Signal-Driven Window-Length Adaptation for Sinusoid Detection in Polyphonic Music

Vishweshwara Rao*, Pradeep Gaddipati and Preeti Rao

Audio processing applications that use short-time signal analysis techniques typically utilize fixed window duration single- or multi-resolution analyses. However, different real-world signal conditions such as polyphony and non-stationarity, manifested as musical accompaniment and pitch-modulations respectively in the context of music content analysis, require varying data window lengths for reliable processing. In this paper, we investigate the use of signal sparsity for adapting analysis window lengths. Adaptive-window analysis driven by different measures of sparsity applied to the local spectrum, such as kurtosis and Gini index, is evaluated and shown to be superior to fixed-window analysis in terms of sinusoid detection and frequency estimation for simulated and real signals. A window main-lobe matching method for sinusoid detection is also shown to be more robust to signal conditions such as polyphony and frequency modulation relative to other methods.

***Index Terms*— Sinusoid detection, signal sparsity, window adaptation**

I. Introduction

SINUSOIDAL models have been widely used in music signal processing to effect accurate extraction of musical attributes such as melody [1] and separation of sources from polyphonic mixtures [2]. The accurate and reliable detection of sinusoids in the polyphonic mixture and their parameters (frequencies and amplitudes) can help reveal underlying harmonic relationships and hence the pitch of each harmonic source, and help in instrument identification and source separation. The challenges posed by the polyphonic music context are the closely spaced frequency components due to the presence of several instruments/voices and the often rapidly time-varying nature of the voice/instrument harmonics. A prominent example is a pitch-modulated singing voice in a polyphonic background.

A commonly used first step in sinusoidal modeling is the computation of the short-time spectrum (Fourier transform) of the input signal which typically provides the base representation for sinusoid detection and parameter estimation. The duration and shape of the short-time analysis window have an important influence on the achievable accuracy of the subsequent processing. For example, longer window durations serve to obtain enhanced resolution of individual harmonics of multiple, simultaneously present, pitched sources [1], [3]. On the other hand, when the lead instrument/voice is heavily ornamented e.g. by the use of extensive vibrato or culture-specific musical ornamentation, manifested as large and rapid pitch modulations, the use of long windows usually results in a distortion of the voice harmonics, especially at higher frequencies. Motivated by the correspondingly larger frequency modulations seen in the higher harmonics, sometimes window durations are systematically reduced across frequency bands spanning the spectrum to obtain a multi-resolution analysis [4]. However the analysis parameters are essentially fixed in time and not signal-adaptive. The primary aim of this work is to investigate the possibility of automatically varying the analysis window-length in order to maximize the accuracy of sinusoidal modeling in the polyphonic music context. We consider the specific task of sinusoid identification and frequency estimation for the singing voice in polyphonic music. We investigate the use of certain easily computable mathematical measures of signal sparsity for optimal window-length selection as determined by the measured accuracy of sinusoid detection.

This paper is organized as follows. Section II briefly reviews some known methods of sinusoid identification and comparatively evaluates them using simulated signals suited to the polyphonic music processing context. Section III describes and evaluates different measures for driving window-length adaptation using the same simulated signals. Section IV presents experimental results of using the proposed signal driven window-length adaptation technique to sinusoid identification on real musical signals and compares the system performance with corresponding results from fixed-length windowing. Section V concludes the paper.

II. Evaluation of Sinusoid Identification Methods

Several different approaches to sinusoid detection exist, the most popular of which are the Fourier analysis methods based on the common first step of computing the Fourier spectrum of the windowed signal. We consider Fourier based methods over alternate approaches such as subspace methods for parameter estimation [5], which require prior knowledge about the number of components, and non-linear least-squares based sinusoid detection, which has been shown to not work well for multi-pitch signals [6]. In order to apply Fourier analysis we assume signal stationarity within the analysis duration i.e. the audio signal within each analysis window is modeled by a set of stable sinusoidal components, which have constant amplitude and frequency, and noise. The underlying "sinusoids plus noise" model given by

$$x(n) = \sum_{m=1}^{M} A_m \cos\left(2\pi f_m n + \phi_m\right) + i(n) \qquad (1)$$

where n is the time-index, A_m, f_m and ϕ_m represent the amplitude, frequency and initial phase of the m^{th} sinusoid and M is the number of sinusoids (harmonics) present in the signal. $i(n)$ represents noise or other interference signal.

In the Fourier magnitude spectrum of the windowed signal, the local peaks are potential sinusoid candidates. The task is to distinguish the true sinusoidal candidates from noise and side-lobes arising due to windowing. Sinusoidal components in the Fourier spectrum can be detected based on either their magnitude or phase characteristics [7]. Situations such as closely spaced components due to polyphony and time-varying pitches, however, are expected to influence the reliability of sine identification. To address the non-stationarity arising from time-varying pitches, some frame-level sinusoid parameter estimation methods proposed in the literature track the amplitude, frequency and modulation parameters under certain assumptions on the form of the modulation of the windowed sinusoidal signal [8]-[10]. Constant or first-order AM and linear FM are common assumptions. The influence of neighboring sinusoids in multi-component signals has typically been ignored by assuming that the window length is long enough to make it negligible.

In this section we evaluate three distinct methods of sinusoid identification from the short-time spectrum via simulations that exemplify polyphony and the non-stationarity of the vocal pitch.

Manuscript received August 10, 2010; revised January 27, 2011; re-revised May 23, 2011. The associate editor coordinating the review of this manuscript was Dr. Richard Gaël.

A. Brief Review of Different Methods of Sinusoid Identification

All of the three methods considered here first search the short-time magnitude spectrum for 3-point local maxima. A decision criterion (termed as a "sinusoidality" criterion), based on the spectral properties of the windowed ideal sinusoid, is then applied to the local maximum in order to decide whether it represents a sinusoid (as opposed to a window side-lobe or noise). The first two methods compute the sinusoidality measure from the magnitude spectrum and the third method utilizes the phase spectrum.

The first method employs an amplitude threshold relative to the detected amplitude envelope [2], [11]. The amplitude envelope in this case is obtained by smoothing the magnitude spectrum with a Hamming window shape. The second method, called window main-lobe matching, utilizes a measure of closeness of a local spectral peak's shape to that of the ideal sinusoidal peak. This measure can be computed as the mean square difference [12] or the cross-correlation [13] between the local magnitude spectrum and that of the analysis window main lobe.

Phase-based sinusoidality criteria exploit the phase coherence of sinusoidal components by computing different instantaneous frequency (IF) estimates from phase spectra in the vicinity of the local maximum. Lagrange [14] has demonstrated the theoretical equivalence of different IF estimation methods, which earlier were experimentally shown to perform similarly [7]. We consider a version of the bin-offset method, in which the IF is computed from the derivative of the phase, further modified by Dressler to "weighted bin-offset" for the polyphonic context [4].

Implementation details and parameters used for each of the methods are provided in the Appendix. For the magnitude based methods [11], [12], the frequency estimate of the sinusoid is further refined using parabolic interpolation [7]. Refinement of the sinusoidal frequency is inherent in the weighted bin-offset method.

B. Description of Simulated Signals

We use three simulated signals across evaluations, all sampled at 22.05 kHz. The first two signals, described next, follow the model described in Eq. (1). The first signal is a representation of a vocal utterance. The vocal signal is a vowel /a/ generated using a formant synthesizer [15] at a constant pitch of 325 Hz with harmonics up to 4 kHz (M=12). The second signal represents the polyphonic case by adding a relatively strong harmonic interference to the previous voice-only signal. The interference signal $i(n)$ is a complex tone (also a sum of sinusoids with constant frequency and amplitude) with 7 equal-amplitude harmonics and a pitch of 400 Hz. The signals are added at a Signal-to-Interference Ratio (SIR) of 0 dB. The equal-amplitude interference harmonics are, in general, stronger than the vowel harmonics that roll-off.

The third signal represents the time-varying nature of the voice pitch and does not fit the signal model of Eq. (1), especially for long analysis windows. This is a vocal utterance with no interference (same as the first signal), but the pitch of the vowel now contains vibrato leading to non-stationary harmonics. Vibrato for singing is described as a periodic, sinusoidal modulation of the phonation frequency [16]. The pitch of the vibrato signal is given as

$$f_{vib}(n) = f_{base} \cdot 2^{\left(\frac{A \cdot \sin(2\pi f_r \cdot n / F_S)}{1200}\right)} \qquad (2)$$

where f_{base} is the base frequency (325 Hz), A is half the total vibrato extent, f_r is vibrato rate and F_S is the sampling frequency. The vibrato extent and rate we have used here are 100 cents and 6.5 Hz respectively; these are typical values [16]. The spectrograms of the simulated signals (each 3 sec long) are shown in Fig. 1. In all cases

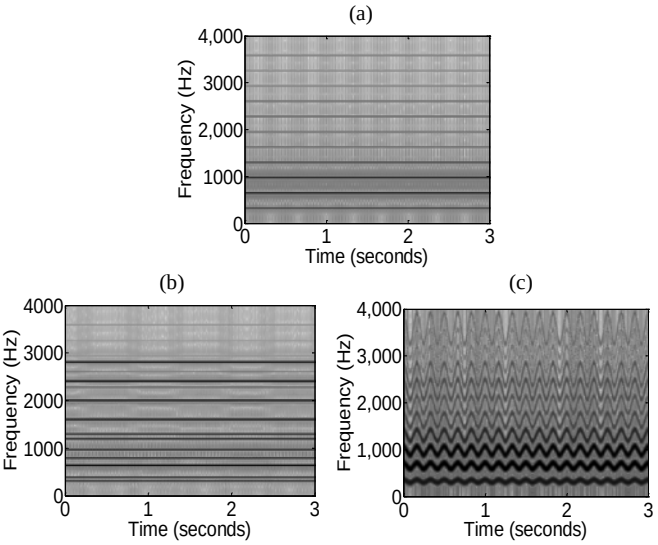

Fig. 1. Spectrograms of (a) synthetic vowel /a/ at pitch 325 Hz, (b) mixture of previous synthetic vowel and harmonic interference at pitch 400 Hz (7 equal amplitude harmonics) added at 0 dB SIR, (c) synthetic vowel with base pitch 325 Hz and vibrato (extent 1 semitone and rate 6.5 Hz).

the magnitude and phase spectra are computed every 10 ms using a fixed 92.9 ms window (2048 point DFT) unless otherwise mentioned.

C. Evaluation

The evaluation criteria used for the sinusoid identification methods are recall, precision and standard deviation of the frequency error from expected (ground truth) harmonic frequency locations. Recall is defined as the ratio of the number of correctly detected sinusoids to the true number of sinusoids present. Precision is the ratio of the number of correctly detected sinusoids to the total number of detected sinusoids. For each frame of the test signal a set of detected sinusoids (frequencies and amplitudes) is computed as those local spectral maxima that have satisfied the particular sinusoidality criterion for that method. Then the n^{th} harmonic of the target signal, with known pitch $f0$, with frequency f_n=$n.f0$, is said to be correctly detected if at least one measured sinusoid, with estimated frequency $f_n^{'}$, satisfies

$$\left| f_n - f_n^{'} \right| < \min\left(0.03\, f_n,\ 50 \text{ Hz}\right) \qquad (3)$$

If more than one measured sinusoid satisfies the above validation criterion, only that sinusoid with the smallest value of $\left| f_n - f_n^{'} \right|$ is labeled as correctly detected. All other detected sinusoids, including those that do not satisfy the validation criterion for *any* expected harmonic, are labeled as false alarms. So only a single measured sinusoid can be assigned to an expected harmonic. For the simulated polyphonic case, we specifically exclude the detected harmonics of the interference signal, representing musical accompaniment, from the list of false alarms. This is done by first computing the number of correct sinusoid detections for the *interference* signal, after applying the above validation criterion, and subsequently subtracting this number from the total number of false alarms for that frame.

The frequency error for the n^{th} expected harmonic with frequency f_n is given as

$$FE_n = f_n - f_n^{'} \quad \text{; if a sinusoid is detected for } f_n$$
$$= \quad 0 \quad \text{; otherwise}$$

We then compute the standard deviation (σ_{FE}) of the FE for all correctly detected harmonics for all analysis time-instants.

TABLE I Performance (RE – Recall (%), PR – Precision (%), σ_{FE} – Frequency error (Hz) of different sinusoid detection methods for different simulated signals.

SIGNAL		AMPLITUDE ENVELOPE	MAIN-LOBE MATCHING	BIN OFFSET
Clean vowel (92.9 ms)	RE	100.0	100.0	93.3
	PR	100.0	100.0	93.9
	σ_{FE}	0.3	0.3	0.5
Vowel + Interference (92.9 ms)	RE	72.7	98.8	75.7
	PR	99.5	100.0	88.7
	σ_{FE}	15.7	15.3	14.4
Vowel with vibrato (23.2 ms)	RE	73.7	89.3	62.0
	PR	67.2	97.2	75.8
	σ_{FE}	8.7	8.4	13.1

D. Results

The sinusoid detection performances of the different methods for the different simulated signals within a 0 to 4 kHz frequency band appear in Table I. For each case we have reported that performance (recall & precision) that maximized the F-measure given by

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

Note that for the case of the vibrato vowel we have used a reduced window length (23.2 ms) rather than the 92.9 ms window used for the other simulated signals. The data window length is reduced to decrease the effect of signal non-stationarity within the window; all three methods showed very poor results with a 92.9 ms window for the vibrato case. In all cases, the DFT size is retained at 2048 points by zero-padding the windowed signal if required.

It is observed from Table I that the main-lobe matching method is more robust to harmonic interference and pitch modulation than the other two methods. The superiority of main-lobe matching to other *single-frame* sinusoid identification methods has also been previously observed by Wells [9]. Here we note that the main-lobe matching method is also superior to the weighted bin-offset method, which relies on the phase computed from the present and *previous* analysis frame. The amplitude method suffers from distortions in the computation of the amplitude envelope itself for the polyphonic and non-stationary signals but performs well for the clean signal. The weighted bin-offset method is prone to lower recall and precision even for the clean signal due to increased distortion in the phase spectrum of the weaker amplitude harmonics [4]. The frequency error metrics for all methods are similar in the case of clean and polyphonic signals. For the vibrato signal however, this is higher for the weighted bin-offset method. Since the frequency error is only computed for detected harmonic sinusoids, this indicates that the phase-spectrum is more severely affected by frequency modulation relative to the magnitude spectrum.

Table I provided results only for selected window lengths. As discussed in Section I the choice of window length is expected to influence the reliability and accuracy of sinusoid detection and frequency estimation for the different signal conditions. In the following section we investigate possible performance gains from window-length adaptation based on measures of signal sparsity. In the interest of space, only the results obtained with the window main-lobe matching method are reported although the important trends hold across sinusoid identification methods.

III. Evaluation of Different Measures of Sparsity for Window Length Adaptation

In order to obtain the most accurate sinusoid detection, it is necessary to choose the window length so as to minimize the biasing of the computed sinusoidality measure due to the presence of pitch modulations and interfering components. These two non-idealities impose opposing constraints on the window length.

Signal-driven window-length adaptation has been previously used in audio coding algorithms, such as MPEG I and AAC, for discriminating between stationary and transient audio segments [17]. However, in the context of singing voice analyses, the only common signal-driven adaptive window-length analysis has been pitch-adaptive windowing based on previous detected pitch [18]. This loses its relevance in polyphonic music where multiple pitched instruments co-occur. Jones [19] used a kurtosis measure to adapt the window used in the computation of time-frequency (t-f) representations of non-stationary signals. However the evaluation was restricted to visual comparison of t-f representations of complicated signals. Goodwin [20] used adaptive time segmentation in a signal modeling and synthesis application. The method has a very high computational cost since the window adaptation is based on minimizing the actual reconstruction error between the original and synthesized signals.

In this section we investigate the use of some easily computable measures for automatically adapting window lengths to signal characteristics in the context of our application. Most of these measures have been previously proposed as indicators of signal sparsity [21]. Based on the hypothesis that a sparse short-time spectrum, with its more "concentrated" components, would facilitate the detection and estimation of the signal harmonics, we apply the different sparsity measures to the task of window adaptation. We review five different measures of signal sparsity tested in this work – L2 norm (L2), normalized kurtosis (KU), Gini Index (GI), spectral flatness (SF) and the Hoyer measure (HO). Of these SF has been widely used for driving window switching in audio coding algorithms. The performances of the different sparsity measures are evaluated in terms of the achieved accuracies of sinusoid detection and frequency estimation on the previously used set of simulated signals. Due to its superior performance as seen in Section II, only window main-lobe matching sinusoid detection is considered here.

A. Description of Different Measures

For the magnitude spectrum, $X(k)$ for frequency bin "k" of N bins, the definitions of different sparsity measures are given below.

1) ℓ^2 *norm* [21]

$$L2 = \sqrt{\sum_k X^2(k)} \tag{5}$$

2) Normalized Kurtosis [21]

$$KU = \frac{\frac{1}{N}\sum_k \left|X(k) - \overline{X}\right|^4}{\left(\frac{1}{N}\sum_k \left|X(k) - \overline{X}\right|^2\right)^2} \tag{6}$$

where $\overline{X}$ is the mean spectral magnitude value.

3) Gini Index [21]

The magnitude spectral coefficients $X(k)$ are first sorted in ascending order to give the ordered set $X_{(k)}$. The Gini Index is then given as

$$GI = 1 - 2\sum_k \left(\frac{X_{(k)}}{\|X\|_1}\left(\frac{N - k + 0.5}{N}\right)\right) \tag{7}$$

where $\|X\|_1$ is the ℓ^1 norm of $X(k)$.

4) Hoyer measure [21]

The Hoyer measure is a normalized version of $\frac{\ell^2}{\ell^1}$, and is defined as

$$HO = \left(\sqrt{N} - \frac{\sum_k X(k)}{\sqrt{\sum_k X^2(k)}} \right) \left(\sqrt{N} - 1 \right)^{-1} \quad (8)$$

5) Spectral Flatness

Spectral flatness has been used as a measure of tonality of a signal in perceptual audio coding [22]. Here we use it as an indicator of signal sparsity; the more peaky the spectrum of a signal, the more sparse it is. Spectral flatness is defined as the ratio of geometric mean of the power spectrum to the arithmetic mean of the power spectrum, and is given as

$$SF = \frac{\sqrt[N]{\prod_k X^2(k)}}{\frac{1}{N}\sum_k X^2(k)} \quad (9)$$

B. Window-length adaptation

Each of the previously mentioned sparsity measures is individually used in a window-length adaptation scheme described next. For each frame of audio we would like to apply that window-length that maximizes signal sparsity, anticipating that this would improve sinusoid detection. For a particular analysis time instant this amounts to selecting that window length among the set {23.2, 46.4 and 92.9 ms} that either maximizes KU, HO and GI or minimizes L2 and SF. As we expect increased signal non-stationarity at higher frequencies, we compute fixed and adapted window analyses separately across three frequency bands, viz. 0–1.5 kHz, 1–3 kHz and 2.5–4 kHz.

The implementation of the adaptive window representation in our evaluation involves the initial computation of the full-band spectral representation using each of the three window lengths. Note that the analysis time instants are fixed (at frame-centers) by the use of a fixed hop (10 ms). For all window lengths we use a fixed 2048 point DFT. For the 23.2 and 46.4 ms windows this involves zero-padding the windowed signal. Then for a given frequency band we compute a sparsity value from the frequency bins corresponding to the desired frequency range for each window-length representation. We select that window length that maximizes the sparsity measure for that frequency band and use it in the subsequent sinusoid detection step. The window main-lobe spectra for each of the three window lengths are pre-computed for use in the main-lobe matching method.

C. Evaluation and Results

We use the same evaluation metrics used in the previous section i.e. recall, precision and standard deviation of frequency error, and also the same simulated signals i.e. polyphony and vibrato cases. The evaluation metrics are computed for different fixed-window analyses (23.2, 46.4 and 92.9 ms), for a fixed multi-resolution (MR) analysis i.e. using progressively shorter windows for higher frequency bands, and for the different cases of sparsity-driven adaptive windowing.

The results for the above evaluation for the polyphonic and vibrato simulated signals are shown in Table II and Table III respectively. From these tables it can be seen that, rather than the multi-resolution analysis, the 92.9 ms and 23.2 ms fixed frame analyses consistently give the best performance across all bands for the polyphonic and vibrato simulations respectively. For the vibrato signal the disparity in performance across the different windows is more significant in the higher frequency band since the extent of non-stationarity in the signal is proportionately higher in this band. Of the different adaptive cases, the normalized kurtosis and Hoyer measure are observed to closely capture the longer-window superiority for the polyphonic signal and the shorter-window superiority for the vibrato signal across all frequency bands. A large difference in the performance of the sparsity measures is observed for the vibrato signal, especially in the highest frequency band.

TABLE II Performance of window main-lobe matching method (RE – Recall (%), PR – Precision (%), σ_{FE} – Frequency error (Hz)) for different fixed windows (23.2, 46.4, 92.9 ms & MR – multi-resolution) and sparsity measure (L2 norm, KU – Kurtosis, GI – Gini Index, SF – Spectral flatness and HO - Hoyer) driven adapted windows for simulated *polyphonic* signal.

Band	0-1.5 kHz			1-3 kHz			2.5-4 kHz		
	RE	PR	σ_{FE}	RE	PR	σ_{FE}	RE	PR	σ_{FE}
Fixed and Multi-resolution analysis									
23.2 ms	50.0	100.0	2.1	36.8	80.0	27.2	40.9	98.0	13.1
46.4 ms	100.0	100.0	0.4	78.1	100.0	23.7	62.7	99.1	3.4
92.9 ms	100.0	100.0	0.1	98.6	100.0	18.1	96.6	100.0	0.3
MR	100.0	100.0	0.1	78.1	100.0	23.7	40.9	98.0	13.1
Signal sparsity-driven adaptive windowing									
L2	100.0	100.0	0.1	98.6	100.0	18.1	96.6	100.0	0.3
KU	100.0	100.0	0.1	98.6	100.0	18.1	96.6	100.0	0.3
GI	100.0	100.0	0.1	97.4	100.0	18.1	89.4	100.0	0.8
SF	100.0	100.0	0.2	97.1	100.0	18.1	86.8	100.0	0.5
HO	100.0	100.0	0.1	97.5	100.0	18.1	92.8	100.0	0.3

TABLE III Performance of window main-lobe matching method (RE – Recall (%), PR – Precision (%), σ_{FE} – Frequency error (Hz)) for different fixed windows (23.2, 46.4, 92.9 ms & MR – multi-resolution) and sparsity measure (L2 norm, KU – Kurtosis, GI – Gini Index, SF – Spectral flatness and HO - Hoyer) driven adapted windows for simulated *vibrato* signal.

Band	0-1.5 kHz			1-3 kHz			2.5-4 kHz		
	RE	PR	σ_{FE}	RE	PR	σ_{FE}	RE	PR	σ_{FE}
Fixed and Multi-resolution analysis									
23.2 ms	97.1	100.0	1.4	90.0	97.9	6.4	82.7	98.0	8.1
46.4 ms	97.4	100.0	2.3	56.2	96.3	9.4	46.9	81.7	15.1
92.9 ms	64.8	100.0	6.2	54.0	86.7	17.8	48.8	52.9	18.9
MR	64.8	100.0	6.2	56.2	96.3	9.4	82.7	98.0	8.1
Signal sparsity-driven adaptive windowing									
L2	65.3	100.0	5.4	55.4	80.7	13.0	48.3	82.4	16.3
KU	96.7	100.0	2.7	86.5	94.3	7.0	73.0	91.4	9.2
GI	72.4	100.0	4.1	50.9	84.6	9.3	56.2	79.8	11.5
SF	89.3	95.8	3.3	77.3	91.2	7.1	47.5	81.6	15.3
HO	96.9	100.0	2.9	63.9	92.5	8.1	60.2	88.6	10.7

IV. EXPERIMENTS WITH REAL SIGNALS

From the results of the previous section it seems that using a signal sparsity-driven adaptive window analysis should lead to better sinusoid identification across varying signal conditions of polyphony and non-stationarity (in terms of pitch modulation) as compared to a multi-resolution approach. Since the singing voice is the dominant

source in vocal music, we expect that the above method should show good sinusoid identification performance for real music signals as well, which is investigated next.

A. Dataset Description

We use two datasets, sampled at 22.05 kHz, each of about 9.5 minutes duration of which the singing voice is present about 70 % of the time. The first dataset contains excerpts of polyphonic recordings of 9 Western pop songs of singers such as Mariah Carey and Whitney Houston, who are known for using extensive vibrato in their singing. The second dataset contains 5 Indian classical vocal music recordings. Indian classical singing is known to be replete with pitch inflections and ornaments, and the instrumental accompaniment comprises pitched instruments as well as percussion. The ground-truth vocal pitch is detected at 10 ms intervals throughout the singing segments using a semi-automatic melody extraction tool based on a state-of-the-art melody extraction algorithm [23], [24].

B. Evaluation and Results

We evaluate the performance of the window main-lobe matching based sinusoid detection method for fixed multi-resolution and adaptive window-length analysis for the real signals. In the evaluation we only compute recall using the expected harmonic locations computed from the ground-truth voice-pitch. Although we could compute the precision as well from the number of false positives for the real signals, this would not be indicative of sinusoid detection performance since there could be various simultaneously present accompanying musical instruments which also have harmonic spectra. Recall is only computed during active frames i.e. those for which the singing voice is present.

The results of the above evaluation for the Western pop and Indian classical datasets are presented in Fig. 2. Adaptive windowing improves upon the performance of the fixed multi-resolution analysis as seen from the increased recall. Overall it can be seen that the kurtosis-driven window adapted sinusoid detection gives better performance than any fixed or adaptive window analysis method across the datasets with the Hoyer measure closely following.

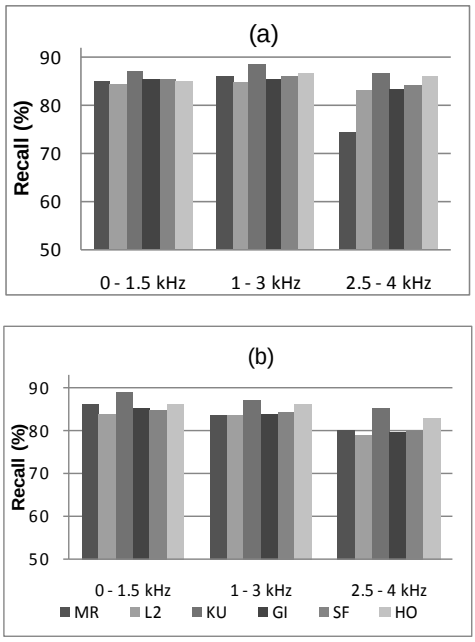

Fig.2. Performance of window main-lobe matching for multi-resolution (MR) and sparsity measures (L2 norm, KU – Kurtosis, GI – Gini Index, SF – Spectral flatness and HO – Hoyer measure) driven adapted windows for different frequency bands for (a) Western pop and (b) Indian classical data.

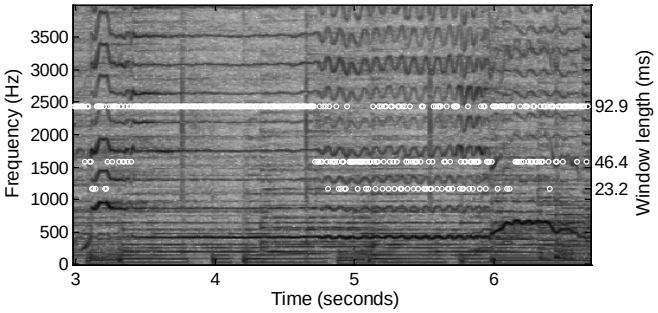

Fig. 3. Spectrogram of an excerpt of Whitney Houston's "I will always love you". White circles represent window choice (92.9, 46.4 or 23.2 ms) driven by maximization of kurtosis in the 2.5-4 kHz frequency band.

An example of the window adaptation using kurtosis for the highest frequency band is shown for an excerpt from the Western pop dataset in Fig. 3. Here it can be seen that during the stable note (between 3 and 5 sec) the measure is maximized for the longest window but during the vibrato region (between 5 and 6 sec) the measure frequently favors lower window lengths. Further, during vibrato the longer windows are selected in frames corresponding to the peaks and valleys of the vibrato cycle, and shorter windows are chosen during the vibrato mean crossings where the rate of frequency variation is highest.

V. Discussion and Conclusion

The observed performance improvements from sparsity driven window-length adaptation suggest that certain sparsity measures do indeed serve to usefully quantify spectrum shape deviation from that of an ideal sinusoid. A simple example, provided next, demonstrates the sensitivity of sparsity measures to signal non-stationarity. Consider linear chirp pure tones with fast and slow chirp rate. Let the slow rate equal to one-eighth the fast chirp rate, and both belong within the typical range of voice pitch modulations (e.g. vibrato). For each of the chirps we plot different sparsity measures (KU, GI and SF) versus window length, varying from 20 ms to 90 ms in steps of 10 ms, in Fig. 4. We see that all three sparsity measures show the intuitively expected concave form, attaining a single maximum at a finite window length which itself decreases as the chirp rate increases. We observe that KU is most sensitive to chirp rate. We have not plotted the HO and L2 measures since the former shows similar trends as KU and the latter does not show any sensitivity to changing chirp rates but continues to increase in value with window length. A closer inspection of the dependence of computed sparsity on spectrum shape revealed that the GI is affected by the shape of the main-lobe as well as the side-lobe roll-off whereas the KU reflects main-lobe spread mainly with the low amplitude side-lobes scarcely affecting the 4^{th} power average in Eq. (6). This explains, in part, the superiority of KU in the sinusoid detection context inspite of the general superiority of GI as a sparsity measure [21].

In summary, sparsity driven window-length adaptation consistently results in higher sinusoid detection rate and minimal frequency estimation error when compared with multi-resolution fixed-window analysis in the context of sinusoid detection of the singing voice in polyphonic music. Normalized kurtosis applied to the local magnitude spectrum is found to outperform alternate measures of signal sparsity such as the L2 norm, Gini Index, spectral flatness and Hoyer measure, for various signal conditions such as polyphony and non-stationarity (manifested as pitch vibrato) in simulated and real music signals. Another result of this work is that the window main-lobe matching sinusoid detection method outperforms an amplitude envelope and phase-based sinusoid detection method for the above signal conditions. Future work will

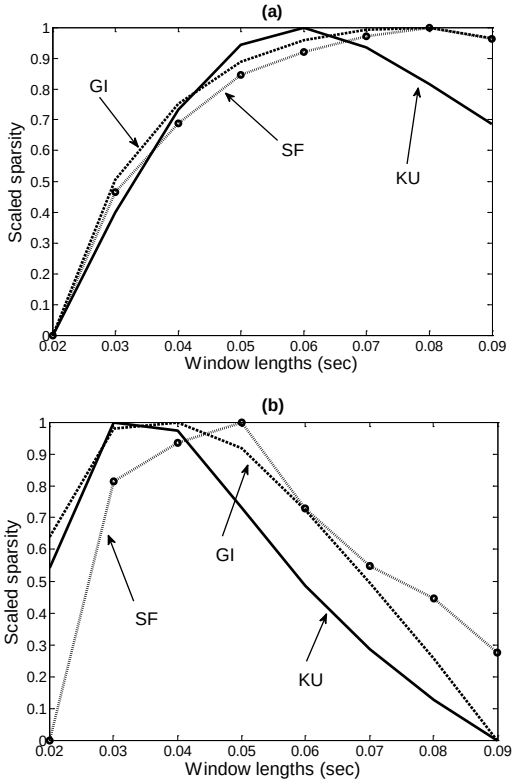

Fig. 4. Scaled sparsity values (KU, GI and SF) computed for different window lengths for a pure-tone chirp for (a) slow and (b) fast chirp rates.

investigate the extension of adaptive windowing to non-stationary sinusoidal modeling methods [8], [10] and to end applications such as singing voice detection and melody extraction.

APPENDIX

The implementation details and parameters for each of the sinusoid identification methods are given here. Two of the methods utilize the magnitude spectrum only while the third one exploits the phase spectrum properties of sinusoids. The inputs to all the methods is the magnitude spectrum $X(k)$ of the signal. All methods first search the short-time magnitude spectrum for 3-point local maxima to which they apply specific sinusoidality criteria.

A. *Amplitude Envelope Thresholding* [11]

This method involves finding a frequency dependent amplitude threshold. The amplitude envelope of the magnitude spectrum $X(k)$ is first obtained by convolving it with a Hamming window $H(k)$ in the frequency domain, as given below

$$A(k) = X(k) \otimes H(k) \tag{10}$$

where $H(k)$ is a normalized Hamming window of length $1+N/64$ frequency bins. Here N is the number of points in the DFT. The length of the Hamming window used for computing the amplitude envelope is suitably reduced when using shorter windows because the amount of smoothing required for computation of an accurate envelope is lesser for shorter window durations. Next $A(k)$ is flattened as follows

$$E(k) = \left(A(k)\right)^{c} \tag{11}$$

where c is a compression factor. Smaller values of c lead to a flatter envelope. The value $c = 0.8$ works well in our implementation. Then a threshold height is computed as

$$\eta = K \cdot \overline{X}^{(1-c)} \tag{12}$$

where $\overline{X}$ is the mean spectral amplitude and K is a constant (0.7). The final threshold is given as $M\,\eta\,E(k)$, where M is chosen such that the threshold is L dB below $\eta\,E(k)$. All local maxima in $X(k)$ above this final threshold value are labeled as detected sinusoids. The sinusoidal frequency estimate is refined by parabolic interpolation [7]. The value of M is varied to obtain different points on the precision-recall curve.

B. *Window Main-lobe Matching* [12]

This method is based on matching the main-lobe of the window transform to the spectral region around local maxima. The deviation of the ideal window main-lobe magnitude-spectrum shape $W(k)$, centered around the frequency-bin corresponding to a local maxima location in the magnitude spectrum $X(k)$, to the spectral region around this local maxima is computed as an error function, given as

$$\varepsilon = \sum_{a}^{b}\left[X(k) - |A|W(k)\right]^2 \quad \text{where } A = \frac{\sum_{a}^{b} X(k)W(k)}{\sum_{a}^{b} W^2(k)} \tag{13}$$

Here A is a scaling factor that minimizes ε and $[a, b]$ is the interval of the main-lobe width around the local maximum. This error is normalized with the signal energy as follows

$$\xi = \frac{\varepsilon}{\sum_{a}^{b} X^2(k)} \tag{14}$$

The sinusoidality criterion, in this case a measure of the closeness of shape of the detected peak and the ideal main-lobe, is now defined as $S = 1- \xi$. Local maxima for which S lies above a predefined threshold are marked as sinusoids. Note that the shape of the ideal main lobe $W(k)$ changes with change in window length. The sinusoid frequency estimate is refined by parabolic interpolation [7]. The threshold value on S is varied to obtain different points on the precision-recall curve.

C. *Weighted Bin Offset Method* [4]

This method applies thresholds to the bin offset κ, which is the deviation of the sinusoid's instantaneous frequency (IF) from the bin frequency of the local maxima. The bin offset at bin k is given by

$$\kappa(k) = \frac{N}{2\pi L}\, princ\arg\left[\phi_l(k) - \phi_{l-1}(k) - \frac{2\pi L}{N}k\right] \tag{15}$$

where $\phi_l(k)$ is the phase spectrum of the l^{th} frame, N is the number of DFT points, L is the hop length and princarg maps the phase to the $\pm\pi$ range. Local maxima are marked as detected sinusoids if

$$\kappa(k) < 0.7R \;,\; \left|\kappa(k) - \kappa(k+1) - 1\right| < 0.4 \cdot \frac{A_{peak}}{X(k+1)} \quad \text{and}$$
$$\left|\kappa(k) - \kappa(k-1) + 1\right| < 0.4 \cdot \frac{A_{peak}}{X(k-1)} \tag{16}$$

where A_{peak} is the instantaneous magnitude of the local maxima, which is computed by applying bin-offset correction to the window transform. The value of R is varied to obtain different points on the precision-recall curve.

The bin-offset value is used to refine the sinusoidal frequency estimate $f(k)$, for sampling frequency f_S, using

$$f(k) = (k + \kappa(k))\frac{f_S}{N} \tag{17}$$

REFERENCES

[1] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[2] Z. Duan, Y. Zhang and C. Zhang, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 16, no. 4, pp. 766–778, May 2008.

[3] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[4] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. 9th Intl. Conf. on Digital Audio Effects (DAFx-06)*, pp. 247–252, Montreal, 2006.

[5] R. Badeau, G. Richard and B. David, "Fast and stable YAST algorithm for principal and minor subspace tracking," *IEEE Trans. Sig. Process.*, vol. 56, no. 8, pp. 3437–3446, Aug. 2008.

[6] M. Christensen, P. Stoica, A. Jakobsson and S. Jensen, "Multi-pitch estimation," *Sig. Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[7] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. 5th Intl. Conf. on Digital Audio Effects (DAFx-02)*, pp. 51–58, Hamburg, Germany, 2002.

[8] M. Betser, P. Collen, G. Richard and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Trans. Sig. Process.*, vol. 56, no. 2, pp. 505–517, Feb. 2008.

[9] J. Wells and D. Murphy, "A comparative evaluation of techniques for single-frame discrimination of non-stationary sinusoids," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 498–508, Mar. 2010.

[10] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *Proc. 11th Intl. Conf. on Digital Audio Effects (DAFx-08)*, pp. 281–288, Espoo, Finland, Sept. 2008.

[11] M.R. Every, "Separation of musical sources and structure from single-channel polyphonic recordings," *Ph. D. dissertation*, Dept. Electron., Univ. York, York, U.K., 2006.

[12] D. Griffin and J. Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech and Sig. Process.*, vol. 36, no. 8, pp. 1223–1235, 1988.

[13] M. Lagrange, S. Marchand and J-B Rault, "Sinusoidal parameter extraction and component selection in a non-stationary model," in *Proc. 5th Intl. Conf. on Digital Audio Effects (DAFx-02)*, pp. 59–64, Hamburg, Germany, Sep. 2006.

[14] M. Lagrange and S. Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods, *J. Audio Engg. Soc.*, vol. 55, no. 5, pp. 385–399, 2007.

[15] M. Slaney, "The auditory toolbox," Interval Research Corporation, Tech. Rep. #1998-010, 1998.

[16] J. Sundberg, "A rhapsody on perception," *The Science of the Singing Voice*, Northern Illinois University Press, 1987.

[17] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. of the IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[18] K. Kim and I. Hwang, "A multi-resolution sinusoidal model using adaptive analysis frame," in *Proc. EUSIPCO*, 2004.

[19] D. Jones and T. Parks, "A high-resolution data-adaptive time-frequency representation," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 38, no. 12, pp. 2127–2135, Dec. 1990.

[20] M. Goodwin, "Adaptive signal models: Theory, algorithms and audio applications," *Ph. D. dissertation*, MIT, 1997.

[21] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Info. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[22] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[23] S. Pant, V. Rao and P. Rao, "A melody detection user interface for polyphonic music," in *Proc. Ntnl. Conf. Communications*, Chennai, India, Jan. 2010.

[24] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment for polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2145–2154, Nov. 2010.