SINUSOIDAL EXTRACTION USING AN EFFICIENT IMPLEMENTATION OF A MULTI-RESOLUTION FFT

Karin Dressler

Fraunhofer Institute for Digital Media Technology
Langewiesener Str. 22, 98693 Ilmenau, Germany
dresslkn@idmt.fraunhofer.de

ABSTRACT

This paper provides a detailed description of the spectral analysis front-end of a melody extraction algorithm. Our particular approach aims at extracting the sinusoidal components from the audio signal. It includes a novel technique for the efficient computation of STFT spectra in different time-frequency resolutions. Furthermore, we exploit the application of local sinusoidality criteria, in order to detect stable sinusoids in individual FFT frames. The evaluation results show that a multi resolution analysis improves the sinusoidal extraction in polyphonic audio.

1. INTRODUCTION

The presented approaches to sinusoidal identification are part of a melody extraction algorithm, which aims at the transcription of the predominant voice out of polyphonic real-world music. The most relevant melody information can be found in the sinusoidal components of the audio signal, thus it is very desirable to divide the signal into a deterministic signal component plus noise [1, 2]. Only partials which originate from periodic sound are processed further. This way we particularly reduce the computational cost of our pitch estimation method, which strongly depends on the number of peaks to be analysed.

Often, sinusoidal extraction methods, which are designed to work with monophonic audio or artificial test signals, fail to produce satisfactory results with polyphonic audio signals. The presented approaches are adapted to the specific demands of this challenging task and – consequently – will also be validated against the melody extraction results.

Another problem we address, is the choice of the proper time-frequency resolution of the spectral analysis method. If the spectral representation exhibits constant frequency resolution, almost no change in frequency is observed for the low fundamental of an frequency-modulated tone, but vivid dynamics are noted for its higher harmonics. To cover fast signal changes, we have to increase the analysis bandwidth, but at the same time we have to maintain an adequate discrimination of concurrent sounds. A solution to this conflict lies in analysis methods which provide a more ore less logarithmic frequency scale.

Unfortunately such techniques are often computationally expensive, so the Fast Fourier Transform remains the tool of choice in time-critical applications. In need of good frequency resolution, long FFT windows have to be applied – accepting a distorted spectrum for faster changing signal components. In practice, local sinusoidality criteria for the detection of sinusoids often fail in high frequency bands, because they depend on the shape of the spectral window function. The same is true for the estimation of the instantaneous frequencies and magnitudes. Masri presented a method for the identification of non-stationary sinusoids for well defined types of spectral distortion, but in real-world signals the frequency modulation will not follow such idealised trajectories [3].

A good compromise is a multi-resolution analysis based on the FFT algorithm. A prominent example is the application of a multirate filter bank in combination with the FFT used by Goto [4]. Another straight forward idea is the calculation of the FFT with different window lengths, resulting in different time-frequency resolutions. Essentially the multi-resolution FFT (MR FFT) is an efficient implementation of this idea.

2. MULTI-RESOLUTION FFT

Given a sequence of data samples $x[n]$ the Short Time Fourier Transform is defined as $X_l[k]$:

$$X_l[k] = \sum_{n=0}^{M-1} x[n + lL] \cdot w_N^{-kn}, \qquad l = 0, 1, ... \text{ and } k = 0, 1, ..., N-1 \tag{1}$$

where

N	is the number of STFT points
L	is the time advance of the data frame (hop-size)
M	is the size of the data frame
l	is the number of the data frame
k	is the frequency bin number
w_N	is the N^{th} primitive complex root of unity

The values of N, L and M are the control parameters of the STFT, which determine certain characteristics of the spectrogram representation: the spacing of the discrete time-frequency grid of the spectrogram depends on the sampling rate f_s, the number of STFT points N and on the time advance of the data window L, which is also called hop-size. The grid spacing is determined by $\Delta f_{\text{grid}} = \frac{f_s}{N}$ and $\Delta t_{\text{grid}} = \frac{L}{f_s}$.

The grid spacing is not necessarily the time and frequency resolution we obtain from the spectrogram. The frequency resolution (the ability to distinguish two closely spaced frequencies from the original input signal) and also the time resolution is determined by the sampling rate, the size of the data window M and also by the shape of the window function, which will not be under consideration here. The resolution is given by $\Delta f = \frac{f_s}{M}$ and $\Delta t = \frac{M}{f_s}$.

If we use zero-padding, the frequency resolution is smaller than the spacing between the frequency bins, because the used data frame is smaller than the number of STFT points ($M < N$). If we use overlapping data frames ($L < M$), the time resolution does not increase, but nevertheless more STFT frames are included on the time axis. The additional information is obtained by interpolation.

The MR FFT determines the different resolutions by changing the data frame size M only. The hop-size L, the number of STFT points N and – in consequence – the spacing of the time-frequency grid remain unchanged among the different spectrograms.

The basic idea of the MR FFT is derived from the fact, that the summation operation is associative, thus we are allowed to split summations into simpler sums. In a reformulation of equation (1) (for clarity with $M = N$), we split the original sum with length N into N/L sums of length L (hop-size). Hereafter we can again sum the partial sums, and the result is of course the same:

$$\begin{aligned} X_l[k] &= \sum_{n=0}^{N-1} x[n+lL] \cdot w_N^{-kn} \\ &= \sum_{c=0}^{\frac{N}{L}-1} \sum_{n=cL}^{(c+1)L-1} x[n+lL] \cdot w_N^{-kn}. \end{aligned} \tag{2}$$

The inner sum in equation (2) can be expressed as a (time-shifted) zero-padded STFT of the data sequence $x_c[n]$:

$$X_c[k] = \sum_{n=0}^{N-1} x_c[n] \cdot w_N^{-kn}, \quad k = 0, 1, ..., N-1, \tag{3}$$

with

$$x_c[n] = \begin{cases} x[n+lL], & \text{for} \quad cL \le n < (c+1)L; \\ 0, & \text{elsewhere;} \end{cases}$$

where c is a circular counter related to the data frame number l by $c = l \bmod \frac{N}{L}$.

This transform can be computed by an FFT algorithm. The resulting complex Fourier coefficients are stored in a circular buffer of the dimension $[N/L, k_{\text{max}}]$, since one elementary transform is used in N/L calls of the MR FFT method and only up to k_{max} frequency bins may be of interest for the subsequent analysis.

The FFT spectra $X_c[k]$ form the basis of the MR FFT: all different resolutions can be calculated as a summation of the elementary transforms. Summing up to N/L neighboring elementary transforms increases the frequency resolution from f_s/L to f_s/N with the increasing number of summands r. In order to comply with the condition for windowing in the frequency domain (see section 2.1), the number of summands r is restricted to certain values, because the fraction $N/M = N/(rL)$ has to be an integer value. For example, if $N = 2048$ and $L = 256$, the sum of $r = 1, 2, 4, 8$ elementary transforms is possible – resulting in four different spectrogram resolutions with $M = 256, 512, 1024, 2048$.

While the magnitudes of the summed spectrograms are immediately valid, the phase of the complex Fourier coefficients has to be corrected in order to make windowing in the frequency domain possible. The phase error is due to the time-shift of the data which introduces a phase shift in the frequency domain according to the shifting theorem of the DFT:

$$x[n+L] \overset{N}{\longmapsto} X[k] \cdot w_N^{kL}. \tag{4}$$

The angle of the phase shift is dependent on the frequency of the designated frequency bin k and the circular counter c, which is related to the data frame number l as defined in (3). Fortunately this effect can be cancelled by multiplying the phase-shifted spectrum $X_r^*[k]$ with a twiddle factor as follows:

$$X_r[k] = X_r^*[k] \cdot w_N^{-k\, c_{\text{min},r} L}, \quad r = 1, 2, 4, ..., N/L, \tag{5}$$

where r is the number of summed elementary transforms $X_c[k]$ and $c_{\text{min,r}}$ is the circular counter index of the smallest frame number $l_{\text{min,r}}$ of the summed elementary transforms:

$$c_{\text{min,r}} = l_{\text{min,r}} \bmod \frac{N}{L}. \tag{6}$$

2.1. Windowing in the frequency domain

It is obvious that we cannot use time domain windowing with the MR FFT. But rather than applying the window in the time domain, we always have the option to perform frequency domain windowing, because the transform of a product is equivalent to the convolution of the two corresponding transforms. Admittedly, convolution is a time consuming operation, and it is only an alternative if the discrete spectrum of the window function is a short sequence of convolution coefficients. Fortunately some common windows have this desired property. The temporal weightings of interest have the general form:

$$\begin{aligned} h[n] &= \sum_{m=0}^{M/2} (-1)^m a_m \cos\left[\frac{2\pi}{M} mn\right] \\ &= a_0 - a_1 \cos\left(\frac{2\pi}{M} n\right) + a_2 \cos\left(\frac{2\pi}{M} 2n\right) - \\ &\quad a_3 \cos\left(\frac{2\pi}{M} 3n\right) + ..., \qquad n = 0, 1, ..., M-1, \end{aligned} \tag{7}$$

and

$$\sum_{m=0}^{M/2} a_m = 1,$$

where M is the size of the data window and a_m are real constants [5]. Since the most important windows of this form have $a_m \neq 0$ only for small m, equation (7) is reduced to a few terms.

For any K nonzero coefficients a_m, the continuous spectral window function $H(\omega)$ consists of a summation of $2K-1$ weighted Dirichlet kernels:

$$H(\omega) = \sum_{m=0}^{M/2} (-1)^m \frac{a_m}{2} \left[D\left(\omega - \frac{2\pi}{M} m\right) + D\left(\omega + \frac{2\pi}{M} m\right) \right], \tag{8}$$

where $D(\omega)$ is the Dirichlet kernel as given in

$$D(\omega) = \left(+j\frac{\omega}{2}\right) \frac{\sin\left(\frac{M}{2}\omega\right)}{\sin\left(\frac{1}{2}\omega\right)}. \tag{9}$$

The Dirichlet kernel is implicitly available through the STFT with a rectangular window. So for the discrete case and if $M = N$ equation (8) simplifies as indicated in:

$$X[k]|_{\text{win}} = \sum_{m=0}^{M/2} (-1)^m \frac{a_m}{2} \left(X[k-m] + X[k+m]\right), \quad k = 0, 1, ..., N-1. \tag{10}$$

That is the reason why these windows are especially useful for frequency domain windowing, because they can be described by a short $(2K-1)$ sequence of convolution coefficients. For example the Hann and Hamming windows possess only three nonzero coefficients, the Blackman window five:

- Hann: $a_0 = 0.5$, $a_1 = 0.5$
- Hamming: $a_0 = 0.53836$, $a_1 = 0.46164$
- Blackman: $a_0 = 0.42$, $a_1 = 0.5$, $a_2 = 0.08$

Other windows of this form with very good sidelobe behavior are described in [5].

Yet, we have to pay attention to the fact that equation (10) only holds for the special case $M = N$. In order to apply frequency windowing to the transform of zero-padded data we have to refine:

$$X[k]|_{\text{win}} = \sum_{m=0}^{M/2} (-1)^m \frac{a_m}{2} \left(X\left[k - m\frac{N}{M}\right] + X\left[k + m\frac{N}{M}\right] \right) \quad k = 0, 1, ..., N-1. \tag{11}$$

The term $m\frac{N}{M}$ in (11) must be an integer, because we only know the Fourier coefficients $X[k]$ at discrete bin locations k. Hence the number of possible spectrogram resolutions is reduced to a sub set by the condition $(\frac{N}{M} = \frac{N}{rL}) \in \mathbb{N}$, where r is the number of summed elementary transforms.

2.2. Algorithmic Complexity

The algorithmic complexity of an N-point FFT is usually quantified as $C_{\text{FFT}} = O(N \log N)$, yet, if we use overlapping FFT windows only hopsize L new data samples are processed within one FFT frame.

While the computation of the MR FFT elementary transforms $X_c[k]$ has also complexity C_{FFT} per L samples, the additional effort depends on the required frequency range of the distinct spectrogram resolutions. If we compute all MR FFT spectrograms up to the highest frequency bin k_{max}, we have to perform $\left(\frac{N}{L} - 1\right) k_{\text{max}}$ complex additions (due to the summation of the elementary transforms), plus k_{max} complex multiplications for every spectrogram resolution. Furthermore, we require additional storage for the intermediate results and the twiddle factors. If we compute the multiple resolutions (including zero padding) using the standard FFT, the computational cost is C_{FFT} multiplied with the number of resolutions. If a Hann window is applied to a real data sequence of length N in the time domain, N real multiplies have to be performed. In the frequency domain, the computational complexity can be identified as $6k_{\text{max}}$ real multiplications and $3k_{\text{max}}$ complex additions .

2.3. Implementation

Our approach to sinusoidal extraction has been successfully implemented in a melody extraction algorithm. For audio data sampled at $f_s = 44.1$kHz, we employ a Multiresolution FFT with $N = 2048$ and $L = 256$, resulting in four distinct spectrogram resolutions. While the best time resolution of 5.8 ms is obtained with the elementary transform ($M = 256$), the highest frequency resolution is achieved by the summation of all elementary transforms ($M = 2048$) and amounts to 21.5 Hz. The spectrogram with the most accurate frequency representation is used in the low frequency region, or to be exact, in the first six critical bands of the Bark scale. Accordingly, every other resolution covers five critical bands up to the maximum frequency $f_{\text{max}} = 5000$ Hz, i.e. $k_{\text{max}} = 232$.

3. SINUSOIDAL IDENTIFICATION

Since sinusoidal components of the audio signal contain the most relevant information about the melody, a sinusoids plus noise analysis is performed on the spectral data. The underlying idea of this technique is, that an audio signal can be divided into stable partials originating from periodic sound and a noise component [1]. Only partials which (probably) result from a deterministic signal are used for further melody analysis; stochastic components are neglected.

The most common criterion for detecting a sinusoid is the spectral peak. While peak picking is very robust against noise and distortion and also works in dense spectra, it produces a high number of false positives due to spurious peaks. That is why additional criteria are employed; for example the continuity of sinusoidal trajectories over time, or the concordance of the extracted sinusoidal components with harmonic patterns [1].

Since the explicit identification of continuous sinusoidal trajectories is not a precondition for our frame-wise pitch estimation method, we aim to identify sinusoidals by distinct spectral features in one frame alone.

3.1. Estimation of Instantaneous Frequency and Magnitude

There are many methods for the estimation of the instantaneous frequency (IF) and magnitude from Fourier coefficients. Keiler and Marchand compared some of the most popular ones in [6]. They found that methods which are in some way based on the phase information of the FFT give the best results regarding frequency resolution. This property is extremely important for the analysis of polyphonic music.

We apply the well-known phase vocoder method proposed by Flanagan and Golden for the IF extraction and compute the instantaneous frequency $f_i[k]$ from the phase difference $\Delta\phi[k]$ of successive phase spectra as follows [7]:

$$f_i[k] = (k + \kappa[k]) \frac{f_s}{N}, \tag{12}$$

with:

$$\kappa[k] = \frac{N}{2\pi L} \,\text{princarg}\left[\phi_l[k] - \phi_{l-1}[k] - \frac{2\pi L}{N} k\right],$$

where *princarg* is the principal argument function mapping the phase to the $\pm\pi$ range. The bin offset κ denotes the deviation of the partial's IF from the bin frequency expressed in the unit bin. If the estimated bin offset of a peak is less than $\pm 1/2$, we can say that the instantaneous frequency of the peak corresponds to the bin frequency. In order to estimate valid IF over a range of frequency bins with the phase vocoder method, the use of overlapping STFT windows (or zero-padding) is required, because otherwise the phase difference between frames might exceed 2π. The maximum bin offset which can be computed with this method is $\frac{N}{2L}$.

The instantaneous magnitude of the sinusoidal peak is estimated from the local maximum $|X[k]|$ and its bin offset $\kappa[k_{\text{peak}}]$ as follows:

$$A_{\text{peak}} = \frac{1}{2} \frac{|X[k]|}{W_{\text{Hann}}\left(\frac{N}{M}\kappa[k_{\text{peak}}]\right)}, \tag{13}$$

where W_{Hann} is the Hann window kernel:

$$W_{\text{Hann}}(\kappa) = \frac{1}{2} \frac{\text{sinc}\left(\frac{M}{N}\pi\kappa\right)}{1 - \left(\frac{M}{N}\kappa\right)^2}.$$

3.2. Bin Offset Criterion

Charpentier proposed a sinusoidality criterion for speech processing in [8], which is derived from local characteristics of the phase spectrum, or more precisely, the instantaneous frequencies of neighboring frequency bins. By using a local criterion we avoid including restrictive assumptions about the harmonic structure of the signal and leave the observation of the temporal continuity to a processing level subsequent to the pitch estimation.

Using the sinusoidality criterion suggested by Charpentier we can verify a given spectral peak[1] at bin k by the two conditions:

$$\kappa[k] \;\leq\; \frac{1}{2} \tag{14}$$

and:

$$(f_i[k] \approx f_i[k-1]) \;\wedge\; (f_i[k] \approx f_i[k+1]). \tag{15}$$

Equation (15) may be expressed in terms of the bin offset κ:

$$\kappa[k] \approx \kappa[k \pm 1] \pm 1. \tag{16}$$

For sinusoidal peaks with a stationary frequency and amplitude these conditions hold, but for the most noisy peaks they do not. Since we cannot expect that all sinusoids are ideally stationary, we allow some error in the actual implementation considering peaks with:

$$\kappa[k] \;<\; 0.7 \tag{17}$$

and:

$$|\kappa[k] - \kappa[k \pm 1] \mp 1| < 0.4 \tag{18}$$

as sinusoidal.

3.3. Weighted Bin Offset Criterion

While Charpentier's criterion works very well with monophonic audio, we face a more challenging situation within polyphonic audio, where we find a higher number of concurrent harmonics and additional noise through percussive instruments. Effectively, the phase spectrum is more distorted and the calculated IF are often not very reliable. This is especially true for Fourier coefficients with a weaker magnitude – often close to the minima of the spectral window function. That is why we have to relax the bin offset criterion further. Hence, the estimated frequency error (the difference between instantaneous frequencies of two frequency bins) is weighted according to the instantaneous magnitude A_{peak} of the sinusoid and the respective magnitude of the neighboring Fourier coefficients $|X[k \pm 1]|$:

$$|\kappa[k] - \kappa[k \pm 1] \mp 1| \;<\; 0.4\,\frac{A_{\text{peak}}}{|X[k \pm 1]|}. \tag{19}$$

Furthermore, the IF of the sinusoidal peak may deviate more from the corresponding bin frequency:

$$\kappa[k] \;<\; 0.7(r+1)\,, \tag{20}$$

where r is the MR FFT resolution parameter.

[1] Within a specific MR FFT resolution a spectral peak can only be verified if the neighboring bins are of the same resolution. Hence the different resolutions must overlap by one bin.

3.4. Masking Criterion

Unlike the before-mentioned criteria the masking criterion is a method to exclude non audible peaks – sinusoidal or not – from further processing. We use a very simplified implementation of simultaneous and temporary masking, which by far does not reach the complexity of models used in modern lossy audio coders, as for example the AAC codec [2] [9].

Simultaneous masking is a property of the human auditory system where certain maskee sounds are not audible in the presence of concurrent masker sounds. The spread masking across critical bands is very basically modeled as triangular spreading function $\text{SF}(z)$ with slopes of $+25$ and -10 dB on the normalised Bark scale:

$$\text{SF}(z) = \begin{cases} 10^{25z/20}\,, & \text{for} \quad z \leq 0 \\ 10^{-10z/20}\,, & \text{for} \quad z > 0 \end{cases} \tag{21}$$

The resulting approximate spread spectrum function $\text{SSF}[i]$ is computed with a resolution of $1/3$ Bark, so that the critical band partition index i corresponds to the Bark value z with $z = i/3$.

Temporal masking is the characteristic of the auditory system where the masker sound makes inaudible other sounds which are present immediately preceding or following the stimulus. We only take into account the much more pronounced effect of forward masking (masking that obscures a sound immediately following the masker) which is computed as given in:

$$\text{TS}[i] = 0.4\,\text{TS}_a[i] + 0.6\,\text{TS}_b[i]\,, \tag{22}$$

with:

$$\text{TS}_a[i] \leftarrow t_a\,\text{TS}_a[i] + (1-t_a)\,\text{SSF}[i] \quad \text{and}$$
$$\text{TS}_b[i] \leftarrow t_b\,\text{TS}_b[i] + (1-t_b)\,\text{SSF}[i].$$

The parameters t_a and t_b are time constants which determine the exponential growth and decay of $\text{TS}_a[i]$ and $\text{TS}_b[i]$:

$$t_a = 0.5^{\Delta t_{\text{grid}}/5ms} \quad \text{and} \quad t_b = 0.5^{\Delta t_{\text{grid}}/70ms}.$$

Although the resulting masking threshold depends on the tonality of the masker, our approach uses a constant masking threshold $\text{M}[i]$ which is 15 dB below the maximum between the values of $\text{SSF}[i]$ and $\text{TS}[i]$:

$$\text{M}[i] = \max(\text{SSF}[i], \text{TS}[i]) \cdot 10^{15/20}. \tag{23}$$

4. EVALUATION

4.1. Multiresolution FFT

Figure 1 shows a comparison of the analysis results obtained with either the MR FFT or the FFT with constant frequency resolution. Spectrogram (a) shows the FFT together with simple peak picking. The high number of spurious peaks is obvious. The number of peaks is gradually reduced in spectrogram (b), which illustrates the spectral peaks obtained by the MR FFT analysis. The decreasing number of peaks for the high frequency regions is due to a masking effect, which can be explained by the wider main lobe of the spectral window function with decreasing frequency resolution. As a consequence we will of course loose selectivity in the

[2] The proposed masking model is optimized for sinusoidal peak identification and efficiency. It is not useful for encoding audio.

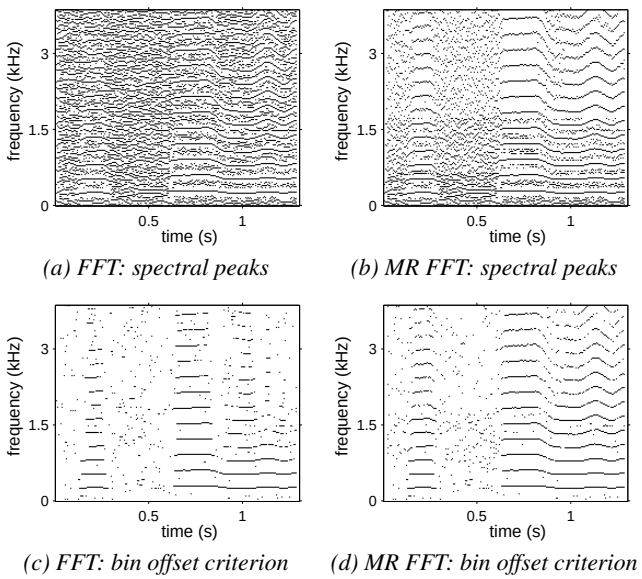

Figure 1: *Comparison of spectral analysis using either FFT or MR FFT.*

higher frequency regions and concurrent signal components may not be resolved. Besides the fact that the human auditory system also has a limited, frequency dependent, resolution, the MR FFT offers additional advantages.

This is especially true for the instantaneous frequency estimation. If the signal's frequency is changing in time an accurate measurement of frequency should be as local as possible [10]. This implies that the analysis window size should be as local as possible. There is always a trade-off between this claim and the wish to discriminate concurrent signal components. In polyphonic music, we find a mixture of voices in the low and middle frequencies, while the harmonics of the leading voice dominate the higher spectral bands [4]. Thus a good frequency resolution is required mostly in the low frequency regions, where the harmonics exhibit a quasi stationary frequency compared with the FFT filter bandwidth. With increasing harmonic number frequency modulation of the partials becomes more evident, so for higher harmonics the stationarity criterion is often violated. The MR FFT analysis offers the possibility to adapt the frequency resolution accordingly and considerably improves the IF estimation for the higher frequency regions.

Another advantage of the MR FFT lies in the improvement of the sinusoidal detection. Whenever a sinusoid is not stationary within one FFT frame, the corresponding spectral peak becomes distorted. Since both the Charpentier bin offset criterion and the weighted bin offset criterion require a more or less undistorted phase spectrum, such sinusoidal peaks will not be identified. This effect may be observed in spectrogram (c), which shows the application of the Charpentier bin offset criterion on the ordinary FFT spectrum. Indeed most of the unwanted spectral peaks disappear in the lower frequency regions, but at the same time we observe the deletion of high harmonics with a rapidly changing frequency. Finally spectrogram (d) shows the sinusoidal identification based on the MR FFT analysis, where high harmonics are correctly identified as deterministic partials. We see that the MR FFT accounts for a significant improvement of the sinusoidal detection.

4.2. Sinusoidal Identification

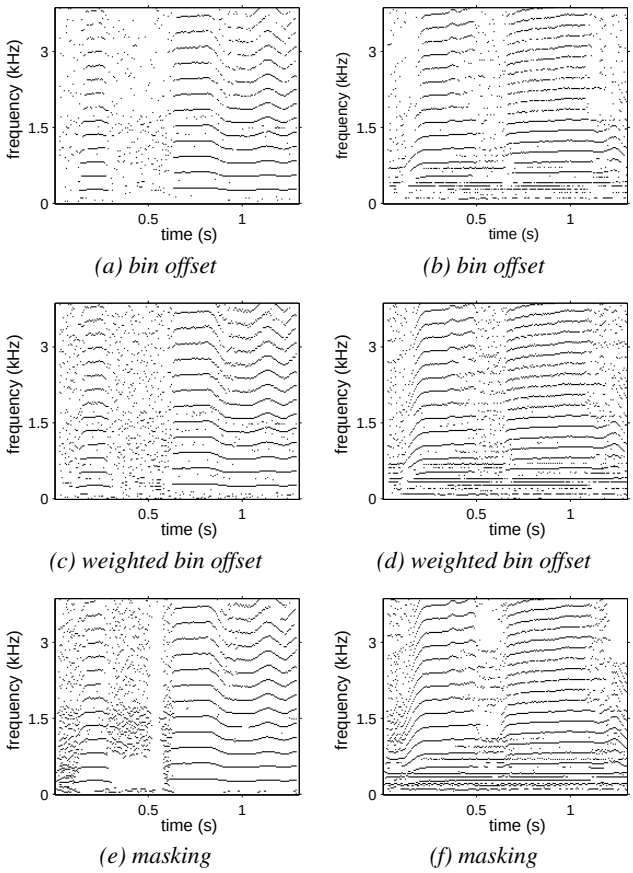

Figure 2: *Analysis results for the distinct sinusoidality criteria using the MR FFT. Monophonic audio: figures (a), (c) and (e). Polyphonic audio: figures (b), (d) and (f).*

Figure 2 allows us to compare qualitatively the analysis results of the proposed criteria. All examples have been computed using the MR FFT. The images on the left hand side show the results for monophonic audio (the sung word 'history'). We see that all methods perform well on the monophonic example. However, the bin offset criterion inspired by Charpentier's method causes the lowest number of false positives (peaks which have been erroneously identified as sinusoidal). At the same time it maintains most of the valid sinusoidal peaks. Clearly, this criterion is very suitable for monophonic analysis – for example speech processing.

Yet, the task of melody extraction naturally implies the analysis of polyphonic audio. The sinusoidal detection should not only be quiet robust against noise and distortion, it should also give adequate results for dense spectra consisting of many simultaneous sounds, which for example can be found in many pieces of rock music.

We have chosen a short excerpt from the file pop1.wav , which is included in the ISMIR2004 melody contest test set[3], to illustrate

[3]The ISMIR 2004 melody test set with reference transcriptions is avail-

Voicing Detection	Voicing False Pos.	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
81.8%	17.3%	68.1%	71.4%	**71.4**%

Table 1: *MIREX 2005 evaluation results.*

the performance of the distinct sinusoidality criteria with polyphonic music. This piece of music exhibits a rather moderate degree of polyphony, featuring a predominant singing voice and a light accompaniment without percussion. Nevertheless, it reveals the main drawback of the two phase dependent criteria: if two sinusoidals are very close to each other in frequency, the frequency responses interfere and the sinusoidal peak will not be identified due to a distorted phase. Even if the frequency estimate of the actual peak remains reliable, the estimated bin frequencies (or bin offsets) of the neighboring bins are often more affected. This is due to their weaker magnitude and of course they may be closer to the disturbing sinusoidal.

This effect can clearly be noted in spectrogram (b), where the harmonics in the lower frequency regions almost disappear. The weighted bin offset criterion allows a bigger offset error for weaker magnitudes – as we can see in spectrogram (d) the sinusoidal detection is improved noticeable. This criterion was utilized in our submission to the MIREX2005 Melody Contest (see section 4.3). It proofed suitable for the extraction of the predominant voice from musical audio with a moderate accompaniment, since the melody line still can be reconstructed from the higher harmonics. However, the bass line as well as other accompanying instruments often cannot be extracted reliably, even if they can be easily discriminated by human listeners.

Finally spectrogram (f) displays the most general solution to sinusoidal identification, which is based on psychoacoustic masking principles. In comparison with the bin offset criteria the masking criterion produces a higher number of false positives. However, this criterion guaranties a robust identification of all kinds of music and nonetheless significantly reduces the number of spectral peaks.

4.3. MIREX 2005 Audio Melody Extraction Contest

The aim of the MIREX Audio Melody Contest is to evaluate different approaches to extracting the main melody from polyphonic audio[4]. The MIREX 2005 dataset contains 25 phrase excerpts of 10-40 seconds length from different genres.

Our submission to the MIREX Audio Melody Contest used the MR FFT analysis as described in section 2 together with the weighted bin offset criterion introduced in section 3.3. Reaching 71.4% our algorithm has performed best on Overall Accuracy with a significant difference to other submissions. Of course the result should not be attributed to the spectral analysis front-end alone, but at least we can say that an FFT-based analysis does not contradict with a good melody extraction performance.

During the MIREX evaluation the execution time for the entire melody analysis has been measured as approximately 20 times faster than real-time on an Intel® Pentium® 4 3.0 GHz CPU system with 3 GB RAM – the fastest runtime among all ten submissions. Despite this encouraging result, we want to emphasise that the MIREX evaluation did not attach importance to the execution time, and the measures should be recognised as rough indicators for algorithm efficiency, if at all.

able at http://www.iua.upf.es/mtg/ismir2004/contest/melodyContest/FullSet.zip

[4] A detailed description of the MIREX 2005 evaluation procedure and the results can be found online at http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html

5. SUMMARY

The MR FFT has proved to be useful for the efficient analysis of polyphonic audio – especially, if good frequency resolution as well as good time resolution is needed. The multi-resolution approach significantly improves the IF estimation and the detection of sinusoidal components by simple local sinusoidality criteria. As an alternative to the phase dependent criteria we proposed a masking criterion which is based on a very simple psychoacoustic model. This criterion has proved robust in all kinds of polyphonic music.

The presented spectral analysis front-end successfully deals with the the dynamics of the singing voice or a lead instrument, but of course it can be applied to various problems in audio analysis.

6. REFERENCES

[1] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, USA, 1989.

[2] A. Röbel, M. Zivanovic, and X. Rodet, "Signal decomposition by means of classification of spectral peaks," in *Proc. Int. Comp. Music Conf. (ICMC'04),* Miami, USA, 2004, pp. 446–449.

[3] P. Masri and A. Bateman, "Identification of nonstationary audio signals using the FFT,with application to analysis-based synthesis of sound," in *Proc. IEE Colloquium Audio Eng., Digest No. 1995/96*, 1995, pp. 11/1–11/6.

[4] 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*, vol. 43, no. 4, pp. 311–329, 2004.

[5] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-29, no. 1, pp. 84–91, 1981.

[6] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. Int. Conf. on Digital Audio Effects (DAFx-02),* Hamburg, Germany, 2002, pp. 51–58.

[7] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[8] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'86),* Tokyo, Japan, 1986, pp. 113–116.

[9] ANSI, *Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)*, INCITS/ISO/IEC 13818-7, 2003.

[10] M. S. Puckette and J. C. Brown, "Accuracy of frequency estimates using the phase vocoder," *IEEE Trans. Speech and Audio Proc.*, vol. 6, no. 2, pp. 166–176, 1989.