TIME-FREQUENCY ANALYSIS OF MUSICAL SIGNALS USING THE PHASE COHERENCE

Alessio Degani, Marco Dalai, Riccardo Leonardi and Pierangelo Migliorati,

DII, Signals and Communication Lab
University of Brescia
Brescia, ITALY
alessio.degani@ing.unibs.it, marco.dalai@ing.unibs.it,
riccardo.leonardi@ing.unibs.it, pierangelo.migliorati@ing.unibs.it

ABSTRACT

In this paper we propose a technique based on the phase evolution of the Short Time Fourier Transform (STFT) for increasing the spectral resolution in the time-frequency analysis of a musical signal. It is well known that the phase evolution of the STFT coefficients brings important information on the spectral components of the analysed signal. This property has already been exploited in different ways to improve the accuracy in the estimation of the frequency of a single component. In this paper we propose a different approach, where all the coefficients of the STFT are used jointly to build a measure of how likely all the frequency components are, in terms of their phase coherence evaluated in consecutive analysis window. In more detail, we construct a phase coherence function which is then integrated with the usual amplitude spectrum to obtain a refined description of the spectral components of an audio signal.

1. INTRODUCTION

Time-frequency analysis is a central tool in most of the applications of audio/music signal processing, Music Information Retrieval algorithms [1] and audio coding systems. The most common used tool for this purpose is the Short Time Fourier Transform (STFT) [2] which is the non-stationary counterpart of the Discrete Fourier Transform (DFT). The STFT decomposes the discrete signal in partially overlapping frames, and it expands each of these frames in the discrete Fourier basis [3]. It thus provides a time varying discrete-frequency content description of the signal. Usually, only the amplitude spectrum of the STFT is taken into account. In some applications which require both good frequency accuracy and good time localization, this may not suffice. To overcome this issue, some additional processing may be required to increase the frequency resolution (see [4]) or to add a-priori information (see [5]). One possible approach consists in using the phase of the STFT to improve the frequency resolution. Some works in the literature propose specific techniques to refine the frequency estimation by using the phase evolution of STFT coefficients [6, 7, 8]. These methods, however, operate on a coefficient-wise basis to improve the frequency estimation of a single sinusoid in a local frequency interval. In particular they do not allow a global exploitation of the full phase spectrum evolution to blindly enhance the frequency analysis over the entire range. This is instead the aim of other approaches based on the reassignment method first proposed in [9]. This technique corrects the information contained in the spectrogram by "moving" the energy in the time-frequency plane according to phase information (see [10]) for more details.

In this paper, we propose a different approach where phase information is used to assign a "coherence score" to spectral amplitude components. As for other works based on the phase evolution of the STFT, the underlying idea goes back to Flanagan and Golden [11] (see [12, 13] for recent advances). Here, however, we propose a technique for combining the phase evolution of different coefficients in order to obtain a function $\mathcal{X}_m(f)$, that we call Phase Coherence Function (PCF), which measures the likelihood of the presence of a sinusoidal component at the unquantized frequency f at time instant m. The function $\mathcal{X}_m(f)$ is computed using only the phase information, and we then combine it with the STFT amplitude spectrum to obtain a refined spectral analysis of the signal. The main difference between our method and other available techniques is that our method does not try to move the components from one frequency to another, but rather assigns coherence scores to components. In particular, the function $\mathcal{X}_m(f)$ takes on positive (respectively, negative) values for those f that are likely (unlikely) to be present in the signal according to the phase evolution of nearby coefficients of the spectrogram. The "coherence score", furthermore, is computed in a way which inherently takes into account the issue of the phase unwrapping which often constitutes a problem in many of the methods mentioned above.

The paper is structured as follows. In Section 2 we give the basic notions on the STFT and we introduce the key idea of the phase coherence. In Section 3.1 we show how

to combine the information given by different coefficients to obtain the PCF. In Section 4 we present the experimental results and we discuss the possible applications of our technique.

2. TIME-FREQUENCY ANALYSIS

2.1. Short Time Fourier Transform

As we mentioned in Section 1, the classic time-frequency analysis is performed using STFT. The N-terms STFT, at time frame m, of a discrete signal $x[n]$ is defined as

$$X_{m,k} = \sum_{n=0}^{N-1} x[n + \tau m] \cdot w[n] \cdot e^{-j2\pi \frac{k}{N} n}, \qquad (1)$$

where $k = -N/2 + 1, \ldots, N/2$, τ is the hop size (in samples) from two subsequent frames and $w[n]$ is the windowing function. The STFT is a complex valued function which can be equivalently described in terms of its amplitude $|X_{m,k}|$ and its phase

$$\Phi_{m,k} = \measuredangle X_{m,k}. \qquad (2)$$

If $x[n]$ is sampled at frequency F_s, the frequency resolution of the STFT is given [14] by the expression

$$\Delta_k = \frac{F_s}{N} \qquad (3)$$

that can be seen as the width of the frequency interval associated to each coefficient $X_{m,k}$ ignoring the windowing effects and thus it is related to the STFT accuracy in positioning the spectral components of the signal. In this paper we assume, where not otherwise specified, $F_s = 22050$ Hz, $N = 4096$ without zero-padding, $\tau = 1024$ samples and $w[n]$ is the Hanning analysis window. We will see that, under certain hypothesis on the structure of the analysed signal, we can partially overcome this limit. If we are interested in the detection of the frequency location of the short term sinusoidal components in an audio signal, we can exploit the phase evolution of two consecutive frames of the STFT to increase the frequency resolution of a time-frequency representation. This basic idea is shared by all the frequency estimation methods that uses the phase spectrum.

2.2. Phase evolution of the STFT

In this section, we introduce the principle at the base of the coherence measure that will be described in the next section. The key point is in the phase evolution of the STFT coefficients of a pure sinusoidal signal. For the sake of simplicity, we consider complex exponential functions; the effect on real signals will then be intuitively derived from this analysis. Consider then a signal $x[n]$ assumed to be a sampled version of a signal $x(t) = e^{j2\pi F_0 t}$ at sampling frequency F_s, that is

$$x[n] = e^{j2\pi f_0 n}, \qquad (4)$$

where $f_0 = F_0/F_s$ is the normalized frequency. In the continuous domain, if we let $X(F) = \mathcal{F}\{x(t)\}(F)$ be the Fourier transform of $x(t)$, we know that

$$\mathcal{F}\{x(t + t_0)\}(F) = e^{j2\pi F t_0} \mathcal{F}\{x(t)\}(F). \qquad (5)$$

Since our signal is a pure exponential, however, its Fourier transform is a Dirac delta function and thus we may as well write

$$\mathcal{F}\{x(t + t_0)\}(F) = e^{j2\pi F_0 t_0} \mathcal{F}\{x(t)\}(F). \qquad (6)$$

Since the STFT is a sliding window discrete version of the Fourier transform, intuition suggests that its coefficients in (1) evolve for varying m ruled by this property of the Fourier transform. Since a pure sinusoidal function, in general, affects different coefficients due to the windowing effect, one may be induced to expect eq. (5) to hold rather than eq. (6). This is not the case, however, since it is easily checked that

$$X_{m,k} = \sum_{n=0}^{N-1} e^{j2\pi f_0 (n+\tau m)} w[n] \cdot e^{-j2\pi \frac{k}{N} n} \qquad (7)$$

$$= e^{j2\pi f_0 \tau m} \sum_{n=0}^{N-1} e^{j2\pi f_0 n} w[n] \cdot e^{-j2\pi \frac{k}{N} n} \qquad (8)$$

$$= e^{j2\pi f_0 \tau m} X_{0,k}. \qquad (9)$$

That is, the k-th coefficient evolves as a complex exponential function of m with frequency f_0, regardless of the value of k.

This relation holds exactly for all k for a complex exponential signal $x[n]$. For a real sinusoidal function $x[n] = \cos(2\pi f_0 n)$, one can use Euler's relation to write $x[n]$ as a composition of exponential functions with frequency $\pm f_0$. It is then seen that, if the windowing function $w[n]$ has a discrete transform that decreases sufficiently fast in k, the above analysis shows that the k-th coefficient evolve as a complex exponential in m with frequency $\pm f_0$ if k/N is sufficiently close to $\pm f_0$. Since we always work with real signals in the audio frequency range, we may only consider the positive frequency f_0.

So, in presence of a pure tone at frequency f_0, one should expect the coefficients with k/N in the neighbourhood of f_0 to evolve as exponential functions of m with frequency f_0. It is this property that can be used to measure the likelihood of having a sinusoidal component at some given frequency by only considering the phase evolution of the STFT coefficients. In the next section, we propose a method to jointly perform such an analysis over different coefficients to "test" different possible frequency components.

3. PHASE COHERENCE MEASURE

3.1. Coherence measure

The analysis in the previous section shows that in the presence of a pure sinusoid we can predict $X_{m,k}$ from $X_{0,k}$ according to (9). In practice, we will only need to consider two adjacent frames in the STFT and in this case we can say that, given $X_{m,k}$ and τ, we can write the one step forward prediction for $X_{m+1,k}$ as

$$\hat{X}_{m+1,k} = X_{m,k} \cdot e^{j2\pi f_0 \tau}. \tag{10}$$

Equation (10) gives the ideal evolution of the k-th coefficient when the signal is a pure esponential at frequency f_0. Since we do not know f_0 but rather measure the true coefficient phases, we can say that a frequency f_0 is compatible with the measured phase evolution if

$$\Phi_{m+1,k} - \Phi_{m,k} - 2\pi f_0 \tau = 0 \mod (2\pi). \tag{11}$$

This equation is often used, with a fixed value of k, as a way to extract the value of f_0 from the knowledge of the other terms. This approach has however some problems. First, the fact that the equation only holds (mod 2π) leads to the problem of the phase unwrapping. That is, the frequency f_0 is not certain even in the ideal case, since it can only be determined up to multiples of $1/\tau$. Second, we should not expect to have exact equality in (11), since our real signal will not be an ideal sinusoid, but a combination of sinusoidal components usually affected by noise. Hence, when considering (11) for different k values, different noisy estimates for f_0 are obtained. Here, we suggest a different approach that does not try to estimate one single f_0 from eq. (11), but rather uses that equation to test whether a frequency f is compatible with the phase evolution of the coefficient k. In our test, moreover, we chose to adopt a "soft" approach defining a coherence measure that is function of three variables m, k and f. More precisely, setting $\Delta\Phi_{m,k} = \Phi_{m+1,k} - \Phi_{m,k}$, we define a coherence measure as given by the expression

$$C_{m,k}(f) = \cos(\Delta\Phi_{m,k} - 2\pi f \tau) \in [-1, 1]. \tag{12}$$

It is easy to see that for all f that exhibit coherent phase evolution we have $C_{m,k}(f) = 1$. Conversely, $C_{m,k}(f) = -1$ for all those f for which $\Delta\Phi_{m,k} - 2\pi f \tau = \pi \mod (2\pi)$, which means that we have phase opposition between the predicted coefficient and the measured one.

It may be useful to note here that we can rewrite (12) in terms of the cross-spectral components as follows

$$C_{m,k}(f) = \Re\left\{ \frac{X^*_{m,k}}{|X_{m,k}|} \cdot \frac{X_{m+1,k}}{|X_{m+1,k}|} \cdot e^{-j2\pi f \tau} \right\}, \tag{13}$$

where $\Re\{\cdot\}$ denotes the real part and $(\cdot)^*$ the complex conjugate. Thus, our coherence function is a measure of the contribution given by the frequency bin k in the the "modified" cross-correlation between the two frames of the signal, according to the fact that equation (6) holds in place of the usual (5).

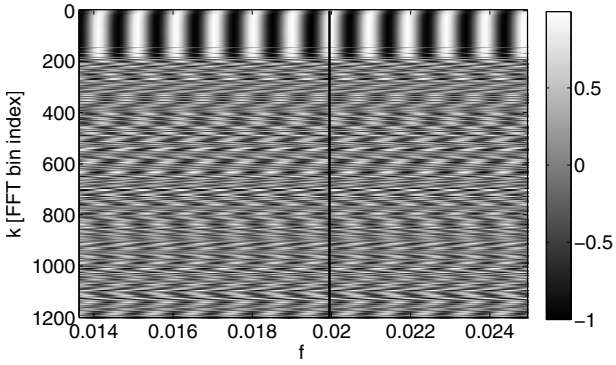


Figure 1: $C_{m_0,k}(f)$ for $x_t[n]$ at m_0-th time frame. The black solid line is the reference for $f = \frac{440}{F_s}$.

For the sake of simplicity we look at the coherence measure for a fixed $m = m_0$ and $k = k_0$. From (12), we know that for a given m and k, $C_{m,k}(f)$ is a sinusoidal function of the variable f with a "frequency" τ, since $\Delta\Phi_{m_0,k_0} = \Delta\Phi$ is a constant value. This function has local maxima f_M in

$$f_M(n) = \frac{n}{\tau} + \frac{\Delta\Phi}{2\pi\tau}, \forall n \in \mathbb{Z}. \tag{14}$$

If we consider the maximum obtained for $n = 0$, we find that $N \cdot f_M(0)$ gives an instantaneous frequency in the range $[k_0 - 1/2, k_0 + 1/2]$ as defined in [11]. This is the frequency which is usually selected by other methods based on the phase evolution of the STFT. Here, however, we will not select this frequency *a-priori* and we will instead use the whole function $C_{m,k}(f)$ over different values of k to test a generic f value.

Considering a test signal defined as $x_t[n] = \sin[2\pi \frac{440}{F_s} n]$, the phase coherence measure of $x_t[n]$ at the time frame $m = m_0$ is shown in Fig. 1. We can easily see that at the top of the graph there is a group of bins k in which $C_{m_0,k}(f)$ shares the same phase, that suggest, as seen in Section 2.2, that the neighbours of the STFT evolves with an identical phase difference due to the presence of a stationary tone. In this "coherence band", the local maxima of $C_{m_0,k}(f)$ are located corresponding to $f = \frac{440}{F_s}$.

3.2. Coherence function

Our aim is to refine the amplitude spectrum using the coherence measure in order to obtain a better resolution in the localization of sinusoidal components. Intuitively not all the values of $C_{m_0,k}(f)$ are of practical interest. More precisely, for a given $f = f_0$, it is clear that only the neighbouring values of the discrete frequency $k_0 = N \cdot f_0$ give a reliable

phase coherence measure. The spreading of the frequency components due to the windowing effects of the STFT, together with the propagation of the phase coefficients described in Section 2.2, suggests in fact that in presence of a component at frequency f_0, in the neighbours of $N \cdot f_0$, the coefficients evolve according to f_0. Far from this region, instead, the coefficients evolve independently from this component. We automatically consider only the relevant coefficients by choosing a weighting procedure that uses the amplitude spectrum of the analysis window to weight the coefficients of the phase coherence measure around a specified k_0. The weighting coefficients are calculated as the unity energy amplitude spectrum of the analysis window modulated at the normalised frequency f as follows

$$W_k(f) = \frac{\left| DFT_k^N \left[w[n] \cdot e^{j2\pi f n} \right] \right|}{\sqrt{\sum_{n=0}^{N-1} |w[n]|^2}}, \forall f \in [0, \tfrac{1}{2}], \quad (15)$$

where $DFT_k^N[\cdot]$ is the Discrete Fourier Transform using N samples.

Now we can define the Phase Coherence Function (PCF) as a weighted sum of the coherence measure as

$$\mathcal{X}_m(f) = \sum_k W_k(f) \cdot C_{m,k}(f), \forall f \in [0, \tfrac{1}{2}]. \quad (16)$$

The PCF is a phase coherence indicator between time frame m and $m+1$ for all f. In Fig. 2 it is shown the PCF around $f \cdot F_s = 440$ Hz, for the test signal $x_t[n]$ previously defined, at a given time frame $m = m_0$.

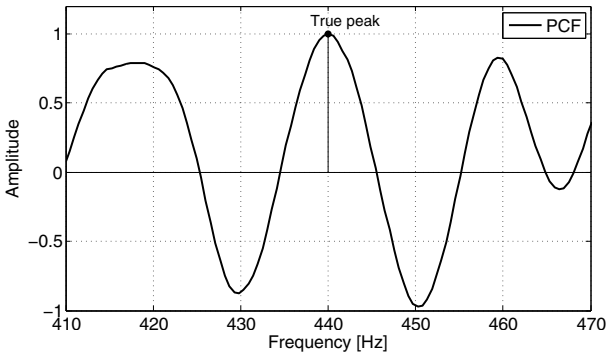


Figure 2: Plotting of $\mathcal{X}_{m_0}(f)$ for $x_t[n]$ at the m_0-th time frame. The Frequency axis shows the non-normalized frequencies in Hz.

By looking at Fig. 2, we can notice high values of coherence also in $f \cdot F_s \neq 440$ Hz. This is due to the periodicity of $C_{m_0,k_0}(f)$; two local maxima are obtained near $f \cdot F_s = 440$ Hz at a distance F_s/τ since $C_{m_0,k_0}(f)$ has period $1/\tau$. However, the two local maxima show different amplitudes since the weighting coefficients $W_k(f)$ change there with respect to the true central frequency.

The obtained function gives a measure of the likelihood that the signal contains a pure sinusoidal component at each frequency f by only considering the phase of the STFT. Taking one step further, a more useful representation of the signal is obtained by combining this phase information with the amplitude spectrum. We define the Phase Coherence Function Weighted Modulus (PCFWM) as

$$\bar{\mathcal{X}}_m(f) = \sum_k |X_{m,k}| \cdot W_k(f) \cdot C_{m,k}(f), \forall f \in [0, \tfrac{1}{2}]. \quad (17)$$

The PCFWM is an amplitude spectrum-like representation with improved localization of the spectral peaks of pure tones. However, it is not strictly an amplitude spectrum because negative values of the PCFWM may occur. A negative phase coherence is measured when the phase difference between two consecutive time frames approaches $\pm\pi$.

The advantages of this combination can be seen in Fig. 3. The secondary peaks in the coherence function are strongly attenuated by the amplitude since no spectral energy is present there. Furthermore, the negative peaks of the coherence function fall in a region where the spectrum does take relevant values. Those negative values of the phase coherence indicate that the energy contained in this coefficients of the STFT is in some sense "spurious", and this is due to pure components in nearby frequencies.

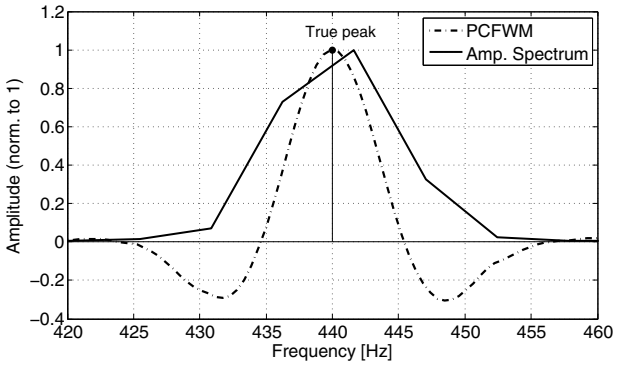


Figure 3: Amplitude spectrum (solid line) and $\bar{\mathcal{X}}_{m_0}(f)$ (dashed line) of $x_t[n]$. Plots are scaled to unit amplitude.

4. RESULTS AND APPLICATIONS

Our method find its main application in signal processing tasks that requires blind but accurate frequency localization of pure sinusoidal components of a signal. As shown in Fig. 3, in presence of a single component our method leads to a sharper lobe in the frequency analysis, and it allows for a more precise estimation of the peak position, if desired, ensuring that the results of [6, 7, 8] are recovered. Fig. 4 shows then the advantage when more than one components

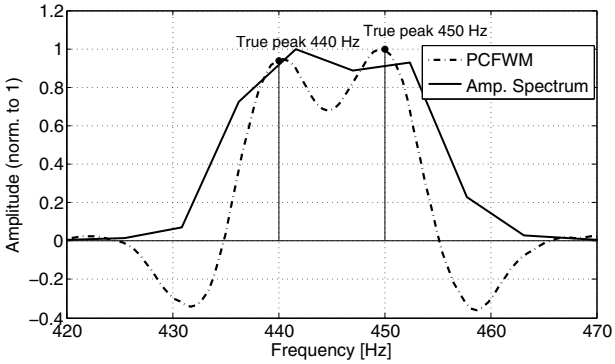


Figure 4: Amplitude spectrum (solid line) and $\bar{\mathcal{X}}_{m_0}(f)$ (dashed line) of $x_d[n] = \sin[2\pi\frac{440}{F_s}n] + \sin[2\pi\frac{450}{F_s}n]$. Plots are scaled to unit amplitude.

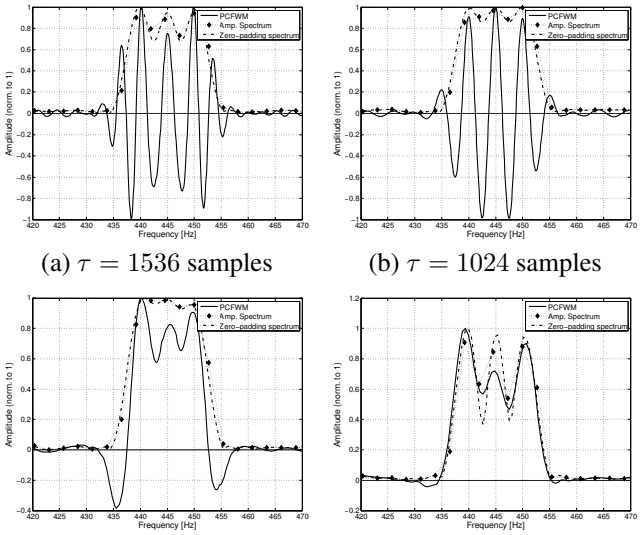
are present. Here, two pure sinusoids with very close frequencies are analysed. Our method does not assume any *a-priori* knowledge on the number of components.

Figure 5 and 6 show the effect of the parameter τ. Here, we used a noisy synthetic sound with three sinusoids with frequency 440 Hz, 445 Hz and 450 Hz. In these figures, the amplitude spectrum, the interpolated amplitude spectrum and the PCFWM are compared. The interpolated amplitude spectrum is calculated using zero-padding during the FFT in order to obtain the same frequency resolution of the PCFWM.

It is important to keep in mind that a negative value of the PCFWM at a frequency f indicates that a pure sinusoidal function at that frequency is very unlikely to be present in the signal. Hence, the alternation of large positive and negative peaks allows us to give a sharp estimation of true peak positions. Due to the coherence measure adopted according to equation (12), the value of τ determines how fast this positive and negative peaks alternate. This however also impacts on the number of peaks with a large positive value that are generated around each single sinusoidal component in the signal. Figures 5 and 6 show this trade-off. The parameter τ sets a trade-off between how narrow the peaks in the $\bar{\mathcal{X}}_{m_0}(f)$ are and the number of "false positive" peaks. In these examples $\tau = 1024$ samples (50% of overlap) is a good compromise between resolution and "false" peak detection.

5. CONCLUSIONS

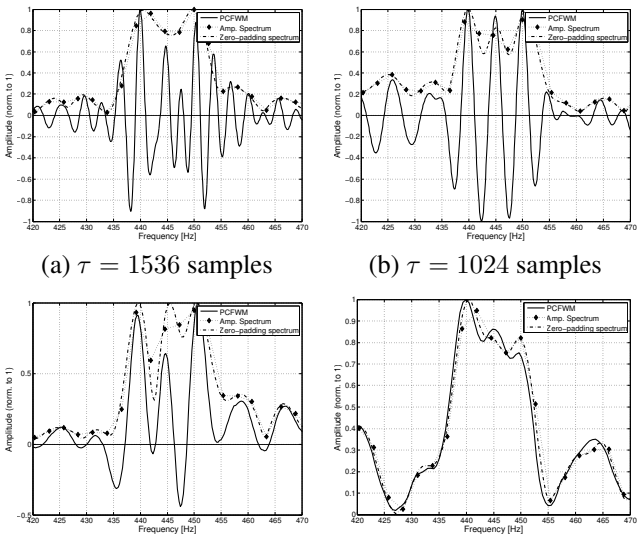
In this paper we have introduced a novel technique that combines the amplitude spectrum and the phase coherence measure in order to refine the time-frequency representation of musical signals. We have demonstrated how this method can improve the frequency localization of short term stationary sinusoid in a audio signal. Since musical signal



(a) $\tau = 1536$ samples (b) $\tau = 1024$ samples

(c) $\tau = 512$ samples (d) $\tau = 256$ samples

Figure 5: Comparison of amplitude spectrum, interpolated (zero-padding) spectrum and $\bar{\mathcal{X}}_{m_0}(f)$ for different τ, calculated on a signal with three sinusoids at frequency 440 Hz, 445 Hz and 450 Hz plus noise with $SNR = 10$ dB. In this example $N = 2048$ and $Fs = 5513$ Hz.



(a) $\tau = 1536$ samples (b) $\tau = 1024$ samples

(c) $\tau = 512$ samples (d) $\tau = 256$ samples

Figure 6: Comparison of amplitude spectrum, interpolated (zero-padding) spectrum and $\bar{\mathcal{X}}_{m_0}(f)$ for different τ, calculated on a signal with three sinusoids at frequency 440 Hz, 445 Hz and 450 Hz plus noise with $SNR = -10$ dB. In this example $N = 2048$ and $Fs = 5513$ Hz.

are composed primarily by notes and tones, our technique brings benefits for time-frequency analysis of this kind of signals, when accurate frequency measures are needed.

6. REFERENCES

[1] Karin Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution fft," *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, 2006.

[2] Leon Cohen, *Time-Frequency Analysis*, Prentice Hall PTR, Upper Saddle River, New Jersey 07458, 1994.

[3] Elias M. Stein and Rami Shakarchi, *Fourier Analysis: an introduction*, Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2002.

[4] Xavier Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, A. Picialli C. Roads, S. Pope and G. De Poli, Eds., chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122. Swets & Zeitlinger Publishers, 1997.

[5] Ralph O. Schmidt, "Multiple emitter location and signal parameter estimation," *Proc. RADC Spectrum Estimation Workshop*, pp. 243–258, 1973.

[6] Kevin M. Short and Ricardo A. Garcia, "Signal analysis using the complex spectral phase evolution (cspe) method," *120th AES Convention, Paris, France*, 2006.

[7] Bertrand Gottin, Irena Orovic, Cornel Ioana, Srdjan Stankovic, and Jocelyn Chanussot, "Signal characterization using generalized "time-phase derivatives" representation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3001–3004, 2009.

[8] Mathieu Lagrange and Sylvain Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods," *Journal of AES*, vol. 55, no. 5, 2007.

[9] Kunihiko Kodera, Roger Gendrin, and Claude Villedary, "Analysis of time-varying signals with small bt values," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 64–76, 1978.

[10] Francois Auger and Patrick Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, 1995.

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

[12] Alexis Moinet and Thierry Dutoit, "Pvsola: A phase vocoder with synchronized overlap-add," *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, 2011.

[13] Sebastian Kraft, Martin Holters, Adrian von dem Knesebeck, and Udo Zölzer, "Improved pvsola time-stretching and pitch-shifting for polyphonic audio," *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, 2012.

[14] Udo Zölzer, *DAFX - Digital Audio Effects*, John Wiley & Sons, LTD, Baffins Lane, Chichester, West Sussex, PO 19 1UD, England, 2002.