Miroslav Zivanovic,[*] Axel Röbel,[†] and Xavier Rodet[†]
[*]Universidad Pública de Navarra
Campus Arrosadía, 31006
Pamplona, Spain
miro@unavarra.es
[†]Institut de Recherche et Coordination Acoustique/Musique
1 Place Stravinsky
75004 Paris, France
{axel.roebel, xavier.rodet}@ircam.fr

Adaptive Threshold Determination for Spectral Peak Classification

The decomposition of audio spectra into sinusoids, transients, and noise can serve as a useful tool for improving the results of parameter estimation or signal manipulation applications. As has been shown for the case of transient detection (Röbel 2003) and sinusoidal and noise discrimination (Zivanovic, Röbel, and Rodet 2004), the classification of spectral peaks is a beneficial step toward the identification of these signal components. Such a classification scheme that makes optimal use of the information provided by spectral peaks can then be used to achieve a robust segmentation into higher-level signal components, for example, partials or unvoiced regions. Unlike the perceptual audio segmentation (Painter and Spanias 2005), which attempts to maximize the matching between the auditory excitation pattern associated with the original signal and the corresponding auditory excitation pattern associated with the modeled signal, we base our classification purely on signal characteristics.

The basis for spectral peak classification is an adequate choice of criteria that would best describe sinusoidal and noise spectral peaks of audio signals. Ideally, those criteria (from now on, *descriptors*) would be able to precisely detect the nature of each peak in the spectrum and thus provide a complete separation between the corresponding peak classes in the descriptor domains. Consequently, the decision boundary for the classification process would be unambiguous, and no misclassfication of spectral peaks would occur. This scenario, however, is purely hypothetical as the peaks corresponding to sinusoids (partials) in the spectra of real-world signals are usually subject to additive noise and some type of modulation. In these cases, the descriptor distributions of the different peak classes overlap, and the optimal determination of the decision boundaries depends on the specific application.

The peak classification method proposed in Zivanovic, Röbel, and Rodet (2004) uses descriptors that were designed to adequately characterize nonstationary sinusoidal signals. These descriptors have proven to lead to better classification performance than other approaches devoted to sinudoidal detection/estimation (Thompson 1982; Rodet 1997). It was also shown in Zivanovic, Röbel, and Rodet (2004) that the peak classes can be characterized by distributions in the descriptor domains, similar to probability density functions. Once the distributions have been generated, a simple decision tree can be derived that allows the classification of spectral peaks into sinusoids, noise, and sidelobes.

The peak classification method has been used successfully in a number of applications. As examples we mention polyphonic fundamental-frequency detection (Yeh, Röbel, and Rodet 2005), adaptive noise-floor determination (Yeh and Röbel 2006), and voiced/unvoiced frequency boundary determination. Another interesting application lies in the pre-selection of the sinusoidal peaks to reduce the number of candidate peaks considered for partial tracking in additive analysis. A reliable classification of noise peaks could reduce the number of incorrect connections, and, for probabilistic approaches like that described by Depalle, Garcia, and Rodet (1993), it would considerably reduce the computational cost.

The major problem with the classification scheme of Zivanovic, Röbel, and Rodet (2004) is the control of the classification boundaries (classification thresholds) that generally need adaptation for the specific problem at hand. Another problem is that the descriptor boundaries of the different classes will depend on the analysis window that is used. Up

Computer Music Journal, 32:2, pp. 57–67, Summer 2008

to now, no high-level control parameter existed that would allow a user to adjust the sensitivity of the algorithm in an intuitive manner. There are two signal parameters that directly affect the classification boundaries. The first is the maximum modulation depth and period of the sinusoids. The second is the minimum amplitude of the sinusoids above the noise floor. Both parameters influence the boundaries of the sinusoidal class, and accordingly, both can be used to control the decision boundaries. The problem using the modulation limits as a control parameter is the fact that the modulation is not a single parameter, but rather a parameter vector of at least four dimensions (i.e., period and depth for both amplitude and frequency modulation). Therefore, it cannot be used to provide an intuitive control of the classification boundaries. On the other hand, the sinusoidal peak amplitude above the noise floor is a single parameter that for a given modulation limit would allow us to control the complex decision thresholds rather intuitively.

Accordingly, in this article we investigate the relation between the peak amplitude above the noise floor and the descriptor boundaries for the class of sinusoidal peaks. The descriptors are defined and their properties discussed thoroughly in Zivanovic, Röbel, and Rodet (2004), but for the sake of clarity, we give a brief description of the most prominent characteristics in the next section. For the sinusoidal model described in the subsequent section, we define the space of sinusoidal components by selecting particular limits of the amplitude and frequency modulation rate and depths, as well as the modulation laws. Then, we present the descriptor distributions for the different signal classes, and we establish the mathematical model for the descriptor limits of the sinusoidal class as a function of the peak amplitude level above the noise floor. Finally, we show that the threshold model successfully adapts to the limits of the distributions of sinusoidal peaks for different types of analysis windows.

Spectral Peak Descriptors: A Summary

We define a spectral peak, the elementary classification object, as the normalized energy spectral density between two neighboring minima in the discrete Fourier transform (DFT) modulus $|X(k)|$ of the signal $x(n)$, multiplied by the analysis window. The spectral peak descriptors proposed in Zivanovic, Röbel, and Rodet (2004) are the normalized bandwidth descriptor, the normalized duration descriptor, and the frequency coherence descriptor. The first two are well suited to distinguish between sinusoidal and noise peaks, and the third can be used to detect the side-lobe structure that is an artifact of the windowing process.

Normalized Bandwidth Descriptor (NBD)

Energy distribution along the frequency grid provides useful information for identifying the nature of the signal related to a given spectral peak. Taking $X(k)$ as the DFT of the windowed signal and L as the number of samples in the spectral peak, we define the NBD as a function of mean frequency $\bar{k}$ (in bins) and a root-mean-square bandwidth BW_{rms}:

$$NBD = \frac{BW_{rms}}{L} = \frac{1}{L}\sqrt{\frac{\sum_k \left(k-\bar{k}\right)^2 |X(k)|^2}{\sum_k |X(k)|^2}} \quad (1)$$

$$\bar{k} = \frac{\sum_k k|X(k)|^2}{\sum_k |X(k)|^2} \quad (2)$$

These sums are performed over the L bins in the peak under consideration.

Normalized Duration Descriptor (NDD)

As with mean frequency and bandwidth, the mean time and root-mean-square duration give a rough idea of the distribution of the signal related to a spectral peak along the time grid. The time duration for continuous signals is defined in Cohen (1995) as the standard deviation of the time with respect to the mean time. For discrete signals, the following expressions characterize the duration T_{rms} and mean time $\bar{n}$, respectively:

$$T_{rms} = \sqrt{\sum_n \left(n-\bar{n}\right)^2 |x(n)|^2} \quad (3)$$

$$\bar{n} = \sum_n n\left|x(n)\right|^2 \quad (4)$$

where $|x(n)|^2$ is the normalized signal energy. It is also shown in Cohen (1995) that, from the duality of the Fourier transform, both mean time and duration can be expressed in terms of the spectrum. This important feature permits us to describe individual spectral peaks through the parameters generally employed in the time domain. Considering M to be the size of the analysis window, for discrete spectra the NDD can be obtained as

$$\mathrm{NDD} = \frac{T_{rms}}{M} = \frac{1}{M}\sqrt{\frac{\sum_k \left(A'(k)^2 + \left(g_d(k) + \bar{n}\right)^2\right)\left|X(k)\right|^2}{\sum_k \left|X(k)\right|^2}} \quad (5)$$

$$\bar{n} = -\frac{\sum_k g_d(k)\left|X(k)\right|^2}{\sum_k \left|X(k)\right|^2} \quad (6)$$

where $g_d(k)$ is the group delay, and $A'(k)$ is the frequency derivative of the continuous magnitude spectrum. The group delay $g_d(k)$ is defined to be the derivative of the phase spectrum with respect to frequency. For a single bin of the DFT spectrum, it equals the mean time (according to Cohen 1995) and specifies the contribution of this frequency to the center of gravity of the signal related to the spectral peak. This property of the group delay is also used in Auger and Flandrin (1995) to derive the time reassignment operator, which together with the frequency reassignment attempts to improve signal localization in the time-frequency plane. According to Auger and Flandrin (1995), the group delay can be calculated efficiently as

$$g_d(k) = -\mathrm{real}\frac{X_t(k)X^*(k)}{\left|X(k)\right|^2} \quad (7)$$

where $X_t(k)$ is the DFT of the signal using a time-weighted analysis window. It can also be shown that $A'(k)$ is the imaginary counterpart of the group delay in Equation 7:

$$g_d(k) = -\mathrm{imag}\frac{X_t(k)X^*(k)}{\left|X(k)\right|^2} \quad (8)$$

As with the NBD, all the summations are performed over all bins in the spectral peak.

Frequency Coherence Descriptor (FCD)

The frequency-reassignment operator for constant-amplitude chirp signals points exactly onto the frequency trajectory of the chirp at the position of the center of gravity of the windowed signal. The frequency offset Δ_ω between the frequency at the center of a DFT bin and the reassigned frequency in radians is given by

$$\Delta_\omega(k) = \mathrm{imag}\frac{X_{dt}(k)X^*(k)}{\left|X(k)\right|^2} \quad (9)$$

where $X_{dt}(k)$ is the DFT of the signal windowed by the time derivative of the analysis window. The frequency coherence descriptor is defined as a minimum absolute frequency offset $\Delta_\omega(k)$ for all the bins belonging to that peak:

$$\mathrm{FCD} = \frac{N}{2\pi}\min_k\left|\Delta_\omega(k)\right| \quad (10)$$

where N is the number of bins in the DFT. The normalization factor in Equation 10 ensures that the descriptor is expressed in terms of DFT bins.

Sinusoidal Model and Peak Distributions

To classify a sinusoidal component, we must define what we consider to belong to the sinusoidal class. As is common for sinusoidal modeling, we represent a sinusoidal component as a sinusoid with slowly varying amplitude and frequency parameters (McAulay and Quatieri 1986; Serra and Smith 1990). For an investigation into the properties of the spectral-peak classes, this requirement is not sufficient. To completely define the space of sinusoidal components, we must select concrete limits of the amplitude and frequency-modulation rate and depths, and we must specify a concrete form of the modulation laws.

For the present application, there exists an obvious constraint for the modulation that is related to

the fact that the spectrum of the sinusoidal component must contain a dominant main lobe. Otherwise, the investigation of an individual spectral peak cannot provide us with sufficient information about the underlying sinusoid. Accordingly, the modulation rate and depth should be limited such that a dominant main lobe is present in the Fourier spectrum of each sinusoidal component. Because frequency and time resolution are related to the window size and shape, the modulation limits depend on these two variables. A simple solution to ensure the modulation constraint for all window sizes is to determine the maximum modulation that respects the constraint for a given window size and to change the worst-case modulation rate proportionally with the window size.

As the next step, we must define the worst-case signal, which will be used to derive the descriptor limits of the sinusoidal class. From the wide range of possible modulation laws, we have chosen the sinusoidal amplitude and frequency modulation in white Gaussian background noise as our worst-case reference signal. The choice is motivated by the fact that a range of FM and AM conditions can be covered. If the window size is small compared to the vibrato rate, for example, it is easy to see that the vibrato signal approximately creates linear FM and AM. Recent investigations (Arroabarren, Rodet, and Carlosena 2006; Verfaille, Guastavino, and Depalle 2005) have shown that, for real-world vibrato signals, the AM and FM will generally not be phase-synchronous. Accordingly, the worst-case signal model exhibits arbitrary phase relations between the amplitude and frequency modulation. Because part of the AM is induced by the FM and the resonator filter of the sound source, the dominant AM rate may either be the same as the FM rate or twice as high. As the latter case is more critical, we chose it for our worst-case signal scenario.

In view of the aforementioned discussion, the following mathematical expression for the sinusoidal model is proposed:

$$x(n) = \cos\left[2\pi F_0 n + A_{FM}\sin\left(2\pi F_{FM} n + \alpha\right)\right] \times \left[1 + A_{AM}\cos\left(2\pi F_{AM} n + \beta\right)\right] + r(n) \qquad (11)$$

where $r(n)$ is additive Gaussian noise. According to the previous discussion, we set $F_{AM} = 2F_{FM}$. The frequency vibrato rate F_{FM} must be selected such that the spectrum always contains a significant main lobe, which is ensured by $F_{FM} = 1/(4.2M)$. Accordingly, the window covers less than one-fourth of the FM vibrato period. The values for the amplitude and frequency modulation depth have been chosen as $A_{AM} = 0.5$ and $A_{FM} = 10$.

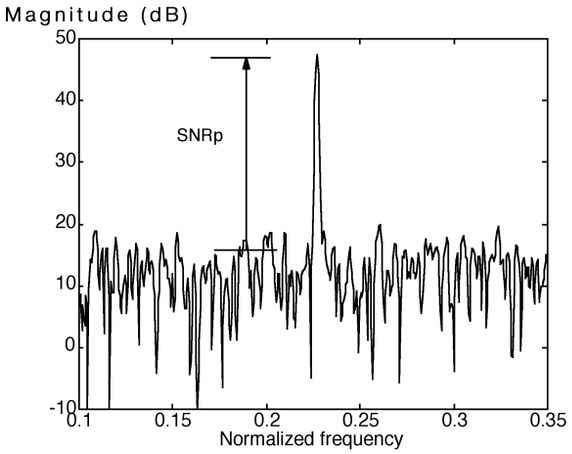
These values ensure a dominant-peak main lobe for arbitrary phase angles (α and β). The window length M, the sinusoidal frequency F_o, and the sample rate R do not impact the results. The size of the DFT N is chosen to assure that the "picket-fence" effect has minimal impact on a peak representation in the discrete spectrum. For completeness, we note the values that we used for the following investigation into the descriptor distributions: $M = 40$ msec, $N = 4{,}096$, $F_o = 880$ Hz, and $R = 44{,}000$ samples/sec).

It is clear that the present worst-case signal does not cover all modulations that may be encountered in a real-world setting, even if we respect the fact that a dominant main lobe is required to detect a modulated sinusoid. The explicit inclusion of time-varying sinusoids into the model will nevertheless lead to a classifier that has significant advantages in real-world situations involving time-varying sinusoids.

Because the part of the sinusoidal peak that can be observed changes with the variance σ_r^2 of the background noise level $r(n)$ the peak descriptors will not only change with the modulation, but also with the signal-to-noise (SNR) ratio. For multi-component signals, the global SNR does not provide meaningful insight, and therefore, we use the *peak signal-to-noise ratio* (SNR_p) as our noise-level parameter. The SNR_p indicates the sinusoidal peak power level in dB over the noise floor (see Figure 1) and it presents a convenient parameter to control the limits of the sinusoidal class.

To experimentally create the descriptor distributions, we proceed as follows. For the noise class distributions, we calculate the descriptors for all spectral peaks in the DFT of white Gaussian noise processes using an analysis window of size M. For the sinusoidal class, we create a grid of phase values

Figure 1. Illustration of the parameter SNR_p (peak signal-to-noise ratio).



covering all combinations α and β over the range –π to π, and we set $\sigma_r^2 = 0$. Then, we calculate the descriptor values for the largest peak in each frame. This gives us the distributions for an infinite SNR_p. The side-lobe distributions are calculated from all but the strongest spectral peak in the spectrum of the worst-case sinusoid. The resulting descriptor distributions are normalized by the maximum value and shown in Figure 2 for the Hanning window.

As we can see from Figure 2, the NBD distributions for modulated noise-free sinusoidal peaks and for noise peaks do not overlap at all, making them a very good candidate for sinusoidal and noise separation. The sine and noise distributions for the NDD significantly overlap, but the sinusoidal distribution covers only a small range of descriptor values. This fact will be used to refine the sine/noise separation done by the NBD for signals of finite SNR_p, as explained in the next section. Both descriptors do not allow one to distinguish the side lobes from real signal components. For this task, the FCD descriptor is extremely efficient. Note that in Figure 2, the maximum of the side-lobe distribution is to be interpreted as a cumulus of all the side-lobe FCD values distributed out of the current axis range.

Classification Strategy

The peak-classification algorithm, based on the proposed peak descriptors, is established through a

Figure 2. Normalized distributions (bandwidth [NBD], duration [NDD], and frequency coherence [FCD]) for three peak classes (sine, noise, and side-lobe) in the descriptor domain, with $\sigma_r^2 = 0$ and using a Hanning window.

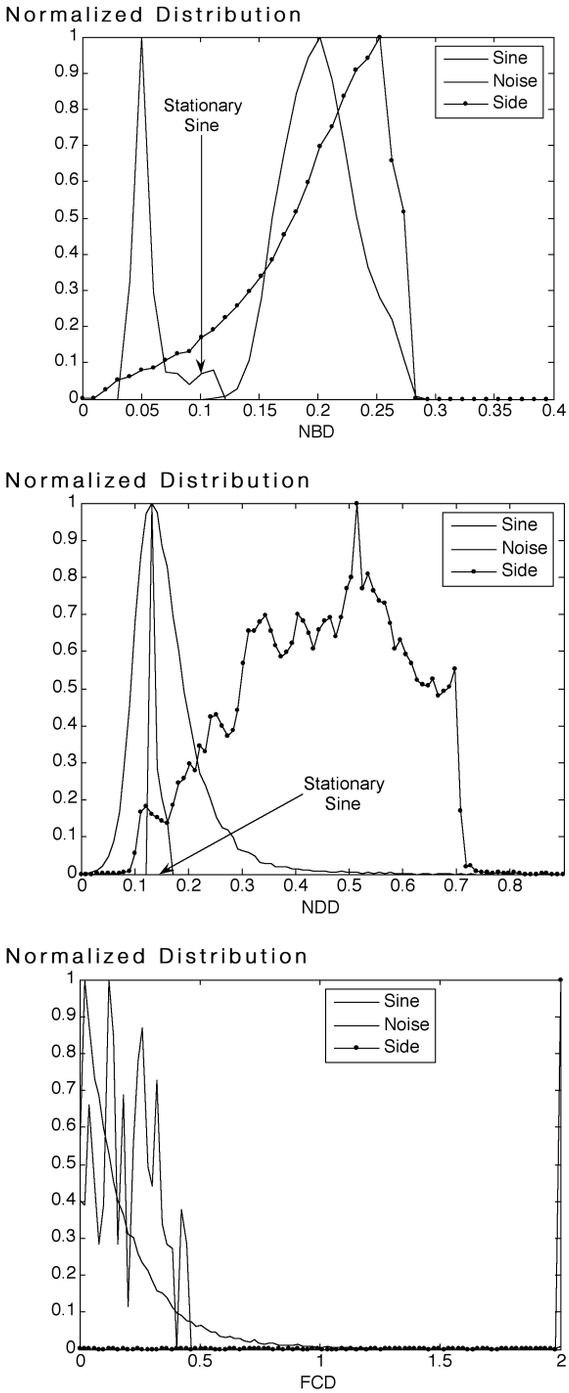


Table 1. Peak Classification Thresholds for Infinite SNR_p Computed Using a Hanning Window

Peak Classes	*Descriptor Values*
side-lobe / non-side-lobe	FCD ≥ N/M
sine / noise	NBD ≤ 0.13 and 0.13 ≤ NDD ≤ 0.16

two-level decision tree as follows. In the first level, the side-lobe and non-side-lobe classification is performed. In the second level, the peaks previously declared as non-side-lobes are classified as sinusoids and noise. The thresholds for both levels of classification are obtained by analyzing the distributions shown in Figure 2. For infinite SNR_p, the classification could be obtained by simply using FCD and NBD thresholds to perfectly separate all three peak classes. Note that only in this particular case, the NBD attains almost perfect sine/non-sine classification; therefore the contribution of the NDD is negligible. The classification scheme that is used for infinite SNR_p is shown in Table 1.

For a finite SNR_p, the sinusoidal distributions experience a spread proportional to the noise level in the worst-case signal. In particular, the NBD sinusoidal distribution extends towards the right, whereas the NDD sinusoidal distribution spreads in both directions. The sinusoidal NBD distribution overlaps partially with the noise NBD distribution, which means that the NBD can no longer perfectly separate the peak classes. To reduce this ambiguity, we use the NDD. As mentioned previously, the sinusoidal NDD distribution covers only a small range of descriptor values. Hence, by considering only the peaks within the limits of the sinusoidal NDD distribution as sinusoids, we can eliminate some of the noise peaks previously classified as sinusoids and thus refine the initial sine/noise classification.

It is important to understand that a decreasing SNR_p will modify the limits of the sinusoidal distribution in a manner similar to that of an increase in the modulation parameters. Therefore, the minimum SNR_p can be used to control the decision thresholds in a rather intuitive manner.

To keep track of the limit values of the sinusoidal distributions, we would need to regenerate all the sinusoidal distributions every time the minimum SNR_p that is selected by the user is changed. As shown subsequently, however, the experimental evaluation of the distribution limits can be avoided thanks to a simple approximate formula that expresses the relationship between the parameter SNR_p and the margins of the sinusoidal peak distributions in the descriptor domain. These can be used to adapt the classifier to the selected SNR_p. The thresholds to be adapted are the right margin of the NBD sinusoidal distribution and both margins of the NDD sinusoidal distribution. As for the FCD, the threshold can be held fixed thanks to the good side-lobe separation from the rest of the peak classes.

Modeling SNR_p Dependency

The relation between the classification threshold and the SNR_p is rather complex, and to be able to achieve a model of these relations, the problem requires a number of simplifications. We first experimentally determine the signal pattern that is related to the descriptor limits for infinite SNR_p. Then, we develop a simplified model of the effect of the additive noise to be able to achieve a mathematical formulation of the threshold dependency on the SNR_p. The relation does not take into account the fact that the signal pattern at the descriptor limits may depend on the SNRp.

NBD Threshold

Recall that the NBD is the ratio of the peak bandwidth to the peak width. As described above, we must first determine the sinusoidal signal that will give rise to the maximum value of the descriptor $NBD_{max} = BW/L$. This can be done by means of a straightforward search over the two-dimensional grid of phase values α and β for a given analysis window. (See Table 2 for a listing of some prominent analysis windows.)

The presence of noise will affect both BW and L. It is clear that L will decrease, because the peak local minima approach the peak maximum in terms of magnitude. In a simple approximation, we can

Table 2. Phase Values of the Sinusoidal Model Corresponding to NBD_{max} for Various Analysis Windows

Window	α_{max}	β_{max}
Hanning	0.75π	0.50π
Blackman	0.75π	0.55π
Hamming	0.70π	0.45π

assume that *BW* will stay roughly constant, because the peak shape around the maximum is only slightly affected by additive noise. Accordingly, we can assume that the NBD_{max} is a function of *L* solely, which in turn depends on the SNR_p. Practically, for the given α_{max} and β_{max}, we calculate the spectrum of the sinusoidal signal only once and store it in memory. Then, the NBD threshold can easily be calculated by taking into account only the DFT bins of the main lobe that lie above the noise floor given by SNR_p. The validity of this simple approximation is checked in the next section of this article by comparing its values to those obtained by measuring NBD_{max} for different SNR_p and different analysis windows.

NDD Threshold

The sinusoidal model in Equation 11 is herein simplified to investigate NDD thresholds. More specifically, the FM case can be disregarded because it does not modify the NDD of a sinusoid. Hence,

$$x(n) = \cos\left(2\pi F_0 n\right) \times \left[1 + A_{AM} \cos\left(2\pi F_{AM} n + \beta\right)\right] + r(n) \quad (12)$$

The phase β that gives rise to the minimum and maximum values of the NDD descriptor for the signal in Equation 12 and after applying the analysis window can be calculated numerically. The solution shows that the maximum value is obtained when the minimum of the AM envelope is located in the signal center. The minimum of the NDD is obtained for a phase β that places the AM envelope maximum close to the window center. Owing to the interactions between the analysis window and the envelope, the AM envelope is not exactly aligned with the window center. To simplify the discussion (and because all values of β in the range $-\pi \le \beta \le 0$ result in a variation of the NDD of less than 1%), we use the signal pattern with AM envelope maximum in the window center for the following discussion.

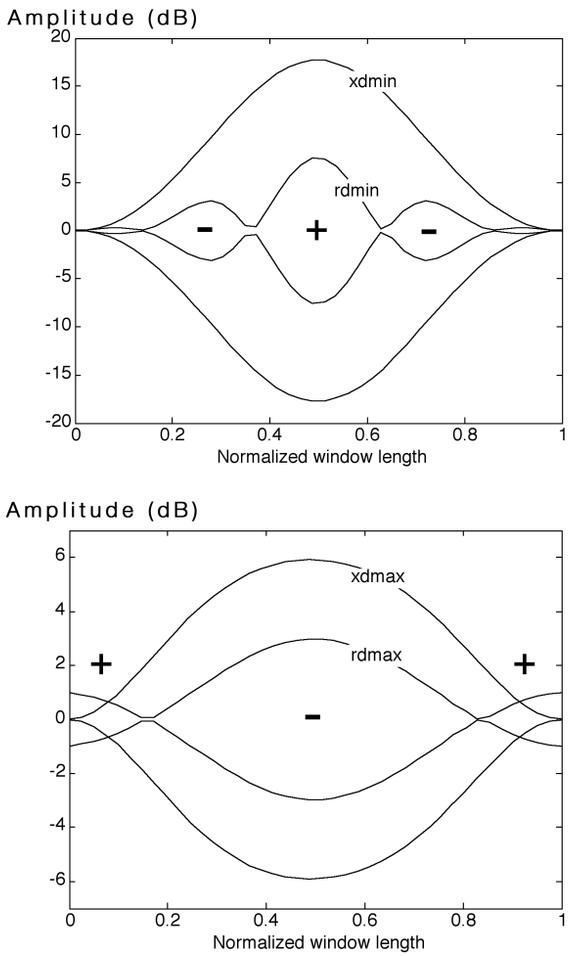


Figure 3. Envelopes of signal patterns and noise patterns corresponding to the NDD thresholds for SNR_p = 10 dB. Sign symbols mark the carrier phase relationship between the waveform; a Hanning window is used.

Accordingly the approximate signal patterns for the shortest and longest signal in terms of the NDD are given by

$$x_{d\min} = x\left(n;\beta = -0.5\pi\right) w(n)$$
$$x_{d\max} = x\left(n;\beta = 0.5\pi\right) w(n) \quad (13)$$

where $w(n)$ is the analysis window. The envelopes of the signal patterns $x_{d\min}$ and $x_{d\max}$ for the Hanning window are displayed in Figure 3. For finite SNR_p, the patterns in Equation 13 are superposed to a

Table 3. Coefficient Values for Modeling the m_{min} and m_{max} Dependency on SNRp

Window	$a_o\,(r_{dmin})$	$a_1\,(r_{dmin})$	$a_2\,(r_{dmin})$	$b_o\,(r_{dmax})$	$b_1\,(r_{dmax})$	$b_2\,(r_{dmax})$
Hanning	0.0174	–0.5770	10.6280	–0.0006	0.1211	0.8279
Blackman	0.0081	–0.3903	9.0630	–0.0022	0.1472	0.7083
Hamming	0.0003	–0.2816	2.4716	–0.0037	0.1615	0.7230

narrow-band Gaussian noise. Owing to the small bandwidth of the signal peak, the effective noise bandwidth is rather small. For each SNR_p, there exist two noise signal patterns, $r_{d\min}$ and $r_{d\max}$, that will maximally increase and decrease, respectively, the NDD_{max} and NDD_{min} values. We use a simple signal model consisting of an amplitude-modulated carrier as basis for our noise model. The noise model is band-limited (reflecting the bandwidth of the spectral peak) but not necessarily time-limited. Owing to the small bandwidth, the noise pattern may extend out of the signal window. Because we do not want to take into account the length of the DFT for the simple model here, we limit the noise signal to the time segment of the analysis window.

To reduce NDD_{min}, $r_{d\min}$ should narrow the width of the central maximum of $x_{d\min}$. To achieve this, $r_{d\min}$ must be in phase with $x_{d\min}$ around the window's center and in counter-phase otherwise. Because a strong amplitude at the window boundaries would always enlarge the NDD, we additionally assume that the noise pattern $r_{d\min}$ has the analysis window applied.

On the contrary, $r_{d\max}$ must be in counter-phase with $x_{d\max}$ around the window's center and in phase close to the window edges. The resulting waveform would have the energy more uniformly distributed along the analysis window and thus larger NDD_{max}. Here, $r_{d\max}$ must not be tapered in order to contribute significantly to the energy concentration in $x_{d\max}$ around the window edges.

According to this discussion, we used the following model for the narrow-band Gaussian noise patterns:

$$r_{d\min}(n) = A\cos(2\pi F_o n)\left[1 + m_{\min}\cos(4\pi n / M)\right]w(n)$$

$$r_{d\max}(n) = -A\cos(2\pi F_o n)\left[1 - m_{\max}\cos(2\pi n / M)\right] \qquad (14)$$

The noise patterns are therefore sine-modulated waveforms. The modulation frequencies are different because the bandwidth of the peaks related to NDD_{min} and NDD_{max} are different. They have been selected such that they obey a simple relation to the window size. Note that the exact frequency values are not critical for the model and that the frequencies do not depend on the SNR_p.

The modulation indices m_{min} and m_{max} must be greater than unity to ensure the phase change of π in the crossover between contiguous modulation lobes. Both amplitude A and modulation indices are a function of the SNR_p. A determines the total energy of each pattern, and m_{min} and m_{max} control the distribution of that energy along the analysis window. The amplitude is simply a scaling factor that ensures the most of the spectral energy of the noise patterns lies SNR_p dB under the main lobe of the worst-case signal. The values for the modulation indices are more difficult to estimate, as they change in a non-linear fashion with the SNR_p. To obtain a mathematical model, we used the signal in Equation 13 and a wide range of SNR_p settings and have experimentally determined the maximum and minimum NDD as well as the values for m_{min} and m_{max} that would best match the experimental data.

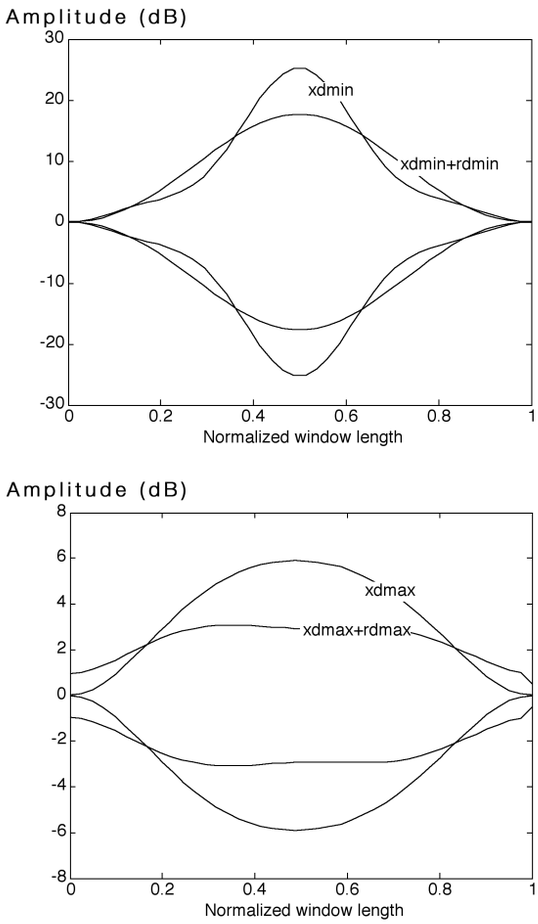
Finally, we derived a second-order polynomial representation of the modulation indices by means of adapting a second-order polynomial to the set of modulation indices. For various types of analysis windows, the resulting functions are given by

$$m_{\min} = \sum_i a_i SNR_p^i \;,\; m_{\max} = \sum_i b_i SNR_p^i$$

with the corresponding coefficients given in Table 3. For the Hanning window, the envelopes of the corresponding noise patterns for SNR_p = 10dB are shown in Figure 3, and the envelopes of the resulting waveforms after the superposition are shown in Figure 4.

Note that the energy distributions of the signal

Figure 4. Resulting envelopes after the superposition of the signal patterns to the corresponding noise patterns for SNR_p = 10 dB, using a Hanning window.



patterns have indeed been modified as in the aforementioned explanation. In practical applications, the signal patterns are calculated only once, whereas the noise patterns are recalculated each time the SNR_p or type of analysis window is changed such that the new thresholds can be obtained. In the following section, we show the behavior of this model with respect to the measured NDD_{min} and NDD_{max} for different SNR_p and various analysis window types.

Experimental Results

Next, we check the validity of the proposed adaptive threshold-selection algorithm. For different types of analysis windows and for a wide range of SNR_p values, the decision thresholds NBD_{max}, NDD_{min}, and NDD_{max} were generated from the corresponding models and compared to their respective measured values. The measured values are obtained from the Gaussian noise added to the sinusoidal model in the proportion established by the SNR_p. The approximation errors are calculated as a difference between the measured and modeled values and are shown in Figure 5. Generally, the approximation errors are larger for smaller values of SNR_p.

In the case of the NBD_{max} and the NDD_{min} thresholds, the experimentally obtained errors show a systematic trend, which could be used to refine the model. For the NDD thresholds, the error generally lies in overestimating the change of the boundaries that corresponds to the SNRp. For the NBD threshold, the threshold change is underestimated. The overall approximation error is obtained by evaluating the correlation coefficient R between the measured and approximated curve for each threshold and various analysis windows. From Table 4, we can see that in almost all situations, the correlation coefficient is above 0.95, which can be considered a very good approximation. Also, note that the largest approximation errors are found in the NBD_{max}-thresholding domain when using the Hanning window. On the contrary, the Blackman window thresholding adapts well to the corresponding curve of measured threshold values.

Conclusions

In this article, we presented a new adaptive threshold-selection algorithm that can be used for classification of spectral peaks. By means of the set of peak descriptors from previous work and a herein-proposed compact sinusoidal model related to the analysis window, the limit values for the distributions of sinusoidal peaks in the descriptor domain can be explicitly obtained. Next, the variations of those limit values, owing to the presence of noise in the sinusoidal model, are characterized in a deterministic fashion through only one parameter that we refer to as the peak

Figure 5. Approximation errors calculated as a difference between the measured and modeled values for each SNR_p using various analysis windows. Values in the legend correspond to infinite SNR_p.

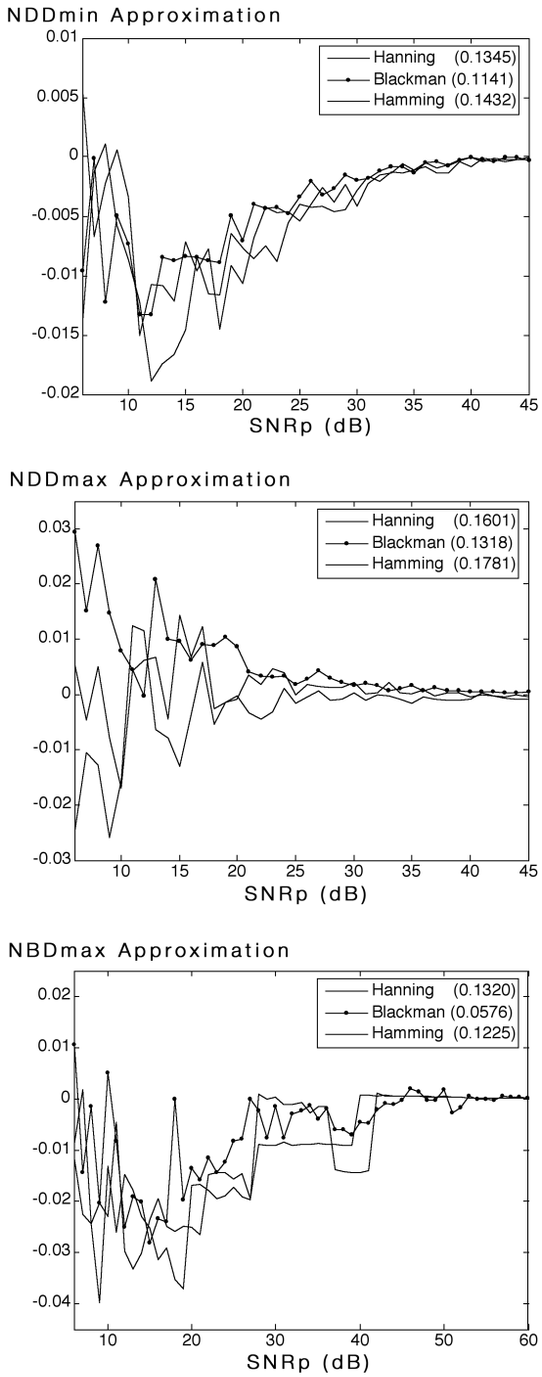


Table 4. Correlation Coefficients Calculated Between the Measured and Approximation Threshold Curves for Various Analysis Windows

	Hanning	*Blackman*	*Hamming*
$R(NDD_{min})$	0.9567	0.9685	0.9604
$R(NDD_{max})$	0.9799	0.9792	0.9840
$R(NBD_{max})$	0.9139	0.9885	0.9585

signal/noise ratio. By means of this control parameter, the descriptor limits of the classification algorithm can be adapted intuitively to increase or decrease the tolerance of the classifier with respect to noise level and modulation. The approximation accuracy given through the correlation coefficient is shown to be large for different types of analysis windows.

At the present state, the new threshold-selection method provides a control precision that can be considered sufficient for interactive control of a classification algorithm. Further investigation will be concerned with enhancing the threshold models to reduce approximation errors and improve precision. Also, we expect to improve the quality of the final decision in the peak-classification process by investigating different classification strategies and classifiers.

Acknowledgments

The first author would like to gratefully acknowledge the financial support of the Universidad Publica de Navarra, Spain.

References

Arroabarren, I., X. Rodet, and A. Carlosena. 2006. "On the Measurement of the Instantaneous Frequency and Amplitude of Partials in Vocal Vibrato." *IEEE Transactions on Speech and Audio Processing* 14(4):1413–1421.

Auger, F., and P. Flandrin. 1995. "Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method." *IEEE Transactions on Signal Processing* 43(5):1068–1089.

Cohen, L. 1995. *Time-Frequency Analysis.* Englewood Cliffs, New Jersey: Prentice Hall.
Depalle, P., G. Garcia, and X. Rodet. 1993. "Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models." *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Vol. I.* New York: Institute for Electrical and Electronics Engineers, pp. 242–245.
McAulay, R., and T. Quatieri. 1986. "Speech Analysis Synthesis Based on a Sinusoidal Representation." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34(4):744–754.
Painter, T., and A. Spanias. 2005. "Perceptual Segmentation and Component Selection for Sinusoidal Representation of Audio." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 13(2):149–162.
Röbel, A. 2003. "A New Approach to Transient Processing in the Phase Vocoder." *Proceedings of the 6th International Conference on Digital Audio Effects.* London: Queen Mary, University of London, pp. 344–349.
Rodet, X. 1997. "Musical Sound Signal Analysis/Synthesis: Sinusoidal + Residual and Elementary Waveform Models." *Proceedings of the IEEE Time-Frequency and Time-Scale Workshop '97.*
Serra, X., and J. O. Smith. 1990. "Spectral Modeling and Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition." *Computer Music Journal* 14(4):12–24.
Thompson, D. J. 1982. "Spectrum Estimation and Harmonic Analysis." *Proceedings of the IEEE* 70(9):1055–1096.
Verfaille, V., C. Guastavino, and P. Depalle. 2005. "Perceptual Evaluation of Vibrato Models." *Proceedings of the 2005 Conference on Interdisciplinary Musicology.* Montreal: Observatoire International de la Création Musicale and University of Montreal, pp. 149–151.
Yeh, C., A. Röbel, and X. Rodet. 2005. "Multiple Fundamental Frequency Estimation Of Polyphonic Music Signals." *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing, Vol. III.* New York: Institute of Electrical and Electronics Engineers, pp. 225–228.
Yeh, C., and A. Röbel. 2006. "Adaptive Noise Level Estimation." *Proceedings of the 9th International Conference on Digital Audio Effects.* Montreal: McGill University, pp. 145–148.
Zivanovic, M., A. Röbel, and X. Rodet. 2004. "A New Approach to Spectral Peak Classification." *Proceedings of the 12th European Signal Processing Conference.* Vienna: Vienna University of Technology, pp. 1277–1280.