shows the lag positions and the peak resulting from this approximation. Fig.4 shows a result.

3.6. Voiced Activity Detection

For the determination of voiced and unvoiced regions, a simple post processing procedure has been used to clean the track from frames which are falsely labeled as voiced. This has been also a major field of study for pitch detection algorithms, specially with the use of speech signals. A segment is defined as voiced if, based on the source-filter model of speech production, the signal to excite the vocal tract filter is of a periodic nature.

Thus the pitch candidates which lie below the threshold described in Eq. 7 are not taken in consideration. Secondly, candidates whose frame's root mean square (RMS) energy is below a threshold $\Gamma_E = 0.705 \times 10^{-3}$ are also not considered and taken as unvoiced. Lone pitches in frames which are separated more than 20 lags from the previous one are also considered as unvoiced. Moreover, pitch candidates are only considered if the frequency positions in the ACF are separated at least 15 lags. Lone candidates which spawn over three consecutive lags, are also discarded. This results in a continuous pitch track.

In summary, to find voiced and unvoiced regions, a frame b is voiced if

(a) $$\mathrm{RMS}\{x_b(n)\} \geq \Gamma_E, \tag{15}$$

and

(b) $$r_{xx}(m) > q_r(m), \tag{16}$$

and if

(c) $$|m_{b-1} - m_b| \leq 15 \tag{17}$$

which is the maximum lag so that the exchange between the candidates does not surpass this threshold.

Fig. 6 shows the results before and after removal of false positives for a saxophone and a bassoon recording from the Bach 10 dataset.

Finally, the algorithm can be summarized for each particular window at a frame b as follows:

1. Extract a signal segment $\hat{x}_b = x(n) \cdot w(n)$, where $w(n)$ is a hanning window of length N_w.
2. Calculate the autocorrelation function $r'_{xx}(m)$ according to Sec.3.1.
3. Calculate the AMDF as in Eq. 3
4. At each lag m, calculate the weighted function $\hat{r}_{xx}(m)$ as in Eq. 5
5. Normalize the weighted ACF $\hat{r}_{xx}(m)$ with respect to its maximum
6. Get a pitch candidate vector from the autocorrelation peaks and sort it according to Sec.3.2
7. Confirm pitch candidates according to the algorithm described in Sec.3.3, so that a new peak vector

$$\mathring{\mathbb{P}} = \{m_\kappa, \ldots, m_K\}, \tag{18}$$

is obtained, where m_1 corresponds to the estimated pitch lag, so that the estimated f_0 of frame b is $\hat{f}_0 = \frac{f_s}{m_1}$.

8. Run the tracking algorithm of Sec.3.4 for detecting discontinuities.
9. Determine voiced and unvoiced frames.

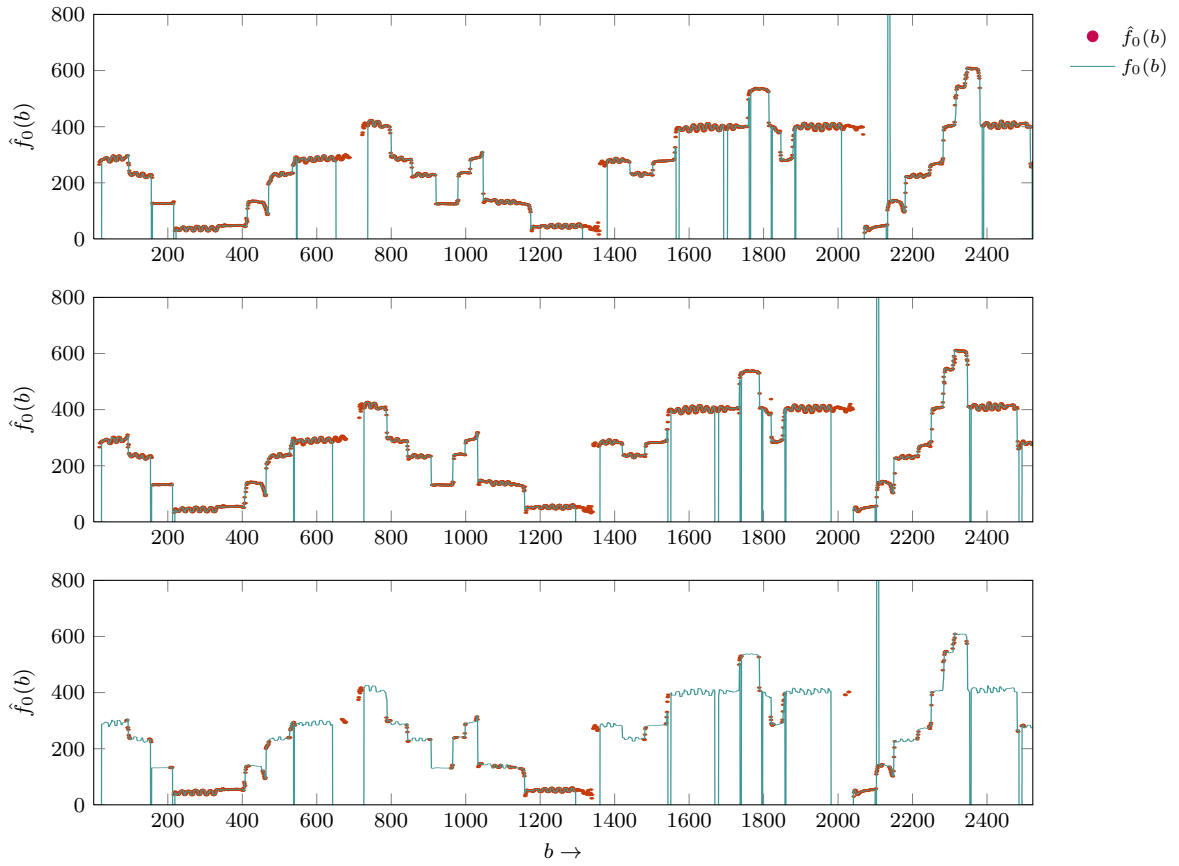

Figure 5: Pitch track for a violin recording of Bach's 'Ach Gott und Herr'. The bottom plot shows a result after selecting the first candidate of the ACF without the spectral modification decision. The middle one shows the same result but with a smaller window $N_w = 1024$. The uppermost plot shows the result with spectral modification with the same parameters as the bottom plot.

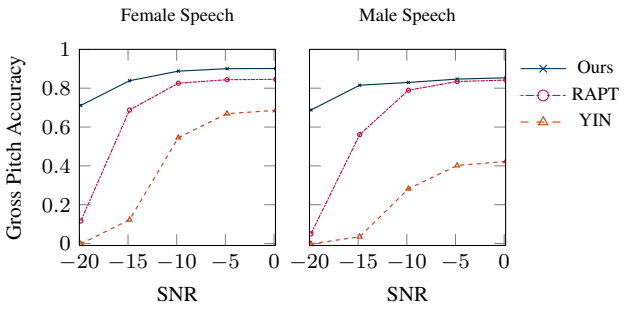

Figure 7: GPA for different SNR levels using the presented algorithm. The result shows an improvement on the pitch accuracy in voiced regions.

4. RESULTS AND EVALUATION

For the evaluation of the performance of the algorithm, two databases are used: The PTDB-TUG database, consisting of 10 female and 10 male speakers, each pronouncing 236 utterances from the TIMIT set, and the Bach 10 Chorales dataset which consists of stems for 10 different Bach songs which are recorded for violin, clarinet, saxophone and bassoon. The database also contains a midi set, but the proposed algorithm is not tested on midi data. The previously presented algorithm is thus evaluated on the monophonic stems of the Bach 10 dataset. For both databases, the original ground truth data is used as evaluation criterium. Results are evaluated with the YIN algorithm and the RAPT and PEFAC implementations found in the voicebox[17].

The overall accuracy is given by the F_0 Frame Error (FFE), which calculates the accuracy given a particular constrain among all the data and all the frames in the signal. For the evaluation, an error of 20%, 8% and $10\,Hz$ within the ground truth is chosen for all cases. Moreover, the results are also compared taking the Gross Pitch Error (GPE), which is the proportion of voiced frames in both the ground truth (GT) and the result, and the Fine Pitch Error (FPE) where the standard deviation of the distribution of relative error values is taken into account [18, 19]. Results are shown for the Bach 10 database in Table 1 and for the PTDB-UG in Table 2.

The algorithm achieves around 97% frame accuracy over the Bach10 dataset and around 90% frame accuracy for the speech examples of the PTDB, relative to 20% of the ground truth values provided for both datasets. For the Bach 10 evaluation a window of $N_w = 1536$ is used which gives a ground pitch error of $3,4\%$ and a F_0 frame error of about $8,4\%$ within 10 Hz for the whole of

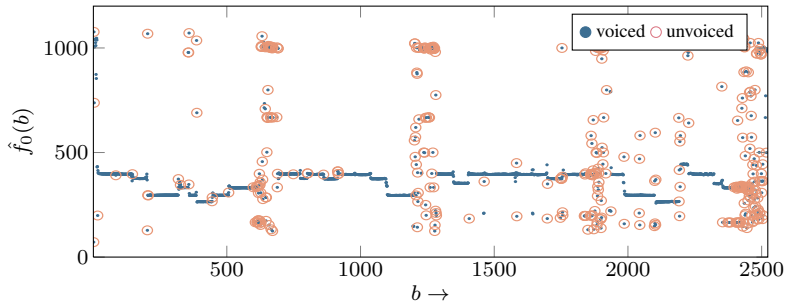

Figure 6: Voiced/Unvoiced detection is performed by applying a simple rule which combines the energy in the frame, the harmonic ratio, and the continuity of the fundamental track.

	violin			clarinet			sax			basoon			average		
Algorithm	GPE	FPE	FFE	GPE	FPE	FFE	GPE	FPE	FFE	GPE	FPE	FFE	GPE	FPE	FFE
Mod.	0.0314	1.7011	0.1107	0.0298	1.1313	0.0662	0.0209	1.1574	0.0478	0.0526	0.7795	0.0955	0.0337	1.1923	0.0801
8%	0.0185	0.5683	0.0199	0.0190	0.6052	0.0180	0.0150	0.6012	0.0155	0.0513	0.5618	0.0559	0.0260	0.5841	0.0273
20%	0.0156	0.8284	0.0179	0.0165	0.8369	0.0166	0.0122	0.8992	0.0142	0.0477	0.9681	0.0546	0.0230	0.8831	0.0258
$N_w = 1536$	0.0315	1.7293	0.1098	0.0301	1.1598	0.0631	0.0256	1.2116	0.0526	0.0508	0.8810	0.1111	0.0345	1.2454	0.0842
8%	0.0170	0.6175	0.0188	0.0183	0.5958	0.0183	0.0180	0.6667	0.0199	0.0492	0.6277	0.0753	0.0256	0.6269	0.0331
20%	0.0131	0.9566	0.0164	0.0151	0.9179	0.0164	0.0119	1.1938	0.0163	0.0421	1.2224	0.0720	0.0205	1.0727	0.0303
RAPT	0.0405	1.8504	0.1277	0.0343	1.1431	0.0711	0.0352	1.2343	0.0612	0.0693	0.9417	0.1090	0.0449	1.2924	0.0922
8%	0.0236	0.6702	0.0210	0.0223	0.6129	0.0198	0.0253	0.7280	0.0225	0.0675	0.6858	0.0602	0.0347	0.6742	0.0309
20%	0.0177	1.1332	0.0157	0.0181	1.0013	0.0161	0.0174	1.3514	0.0154	0.0573	1.4394	0.0512	0.0276	1.2313	0.0246
YIN	0.0555	2.286	0.1118	0.0340	1.3710	0.0561	0.0529	0.9837	0.0696	0.1510	0.4114	0.1631	0.0734	1.2622	0.1001
8%	0.0503	0.5167	0.0449	0.0281	0.5109	0.0250	0.0505	0.4570	0.0449	0.1508	0.2572	0.1344	0.0699	0.4353	0.0623
20%	0.0498	0.6102	0.0444	0.0272	0.6164	0.0241	0.0493	0.6102	0.0438	0.1504	0.3490	0.1340	0.0692	0.5387	0.0616
PEFAC	0.8670	1.7155	0.8055	0.6378	0.9186	0.5784	0.1250	0.9773	0.1338	0.1205	0.7539	0.1306	0.4376	1.0913	0.4121
8%	0.8648	0.7488	0.7698	0.6360	0.4417	0.5658	0.1202	0.5601	0.1069	0.1185	0.5756	0.1058	0.4348	0.5816	0.3871
20%	0.8594	4.478	0.7651	0.6310	1.999	0.5613	0.1083	1.6454	0.0964	0.1074	1.5181	0.0959	0.4265	2.4101	0.3797

Table 1: Results based on absolute error for the Bach 10 database. Comparison is done with the AMDF weighted ACF, with the exchange turned on.

the dataset. The error is higher on the bassoon stems, and improves to about 8% when the spectral modification algorithm is used. The only parameter which is being changed is the window length for the calculation.

For the PTDB database in Table 2, the accuracy is also presented using two different windows, at $N_{w1} = 2048$ and at $N_{w2} = 4096$ with spectral modification. The algorithm performs well over the other for the female speech samples, although for the smaller window size the algorithm finds problems in detecting low frequency fundamentals. For a higher window length, the algorithm performs well with spectral modification achieving over 92% accuracy for the F_0 frame error and around 89% ground pitch accuracy within 10 Hz of the ground truth.

Results under additive white gaussian noise (AWGN) at different signal to noise ratios (SNR) also show an advantage in comparison with the above mentioned algorithms. Fig. 7 show the Gross Pitch Accuracy (GPA) (where GPA $= 1 -$ GPE), for the presented algorithm evaluated under a subset of 20 utterances for the 10 male and 10 female speakers of the PTDB database. By using the presented pitch tracker, an accuracy of over 70% for all the cases, showing an increased improvement in the overall pitch accuracy for voiced regions.

In Fig.5 a result is shown on a violin track for two particular window sizes ($N_{w1} = 1024$ and at $N_{w2} = 2048$). The lower plot shows the detection performed with N_{w2} without spectral modification and the uppermost plot with the decision algorithm. It shows that although the lowest window performs well if the tracking algorithm is used, the accuracy increases using a longer window without diminishing the performance. This also shows the good performance of the spectral modification algorithm at the time of performing a correct detection without the need of further parametrization of the algorithm.

5. CONCLUSION

It has been shown that the algorithm proposed in this paper can be a reliable monophonic pitch detector because it pays attention to several properties in sound signal based on simple heuristics. Unwanted jumps in the pitch track which can occur due to the native timbre characteristics of musical instruments, or due to the resonant frequencies of the vocal tract at the time of uttering particular vowel qualities, can be diminished by the use of a spectral correction function when there exist ambiguities in the output of the weighted autocorrelation signal. Furthermore, the f_0 track is smoothed by the use of a tracking function that resolves the possible disturbances when, for example, a transient between continu-

	female			male		
Method	GPE	FPE	FFE	GPE	FPE	FFE
Mod. $N_{w\,2}$	0.1120	2.2640	0.0750	0.1135	2.1105	0.0805
8%	0.1152	2.1508	0.0715	0.1169	2.0854	0.0798
20%	0.1149	2.1809	0.0715	0.1150	2.1561	0.0797
Mod. $N_{w\,1}$	0.2516	3.8864	0.0997	0.2865	3.8846	0.1251
8%	0.2486	3.7214	0.0960	0.2918	3.8500	0.1243
20%	0.2461	3.8046	0.0959	0.2817	4.0088	0.1241
$N_w = 2048$	0.1063	2.5431	0.0758	0.1329	2.5086	0.1589
8%	0.1069	2.4395	0.0749	0.1375	2.4752	0.1587
20%	0.1057	2.4875	0.0749	0.1346	2.5677	0.1587
YIN	0.3028	2.2002	0.0835	0.5663	1.4291	0.1308
8%	0.3058	2.0779	0.0814	0.5728	1.4050	0.1310
20%	0.3058	2.0781	0.0814	0.5728	1.4057	0.1310
RAPT	0.1523	3.0357	0.1144	0.1523	2.8079	0.0955
8%	0.1520	2.9129	0.1065	0.1536	2.7766	0.0899
20%	0.1503	2.9803	0.1062	0.1496	2.8834	0.0891
PEFAC	0.3931	3.6037	0.2237	0.3586	3.2301	0.1885
8%	0.3827	3.4276	0.2082	0.3609	3.2083	0.1789
20%	0.3819	3.4668	0.2080	0.3562	3.3335	0.1779

Table 2: Performance results of different algorithms for the PTDB-TUG database.

ous notes is present. The fundamental frequency track can thus be reliably extracted by the use of the proposed ACF based algorithm without the need for tuning particular window sizes. Although there exist further possibilities of improvement in detecting the moment where transitions occur and for voiced segment determination, it is possible to find a decision threshold for the application of the presented algorithm. The spectral modification algorithm performs well at the moment of finding these transitions, and can be reliable for improving detection of the fundamental in voiced speech and musical signals.

6. REFERENCES

[1] Bruce Secrest and George Doddington, "An integrated pitch tracking algorithm for speech systems," in *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1983, vol. 8, pp. 1352–1355.

[2] Li Hui, Bei-qian Dai, and Lu Wei, "A pitch detection algorithm based on AMDF and ACF," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[3] Sira Gonzalez and Mike Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 518–530, 2014.

[4] E Chilton and BG Evans, "The spectral autocorrelation applied to the linear prediction residual of speech for robust pitch detection," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 358–361.

[5] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] David Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[7] Sebastian Kraft and Udo Zölzer, "Polyphonic pitch detection by matching spectral and autocorrelation peaks," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1301–1305.

[8] Johannes Böhler and Udo Zölzer, "Monophonic pitch tection by evaluation of individually parameterized phase locked loops," in *19th International Conference on Digital Audio Effects (DAFX16)*, 2016, vol. 68.

[9] Sebastian Kraft, Alexander Lerch, and Udo Zölzer, "The tonalness spectrum: feature-based estimation of tonal components," in *16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland*, 2013, p. 8.

[10] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*. Amsterdam, 1993, vol. 17, pp. 97–110.

[12] Myron Ross, Harry Shaffer, Andrew Cohen, Richard Freudberg, and Harold Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[13] Tom De Mulder, Jean-Pierre Martens, Micheline Lesaffre, Marc Leman, Bernard De Baets, and Hans De Meyer, "An auditory model based transcriber of vocal queries," 2003.

[14] Joseph Carl Robnett Licklider and Irwin Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," *The Journal of the Acoustical Society of America*, vol. 20, no. 1, pp. 42–51, 1948.

[15] Mohan Sondhi, "New methods of pitch extraction," *IEEE Transactions on audio and electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[16] Adrian von dem Knesebeck, *Analyse- und Syntheseverfahren zur automatischen Harmonisierung von Gesang*, PhD dissertation, Helmut Schmidt Universität, 2014.

[17] Mike Brookes et al., "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[18] Wei Chu and Abeer Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3969–3972.

[19] Sofia Strömbergsson, "Today's most frequently used f0 estimation methods, and their accuracy in estimating male and female pitch in clean speech.," in *INTERSPEECH*, 2016, pp. 525–529.