Pitch Estimation by the Pair-Wise Evaluation of Spectral Peaks

Karin Dressler

Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany

kadressler@gmail.com

ABSTRACT

In this paper, a new approach for pitch estimation in polyphonic musical audio is presented. The algorithm is based on the pair-wise analysis of spectral peaks. The idea of the technique lies in the identification of partials with successive (odd) harmonic numbers. Since successive partials of a harmonic sound have well defined frequency ratios, a possible fundamental can be derived from the instantaneous frequencies of the two spectral peaks. Consecutively, the identified harmonic pairs are rated according to harmonicity, timbral smoothness, the appearance of intermediate spectral peaks, and harmonic number. Finally, the resulting pitch strengths are added to a pitch spectrogram.

The pitch estimation was developed for the identification of the predominant voice (e.g. melody) in polyphonic music recordings. It was evaluated as part of a melody extraction algorithm during the Music Information Retrieval Evaluation eXchange (MIREX 2006 and 2009), where the algorithm reached the best overall accuracy as well as very good performance measures.

1. INTRODUCTION

Pitch estimation algorithms have been traditionally discussed primarily in close relation to human perception [1]. Yet, with the growing interest in applications for the automatic transcription of music, new aspects have inspired the research on pitch estimation. The analysis of real world music is a big challenge, as the signal may include many different sound sources. Usually, there is no prior information about the number of sources or their spectral envelopes. Inharmonic spectra may occur, as well as percussive sounds.

Pitch extraction methods which are designed to work with monophonic audio often fail to produce satisfactory results with polyphonic music signals. Even the estimation of the predominant pitch may pose problems, as the method has to be robust against spurious components, the interference of partials from different sounds, and octave ambiguities.

Of course, the human auditory system still plays an important role in recent research on pitch estimation [2, 3, 4, 5]. Methods based on modeling human perception are characterized by the band-wise processing of the audio signal: the signal is analyzed by a filterbank, and then the predominant period is detected in each frequency channel. The period detection within a frequency band may be performed by an autocorrelation function [2, 3, 4] or by the application of the STFT [5]. Finally, the period information of the distinct frequency channels is summed up to obtain a measure of the pitch salience.

However, methods based on the Fourier frequency spectrum prevail in melody extraction and multiple F0 estimation applications [6, 7, 8, 9]. One reason might be the more efficient computation of the spectral analysis. The period detection, which is performed on the spectrogram representation of the audio signal, is often based on the idea of pattern matching. Popular methods include the PreFEst algorithm developed by Goto [10] and the subharmonic summation algorithm proposed by Hermes in [11].

Many of the recent pitch detection algorithms exploit the spectral structure of musical sounds to address the problem of shared harmonics [3, 6, 8, 9]. In polyphonic music, the musical intervals between simultaneously sounding notes usually have a harmonic relationship, so that partials of one tone will be at the same frequency as partials of another tone. It is hard to apply prior knowledge to keep apart partials from distinct audio sources, since there is a huge variety of possible timbres. However, some physical properties of complex tones provide clues to tackle the problem. Such properties include the

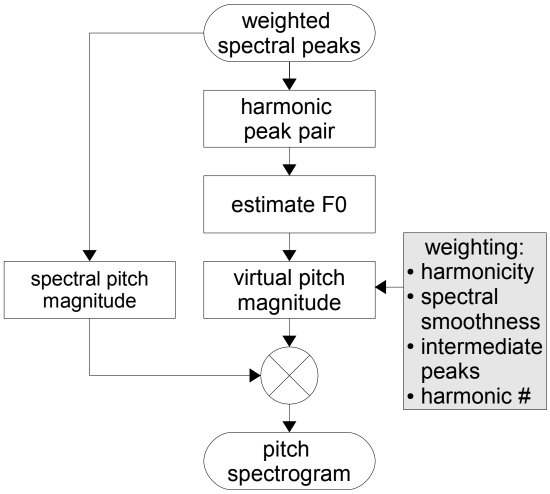

Fig. 1: Overview of Pitch Estimation Algorithm

harmonicity of the tone's partials, the smoothness of the spectral envelope, or the synchrony of the amplitude evolution of the harmonics.

The proposed pitch estimation algorithm is based on the idea of subharmonic summation as described by Terhardt in [12]. Subharmonic summation explains well the perceived pitch of harmonic complex tones and quantitatively predicts a great variety of pitch phenomena. There is one shortcoming of the above method that becomes very apparent in the analysis of polyphonic audio: each spectral peak creates a huge number of candidate virtual pitches, so the situation becomes quite complex if there is more than one note playing simultaneously. In the presented algorithm, the number of possible subharmonics can be reduced considerably by the pair-wise processing of spectral peaks. In order to address the problem of shared harmonics and octave ambiguities, additional measures are introduced. The measures exploit the physical properties of musical sounds, for example the average spectral slope of complex tones, the harmonicity of partials, and the smoothness of the spectral envelope.

2. METHOD

2.1. Overview

The flowchart displayed in figure 1 gives an overview about the pitch estimation method. The input to the algorithm are the magnitude and the instantaneous frequency (IF) of the spectral peaks obtained from a multi resolution spectrogram. Then, each peak magnitude is weighted with its respective IF. As indicated by the leftmost path in the flowchart, the weighted peak magnitude is added as spectral pitch magnitude directly to the pitch spectrogram.

The estimation of the virtual pitch magnitudes includes more processing steps. Consecutively, two spectral peaks at one time are combined into a candidate harmonic peak pair. It is then assumed that both peaks are successive (odd) harmonics (with harmonic numbers 1 and 2, 2 and 3,... as well as 1 and 3, 3 and 5, etc.). Following this assumption, it is possible to calculate the fundamental frequency of the perceived virtual pitch. Some additional weightings are applied, which rate the probability that both peaks are indeed successive (odd) harmonics: 1) the harmonicity weighting rates the frequency relation between spectral peaks, 2) the spectral smoothness criterion determines the maximum supported virtual pitch magnitude, 3) the presence of intermediate spectral peaks reduces the impact of the considered peak pair, and 4) the harmonic number also influences the virtual pitch magnitude. After all peak pairs have been processed, the virtual pitch magnitudes are added to the pitch spectrogram.

2.2. Spectral Analysis

If a partial of a complex tone is not obscured by other harmonics or noise it can be detected as a peak in the magnitude spectrum of the Short Term Fourier Transform (STFT). In the case of polyphonic audio, multiple sound sources play simultaneously and the interference between concurrent sounds becomes more apparent in the Fourier spectrogram. The interference of partials from simultaneously playing notes can be decreased if the frequency resolution of the STFT is increased. However, musical sound changes over time, so long STFT data windows cannot be used to gain a very high frequency resolution. A compromise has to be found between a good frequency resolution and a good time resolution.

Such a compromise could be the use of a multi resolution spectrogram which is obtained from the audio signal by calculating a multi resolution Fast Fourier Transform (MR FFT) [13]. The best frequency resolution ($\Delta f = 21.5$ Hz) is reached for the low frequency components up to approximately 600 Hz. The best time resolution corresponds to a FFT data window length of 5.8 ms for frequencies above 4400 Hz. Due to different amounts of zero padding the resulting STFT frame size and the hop size of the analysis window are 2048 and 256 samples for all STFT resolutions, respectively[1].

The proposed pitch estimation method takes the peaks of the STFT magnitude spectrogram and their respective instantaneous frequencies (IF) as input. For the computation of the pitch spectrogram spectral peaks in the frequency range between 55 Hz and 5 kHz are processed. The lower limit has been set according to the typical frequency range of melody notes, the higher limit denotes the frequency threshold for the induction of a virtual pitch in the human auditory system [14, chapter6]. In order to obtain more stable IF measures, the average frequency of two estimation methods is used, namely the well-known phase vocoder [15] and a method proposed by Charpentier [16].

Terhardt et al proposed the processing of aurally relevant sinusoidal components and the deletion of masked sounds [12]. In previous work [13], we aimed at the explicit identification of sinusoidal components in the music signal. However, the accurate distinction between sinusoidals and noise is a challenging task especially for signals with many concurrent sounds. In order to preserve as much spectral information as possible and since the proposed pitch estimation algorithm is very robust against additional noise, this processing step has been omitted.

If solely the predominant periodicity shall be extracted, the computational efficiency of the algorithm can be increased by setting a magnitude threshold for the spectral peak candidates, a threshold that is 30 dB below the maximum weighted peak magnitude is used for melody extraction.

2.3. Magnitude Weighting

In order to obtain the weighted magnitude A_s for the spectral peak at STFT bin k, its STFT magnitude $|X[k]|$ is multiplied with the peak's instantaneous frequency f_i.

$$A_s[k] = |X[k]| \cdot f_i[k] \qquad (1)$$

This weighting introduces a 6 dB magnitude boost per octave. In effect the weighted signal is proportional to the signal derivative.

The proposed magnitude weighting is based on the spectral structure of musical sounds: The spectral slope of instruments used in music and speech is between - 3 dB and -12 dB per octave [17], the musically most interesting rates of spectral rolloff are between 3 and 9 dB per octave [18]. Sundberg found that the long-term average spectral slope of speech and orchestra music is -6 dB per octave [19]. Hence, the weighting shall equalize the impact of low and high harmonics for the average complex tone in music, which ideally has a spectral slope of -6 dB per octave.

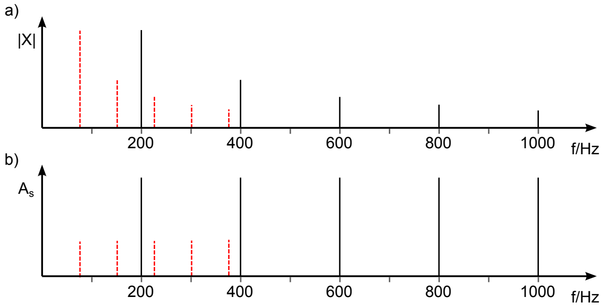

Fig. 2: Magnitude weighting for two tones with a spectral rolloff of 6 dB per octave: a) STFT magnitude b) weighted magnitude

Figure 2 allows a qualitative comparison between the STFT magnitudes and the weighted magnitudes. The example uses two complex tones with 5 harmonics and a spectral slope of -6 dB per octave. After the weighting the harmonics of each tone have equal magnitudes. It can also be noted that the resulting spectrum is not flat. In this respect the proposed weighting differs from spectral whitening methods (for example [3]), as it markedly damps the low frequency bands.

All subsequent processing steps are computed with the weighted spectral magnitude A_s.

2.4. Spectral Pitch Magnitude

The spectral peak itself naturally invokes a pitch perceived at its own instantaneous frequency. So at first the weighted magnitude A_s is added to the pitch spectrogram, if the instantaneous frequency f_i is in the desired pitch frequency range $f_{\min} \leq f_i < f_{\max}$. Since the pitch spectrogram has a logarithmic frequency scale, f_i is converted to a cent value c_i:

$$c_i = 1200 \log_2\left(\frac{f_p}{f_{\text{ref}}}\right) \qquad \text{with} \quad f_{\text{ref}} = f_{\min}. \qquad (2)$$

The minimum pitch frequency $f_{\min} = 55\text{Hz}$ is used as reference frequency. In this case the lowest possible cent

[1] Assuming audio data sampled at 44.1 kHz.

value in the pitch spectrogram is zero. If the frequency resolution of the pitch spectrogram buffer is set to 1 cent, the estimated cent values can be used as indices to the spectrogram.

The spectral pitch magnitude is represented by a Gaussian weighted with A_s:

$$g(c) = A_s e^{-\frac{1}{2}\left(\frac{c-c_i}{35}\right)^2}. \tag{3}$$

The Gaussian reaches half its maximum value with a cent offset of $|c - c_i| \approx 41\,\text{cent}$. The width of the Gaussian has been adjusted experimentally by the evaluation of the melody extraction system [2].

2.5. Virtual Pitch Magnitude

According to Terhardt the formation of virtual pitch can essentially be said to be a process of subharmonic matching [12]. He presumed that each of the spectral pitches evokes candidate virtual pitches at its subharmonic frequencies. The subharmonic frequencies are found by dividing the partial frequency f_i by integer numbers from 1 up to N[3] Basically, the virtual pitch is perceived where most of the candidate virtual pitches of the different spectral peaks match.

The presented approach builds upon the idea of subharmonic matching. Still, contrary to Terhardt, we do not assume virtual pitch candidates at each subharmonic frequency of a spectral peak. Rather, we demand that only (odd) successive harmonics evoke a virtual pitch. This way, the number of candidate virtual pitches can be decreased noticeably, because the considered subharmonic frequencies are derived from the frequency intervals between spectral peaks (see section 2.5.1).

In principle the virtual pitch magnitude is derived from the weighted spectral magnitude of the identified harmonic. However, several additional ratings are introduced that estimate the probability of the virtual pitch. Consecutively, the identified harmonics are rated according to harmonicity, timbral smoothness, the appearance of intermediate spectral peaks, and harmonic number (see sections 2.5.2–2.5.5).

2.5.1. Pair-Wise Subharmonic Summation

In order to detect (odd) successive harmonics, spectral peaks are evaluated pair-wise. Successively, each spectral peak is combined with all other peaks. For each peak pair, it is assumed that both spectral peaks are partials with an (odd) successive harmonic number (harmonic numbers 1 and 2, 2 and 3,... as well as 1 and 3, 3 and 5, etc.). Using the supposed harmonic relationship of the spectral peaks, the most likely harmonic numbers can be derived from their instantaneous frequencies f_{high} and f_{low}.

At first, it is assumed that both peaks are successive harmonics. In this case, the harmonic number h_{low} of the partial with the lower frequency f_{low} is computed as:

$$\frac{h_{\text{low}}}{h_{\text{low}}+1} = \frac{f_{\text{low}}}{f_{\text{high}}} \Rightarrow h_{\text{low}} = \text{round}\left(\frac{f_{\text{low}}}{f_{\text{high}} - f_{\text{low}}}\right). \tag{4}$$

The harmonic number of the partial with the higher frequency is $h_{\text{high}} = h_{\text{low}} + 1$.

Then, the supposed harmonic numbers are calculated assuming odd successive harmonics:

$$\frac{h_{\text{low}}}{h_{\text{low}}+2} = \frac{f_{\text{low}}}{f_{\text{high}}} \Rightarrow h_{\text{low}} = \text{round}\left(\frac{2f_{\text{low}}}{f_{\text{high}} - f_{\text{low}}}\right). \tag{5}$$

Using equation 5, the computed harmonic number is valid only if the rounded result is indeed an odd number. Naturally, the harmonic number of the partial with the higher frequency is $h_{\text{high}} = h_{\text{low}} + 2$.

Because of equation (5) odd harmonics are ”discovered” more often than even harmonics. To avoid an increased impact of odd harmonics in the final pitch spectrogram, all identified harmonic numbers for one peak are at first solely listed. After all possible peak pairs have been evaluated, the computed virtual pitch is added to the pitch spectrogram only once for each harmonic number found.

The virtual pitch frequency f_p is computed individually for each partial by the straightforward division of instantaneous peak frequency and estimated harmonic number, e.g. $f_p = f_i/h$. Experimental results have shown that harmonics with a harmonic number h greater than 20 do not improve the estimation accuracy.

The virtual pitch magnitude is estimated using several weightings which are described in the following sections.

2.5.2. Harmonicity

Let’s imagine a spectral peak pair with the instantaneous frequencies $f_{\text{low}} = 300$ Hz and $f_{\text{high}} = 400$ Hz. According to equation (4) the harmonic number is calculated as $h_{\text{low}} = 300\,\text{Hz}/(400-300)\,\text{Hz} = 3$. Since the values

[2] In order to save compuation time, the Gaussian weightings are precomputed and only 100 values are added to the pitch spectrogram, which has a resolution of 1 cent.

[3] Terhardt sets the maximum harmonic number to 12. In the presented approach, harmonics up to harmonic number 20 are considered.

form an example of an ideal harmonic relation between successive harmonics, the result is exactly the harmonic number and has not be rounded. If we consider another peak pair with the instantaneous frequencies $f_{\text{low}} = 300$ Hz and $f_{\text{high}} = 415$ Hz, the result of equation (4) before rounding is approximately $h^*_{\text{low}} = 2.6$. In this case it may be doubted that both peaks are successive harmonics, because the frequency interval is not close to any ideal harmonic relation.

Most of the evaluated peak pairs are actually not successive (odd) harmonics. The estimated ideal harmonic relation can be a criterion to rule out such peak pairs. The allowed offset between the the estimated frequency interval and the exact harmonic interval is set to 120 cent:

$$1200 \cdot \left| \log_2 \left(\frac{f_{\text{high}}}{f_{\text{low}}} \right) - \log_2 \left(\frac{h_{\text{high}}}{h_{\text{low}}} \right) \right| < 120 \qquad (6)$$

If, for example, the frequency interval between two peaks shows a 100 cent offset from the exact harmonic interval, both peaks will probably not belong to the same sound source. But maybe both peaks are indeed successive harmonics, and only the estimated instantaneous frequencies are erroneous. Anyway, the virtual pitches which are induced by both peaks will not combine to a joint pitch in the pitch spectrogram, because the Gaussian function which is used in the summation has its inflection points at ±35 cents. Hence, the estimated pitch from this peak pair is ambiguous. Nonetheless, such a marked offset is allowed in order to obtain as many valid peak pairs as possible – even though the frequency relation is quasi inharmonic. Very often the ambiguity is resolved by the summation of other harmonics, so that in the end the best matching virtual pitch frequency can be estimated with some reliability [4].

On the logarithmic frequency scale only the lowest neighboring harmonics have very distinct frequency intervals (1200, 702, 498, 386 cent), while for example the intervals between harmonics 14/15 and 15/16 are 119.4 cent and 111.7 cent, respectively. This means that only for the lower harmonics the harmonicity can be an effective criterion to rule out peak pairs. As the frequency intervals between high harmonics are very similar on the logarithmic frequency scale, even small deviations from the ideal harmonic frequencies can lead to a faulty estimation of the harmonic number. That is the reason why the virtual pitch estimates from very high harmonics are not very reliable.

The harmonicity rating r_i is implemented simply as a boolean value – the peak-pair is discarded if the condition given in 6 does not hold[5].

2.5.3. Spectral Smoothness

Most instrument sounds show pronounced peaks (formants) as well as regions with lower energy in their spectral envelope. While it is impossible to make predictions about the spectral power distribution as a whole, a certain smoothness of the spectral shape is observed. The assumed smoothness of the spectral envelope can be used as an additional criterion in the pitch extraction.

As a consequence it is eligible that successive harmonics have more or less the same magnitude. However, some stopped-pipe wind instruments, for example the clarinet or the panpipe, have timbres that mostly contain odd harmonics. In this case it is hard to find partials with successive harmonic numbers. So again, odd successive harmonics have to be considered to estimate the smoothed spectral envelope.

The supported (smoothed) virtual pitch magnitude S_h depends on the magnitudes of the neighboring harmonics. At first, the preliminary support magnitudes S^- and S^+ are estimated separately for each frequency direction. If the current harmonic number h is even, S^- and S^+ are computed from the magnitudes of the harmonic neighbors A_{h-1} and A_{h+1}:

$$\begin{aligned} S^- &= \min(4 \cdot A_{h-1}, A_h) \qquad \text{and} \\ S^+ &= \min(4 \cdot A_{h+1}, A_h). \end{aligned} \qquad (7)$$

If the current harmonic number h is odd, also the odd harmonic neighbors are considered. In this case, the biggest harmonic neighbor from the higher and the lower frequency range is chosen for the calculation:

$$\begin{aligned} S^- &= \min(4 \cdot \max(A_{h-1}, A_{h-2}), A_h) \qquad \text{and} \\ S^+ &= \min(4 \cdot \max(A_{h+1}, A_{h+2}), A_h). \end{aligned} \qquad (8)$$

If a partial has only one harmonic neighbor, the other support magnitude is set to zero.

[4] In the melody extraction algorithm, the harmonics are added to tone objects during the subsequent processing. After the fundamental frequency of the tone object has been estimated, the allowed frequency offset for the inclusion of a harmonic is usually much lower (e.g. the maximum offset is 35+h)

[5] Within the allowed interval the harmonicity may be rated with help of a cosine function: rating 1 is reached for an ideal harmonic relation, rating zero is given at the interval borders.

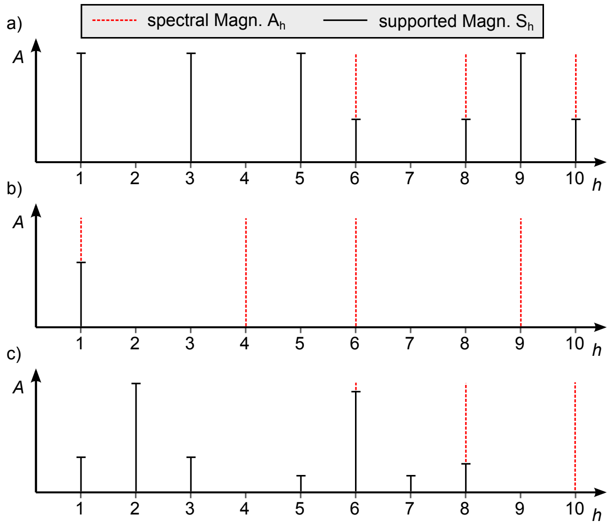

Fig. 3: Combination of harmonic candidates and supported virtual pitch magnitudes: a) combination of 7 partials, which have at least one neighboring (odd) harmonic, b) combination of 5 partials, which have no harmonic neighbors, c) combination of 8 partials which have distinct weighted magnitudes

The final supported magnitude S_h is a weighted sum of S^- and S^+. In the weighted sum, the smaller supported magnitude gets a higher weight. Three conditions are distinguished:

$$S_h = \begin{cases} 0.6A_h + S^+ & \text{if } h = 1 \\ 0.4S^- + S^+ & \text{if } S^- > S^+ \\ 0.4S^+ + S^- & \text{else.} \end{cases} \qquad (9)$$

The estimated magnitude support S_h must not be greater than the weighted spectral magnitude A_h. The constant factors used in equations 7 - 9 have been found empirically. The required support from neighboring harmonics is an important difference to Terhardt's algorithm. If some harmonics are missing or cannot be detected the algorithm outputs may differ drastically. Figure 3 shows the smoothed spectral envelopes for different combinations of sinusoidals. The three examples show how missing or weak harmonic neighbors lead to a reduction of the supported virtual pitch magnitude. If no harmonic neighbors can be identified (as is the case for partials 4, 6 and 9 in figure 3b), no virtual pitch is induced. Still, the implemented timbral smoothing allows a certain degree of variation in the spectral envelope, as can be noted in figure 3c.

2.5.4. Attenuation by Intermediate Peaks

Usually, each spectral peak is combined with a number of different peaks from the lower and higher frequency range. Among the possible peak combinations, the pairings of immediately neighboring spectral peaks are of particular interest for the pitch estimation. Nonetheless, we do not use the order of the peak combination directly as a measure, because the spectrum also includes spurious peaks, which might skew the rating. Rather, the magnitudes of the intermediate spectral peaks are summed up and compared to the magnitudes of the evaluated peak pair. If the noise level is comparatively low, at least the noise peaks will not influence the rating too much.

The rating factor r_m, which represents the attenuation of the virtual pitch magnitude due to intermediate spectral peaks, is given by

$$r_m = \frac{A_{\min}}{A_{\min} + K\sum_i A_i}. \qquad (10)$$

The term $\sum_i A_i$ denotes the sum of all peak magnitudes A_i that exist between the evaluated peaks. The term $A_{\min}$ is the smaller spectral magnitude of the evaluated peak pair, e.g. $A_{\min} = \min(A_{\text{low}}, A_{\text{high}})$. The constant factor K determines the attenuation. We have found empirically that $K = 0.5$ gains the best results in our melody extraction system.

The main benefit of the masking criterion is the prevention of octave errors. Of course, intermediate peaks also occur because of the overlapping spectra of simultaneous sound sources. But certainly, advantage is taken from the fact that often the timbres of different instruments dominate in different spectral regions. For example, the strongest partials of the bass instruments are often found in the low frequency range, while the melody voice usually has strong harmonics in the high frequency regions.

2.5.5. Harmonic Impact

A small, but positive effect is gained if the impact of the higher harmonics is reduced by a small amount. The damping of higher harmonics amounts to only 1 dB per octave. The weighting factor r_h depends on the harmonic number h:

$$r_h = h^{-\frac{1}{20\log(2)}}. \qquad (11)$$

The parameter r_h denotes the harmonic impact.

2.5.6. Estimation of the Rated Virtual Pitch Magnitude

In order to obtain the virtual pitch magnitude A_v, the supported peak magnitude S_h is multiplied with the ratings derived from the harmonicity r_i, the appearance of intermediate spectral peaks r_m, and the harmonic impact r_h:

$$A_v = r_i \cdot r_m \cdot r_h \cdot S_h. \tag{12}$$

The resulting virtual pitch magnitude A_v is added to the pitch spectrogram in the same way as the spectral pitch magnitudes (see section 2.4).

3. RESULTS

3.1. MIREX Audio Melody Extraction

The presented pitch estimation method has been implemented as part of a melody extraction algorithm which was evaluated at the Music Information Retrieval Evaluation eXchange (MIREX) in 2006 and 2009 [20, 21]. On both occasions, the algorithm achieved the best overall accuracy and at the same time stands out due to very short runtimes. Of course, there are more processing steps involved in the extraction of the melody. Nonetheless, the simple tracking of strong pitches in the pitch spectrogram will already produce good results [20].

Table 3.1 shows a brief excerpt of the evaluation results in the years 2006 and 2009. The Raw Pitch measure represents the estimation performance for all voiced frames. This means the evaluation is constrained to the time instants where the melody voice is present. The measure Overall Accuracy requires a voicing detection which is not performed by all systems (affected results are marked by an asterisk).

The two best results for the datasets ADC 2004 and MIREX 2005 are given for comparison, because the ADC database and the development collection of MIREX 2005 is used in a subsequent evaluation[6]. Furthermore, the average results over all databases are presented for the melody extraction task of MIREX 2009.

It should be noted that the results of MIREX 2006 are more significant for the evaluation of the proposed pitch estimation method, because the tone tracking was performed directly on the pitch spectrogram data. The current melody extraction system uses the proposed pitch estimation method to estimate starting points of high level tone objects.

3.2. Frame-Wise Evaluation of the Predominant Pitch Detection

In spite of the promising MIREX results, it is difficult to identify the contribution of the pitch extraction method to the overall performance.

Salamon and Gómez have estimated the potential performance of a a chroma-based pitch salience function by considering an increasing number of salient pitch peaks [7]: presuming an ideal pitch selection process, the melody is identified correctly as soon as one of the peak candidates matches the transcribed reference frequency.

We adopt this idea for the evaluation of the proposed algorithm, however, using modified conditions for the analysis. Since the above-quoted approach uses chroma features, octave errors are not detected. In our evaluation, the reference frequency is not mapped to an octave range. The pitch is identified correctly if the frequency is less than 50 cent away from the ground truth. Moreover, a magnitude threshold which lies 10 dB below the maximum pitch magnitude of the analysis frame is imposed on the pitch candidates.

Figures 5 and 6 show that the estimation accuracy converges towards a ”glass ceiling” with a rising number of pitch candidates. The limiting value for the ADC 2004 database amounts to 93%. It does not differ significantly from the value of about 90% given in [7]. However, an improvement can be noted for the most salient pitch peak (77 versus 71 % in [7]), even though the conditions used for our evaluation are more strict[7].

Another interesting aspect is the impact of the different parameters on the algorithm performance (see figures 5 and 6). The most important individual processing step – apart from the basic algorithm structure – is the magnitude weighting introduced in section 2.3. At first sight, the magnitude weighting seems to be counterproductive, because it in fact takes away power from the fundamental frequency. Yet, during subsequent processing the fundamental frequency will also profit from the strong weighting of the overtones.

And one significant advantage remains: the notes from a potential bass voice are damped while at the same time

[6] Both datasets with reference transcriptions can be downloaded at http://labrosa.ee.columbia.edu/projects/melody/

[7] Unfortunately the estimation results for the MIREX 2005 development collection (MIREX05 train) cannot be compared meaningfully since the reference data has been corrected only recently.

Year	Dataset	Algorithm	Raw Pitch (%)	Overall Acc (%)	Runtime
2006	ADC 2004	Dressler	82.9	82.5	27 s
		Ryynänen & Klapuri	80.6	77.3	440 s
	MIREX 2005	Dressler	77.7	73.2	48 s
		Ryynänen & Klapuri	71.5	67.9	773 s
2009	ADC 2004	Dressler	87.1	86.3	n/a
		Cao & Li 1	85.1	76.6	n/a
		Cancela	82.9	82.5	n/a
	MIREX 2005	Dressler	76.4	74.8	n/a
		Wendelboe	75.0	58.2*	n/a
		Cancela	68.0	66.5	n/a
	all datasets	Dressler	80.6	73.4	24 min
	(unweighted av.)	Tachibana et al	75.1	55.1*	1468 min
		Durrieu et al 1	74.5	66.9	23040 min
		Wendelboe	73.4	55.1*	132 min
		Joo et al	73.3	56.6	3726 min
		Rao & Rao	72.2	65.2	26 min
		Hsu et al 1	66.1	50.5	344 min
		Cancela	64.1	62.9	4677 min
		Cao & Li 1	63.5	52.2	28 min

Fig. 4: Melody Extraction Results of MIREX 2006 and MIREX 2009

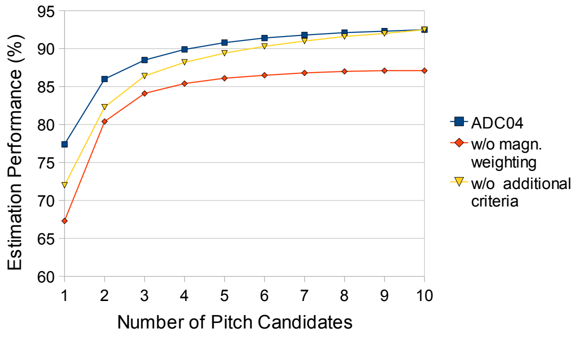

Fig. 5: ADC04: Potential Performance vs Peak Number

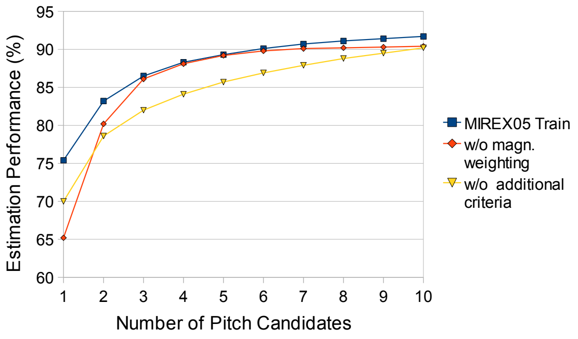

Fig. 6: MIREX05 train: Potential Performance vs Peak Number

the melody notes are boosted. As a consequence the greatest impact of the magnitude weighting is observed if the audio recordings contain musical voices of comparable strength. In fact, the improvement of the estimation accuracy can be attributed rather on the improvement in individual test files than on general characteristics of the data collections. For example, the estimation accuracy is increased markedly for test files with a strong bass voice (for example midi1: +52%, midi4: +58%, train12: +42%, train13: +38%), while the detection of the predominant pitch is slightly worse for files which have additional voices with a higher frequency than the actual melody voice (daisy2: -10%, train10: -8%).

No other criterion or parameter has such a marked effect on the estimation accuracy like the magnitude weighting. Yet, the small contributions of the individual measures sum up to a significant improvement, as can be noted from the yellow curve in the diagram. Here, the spectral smoothing, the attenuation by intermediate peaks, and the harmonic number weighting have been omitted. Instead, the weighted spectral magnitude A_s was added as virtual magnitude to the pitch spectrogram.

Surprisingly, the spectral envelope smoothing has no significant effect on the estimation of the predominant voice in the frame-wise evaluation. This may be contributed to the fact that most of the music pieces tested have a strong melody voice.However, the spectral smoothing plays an important role for the estimation of multiple fundamental frequencies.

4. SUMMARY AND PERSPECTIVES

In this paper we presented a novel approach to pitch detection in polyphonic music. The pair-wise evaluation of spectral peaks results in considerable time savings, because the number of possible virtual pitches can be significantly reduced. Additional ratings have been introduced in order to avoid octave errors and to discriminate peaks/partials from different audio sources.

The MIREX results show that the pitch extraction algorithm works well with different kinds of polyphonic music. Furthermore, the analysis of the pitch estimation front-end reveals that in most cases the predominant voice is identified correctly even without any postprocessing. Another positive characteristic of the proposed method is the very efficient computation of the pitch.

Despite the promising results, it must be noted that many aspects of human pitch perception are not covered by the proposed algorithm. The calculated pitch magnitude does not exactly correspond to the magnitude perceived by humans. In particular the magnitude estimate should not depend substantially on the existence of other audio sources – as it does in the proposed algorithm.

Furthermore, it should be noted that the pitch spectrogram is not a one to one representation of existing musical notes. The pitch strengths can be seen as probabilities of perceiving a predominant pitch. In order to retrieve other (weaker) tones, the effects of the predominant pitch have to be factored out, because considerable pitch strengths occur at integer multiples of its fundamental frequency or in combination with other periodic sound sources.

For the detection of multiple pitches different approaches have been proposed which address the problem of shared harmonics. Such approaches include the iterative detection of the predominant pitch and the subsequent deletion of the tone [2, 3], as well as the joint pitch candidate selection [5, 6, 8]. Preliminary experiments have shown that the presented approach can be quite easily adapted for the iterative detection/cancellation method.

5. REFERENCES

[1] A. de Cheveigné. Pitch perception models. In *Pitch: neural coding and perception*. Springer Verlag, New York, 2005.

[2] A. de Cheveigné and H. Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27:175–185, 1998.

[3] A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, 2003.

[4] R. P. Paiva, T. Mendes, and A. Cardoso. An auditory model based approach for melody detection in polyphonic musical recordings. *Lecture Notes in Computer Science - Computer Music Modeling and Retrieval: Second International Symposium, CMMR 2004*, 3310:21–40, February 2005.

[5] A. P. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Speech and Audio Processing*, 16(2):255–266, 2008.

[6] A. Pertusa and J.M. Iñesta. Multiple fundamental frequency estimation using Gaussian smoothness. In *IEEE International Conference on Acoustics, Speech and Signal Processing: ICASSP 2008*, pages 105-108, 2008.

[7] J. Salamon and E. Gómez. A chroma-based salience function for melody and bass line estimation from music audio signals. In *6th Sound and Music Computing Conference (SMC 2009)*, pages 331–336, Porto, Portugal, 2009.

[8] C. Yeh, A. Robel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1116–1126, 2010.

[9] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2010.

[10] M. Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication (ISCA Journal)*, 43(4):311–329, September 2004.

[11] D. J. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

[12] E. Terhardt. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America*, 71(3):679–687, March 1983.

[13] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, pages 247–252, Montreal, Quebec, Canada, Sept. 2006.

[14] B. C. J. Moore. *An introduction to the psychology of hearing.* Academic Press, San Diego, California, 2003.

[15] J. Flanagan and R. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, September 1966.

[16] F. Charpentier. Pitch detection using the short-term phase spectrum. In *Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing*, pages 113–116, Tokyo, Japan, 1986.

[17] C. D. Tsang and L. J. Trainor. Spectral slope discrimination in infancy: Sensitivity to socially important timbres. *Infant Behavior and Development*, 25(2):183–194, 2002.

[18] M. Mathews. Introduction to timbre. In *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics*, pages 79–87. MIT Press, 2001.

[19] J. Sundberg. Perception of singing. *Speech Transmission Laboratory, Quartery Progress and Status Report (KTH, Stockholm) STL-QPSR*, 20(1):1–48, 1979.

[20] K. Dressler. An auditory streaming approach on melody extraction. In *2nd Music Information Retrieval Evaluation eXchange (MIREX)*, Victoria, Canada, 2006.

[21] K. Dressler. Audio melody extraction for MIREX 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.