Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

Evaluation of Threshold-Based Algorithms for Detection of Spectral Peaks in Audio

Leonardo de O. Nunes,[1] Paulo A. A. Esquef,[2] and Luiz W. P. Biscainho[1]

[1] LPS - PEE/COPPE, UFRJ
Caixa Postal 68504 - Rio de Janeiro, RJ, 21941-972, Brazil

[2] Instituto Nokia de Tecnologia
Av. Torquato Tapajós, 7200 - Manaus, AM, 69048-660, Brazil

lonnes@lps.ufrj.br, paulo@indt.org.br, wagner@lps.ufrj.br

ABSTRACT

Peak picking algorithms play an important role in audio analysis and synthesis methods based on sinusoidal modelling. Peak detection aims at selecting only those peaks corresponding to genuine resonant components present in the signal, while rejecting noise-induced ones. This paper investigates the performance of four threshold-based schemes for peak detection: two-pass split window, low-order autoregressive model, a nonlinear recursive filter and a stochastic spectrum estimator. The results of performance comparisons measured under a common metric are reported for a set of specific experimental setups.

0 INTRODUCTION

Sinusoidal modelling (SM) represents an audio signal as a sum of amplitude and frequency (phase) modulated sinusoids. SM finds use in a variety of audio-related applications, such as speech and audio synthesis, automatic transcription of music, and audio coding. First introduced for speech analysis in [1] and for audio signals in [2] SM has been expanded [3] and modified [4] to suit its different applications.

This paper considers the classic analysis method proposed in [1, 2]. A key-task to carry out a successful SM analysis is the so-called spectral peak picking. In this stage an algorithm has to decide whether a certain spectral peak corresponds to a genuine resonance in the input signal or is a spurious occurrence. The latter can happen due to either intrinsic characteristics of the spectral analysis or additive noise in the input signal.

A common peak picking approach considers an energy-based selection criterion in the frequency domain. For instance, spectral peaks whose energy exceeds a certain threshold can be segregated as genuine peak occurrences.

This paper evaluates a selection of methods that yield frequency dependent threshold curves, within a

peak detection system. A suitable test environment, composed by a test signal generator, a configurable peak detection system, and a performance meter, is created. The objective assessment on how well the investigated methods work can be obtained through performance measurements taken over the peak detection system.

After this introduction, the paper briefly outlines the sinusoidal analysis system studied. Then, it describes the adopted peak picking strategy and its possible processing configurations. In the sequel, the paper explains the structure and setup of the experimental system devised to assess the performance of the tested methods. Finally, the measured performance indicators are presented and discussed.

1 SINUSOIDAL MODELLING

Sinusoidal modelling describes an audio signal $x(t)$ as a sum of L sinusoids, possibly modulated in amplitude- and (phase- or) frequency

$$x(t) = \sum_{l=1}^{L} A_l(t) \sin \Psi_l(t), \tag{1}$$

$$\Psi_l(t) = \Psi_l(0) + \int_0^t \omega_l(u)du. \tag{2}$$

The continuous nature of the amplitude $A_l(t)$ and angular frequency $\omega_l(t)$ leads to a computationally intractable problem. In order to simplify the analysis, Eq. (1) is commonly replaced by a discrete model

$$x[n] = \sum_{l=1}^{L} A_l[n] \sin \Psi_l[n], \tag{3}$$

which can be further considered short-time stationary in amplitude and frequency. That is, for a given partial l, and assuming that $A_l[n]$ and $\Psi_l[n]$ possess much narrower bandwidth than that of the signal under analysis, the approximations $A_l[n] \approx A_l$ and $\Psi_l[n] \approx \Omega_l n + \Psi_l[0]$, where A_l and Ω_l are constant values, hold true during a time interval of N samples.

Fig. 1 illustrates the typical processing stages involved in the analysis part of a sinusoidal modelling system [5].

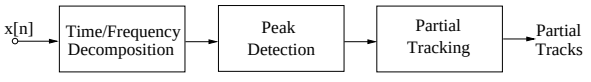

Figure 1: The three essential processing stages of a sinusoidal analysis system.

The 'time / frequency decomposition' stage performs a discrete-time Short Time Fourier Transform (STFT) [3] on the audio signal, i.e.,

$$X[k, m] = \text{STFT}(x[n, m]) \tag{4}$$

$$= \frac{1}{N} \sum_{n=0}^{N-1} w[n]x[n + mH]e^{-jk\frac{2\pi}{N}n}, \tag{5}$$

where $w[n]$ is a window function of length N, e.g., the Hamming window, k is the frequency bin index, m denotes the analysis frame index, and H represents the hop size of the analysis frame along time. The output of the first stage is the magnitude spectrum of $x[n, m]$, which is defined by $S[k, m] = |X[k, m]|$.

The 'peak detection' stage receives $S[k, m]$ as input and selects only those peaks that correspond to stationary sinusoidal components of the signal present in frame m. The amplitude and frequency of the detected peaks can be finely estimated by a number of dedicated methods [2, 6]. Finally, a 'partial tracking' procedure is responsible for matching peaks across consecutive frames, in order to form the so-called partial tracks. These tracks contain information regarding the life-cycle of the sinusoidal components present in the signal under analysis.

Since the computational complexity of the 'partial tracking' algorithm increases with the number of selected peaks, the 'peak detection' stage should output only the most prominent spectral peaks. Another reason for a more judicious selection of spectral peaks is to avoid the formation of too short-living partial tracks due to noise-induced peaks. This way, a more precise model for the sinusoidal components of the signal can be obtained. Next section describes the peak picking strategy adopted in this work.

2 PEAK PICKING STRATEGY

Peak picking in audio spectrum can be carried out by several means, such as via threshold-based techniques [1, 2], analysis-by-synthesis schemes [4], and model-based analysis [7].

The present research focuses on threshold-based peak picking approaches. For that purpose, a general peak picking strategy is employed. It can be divided into two main parts: a spectral pre-processing stage and a peak selection algorithm, as depicted in Fig. 2. The latter will be fixed for all test cases investigated hereafter. On the contrary, the pre-processing stage is the variable element of the peak detection system: the system performance will be assessed w.r.t four different methods for spectral pre-processing. All these methods, as well as the peak selection criterion, are described in the subsequent sections.

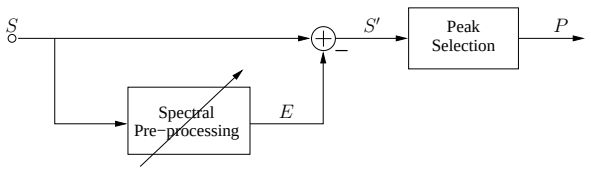

Figure 2: Block diagram illustrating the strategy adopted in the peak detection system.

2.1 Spectral Pre-Processing

2.1.1 Motivation

In threshold-based schemes, the usual selection criterion consists of choosing those spectral peaks that ex-

ceed a pre-defined energy level. While simple, such criterion is ineffective when dealing with general audio signals, which typically exhibit a spectral pattern whose energy decreases with frequency [8]. As a result, genuine low-energy spectral peaks in the high-frequency range may be discarded as spurious occurrences, as seen in Fig. 3.

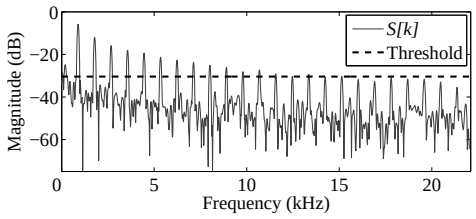

Figure 3: Inadequacy of applying a constant threshold to an observed spectrum of a noisy harmonic signal.

The pre-processing stage attempts to overcome the aforementioned difficulty compensating the observed spectral tilt. In a perhaps oversimplified manner, the spectral tilt can be compensated by an estimate of the colouring profile associated with the additive background noise that corrupts the signal of interest. Alternatively, an equally plain solution resorts to an estimate of the spectral envelope.

The ensuing sections describe four methods that have been designed to provide estimates for both the spectral envelope and the noise spectral profile. Either way, the desired estimates are referred to as $E[k]$.

2.2 Spectral Tilt Estimators

2.2.1 a) Two-Pass Split Window

The TPSW filtering has been originally proposed for noise spectrum estimation in sonar systems [9]. The procedure consists of three stages. In the first stage the sampled magnitude spectrum $S[k]$ of a given input signal is filtered through a split window defined as [9]

$$h^{\text{sw}}[n] = \begin{cases} 0, & |n| < M^{\text{sw}} \\ 1, & M^{\text{sw}} \leq |n| < N^{\text{sw}}, \end{cases} \tag{6}$$

with $0 \leq M^{\text{sw}} < N^{\text{sw}}$, being N^{sw} and M^{sw} positive integers that control, respectively, the length of the split window and the gap size.

The output of the first stage $\tilde{S}[k]$ is then modified according to the following criterion

$$\hat{S}[k] = \begin{cases} S[k], & \text{if } S[k] \leq \alpha\tilde{S}[k] \\ \tilde{S}[k], & \text{if } S[k] > \alpha\tilde{S}[k], \end{cases} \tag{7}$$

where $\alpha \geq 1$ is a parameter related to the ability to reject peaks in the observed spectrum.

In final stage, the modified spectrum $\hat{S}[k]$, which is supposed to be free of prominent peaks, is filtered through a conventional moving average filter, with the same length as that of the split window. The output of this third stage corresponds to the desired estimate, $E^{\text{tpsw}}[k]$. Note that the FIR filters used in first and third stages are normalized for unity DC gain.

In the TPSW procedure, $S[k]$ is considered within the range between from 0 to π. In order to avoid boundary effects during the filtering, it is extended by about 20% of its original size. For that, part of the spectrum at both extremities of $S[k]$ is mirrored. Moreover, filtering delays are compensated in order to guarantee the synchronism between input and output. This is accomplished by taking only the central part of the correlation from the filtering results. In the end, the extensions appended to $S[k]$ are discarded in order to restore its original size.

When it comes to the choice of the split-window parameters, the smoothness of the estimate increases with N^{sw}. The value of M^{sw} should be set appropriately chosen as to yield the window gap as large as the bandwidth of a typical peak present in $S[k]$. The value of α is related to the peak rejection capability. It should be chosen small enough to guarantee that the component $\alpha\tilde{S}[k]$ is below the average magnitude of the peaks and large enough to place the component $\alpha\tilde{S}[k]$ above the noise floor. Typically, choosing $2 \leq \alpha \leq 8$ provides satisfactory results.

2.2.2 b) Low-Order AR Estimation

Autoregressive (AR) models are widely used in audio signal processing, such as vocal tract estimation in speech processing [10, 11].

Here, the idea is to take the spectral *envelope* estimate of a given audio signal for the desired spectral tilt curve. Thus, the procedure consists of fitting a low-order AR model to the time-domain signal $x[n]$ associated with $S[k]$. The desired envelope $E^{\text{ar}}[k]$ is the magnitude spectrum of the estimated AR model.

In mathematical terms, a relaxed assumption is that $s[k]$ is governed by the following AR model

$$x[n] = \sum_{u=1}^{q} a[u]x[n-u] + r[n], \tag{8}$$

where q is the (insufficient) model order, $a[u]$ are the model coefficients, and $r[n]$ is the modeling error.

The model parameters can be estimated via any standard AR estimator, such as the Yule-Walker and Burg methods. Once the model $A(z) = \left[1 - \sum_{u=1}^{q} a_u z^{-u}\right]^{-1}$ is available, the desired spectral envelope is obtained by $E^{\text{ar}}[k] = |A(e^{j\omega_k})|$, where $\omega_k = \frac{2\pi k}{N}$, with N being the length of DFT buffer used in the spectral analysis.

2.2.3 c) Nonlinear Recursive Smoothing Filter

In [12] a nonlinear recursive smoothing filter (NRSF) is proposed for estimating the spectral profile of coloured noise in audio spectra. The filter, devised under the assumption that the power spectrum density of the noise component would vary slowly over frequency, works by limiting in modulus the first derivative (or slew rate) of the spectrum samples, with respect

to frequency. The nonlinear recursive filter that implements the solution is given by

$$E^{\text{nrsf}}[k] = E^{\text{nrsf}}[k-1]\beta^{\text{sign}\left(S[k]-E^{\text{nrsf}}[k-1]\right)}, \qquad (9)$$

where $S[k]$ is the observed magnitude spectrum of the signal under test; $E^{\text{nrsf}}[k]$ is the desired spectrum estimate; $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign}(x) = -1$ if $x < 0$; and β is a constant slightly larger than unity.

The parameter β can be expressed as $\beta = 1 + \lambda$. In theory, λ should be chosen as to exceed the maximum slew rate associated with the power density spectrum of the noise component. In practice, as can be seen from Eq. (9), the value of λ controls the forgetting factor of the filter. Thus, the larger λ, the higher the variance of $E^{\text{nrsf}}[k]$ becomes. Choosing $\lambda = 0.05$ provides a sufficiently smooth $E^{\text{nrsf}}[k]$ [12].

When dealing with recursive schemes, filter initialization should be considered carefully. An improper initialization can bias the initial samples of $E^{\text{nrsf}}[k]$ and thus degrade the overall performance of the algorithm. One possible solution is to extend $S[k]$ at the boundaries, as described earlier in Section 2.1.2.a, and initialize the recursion with $E^{\text{nrsf}}[k-1] = 0$. Of course, the spectrum extension should be long enough for the influence of a wrong initialization be mitigated. Afterwards, the estimated values of $E^{\text{nrsf}}[k]$ outside the original range of $S[k]$ are discarded. Alternatively, $E^{\text{nrsf}}[k-1]$ can be set as the median value among the first C samples of $S[k]$. In both cases, the original frequency range of $S[k]$ is from 0 to π.

2.2.4 d) Stochastic Spectrum Estimation

The stochastic spectrum estimator (SSE) is another non-linear stochastic spectrum estimator that has been introduced by [13].

Assuming $S[k]$ as defined before, the SSE method consists of the four steps described below:

1. Pass $S[k]$ through a three-tap moving average filter, in order to obtain a $S^1[k]$ possibly free of null magnitude samples;
2. Compute $R[k] = \frac{1}{S^1[k]}$;
3. Obtain $R^1[k]$ as a smoothed out version of $R[k]$, by computing a cyclic convolution between $R[k]$ and an N^{sse}-tap moving average FIR filter;
4. Compute the desired estimate as $E^{\text{sse}}[k] = \frac{1}{R^1[k]}$.

As with the TPSW method, the smoothness of $E^{\text{sse}}[k]$ increases with the value of N^{sse}. Moreover, all FIR filters used in the procedure are normalized to force the DC gain equal to 0 dB. Differently from the TPSW filtering, in the SSE scheme $S[k]$ should be considered within the whole range between 0 and 2π, in order to make effective the use of the cyclic convolution. It should be noticed, however, that one could employ $S[k]$ within the range between 0 and π, provided that this $S[k]$ is sufficiently extended at both extremities via the mirroring scheme mentioned in Section 2.1.2a.

2.2.5 Performance Comparison

In order to illustrate the qualitative performance of the previously described methods, they are tried over a test-signal, under the processing setups detailed below.

- Input signal: 2048 samples of a noisy harmonic signal with fundamental frequency equal to 1 kHz, sampled at 44.1 kHz, and windowed by a 2048-sample Hann window;
- S[k]: magnitude of a 2048-point DFT applied to the input signal;
- TPSW filtering: $N^{\text{sw}} = 51$, $M^{\text{sw}} = 8$, and $\alpha = 4$;
- AR method: $q = 10$;
- SSE method: $N^{\text{sse}} = 101$;
- NLRF method: $\beta = 1.01$.

The attained results are summarized in Fig. 4. It can be observed that all methods succeed in catching the overall shape of the spectrum. It is also worth noticing that, with exception of $E^{\text{ar}}[k]$, the remaining spectral envelopes follow closely the local average of the noise floor. The suspended $E^{\text{ar}}[k]$ is nothing to worry about, since those estimates will not be used directly as variable magnitude thresholds for the reference spectrum. Rather, they will serve to compensate the spectral tilt of the reference spectrum, before proceeding to the peak detection stage.

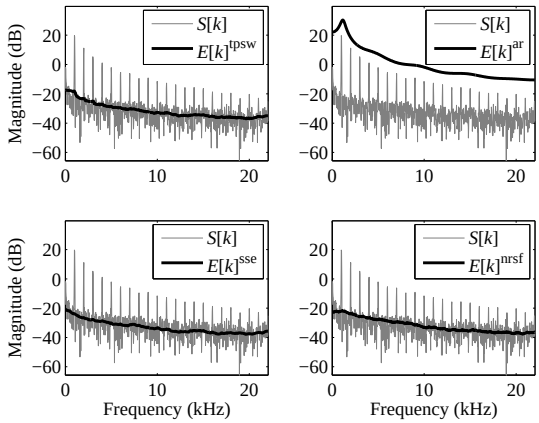

Figure 4: Qualitative performance comparison among the methods under evaluation.

2.3 Peak Selection Criterion

As depicted in Fig. 2, peak selection is performed over $S'[k]$, which is a tilt-compensated version of $S[k]$. Note that the compensation is carried out in the logarithmic scale. Fig. 5 compares an original $S[k]$ and its tilt-compensated version $S'[k]$, in an example where $E^{\text{sse}}[k]$ was used. Now, it is clear that a constant magnitude threshold can be employed to discriminate genuine from spurious spectral peaks.

As regards the adopted peak selection criterion, assume first the set $k \in \{2,\ 3, \ldots,\ (N/2-1)\}$ of $S'[k]$ bin indices. For all elements of k, collect in a sub-set $\mathcal{P}_m$ the indices k_{peak} that satisfy simultaneously the following conditions:

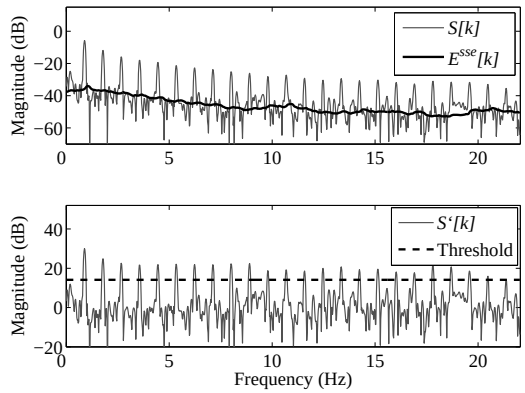

Figure 5: Comparison between the original (top) and tilt-compensated spectra (bottom).

1. $S'[k] > S'[k-1]$;
2. $S'[k] > S'[k+1]$;
3. $S'[k] > d\mu$;

In condition 3, d is an empirically chosen multiplier and μ is an estimate of the standard deviation of the noise observed in $S'[k]$. The selected indices k_{peak} contain the desired bin indices associated with the detected peaks in $S'[k]$.

The first two conditions above define which samples of $S'[k]$ could be qualified as peaks, regardless of being genuine or spurious. The third condition sets the value of the discrimination threshold, which sets the minimum energy a peak should possess to be qualified as a genuine peak.

The value of μ can be obtained by any estimator robust enough to provide reliable estimates for the standard deviation of data series, despite the presence of outliers. It was found experimentally that, when dealing with spectra which are densely populated by genuine peaks (outliers here), the median operator tends to over-estimate the standard deviation of the noise. Fortunately, with exception of the AR-based technique, the very spectrum pre-processing methods presented in Section 2.1.2 are competitive alternatives to aid the estimation of the noise standard deviation.

Among the available options for computing μ, the SSE method was found to be the least affected by the presence of genuine peaks. This is justified not only by the SSE's own formulation, but also experimentally. Thus, the adopted solution was to estimate μ as

$$\mu = \text{mean}(\bar{E}^{sse}[k]), \tag{10}$$

where $\bar{E}^{sse}[k]$ is the curve output by the SSE method to any tilt-compensated spectrum $S'[k]$.

As for the value of d, one assumes that the pre-processing was successful in 'whitening' the noise component, which can be considered Gaussian. Moreover, μ is believed to be a reliable estimate for the standard deviation of the noise, as observed in the frequency domain. In such case, setting $2 \leq d \leq 5$, assures a confidence interval greater than 95% that the spurious peaks will fall below the adopted threshold.

In reality, for a given signal-to-noise ratio (SNR), the larger the number of genuine peaks present in the signal, the less they tend to stand out from the noise floor, due to energy sharing among peaks. This favors the occurrence of detection errors and requires a more careful selection of d. On the contrary, placing the selection threshold is easier when dealing with a few genuine spectral peaks, even for low SNRs.

The aforementioned condition motivates the following strategy to set the value of d.

1. Calibrate d in order to assure a satisfactory detection performance considering a scenario with SNR as low as 10 dB and a spectrum densely populated with genuine peaks;
2. Attribute the previously found value of d to $d_{\min}$;
3. Compute $\rho = 1 \sqcap \left(\frac{\max(S'[k])-\mu}{10}\right)^{0.5}$, where $a \sqcap b$ stands for 'maximum between a and b';
4. Make $d = \rho d_{\min}$.

Although adequate for cases with many spectral peaks, the value of $d_{\min}$ tends to be too low for $S'[k]$ bearing few peaks of interest. As a consequence, the occurrence of false detections is favored.

In the computation ρ, the quantity $(\max(S'[k]) - \mu)$ can interpreted as the available room in magnitude between the spectral maxima and the average level of the noise floor. Thus, if this room is larger than 10 dB, the multiplier $\rho > 1$ contributes to raise the threshold by about half the magnitude room excess. Otherwise, $\rho = 0$ and $d = d_{\min}$. The former situation is likely to happen when there are few genuine peaks in $S'[k]$, avoiding the occurrence of false alarms.

3 SIMULATION SETUP

This section describes the experimental setup used to evaluate how well the peak detection system performs. As depicted in Fig. 6, the test setup consists of a signal generator, whose output $x[n]$ is fed to a STFT analyser, which in its turn provides the magnitude spectrum $S[k]$ to the peak detection system (see Fig. 2). Moreover, a performance meter assesses the peak detection results in quantitative terms, by counting how many of the selected peaks are in fact correctly detected, according to a reference indicator.

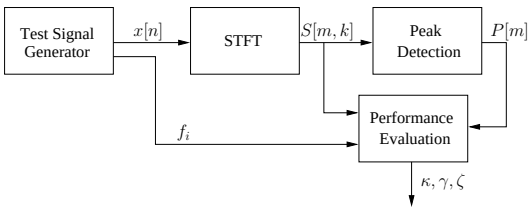

Figure 6: Setup devised to assess the performance of the peak detection system.

3.1 3.1 Test Signal Generation

The test signal, sampled at 44.1 kHz, is composed of two components: one deterministic and another stochastic. The former consists of a sum of harmonic signals with fundamental frequencies $f_{0,i}$. For each $f_{0,i}$, which is chosen randomly between 200 Hz and 1 kHz, with uniform probability, harmonics are generated up to the Nyquist frequency. Moreover, the amplitude of the harmonics can be either constant or decaying over frequency.

As for the zero-mean additive stochastic component, it can be either white or pink noise. In mathematical terms the test signal is defined as

$$x[n] = \sum_{i=0}^{I} \sum_{j=1}^{J_i} a_{ij} \cos\left(\frac{2\pi j f_{0,i} n}{F_s}\right) + \sigma\eta[n]. \qquad (11)$$

The double summation in Eq. (11) represents the deterministic part of the signal, where F_s is the sampling frequency, I is the number of fundamental frequencies in the mixture, and $J_i = \left\lfloor \frac{F_s}{2f_{0,i}} \right\rfloor$, with $\lfloor . \rfloor$ denoting 'the greatest integer less than or equal to', defines the number of harmonics associated with a given $f_{0,i}$. The amplitude of the harmonics a_{ij} is set either to unity or decreases with frequency according to $a_{ij} = \frac{1}{jf_{0,i}}$.

The second part of Eq. (11) represents the noise component, being $\eta[n]$ with $0 \leq n \leq (N-1)$, one realization of a stochastic process having constant or $1/f$ power spectrum density, and σ a variable that controls the noise power in order to force a desired SNR.

3.2 Peak Detection Setup

The computation of $S[k,m]$ and $S'[k,m]$ is identical to the procedures and parameters described as in Section 2.1.3. Those processing parameters were empirically tuned across the different methods as to render fair the comparisons among their performance.

As regards the peak selection algorithm, the main parameter to be set is the multiplier $d_{\min}$ of the noise standard deviation. It was found that $d_{\min} = 2$ is a suitable choice. The length of the second SSE filter was set to $N^{\text{sse}} = 150$.

3.3 Performance Evaluation

Peak detection performance is assessed basically by means of counting the number of correctly detected peaks and that of false alarms. Here, the main issue to consider is the observed *presence* of a given peak, regardless of whether it was detected with precisely estimated magnitude and frequency. In any case, it is necessary to define the conditions upon which a peak can be considered correctly detected.

On the measurement side, the peak observation domain is $S[k,m]$. Thus, the frequency of any observed peak can only lie in one of the available frequency bins, i.e., kF_s/N, for $0 \leq k < N/2$. On the reference side, the frequencies of the peaks in the test signal can be set in a frequency grid as fine as desirable. In order to ensure a meaningful peak detection performance assessment, the reference domain should be made compatible to the measurement domain. The following sections examine the issue in more detail.

3.3.1 3.3.1 Reference Domain Alignment

First, let the set Φ contain all the frequencies ϕ_i associated with the deterministic part of the input signal. The primary goal is to foresee which bins in the observation domain would be more strongly activated by those frequencies ϕ_i. Thus, each element ϕ_i is quantized to the nearest frequency bin, i.e., $\bar{\phi}_i = k_i F_s/N$, for $k_i = \text{round}(\phi_i N/F_s)$, where N is the length of the STFT analysis buffer and F_s the sampling frequency.

Gathering all non-repeated occurrences of k_i in a set $\mathcal{K}$, one can define a preliminary reference vector, whose elements are defined as

$$r[k] = \begin{cases} 1, & \text{if } k \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}, \text{ for } 0 \leq k < \frac{N}{2}. \qquad (12)$$

Note, however, that $r[k] = 1$ may not necessarily coincide with observed peaks of any kind in $S[k,m]$. This is because the position of an observed local maximum in $S[k,m]$, due to a given ϕ_i, is influenced not only by the corrupting noise, but also by the presence of the other ϕ_i.

The key-point here is to build a reference mask that corresponds to peak occurrences truly observable in the measurement domain. It is then evident that frequency quantization alone is an insufficient criterion to align the reference and measurement domains.

A more adequate domain alignment can be created by means of an auxiliary binary vector, the elements of which are defined as

$$c[k] = \begin{cases} 1, & \text{if } k \in O_m \\ 0, & \text{otherwise} \end{cases}, \text{ for } 0 \leq k < \frac{N}{2}, \qquad (13)$$

where O_m is a set with cardinality $|O_m|$ containing all bin indices associated with *observed* peaks in $S[k,m]$, either genuine or spurious. Then, the aligned reference vector is obtained by

$$r_a[k] = r_0[k] \oplus r_{-1}[k] \oplus r_{+1}[k], \qquad (14)$$

with $r_0[k] = r[k] \wedge c[k]$, $r_{-1}[k] = r[k-1] \wedge c[k]$, and $r_{+1}[k] = r[k+1] \wedge c[k]$, for $1 \leq k < (N/2 - 1)$, where the symbols $\wedge$ and $\oplus$ stand for the boolean operations 'AND' and 'XOR', respectively.

The parts of $r_a[k]$ can be interpreted as follows:

- $r_0[k]$ nullifies $r[k]$ if $c[k] = 0$, i.e., when active bins in the reference are not observed peaks;
- $r_{-1}[k]$ moves an active bin in $r[k]$ to the adjacent one on the left side if $c[k] = 0$ but $c[k-1] = 1$;
- $r_{+1}[k]$ moves an active bin in $r[k]$ to the adjacent one on the right side if $c[k] = 0$ but $c[k+1] = 1$.

Table 1: Performance indicators obtained from **Test 1**.

	SNR (dB)														
	0	5	10	15	20	0	5	10	15	20	0	5	10	15	20
TPSW	85.6	98.9	99.3	99.5	99.6	17.5	4.6	0.5	0.3	0.2	0.68	0.94	0.99	0.99	0.99
AR	88.9	99.7	99.8	99.8	99.8	22.5	13.4	1.4	0.2	0.1	0.66	0.86	0.98	1.00	1.00
NRSF	83.3	97.1	99.2	99.3	99.2	15.0	1.7	0.7	0.4	0.29	0.68	0.95	0.99	0.99	0.99
SSE	85.6	97.9	99.2	99.3	99.3	18.1	2.5	0.7	0.5	0.3	0.68	0.95	0.99	0.99	0.99
No method	38.3	63.0	84.9	93.6	98.5	6.3	2.3	0.9	0.4	0.2	0.32	0.61	0.84	0.93	0.98
Metric	γ					ζ					κ				

Table 2: Performance indicators obtained from **Test 2**.

	number of $f_{0,i}$ (I)														
	3	4	5	6	7	3	4	5	6	7	3	4	5	6	7
TPSW	97.2	91.6	85.7	78.0	71.0	3.0	3.5	3.6	3.8	4.0	0.94	0.88	0.82	0.74	0.67
AR	98.3	92.5	86.4	78.4	71.4	4.5	4.3	3.9	3.7	3.5	0.94	0.88	0.83	0.75	0.68
NRSF	91.4	84.6	79.4	72.5	66.4	2.1	3.6	4.5	5.3	5.8	0.89	0.81	0.75	0.67	0.61
SSE	94.3	88.5	82.9	75.5	69.0	2.0	2.9	3.3	3.8	4.1	0.92	0.86	0.80	0.72	0.65
No method	48.4	38.2	33.1	25.6	24.0	2.2	2.8	3.3	3.9	4.5	0.46	0.35	0.30	0.25	0.21
Metric	γ					ζ					κ				

In practice, the indices k for which $r_a[k] = 1$ indicate the reference spectral locations against which to confront those of the detected peaks in $S[k,m]$. Once $r_a[k]$ is obtained, the percentage of correctly detected peaks can be computed as

$$\gamma = \frac{G}{Q}, \tag{15}$$

where $Q = \sum_{k=0}^{\frac{N}{2}} r_a[k]$ is the count of all peaks in the reference vector and $G = \sum_{k=0}^{\frac{N}{2}} g[k]$ is the count of all correctly detected peaks, with

$$g[k] = \begin{cases} r_a[k], & \text{if } k \in \mathcal{P}_m \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

where $\mathcal{P}_m$ is a set of cardinality $|\mathcal{P}_m|$ containing the indices of all detected peaks from $S'[k,m]$.

The percentage of false alarms can be computed as

$$\zeta = \frac{|\mathcal{P}_m| - G}{|\mathcal{O}_m| - Q}, \tag{17}$$

where $|\mathcal{P}_m| - G$ is the count of all incorrectly detected peaks and $|\mathcal{O}_m| - Q$ is the count of all observable peaks in $S'[k,m]$ that should have remained undetected.

In addition, one can define the meter

$$\kappa = \gamma - \zeta, \tag{18}$$

which aggregates the two previous measures into a single performance indicator. Ideally, $-1 \leq \kappa \leq 1$ and a perfect detection is achieved when $\kappa = 1$. A value of κ close to 1 indicates the occurrence of more correctly detected than false alarms. On the contrary, a value of κ close to -1 indicates a larger count of false alarms in comparison to that of correctly detected peaks.

4 RESULTS

All four methods discussed previously were tested with the same test signals, in order to investigate their behaviour under identical conditions. Moreover, the results when no spectral tilt compensation is applied are also calculated and displayed as 'no method'. The simulation setup and its processing parameters are the same as those described in the previous section.

Each test signal was designed to accommodate 10 analysis frames. For each test, 500 realizations were generated. The reported results correspond then to the average performance indicators measured for each frame and for each signal. For convenience the values of γ and ζ are displayed in %.

Test 1 aims at assessing peak detection performance under different SNRs. The chosen test signal consists of a single $f_{0,i}$ ($I = 1$), with $a_{ij} = \frac{1}{j f_{0,i}}$, immersed in pink noise. The SNR is set from 0 to 20 dB in steps of 5 dB. The attained results are seen in Table 1.

The values of γ show that most of the genuine peaks are correctly classified even in low SNR conditions. As expected, the lower the SNR, the higher the values of ζ, showing a tendency to misclassify spurious peaks. All methods yielded similar values of κ, evidencing that all perform equally well, under the tested conditions. By contrast, the systematic lower values of κ when spectral tilt compensation is not applied demonstrate the effectiveness of its use.

Test 2 measures peak detection performance w.r.t. the number of harmonic signals present in the mixture. Setup: test signals containing $3 \leq I \leq 7$ fundamental frequencies, with $a_{ij} = \frac{1}{j f_{0,i}}$, immersed in pink noise, and SNR = 10 dB. Table 2 summarizes the results.

It can be observed that γ tends to decrease with the increase of I. As for ζ, with exception of the AR method, it tends to increase with I. Among the tested methods, the NRSF method achieves the lowest performance whereas the AR scores are the highest, revealing some robustness to the increase of I.

Test 3 strains the methods by replacing pink with white Gaussian noise. This time, $I = 3$, $a_{ij} = \frac{1}{jf_{0,i}}$, and SNR = 10 dB. The results are organized in Table 3. As one could anticipate, the performance of all methods decreases. The AR method showed a high γ value, since $E^{\text{ar}}[k]$ tends to decay with frequency, favoring the detection of more genuine peaks immersed in noise. A side-effect is an also high grade for ζ that reduces the value of κ. Overall, according to κ, the TPSW method achieves the highest performance.

Table 3: Performance indicators obtained from **Test 3**.

	Method				
	TPSW	AR	NRSF	SSE	No method
γ	64.2	76.9	44.1	41.6	38.9
ζ	11.0	31.5	2.1	2.1	0.2
κ	0.53	0.46	0.42	0.39	0.39

5 CONCLUSION

This work investigated the effect spectral tilt compensation on the performance of a threshold-based peak detection system. For this, four spectral tilt estimators were examined. Moreover, an experimental environment, composed of a test signal generator, the configurable peak detection system, and a performance evaluation block, was designed.

The test results clearly favors the use of spectral tilt compensation within a peak detection system. However, none of the methods stand out as contributing to an overall superior peak detection performance. In part, the achieved well-balanced performance among the methods can be attributed to the sensible choice of the processing parameters, as well as to the proposed heuristics used to set the threshold level.

6 ACKNOWLEDGEMENTS

The authors would like to thank *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro* (FAPERJ) for funding this work.

REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[2] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proc. Int. Computer Music Conference*, Champaign-Urbana, 1987.

[3] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Swets & Zeitlinger, 1997.

[4] E. B. George and M. J. T. Smith, "Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 497–516, June 1992.

[5] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[6] F. Keiler and S. Marchand, "Survey of extraction of sinusoids in stationary sounds," in *Proc. 5th Conf. Digital Audio Effects*, Hamburg, Germany, 2002, pp. 51–58.

[7] M. H. Hayes, *Statistical Signal Processing and Modeling*, chapter 6, John Wiley & Sons, Inc., 1996.

[8] J. M. Grey and J. W. Gordon, "Perceptual effects of the spectral modifications of musical timbres," *J. Acoust. Soc. Am.*, vol. 61, pp. 1270–1277, 1978.

[9] W. A. Struzinski and E. D. Lowe, "A performance comparison of four noise background normalization schemes proposed for signal detection systems," *J. Acoust. Soc. Am.*, vol. 76, no. 6, pp. 1738–1742, Dec. 1984.

[10] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 662–677, Apr. 1975.

[11] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[12] M. D. Macleod, "Nonlinear recursive smoothing filter and their ues for noise floor esimation," *IEEE Eletronics Letters*, vol. 28, no. 21, pp. 1952–1953, Oct. 1992.

[13] N. Laurenti, G. De Poli, and D. Montagner, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb. 2007.