Frequency based criterion for distinguishing tonal and noisy spectral components

Maciej Kulesza m_kulesza@sound.eti.pg.pl
Multimedia Systems Department
Gdansk University of Technology
Gdansk, 80-233, Poland

Andrzej Czyzewski ac@sound.eti.pg.gda.pl
Multimedia Systems Department
Gdansk University of Technology
Gdansk, 80-233, Poland

Abstract

A frequency-based criterion for distinguishing tonal and noisy spectral components is proposed. For considered spectral local maximum two instantaneous frequency estimates are determined and the difference between them is used in order to verify whether component is noisy or tonal. Since one of the estimators was invented specially for this application its properties are deeply examined. The proposed criterion is applied to the stationary and nonstationary sinusoids in order to examine its efficiency.

Keywords: tonal components detection, psychoacoustic modeling, sinusoidal modeling, instantaneous frequency estimation.

1. INTRODUCTION

The algorithm responsible for distinguishing tonal from noisy spectral components is commonly used in many applications such a speech and perceptual audio coding, sound synthesis, extraction of audio metadata and others [1-9]. Since the tonal components present in a signal are usually of higher power than noise, the basic criterion for distinguishing tonal from noisy components is based on the comparison of the magnitudes of spectrum bins. Some heuristic rules may be applied to the local spectra maxima in order to determine whether they are noisy or tonal [1]. The other method relies on the calculation of terms expressing peakiness of these local maxima as it was proposed in [10] or level of similarity of a part of spectrum to the Fourier transform of stationary sinusoid, called sinusoidal likeness measure (SLM) [11]. In contrary to the magnitude-based criterions applied to the local spectra maxima, the ratio of geometric to arithmetic mean (spectral flatness measure – SFM) of magnitudes of spectrum bins may be used for tonality estimation of entire signal or for set of predefined bands [4, 5]. Instead of analysis of magnitude spectrum, it is also possible to extract the information related to the tonality of spectral components through comparison of the phase values coming from neighbouring bins as it was proposed in [12]. The method used in MPEG psychoacoustic model 2 employs linear prediction of phase and magnitude of spectrum bins. The tonality measure is then expressed as the difference between predicted values and the ones detected within particular time frame spectrum [1, 13-15]. Also various techniques for separation of periodic components within speech signal and signals composed of two pitched sounds were successfully investigated [3, 16-18].

Recently, it was proved that the tonality of spectral components within polyphonic recordings may be expressed as an absolute frequency difference between instantaneous frequencies of the local spectrum maxima calculated employing two different estimators [19-21]. While the first frequency estimator employs well known technique of polynomial fitting to the spectrum maximum and its two neighbouring bins, the second estimator is hybrid. It involves estimation results yielded by the first mentioned estimator and phase values coming down from three contiguous spectra. This algorithm was successfully combined with psychoacoustic model used in audio coding applications [13]. It was proved that this method allows detecting tonal spectra components even if they instantaneous frequency changes significantly over time. This property of the method is its main advantage over the tonality estimation algorithms commonly used in various applications. Although the efficiency of the mentioned algorithm has been already evaluated using artificial signals and polyphonic recordings, no investigation related to the hybrid frequency estimator and the tonality criterion being the basis for this method has been made. In this article we will focus on the experiments revealing properties of the hybrid frequency estimator and the properties of the tonality criterion employing it. The influence of the analysis parameters as well as influence of the analyzed signal characteristics on tonality estimation efficiency is investigated and deeply discussed. The properties of the hybrid estimator are compared to the properties of the estimator employing polynomial fitting to the spectral bins.

2. CONCEPT DESCRIPTION

For clarity of description, it is assumed here that the analyzed signal contains a single tonal component of constant or modulated instantaneous frequency and variable signal to noise ratio (SNR). A general diagram of the method used in order to examine the proprieties of proposed tonality criterion is shown in Fig.1.

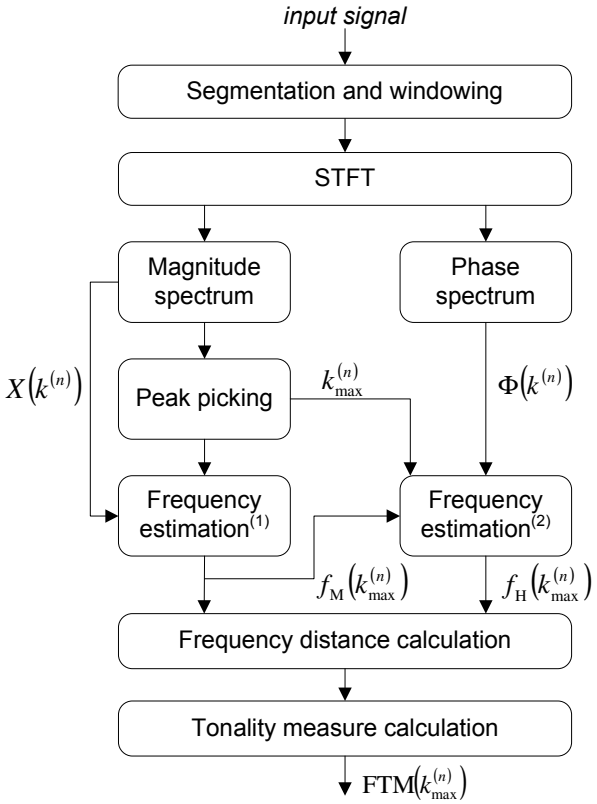

FIGURE 1: General diagram of investigated method for tonality measuring

The input signal is segmented into frames of equal length weighted by the von Hann window in conformity to the short time Fourier transform (STFT) concept [22]. Both the frame length and hop size are the parameters of the method. Moreover, the windowed frame of the signal is zero-padded before applying the FFT. Further, the magnitude and phase spectra denoted as $X\left(k^{(n)}\right)$ and $\Phi\left(k^{(n)}\right)$ are calculated, and spectral bin of highest magnitude is considered to be a candidate for the tonal component. The instantaneous frequency corresponding to the detected spectral component of highest energy (spectrum bin of $k_{\max}^{(n)}$ index) is then estimated using two methods. While the first one employs fitting of polynomial (binomial) to the detected component and its two adjacent bins within magnitude spectrum, the second one is based on the phase and magnitude–spectrum processing [23]. The results of frequency estimation obtained using two above-mentioned methods are denoted in Fig. 1 as $f_{\mathrm{M}}\left(k_{\max}^{(n)}\right)$ and $f_{\mathrm{H}}\left(k_{\max}^{(n)}\right)$. Finally, the absolute frequency difference is calculated and the level of tonality for selected component is assigned to it as a result of normalization of the yielded frequency distance (absolute frequency difference) by the assumed frequency distance threshold. The tonality measure calculated in accordance to the proposed scheme is called frequency-derived tonality measure (FTM).

2.1 Frequency estimator based on magnitude spectrum analysis

Assuming that the local maximum of magnitude spectrum being analyzed corresponds to the tonal component, the straightforward method for its instantaneous frequency estimation employs quadratic interpolation (known as QIFFT) which belongs to the approximate maximum likelihood (ML) estimators [24, 25]. In this approach the magnitude spectrum values of local maximum and two neighboring bins are involved in frequency estimator. The procedure is applied to the log spectrum values as it provides higher precision of frequency estimation in most cases [23, 26]. At the begining the fractional part of spectrum index is determined according to [27]

$$k_{\mathrm{frac}}^{(n)} = \frac{1}{2}\frac{X\left(k_{\max}^{(n)}-1\right)-X\left(k_{\max}^{(n)}+1\right)}{X\left(k_{\max}^{(n)}-1\right)-2X\left(k_{\max}^{(n)}\right)+X\left(k_{\max}^{(n)}+1\right)} \tag{1}$$

where $k_{\max}^{(n)}$ stands for the index of considered spectrum bin (the notation of spectrum bin indices is extended by the time index (number of frame) as superscript), $X\left(k_{\max}^{(n)}\right)$ represents the magnitude spectrum in log scale. The frequency of the spectrum peak detected in the *n*-th frame of signal is then estimated as follows

$$f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) = \frac{k_{\max}^{(n)}+k_{\mathrm{frac}}^{(n)}}{N_{\mathrm{FFT}}} f_{\mathrm{s}} \tag{2}$$

where N_{FFT} is the length of FFT transform and f_{s} is the sampling rate in Sa/s (samples per second) and M in subscript indicates that the instantaneous frequency is estimated basing on magnitude spectrum processing. Since the signal frame is zero-padded before applying the FFT, the zero-padding factor is expressed as

$$Z_{\mathrm{p}} = \frac{N_{\mathrm{FFT}}}{N} \geq 1 \tag{3}$$

where *N* stands for the length of signal frame. The motivation for zero-padding of the signal frame before FFT calculation is the reduction of estimator bias resulting in an improved accuracy of frequency estimation. Basing on experimental results presented in [23], the maximum frequency bias of the QIFFT assuming the von Hann window is up-bounded in the following way

$$f_{\mathrm{Mbias}} \leq \frac{f_s}{N}\left(\frac{1}{4Z_p}\right)^3 \tag{4}$$

For zero-padding factor equal to 2 and frame length equivalent to 32 ms (for instance: f_s=32 kSa/s, N=1024) the bias of considered frequency estimator calculated according to (4) is less than 0.07 Hz. Using zero-padding factor higher than 2 seems to be impractical as it would result in significant increase of the computational complexity, assuring only slight increase of the frequency estimation accuracy. Thus, in investigated method for tonality measuring every frame of the input signal is zero-padded to its doubled length.

2.2 Hybrid frequency estimator

The second estimator suitable for combining with proposed method for tonal components detection and tonality estimation is required to:

- yield inadequate instantaneous frequency values when the spectrum bins involved into the estimator procedure do not correspond to the tonal components (the frequency distance between values obtained using quadratic interpolation and phase-based method should be abnormally high – i.e. higher than half of the frequency resolution of spectral analysis)

- allow of accurate instantaneous frequency estimation of frequency modulated tonal components

Various phase-based instantaneous frequency estimators have been proposed so far [28-32]. Assuming the STFT approach to the signal analysis, one of the straightforward methods for frequency estimation is based on an approach proposed in [28] where instantaneous frequency is computed basing on the phase difference between two successive frame short-term spectra. The hop size H equal to one sample is assumed in this method in order to allow for estimation of instantaneous frequency in full Nyquist band [32]. However, even if the analyzed spectrum maximum corresponds to the component totally noisy, the classic phase-difference estimator (assuming H=1) yields adequate instantaneous frequency estimates because the estimation error is lower than the frequency resolution of spectral analysis. Consequently, the first above-defined requirement for frequency estimator is not met. In order to overcome this problem, the higher hop size of STFT analysis should be used. When the higher hop size is chosen, the phase difference for particular frequency bin can be higher than 2π. In this case, the adequate phase increment cannot be calculated from the phase spectrum, as its values are bounded to ±π and then the phase difference never exceeds 2π. This causes the phase indetermination problem obstructing the instantaneous frequency estimation using classical phase-based method [22, 28, 32, 33]. Furthermore, when the higher hop size is selected the frequency of tonal component may be not longer constant in two successive steps of analysis or even the indices of spectral maxima corresponding to the same tonal component may be different $\left(k_{\max}^{(n)} \neq k_{\max}^{(n-1)}\right)$. Since the instantaneous frequency cannot be accurately determined in this case, the second requirement defined on the beginning of this subsection is not satisfied. Thus, the classical phase-difference estimator was not considered for employing it as an element in our method for tonal components detection. Although some phase-based methods for frequency estimation of nonstationary tonal components were already proposed in [30, 33], the proposed tonality estimation method is based on the dedicated estimator fulfilling the above-defined requirements and optimized for application considered here.

The instantaneous frequency determined by the hybrid estimator is defined as follows

$$f_{\mathrm{H}}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right) + \Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right) \tag{5}$$

where: $f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right)$ is the instantaneous frequency of the spectrum maximum detected within *n*–2 analysis frame using estimator defined in Eq. (2), and $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ is the frequency jump between spectral maxima detected within *n*–2 and next *n* analysis frames estimated using phase-based method.

2.2.1 Phase-based frequency jump estimator

In the investigated method the frequency jump $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ is calculated basing on the phase values detected within three successive spectra. It is assumed that the phase values $\Phi\left(k_{\max}^{(n-2)}\right)$, $\Phi\left(k_{\max}^{(n-1)}\right)$ and $\Phi\left(k_{\max}^{(n)}\right)$ correspond to the same sinusoidal component detected within three contiguous spectra. The second order phase difference is then calculated according to [19]

$$\Delta^2\Phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right) = \Phi\left(k_{\max}^{(n-2)}\right) - 2\Phi\left(k_{\max}^{(n-1)}\right) + \Phi\left(k_{\max}^{(n)}\right) \tag{6}$$

The phase offset which is non-zero in case of frequency modulated tonal components is given by

$$\Delta^2\phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right) = \frac{\pi(N-1)}{Z_{\mathrm{p}}N}\left(k_{\max}^{(n-2)} - 2k_{\max}^{(n-1)} + k_{\max}^{(n)}\right) \tag{7}$$

Finally, the frequency jump can be estimated using following formula

$$\Delta f_{\Phi}\left(k_{\max}^{(n)}\right) = \frac{f_s}{\pi H}\left(\mathrm{princarg}\left(\Delta^2\Phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right)\right) + \Delta^2\phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right)\right) \tag{8}$$

where $\mathrm{princarg}(\varphi) = (\varphi + \pi)\mathrm{mod}(-2\pi) + \pi$ is the function mapping the input phase φ into the ±π range [22]. Further the $\Delta f_{\Phi}\left(k_{\max}^{(n)}\right)$ is updated in order to overcome phase ambiguity problem [19]

$$\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right) = \Delta f_{\Phi}\left(k_{\max}^{(n)}\right) + m\frac{f_{\mathrm{s}}}{H} \tag{9}$$

where *m* is the integer value ensuring that the $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ falls within the maximal and minimal frequency jump range related to the $k_{\max}^{(n)} - k_{\max}^{(n-2)}$ difference [19].

2.3 Tonality measurements

The proposed criterion for distinguishing tonal from noisy components and their tonality measuring is based on the absolute difference between the frequency estimates obtained using the QIFFT method and the hybrid method described in previous subsection. Thus, the frequency distance for particular spectral maximum is given by

$$f_{\Delta}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) - f_{\mathrm{H}}\left(k_{\max}^{(n)}\right) \tag{10}$$

When we combine the estimate (5) with the definition (10) the frequency distance may be expressed by

$$f_{\Delta}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) - f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right) - \Delta f_{\Phi}\left(k_{\max}^{(n)}\right) \tag{11}$$

It is viewable that $f_{\Delta}\left(k_{\max}^{(n)}\right)$ is equal to the difference between frequency jumps derived from the magnitude spectrum analysis and from the phase spectrum analysis, respectively [19]. Let us define a measure based on the frequency distance $f_{\Delta}\left(k_{\max}^{(n)}\right)$ expressing the level of similarity of particular spectrum component to the pure sinusoid

$$\mathrm{FTM}\left(k_{\max}^{(n)}\right)=1-\frac{\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|}{\left|f_{\Delta}\right|_{\mathrm{thd}}} \tag{12}$$

where $\left|f_{\Delta}\right|_{\mathrm{thd}}$ is a frequency distance threshold which is assumed not to be exceeded when the $k_{\max}^{(n)}$ is a tonal spectral component. Tonality measure $\mathrm{FTM}\left(k_{\max}^{(n)}\right)$ is equal to 1 if spectral component considered corresponds to the sinusoid of high SNR and tends to gradually decrease when SNR falls. If $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|\geq\left|f_{\Delta}\right|_{\mathrm{thd}}$ for a particular spectral component, it is treated as a noisy one, and the tonality measure $\mathrm{FTM}\left(k_{\max}^{(n)}\right)$ equal to 0 is assigned to it. The experiments related to the properties of hybrid frequency estimator proposed here together with the criterion for tonal components detection as well as some remarks concerning selection of $\left|f_{\Delta}\right|_{\mathrm{thd}}$ threshold are presented in the following section.

3. EXPERIMENTS

3.1 The performance evaluation of instantaneous frequency estimators

In order to examine the properties of the proposed hybrid estimator, a set of real valued sinusoids with randomly chosen initial phases φ_0 and SNR ranging from 100 dB to -20 dB with 2 dB step were generated. It was assumed that the amplitude of sinusoid is equal to 1 and the power of noise is adjusted in order to achieve a desired SNR in dB according to the formula

$$\mathrm{SNR}[\mathrm{dB}]=10\log_{10}\frac{\sum_{s=1}^{L}x_{\mathrm{t}}^{2}[s]}{\sum_{s=1}^{L}x_{\mathrm{ns}}{}^{2}[s]} \tag{13}$$

where $x_{\mathrm{t}}[s]=a\cos(2\pi\omega s+\varphi_0)$, $\omega=f/f_{\mathrm{s}}$ is the normalized frequency in cycles per sample, $x_{\mathrm{ns}}[s]$ stands for a additive white Gaussian noise (AWGN) realization, *s* is sample number and *L* is the signal length.

For every selected SNR the sinusoids of constant normalized frequencies selected within range from 0.05 to 0.45 with 0.005 step (81 sinusoids) were generated and further analyzed resulting in vector of instantaneous frequency estimates related to the particular SNR. Then, the mean squared error (MSE) of estimates (2) and (5) was calculated basing on frequency estimation results and known apriori frequencies of generated sinusoids. Since this procedure was applied to sets of sinusoids of various SNR, the characteristic revealing frequency estimation errors versus SNR of analyzed sinusoids was obtained. The experiments were carried out for both considered estimators – the hybrid method and the QIFFT method, and the results were compared with lower Cramer-Rao bound (CRB) defining variance of unbiased frequency estimator of real sinusoid in a AWGN [25, 32]

$$\mathrm{var}(\hat{\omega})\geq\frac{12}{(2\pi)^2a^2N\left(N^2-1\right)}10^{-SNR/10} \tag{14}$$

where $\hat{\omega}$ is the normalized estimated frequency in cycles per sample, N is the same as in (3) and a=1 in our experiments.

The sampling rate of analyzed signals was adjusted to 8000 Sa/s, the frame length (von Hann window) was equal to 32 ms (N=256) and the hop size was switched between 32 ms (H=256) and 8 ms (N=64). The characteristics obtained for two above-defined hop sizes of analysis are presented in Fig. 2.

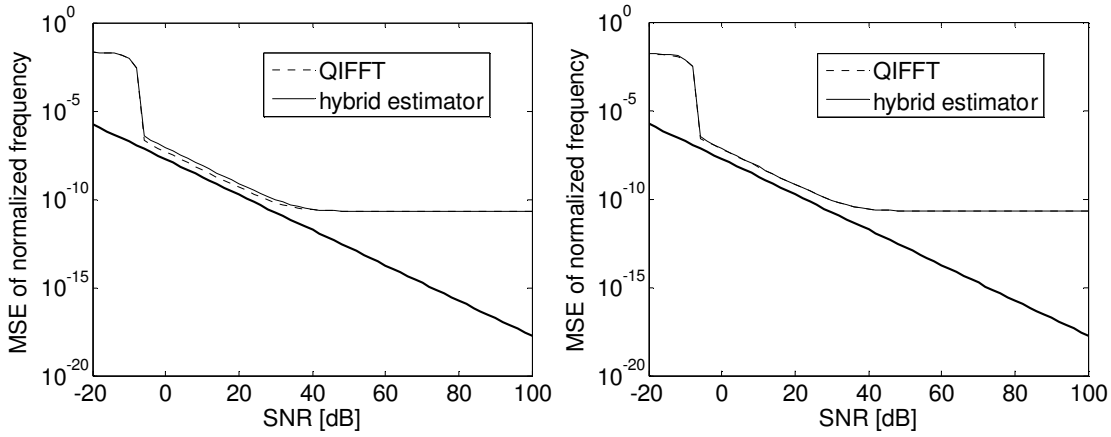

FIGURE 2: Performance of estimators for frequencies in (0.05, 0.45) normalized range for hop size equal to frame length (left), and quarter of frame length (right); Cramer-Rao bound – bold solid line

Since the spectrum bin of maximum energy is considered here to represent sinusoidal component, for lower SNRs the spurious noise peak may be detected instead of it. Thus, the frequency estimation MSEs presented in Fig. 2 are far beyond the CRB when the SNRs of sinusoids are lower than approximately -10 dB [23, 25]. Contrarily, in the SNR range from -10 dB to approximately 30 dB the MSE determined for examined estimators is close to the CRB. Although the curve representing results obtained using the QIFFT is approximately 4 dB above CRB regardless the hop size of analysis, the error of hybrid estimator tends to be slightly higher when the hop size is maximal possible. For SNRs higher than 40 dB the frequency estimation error reveals the bias of concerned estimators, which is related to the assumed zero-padding factor [23, 26].

The influence of the hop size on the estimation error in case of stationary sinusoid of SNR equal to 20 dB and 100 dB and normalized frequency equal to 0.13 is presented in Fig. 3 (sampling rate is the same as in previous experiment). It can be observed from Fig. 3 that the MSE of hybrid estimator is practically identical to the MSE obtained using the QIFFT regardless the hop size of analysis when the SNR is equal to 100 dB (compare results presented in Fig. 3 for the same SNR=100 dB). However, when the SNR is equal to 20 dB, the hybrid estimator performs slightly worse, by approximately 3 dB, than the QIFFT for hop sizes higher than a half of the frame length. For lower hop sizes the difference in performance of both estimators gradually decreases. It can be expected that for shorter hop sizes, the frequency change $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ derived from phase analysis according to (9) tends to have lower influence on the final estimation results. Thus, the shorter the hop size the properties of hybrid estimator are closer to the properties of the QIFFT method. This is not the case when the SNR is equal to -3 dB or lower, because the MSE of hybrid method tends to increase for the hop sizes below approximately a quarter of the frame length and higher than 220 samples. In this hop size range the hybrid method yields occasionally inadequate estimates when the SNR is low resulting in the MSE increase. Therefore, it can be deduced that the hybrid estimator operates most efficiently in the range of the hop size between approximately ¼ to ¾ of the frame length.

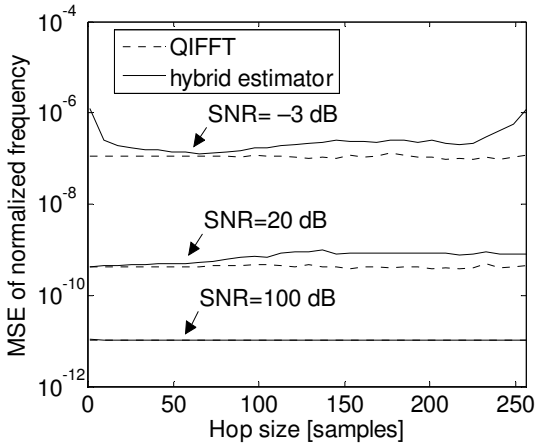

FIGURE 3: Impact of the hop size of analysis on the frequency estimation performance

Further, the MSE of frequency estimation results were determined for sinusoids of constant SNR equal to 100 dB and for normalized frequencies selected within 0.0025 and 0.4975 range with 0.001 step (496 sinusoids, f_s=8000 Sa/s, N=256, H=256). The results of our experiments presented in Fig. 4 indicate that the estimation errors for both considered methods are below 10^{-10} (see also Fig. 2) in almost entire bandwidth. However, when the normalized frequency of considered sinusoid is below approximately 0.005 or higher than 0.495 then the estimation error significantly increases.

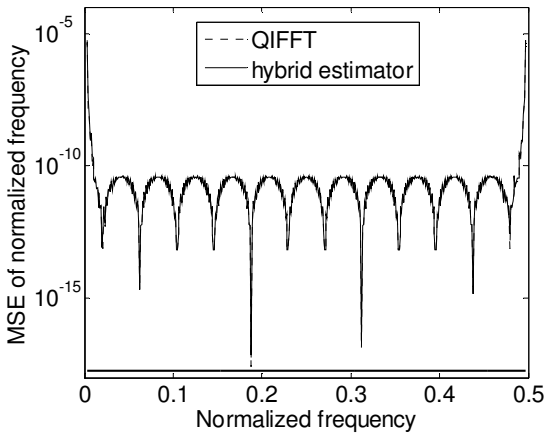

FIGURE 4: Performance of estimators for stationary sinusoids of SNR=100 dB and normalized frequencies selected within 0.0025 and 0.4975 range; Cramer-Rao bound – bold solid line.

Although the characteristics shown in Fig. 4 were determined assuming hop size equal to half of the frame length, they do not alter for other hop sizes. This is expected when considering the MSE obtained for stationary sinusoids of SNR equal to 100 dB presented in Fig. 3.

Since it is assumed that the proposed hybrid estimator should allow estimation of instantaneous frequency of non-stationary tonal components (see subsection 2.2), the set of linearly frequency modulated (LFM) chirps were analysed next [34]. The parameters of the STFT analysis as well as the sampling rate were identical to those used in the previous experiment described in this subsection. The initial normalized frequency of every LFM chirp signal was set to 0.05 and the frequency rates were altered from 0 to f_s/2 per second. The instantaneous frequencies estimated using the QIFFT and hybrid methods were compared with mean frequency values of LFM chirp calculated within a particular frame resulting in the MSE corresponding to the chirps of various

instantaneous frequency slopes. The experiments were carried out for LFM chirps of SNR equal to 100 dB and 20 dB. Although the limitations of the hybrid estimator when the hop size is equal to the frame length have been already revealed (see Fig. 3), in the experiments the hop size of analysis was chosen to be equal to frame length and a quarter of it for comparison purposes. In Fig. 5 the characteristics obtained are shown.

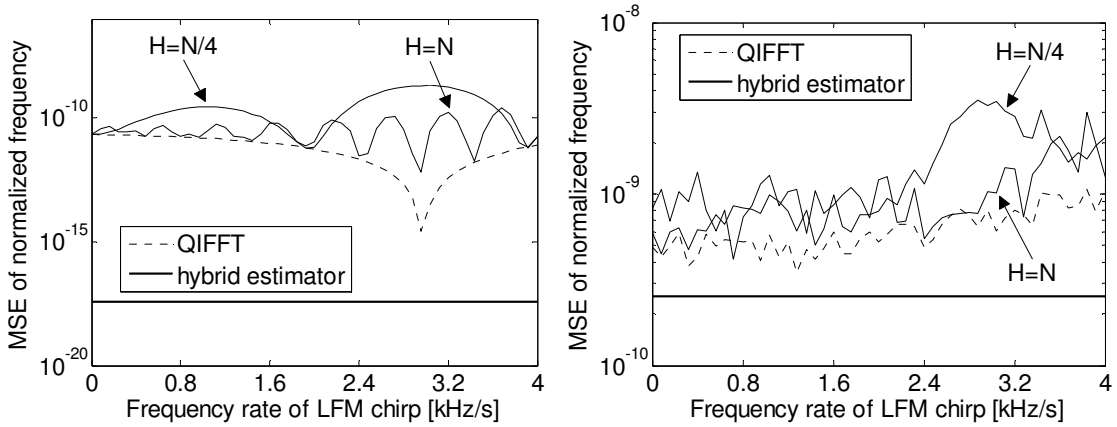

FIGURE 5: Performance of estimators for LFM chirps of various frequency rates and SNR equal to 100 dB (left) and 20 dB (right); Cramer-Rao bound – bold solid line.

When the hop size is equal to the quarter of the frame length the estimation error is higher for some chirps slopes (0.3-0.45) than the errors obtained with hop size equal to the frame length which is especially noticeable when considering characteristics obtained for signals of SNR equal to 100 dB. Furthermore, when SNR is equal to 20 dB (Fig. 5 - right), the errors corresponding to both estimation procedures are still close to the Cramer-Rao bound regardless the linear frequency modulation of analysed sinusoids.

Although the above experiments have confirmed that proposed hybrid estimator operates properly in case of sinusoids of linearly changing frequencies, its properties were also examined in case of non-linearly frequency modulated sinusoids. Thus, the frequency of carrier sinusoid equal to 0.13 (1040 Hz assuming f_s=8000 Sa/a) was modulated using sinusoid of normalized frequency equal to 2.5×10^{-4} (2 Hz). The modulation depth was altered so that the normalized frequency deviation of the carrier was changed between 0 and 0.025 (±200 Hz). Similarly to the experiments with LFM chirps the MSE of frequency estimates were determined for all generated sinusoids of SNR equal to 100 dB and 20 dB. The frame length was adjusted to 32 ms (N=256) and the hop size was switched between 32 ms and 8 ms (H=256, H=64). The results of those experiments are depicted by the curves shown in Fig. 6.

It can be noticed from Fig. 6 that the accuracy of frequency estimation is directly related to the depth of non-linear frequency modulation. The modulation depth seems to have less influence on the MSE for signals of lower SNRs, which is visible when comparing results obtained for sinusoids having the SNR of 100 dB and 20 dB. Additionally, when the framing hop size is short enough the performance of the QIFFT and hybrid estimators tends to be similar to each other.

It was suggested in subsection 2.2 that the desired property of estimator for application considered would be yielding inadequate frequency estimates when spectrum bins used in estimator do not correspond to sinusoidal component. In order to evaluate this property of proposed hybrid estimator, the white noise realization was analysed and in every spectrum the local maximum $k_{\max}^{(n)}$ laying closest to 800 Hz was selected.

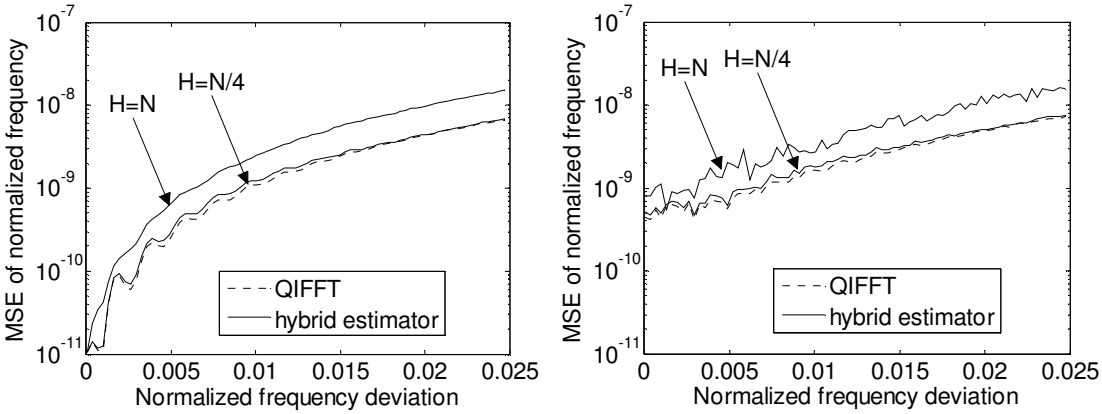

FIGURE 6: Performance of estimators for sinusoids of sinusoidally modulated frequencies of SNR equal to 100 dB (left) and 20 dB (right).

The QIFFT was applied to those peaks as well as the hybrid estimation method was used. Next, the frequencies estimated using these two methods were compared with frequency corresponding to detected local maximum

$$f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)=\frac{k_{\max}^{(n)}}{N_{\mathrm{FFT}}}f_{\mathrm{s}} \tag{15}$$

The absolute frequency differences $\left|f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)-f_{\mathrm{H}}\left(k_{\max}^{(n)}\right)\right|$ and $\left|f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)-f_{\mathrm{M}}\left(k_{\max}^{(n1)}\right)\right|$ calculated for estimation results obtained in every frame of white noise realization (f_s=8000 Sa/s, frame length and hop size equal to 32 ms (N=256, H=256), signal length equal to 2 s) are presented in Fig. 7.

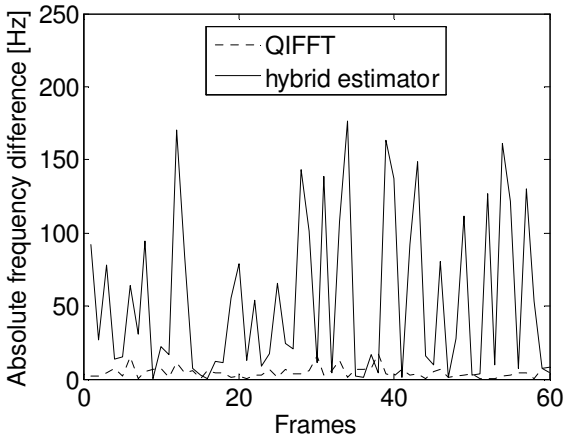

FIGURE 7: Absolute frequency differences between frequency of spectrum bin calculated according to Eq. (20) and estimates obtained using the QIFFT and hybrid method (noisy spectral peaks)

The maximum difference between frequency of spectrum local maximum defined by Eq. (15) and obtained using the QIFFT estimator is bounded to a half of the apparent frequency resolution of spectral analysis. Therefore, the curve depicting results yielded by the QIFFT estimator presented in Fig. 7 never exceeds $f_s/(2N_{\mathrm{FFT}})$=8000/512=15.625 Hz. Contrary to the QIFFT, the instantaneous frequency estimates yielded by the hybrid method are usually totally inadequate and are not bounded to the half of the apparent frequency resolution of spectral analysis. It can be concluded than that proposed hybrid estimator satisfies both requirements defined on the beginning of subsection 2.2, because it allows for frequency estimation of the modulated tonal

components and provides totally inadequate results when the selected spectral maxima do not correspond to the tonal components. Although additional experiments may be carried out in order to examine the properties of proposed hybrid estimator more deeply (i.e. estimation accuracy in case of complex sinusoids, influence of frame length and segmentation window type used, etc.), we have focused here only on the verification of those properties which are of primary importance for considered application.

3.2 Tonality measurements

In order to verify the concept of employing two different instantaneous frequency estimators for tonality measuring, a signal having a frequency modulated sinusoidal component of varying SNR was considered. As the spectrum bin of highest energy may not represent the tonal component when the SNR is very low (see Fig. 2), in our experiments the lowest SNR was adjusted to -3 dB. In the experiment the analysis was applied to the signal sampled at 8000 Sa/s rate, consisting of 24000 samples. The SNR of the sinusoidal components was constant in every segment of 8000 samples and equal to 30 dB, 3 dB and -3 dB, respectively. The instantaneous frequencies of tonal component were estimated within 32 ms frames of the signal (N=256) and the hop size was adjusted to 16 ms (H=128). The spectrogram of analyzed signal together with a curve representing the true pitch of sinusoid, the results of instantaneous frequency estimation employing two considered estimators and the frequency distance calculated according to (10) are presented in Fig. 8.

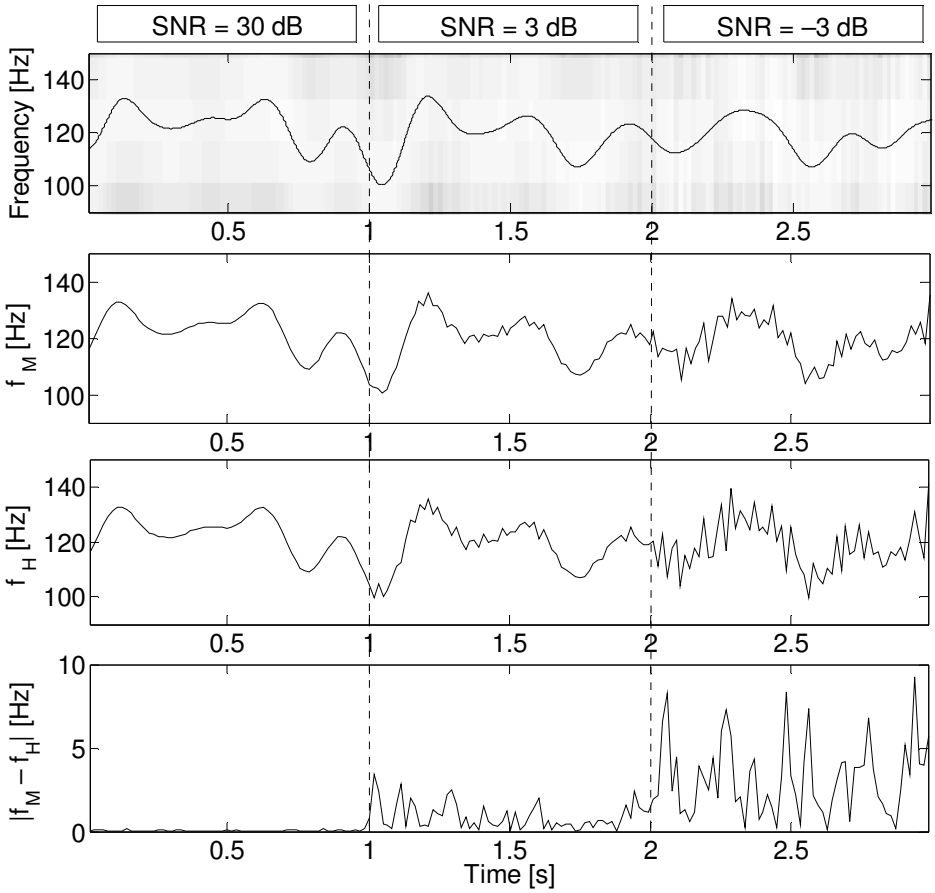

FIGURE 8: Looking from the top: a part of spectrogram together with a curve representing instantaneous frequencies of tonal component, estimated frequencies using the QIFFT and hybrid method, and absolute frequency difference calculated according to (10).

It can be noted that when the SNR is equal to 30 dB the instantaneous frequencies estimated using the QIFFT and hybrid estimates are close to each other resulting in negligible $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ values. However, when the SNR decreases, the $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ distance tends to have a higher mean value. This observation confirms that the absolute difference between frequencies estimated using the QIFFT and the hybrid method can be used as a measure of spectral components tonality [9].

Next, the influence of the hop size on the mean and maximum frequency distance $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ was examined. The single sinusoidal component of -3 dB SNR and constant frequency equal to 800 Hz (sampling rate 8000 Sa/s) was generated and further analysed with the hop size ranging from 0.125 ms (H=1) to 32 ms (H=256) with 1 ms (8 samples) step. For every selected hop size of the STFT analysis the arithmetic mean and maximum value of the vector containing all $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ values corresponding to the considered tonal component was calculated. The results are shown in Fig. 9.

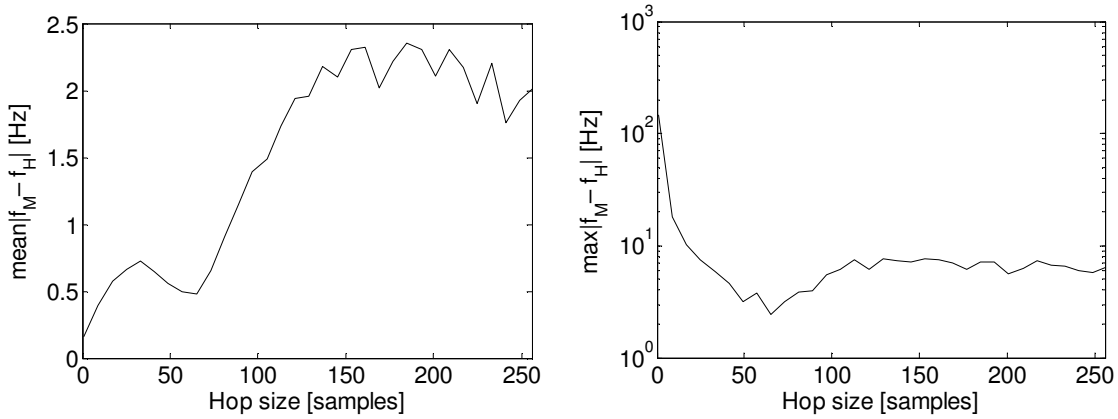

FIGURE 9: The mean (left) and maximum (right) frequency distances $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ obtained for sinusoids of constant frequency and SNR= -3 dB SNR analyzed with various hop sizes

The maximum value of frequency distance is the highest for hop size equal to one sample and decreases while the hop size increases to approximately N/4. This phenomenon is related to the properties of hybrid estimator which yields occasionally inadequate frequency estimates when the sinusoidal component of low SNR is analysed. Additionally, in the above-mentioned hop size range, the mean value of frequency distance is rather low. Thus, taking into account also computational complexity of the algorithm, the hop sizes below quarter of the frame length should be avoided.

Considering hop size range from 60 to about 150 samples it can be observed, that the mean value of $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ rises monotonically and then saturates beyond 2 Hz level. Adequately, the maximum value of frequency distance increases up to about 9 Hz, but saturates for hop size equal to approximately a half of the frame length. While the maximum values seem to be almost constant for higher hop sizes, the mean values tend to even slight decrease for the hop sizes longer than 200 samples. Therefore, the proposed criterion for tonal components detection and their tonality estimation would operate most efficiently when the hop size would be selected within range between ¼ to approximately ¾ of the frame length in the analysis. This observation is coherent with conclusions related to the results presented in Fig. 3. Although the curves presented in Fig. 9 would slightly vary depending on the frequency of analysed sinusoid, their

major character would be retained. Therefore, the presented considerations tend to be valid regardless the frequency of tonal component.

In order to determine the tonality measure of a particular spectral component according to (12) the appropriate value of $|f_\Delta|_{\text{thd}}$ threshold must be selected. This threshold must be at least as high as the maximum value of frequency distance yielded by the algorithm providing that the tonal component is analysed. Since the maximum value of the frequency distance depends on the chosen hop size *H* (see Fig. 9) threshold $|f_\Delta|_{\text{thd}}$ may be selected in accordance to it. However, in the proposed approach it is assumed to be constant regardless the selected *H* value of the STFT analysis. Actually it was selected to be a half of the frequency width corresponding to a bin of zero-padded spectrum

$$|f_\Delta|_{\text{thd}} = \frac{f_s}{2N_{\text{FFT}}} \tag{19}$$

Further, a set of stationary and frequency modulated sinusoids of nominal frequency equal to 120 Hz and SNR values ranging from 100 dB to -20 dB with 2 dB step were generated and analyzed. The frequency deviation of modulated sinusoid was set to 20 Hz and the carrier frequency was modulated using sinusoid of 3 Hz frequency. The sampling rate was equal to 8000 Sa/s, the frame length was selected to be equal to 32 ms (*N*=256) and hop size was adjusted to 16 ms (*H*=128). Since the length of every analysed signal was equal to 3 s, resulting in a vector of FTM values corresponding to the sinusoid of a particular SNR, the arithmetic mean values of these vectors were determined. The results of experiment are presented in Fig. 10.

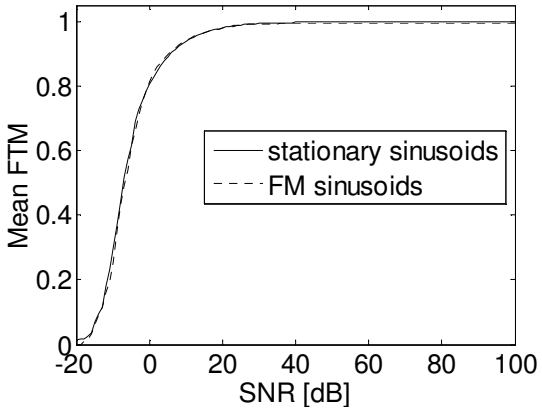

FIGURE 10: Mean values of FTM determined for pure and frequency modulated sinusoids of various SNR

The mean FTM for tonal component of the SNR higher than approximately 40 dB is equal or close to the value of 1, because the instantaneous frequencies estimated using estimators (2) and (5) are almost identical to each other. In the SNR range from 40 dB to -20 dB the mean FTM values gradually decrease indicating lower tonality of the considered spectral component. It can be observed that when the tonal component is totally masked with noise which is the case when SNR is equal to -20 dB, the FTM is close to the value of 0. This confirms that the proposed tonality criterion is efficient in terms of distinguishing tonal from noisy spectral components. Additionally, the curves representing the mean FTM for a pure sinusoid and a frequency modulated one are practically identical to each other indicating that frequency modulation does not affect significantly the tonality measurements provided by the proposed method.

4. CONCLUSIONS

A criterion for distinguishing tonal from noisy spectral components based on a comparison of their instantaneous frequencies estimated using two different methods was proposed and evaluated. Since one of the estimators was specially developed for application considered, the experiments revealing its properties were carried out. It was shown that the proposed hybrid estimator provides satisfactory accuracy of frequency estimation in case of the analysis of pure and modulated sinusoidal components. Regardless the way the tonal components changes its frequency (linearly or periodically) the MSE of the frequency estimation remains below reasonable threshold for the hybrid method. However, it yields inadequate estimation results when the spectral component corresponds to a noise. These two above-mentioned properties of the estimator engineered here were found to be essential for application of the developed tonality criterion (FTM). The experiments revealed that the absolute difference between frequencies estimated using the QIFFT method and the hybrid one is directly related to the SNR of the sinusoids analysed. It was shown that the investigated algorithm operates most efficiently when the hop size of analysis is chosen between ¼ to ¾ of the frame length. The experimental results proved that characteristics of FTM values versus SNR of sinusoidal component are almost identical to each other whenever the sinusoid of constant or modulated instantaneous frequency is analysed. The presented tonality measure may substitute the tonality estimators employed so far in the psychoacoustic models and may be used also in various applications requiring tonal components detection.

5. ACKNOWLEDGMENTS

Research subsidized by the Polish Ministry of Science and Higher Education under grant PBZ MNiSW-02/II/2007 and N N517 378736.

6. REFERENCES

[1] ISO/IEC MPEG, "*IS11172-3 Coding of moving pictures and associated audio for digital storage media up to 1.5 Mbit/s*", Part 3: Audio, Annex D. ISO/IEC JTCI, 1992.

[2] ISO/IEC "*13818-7 Information technology — Generic coding of moving pictures and associated audio information*", Part 7: Advanced Audio Coding (AAC), 4th edition, 2006.

[3] M.G. Christensen, A. Jakobsson, "*Multi-Pitch Estimation*". Synthesis Lectures on Speech and Audio Processing, 5(1):1-160, 2009.

[4] J. D. Johnston, "*Transform coding of audio signals using perceptual noise criteria*". IEEE J. on Selected Areas in Comm., 6:314-323, 1988.

[5] O. Hellmuth, E. Allamanche, J. Herre, T. Kastner, M. Cermer, W. Hirsch, "*Advanced audio identification using MPEG-7 content description*". In proceedings of 111th Audio Eng. Soc. Int. Conf., New York, USA, 2001.

[6] R. J. McAulay, T. F. Quatieri, "*Speech analysis/synthesis based on a sinusoidal representation*". IEEE Transactions on Acoustics, Speech, and Signal Processing, 34:744-754, 1986.

[7] S. Levine, J. O. Smith III, "*Improvements to the switched parametric & transform audio coder*". In proceedings of IEEE Workshop on Application of Signal Processing to Audio and Acoustics, New York, USA, 1999.

[8] T. Painter, A. Spanias, "*Perceptual coding of digital audio*". In proceedings of. of IEEE, 88:451-513, 2002.

[9] S.-U. Ryu, K. Rose, "*Enhanced accuracy of the tonality measure and control parameter extraction modules in MPEG-4 HE-AAC*". In proceedings of 119th Audio Eng. Soc. Int. Conf., New York, USA, 2005.

[10] K. Lee, K. Yeon, Y. Park, D. Youn, *"Effective tonality detection algorithm based on spectrum energy in perceptual audio coder*". In proceedings of 117th Audio Eng. Soc. Int. Conf., San Francisco, USA, 2004.

[11] X. Rodet, "*Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models*". In proceedings of IEEE Time-Frequency and Time-Scale Workshop, Coventry, Grande Bretagne, 1997.

[12] A. J. S. Ferreira, "*Tonality detection in perceptual coding of audio*". In proceedings of 98th Audio Eng. Soc. Int. Conf., Paris, France, 1995.

[13] M. Kulesza, A. Czyzewski, „*Audio codec employing frequency-derived tonality measure*". In proceedings of 127th Audio Eng. Soc. Int. Conf., New York, USA, 2009.

[14] M. Kulesza, A. Czyzewski, "*Novel approaches to wideband speech coding*". GESTS Int. Trans. On Computer Science and Engineering, 44(1):154-165, 2008.

[15] D. Schulz, "*Improving audio codecs by noise substitution*". J. Audio Eng. Soc., 44:593-598, 1996.

[16] P. J. B. Jackson, C. H. Shadle, "*Pitch-scaled estimation of simultaneously voiced and turbulence-noise components in speech*". IEEE Trans. On Speech and Audio Processing, 9:713-726, 2001.

[17] Y. Wang, R. Kumaresan, "*Real time decomposition of speech into modulated components*". J. Acoust. Soc. Am., 119(6):68-73, 2006.

[18] B. Yegnanarayana, C. Alessandro, V. Darisons, "*An iterative algorithm for decomposition of speech signals into periodic and aperiodic components*". IEEE Trans. on Speech and Audio Proc., 6: 1-11, 1998.

[19] M. Kulesza, A. Czyzewski, "*Tonality Estimation and Frequency Tracking of Modulated Tonal Components*". J. Audio Eng. Soc., 57(4):221–236, 2009.

[20] G. Peeters, X. Rodet, "*Signal characterization in terms of sinusoidal and non-sinusoidal components*". In proceedings of Digital Audio Effects (DAFx) Conf., Barcelona, Spain, 1998.

[21] G. Peeters, X. Rodet, "*SINOLA: a new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum*". In proceedings of the Int. Computer Music Conf., Beijing, China, 1999.

[22] U. Zolzer, "*DAFX Digital Audio Effects*". John Wiley & Sons, United Kingdom, 2002.

[23] M. Abe, J. O. Smith III, "*Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks*". In proceedings of 117th Audio Eng. Soc. Int. Conf., San Francisco, USA, 2004.

[24] F. Keiler, S. Marchand, "*Survey on extraction of sinusoids in stationary sound*". In proceedings of the 5th Int. Conf. on Digital Audio Effects (DAFx-02), Hamburg, Germany, 2002.

[25] D. C. Rife, R. R. Boorstyn, "*Single-tone parameter estimation from discrete-time observations*". IEEE Trans. Info. Theory, 20(5):591-598, 1974.

[26] M. Betser, P. Collen, G. Richard, B. David, "*Preview and discussion on classical STFT-based frequency estimators*. In proceedings of 120th Audio Eng. Soc. Int. Conf., Paris, France, 2006.

[27] J.C. Brown, M.S. Puckette, "*A high resolution fundamental frequency determination based on phase changes of the Fourier transform*". J. Acoust. Soc. Am., 94(2):662-667, 1998.

[28] J. Flanagan, R. Golden, "*Phase vocoder*". Bell Syst. Tech. J., 45:1493–1509, 1966.

[29] F.J. Charpentier, "*Pitch detection using the short-term Fourier transform*". In proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 11:113-116, Tokyo, 1986.

[30] S.W. Lang, B.R Musicus, "*Frequency estimation from phase difference*". In proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 4:2140-2143, United Kingdom, 1989.

[31] M.S. Puckette, J.C. Brown, "*Accuracy of frequency estimates using the phase vocoder*". IEEE Trans. On Speech and Audio Processing, 6(2):166-176, 1998.

[32] M. Lagrange, S. Marchand, "*Estimating the instantaneous frequency of sinusoidal components using phase-based methods*". J. Audio Eng. Soc., 55:385-399, 2007.

[33] M. Betser, P. Collen, G. Richard, B. David, "*Estimation of frequency for AM/FM models using the phase vocoder framework*". IEEE Trans. On Signal Processing, 56(2):505-517, 2008.

[34] M. Lagrange, S. Marchand, J.B. Rault, "*Sinusoidal parameter extraction and component selection in non stationary model*". In proceedings of Digital Audio Effects (DAFx) Conf., Hamburg, Germany, 2002.