A NEW APPROACH TO SPECTRAL PEAK CLASSIFICATION

Miroslav Zivanovic and Axel Roebel and Xavier Rodet

Universidad Publica de Navarra, Spain, email: miro@unavarra.es
IRCAM, France email: roebel@ircam.fr,rod@ircam.fr

ABSTRACT

A novel approach to classification of peaks of audio spectra is presented. In extending previous work on detecting transient spectral peaks we here investigate into the classification of sinusoidal and noise peaks. The classification is based on descriptors derived from properties related to time-frequency distributions: mean time, duration, instantaneous frequency and normalized bandwidth. In contrast to existing methods, the descriptors are designed to properly deal with non-stationary sinusoids, which considerably increases the range of applications. The experimental investigation shows superior classification results compared to the standard correlation-based approach.

1. INTRODUCTION

The decomposition of audio spectra in sinusoids, transients and noise is often used to improve the results of parameter estimation and/or signal manipulation applications. In the following paper we are going to investigate into the possibility to classify individual spectral peaks. As has been shown for the case of transient detection [1] the classification of spectral peaks is a beneficial approach to identify 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, e.g. partials or unvoiced region. Complementing the transient peak classification method the present paper will deal with classification into noise and sinusoidal peaks.

There exist few approaches for the classification of spectral peaks. Among them we cite the widely used correlation based measure of sinusoidality [2] and another proposal that is based on the reassigned spectrogram [3]. The former takes the maximum of the complex correlation between the DFT of the analysis window and each peak of the STFT of the signal. If the value is equal to 1, the peak belongs to a noiseless steady-state sinusoid, otherwise it indicates the presence of noise or time-variable components. The letter proposes the classification of the STFT peaks in sinusoids, unresolved sinusoids, transients and noise. Various statistics are calculated for each side of the peak separately and the traditional pattern classification method with a likelihood ratio test is applied to perform the classification.

The shortcoming of both approaches is the underlying assumption of quasi-stationary signals. As shown in the experimental section the performance of the correlation based method severely degrades for non-stationary sinusoids that are present in real world signals. Moreoever, the method presented in [3] uses a probabilistic approach to derive classification thresholds. As long as a probabilistic description of the signal composition is available this will result in optimal performance. This, however, is rarely the case because the probability of noise peaks changes with the size of the analysis window. The larger the window, the more noise peaks will be observed in contrast to the number of sinusoidal peaks which is approximately constant. Due to the conceptual problem with the probabilistic approach, we derive our classification criterion by means of declaring a worst case situation. This situation is characterized by a defined deviation from the stationary noise-free sinusoid.

There exist a number of audio signal processing applications where the classification of spectral peaks could be used. It can be applied as a pre-processing stage to reduce the number of candidate peaks considered for partial tracking in additive analysis. A reliable classification of noise peaks would reduce the number of incorrect connections and for probabilistic approaches [4] it would considerably reduce the computational cost. Also for F0 detection algorithms the impact of noise components could be reduced.

Another domain of application is the voiced/unvoiced segmentation in speech processing. For this case, however, the classification of spectral peaks is not sufficient and needs to be extended to obtain higher level descriptors. This, however, is beyond the topic of the current article.

The paper is organized as follows. In section **2** we define the descriptors that will be used for classification of spectral peaks and discuss the properties of the descriptors if applied to different types of spectral peaks. In the following section **3** we describe the structure of the decision tree and derive the thresholds to be used for classification. We present an experimental result of our classification procedure and demonstrate its superior performance compared to the correlation based peak classification. We conclude the paper with a discussion of the achievements and required further developments.

2. SPECTRAL PEAK DESCRIPTORS

From the many descriptors of spectral peaks that we have studied we have selected four that achieved the best discrimination performance.

2.1 Descriptor definitions

The frequency reassignment operator has been derived in [5] to improve signal localization in the time-frequency plane. For constant amplitude chirp signals it exactly points 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 an DFT bin and the reassigned frequency in rad is given by

$$\Delta_\omega(k) = \text{imag}\,\frac{X_d(k)X^*(k)}{|X(k)|^2}. \qquad (1)$$

Here k specifies the bin index of the DFT. $X(k)$ is the DFT of the signal windowed with the analysis window and $X_d(k)$ is the DFT of the signal windowed with the time derivative of the same window. The operator X^* denotes complex conjugation. To characterize the frequency coherence of a spectral peak we select as descriptor the minimum value of $|\Delta_\omega|$ for all k belonging to this peak and normalize by $\frac{2\pi}{N}$ where N is the size of the DFT. The normalization ensures that the frequency coherence descriptor FCD is invariant with respect the analysis parameters.

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 [6] and specifies the contribution of this frequency to the center of gravity of the signal related to the spectral peak. The mean time is the main feature to detect transient peaks [1]. In the current investigation we found that due to the influence of neighboring peaks the mean time derived from the spectral peak as a whole is not sufficiently robust. Therefore we use a modified version given by

$$t_e = -\frac{g_d(k_{max})|\Delta_\omega(k_{max2})| + g_d(k_{max2})|\Delta_\omega(k_{max})|}{|\Delta_\omega(k_{max2})| + |\Delta_\omega(k_{max})|}, \qquad (2)$$

which characterizes the energy location by means of investigating the peak center only. The indices k_{max} and k_{max2} correspond to the largest and second largest samples in the peak. The weighting by means of the frequency reassignment operator results in the fact that constant amplitude chirp signals will always have a mean time very close to zero even if their frequency trajectory does not exactly pass through a center frequency of a bin. To prevent a dependency of classification results on the analysis parameters we normalize $|t_e|$ by the length of the analysis window to obtain the energy location descriptor ELD. Note that the group delay can be calculated efficiently by

$$g_d(k) = -\,\mathrm{real}\,\frac{X_t(k)X^*(k)}{|X(k)|^2}, \tag{3}$$

where $X_t(k)$ is the DFT of the signal using a time weighted analysis window [5].

The time duration of a signal as defined in [6] is the standard deviation of the time with respect to the mean time interpreting signal energy as distribution. For discrete spectra it can be obtained by means of

$$T = \sqrt{\frac{\sum_k (A'(k)^2 + (g_d(k) - \bar{t})^2)|X(k)|^2}{|X(k)|^2}}, \tag{4}$$

where the sum is performed over the spectral peak under consideration. $\bar{t}$ is the mean time of the signal related to the peak and $A'(k)$ is the frequency derivative of the continuous magnitude spectrum. It can be shown that $A'(k)$ is the imaginary counterpart of the group delay in eq. (3)

$$A'(k) = -\,\mathrm{imag}\,\frac{X_t(k)X^*(k)}{|X(k)|^2}. \tag{5}$$

Similar to the mean time for classification we normalize the time duration T by means of the window size to obtain the duration descriptor DD.

As with mean time and time duration, the mean frequency $\bar{\omega}$ and the bandwidth B give a rough idea of the concentration of the spectral density along the frequency grid. Considering L to be the number of samples in the spectral peak then the normalized bandwidth descriptor NBD can be defined as:

$$\bar{\omega} = \frac{\sum_k k|X(k)|^2}{\sum_k |X(k)|^2}, \tag{6}$$

$$NBD = \frac{B}{L} = \frac{\sum_k (k-\bar{\omega})^2|X(k)^2|}{L\sum_k |X(k)^2|}. \tag{7}$$

As for the duration the summation is done over all the bins in the spectral peak.

2.2 Descriptor properties

For deriving the classification thresholds for the descriptors we rely on the declaration of a worst case scenario. The related test signal is a single AMFM-sinusoid in noise (SNR = 0dB) where both frequency and amplitude change in a sinusoidal fashion. To resemble natural vibrato signals, the period of the frequency modulation is two times the period of the amplitude modulation. The characteristics of the test signal are:

- for amplitude modulation: modulation index 0.5,
- for frequency modulation: 200 Hz of frequency deviation.

The analysis window is a 50ms Hanning window and the frequency modulation period is 100ms. For calculating the DFT we use 4096-point FFT with the sample rate being 44100Hz. This scenario roughly reproduces the analysis conditions for the tenth harmonic of a 333Hz pitch tone under half tone vibrato extent.

In the initial investigation only the two classes, noise and sinusoids, have been taken into account and all but the sinusoid peaks have been considered to be noise. During the initial experiments we found that the noise distributions of the descriptors would change with the SNR. Further investigation revealed that this effect was due to the presence of sinusoidal sidelobes in the noise region. Because sidelobes should not be confounded with sinusoids or noise it was necessary to introduce a further class for sinusoid sidelobes.

The descriptor distributions for the peak classes that have been obtained for the test signal are shown in fig. 1. For the sinusoidal distributions the descriptors were applied only to the largest peak in the spectrum for a total of 1100 time frames. The noise distributions were obtained by analyzing all the peaks in the DFT of a white noise signal. To derive the sidelobe distributions we analyzed all the sidelobe peaks of a stationary noise-free sinusoid. For ease of comparison all distributions are displayed normalized such that their maximum value is equal to one. As the threshold levels we are going to determine aim to preserve fractions of the distributions, this normalization does not affect the results.

After having defined our descriptors we will now shortly discuss the behavior of the descriptors when applied to the different peak classes. The relations between the descriptor values and the signal characteristics are very complicated and can be theoretically explained only for the simple case of a constant amplitude chirp signal. For signal peaks related to noise or complex modulated sinusoids the behavior of the descriptors will be derived from the distributions obtained experimentally.

Because Δ_ω is the frequency location (in bins) of the center of gravity of the band limited signal related to bin k of a DFT spectrum, its minimum, which is the FCD, will always be below 0.5. For the distribution of the FCD of sinusoidal peaks depicted in fig. 1 we observe that the distribution remains limited below 0.5 even for the amplitude and frequency modulated signal used in the worst case scenario. For the noise peaks the distribution is centered around 0 with nearly linear falloff up to 0.5 while the distribution for sidelobe peaks is nearly uniform over a large frequency range (not completely displayed). According to the observed distribution we expect that the FCD achieves a good sidelobe detection but only limited performance for distinction between sinusoids and noise.

The ELD is similar to the mean time and will be close to zero for constant amplitude chirp signals. For amplitude modulation the ELD may increase. However, due to the normalization, its magnitude is always below 0.5. The signals corresponding to isolated sidelobes are not limited to the duration of the analysis window but are confined to the region of the zero-padded analysis window. Therefore, the mean time extends over larger range (not displayed). Due to the strong variations of the ELD distribution for sinusoidal peaks with the modulation parameters, it is hard to expect a good performance for discrimination between sinusoids and noise. Nevertheless, we expect that this descriptor achieves a good detection of sidelobe peaks.

Considering the DD we know that for constant amplitude chirp signals it will always be close to the duration of the analysis window itself. For amplitude modulation the DD distribution of the sinusoidal peaks will spread and move its center thus covering a considerable part of the DD distribution of the noise peaks. As explained in the discussion of the ELD sidelobe related signals extend outside the analysis window and therefore have systematically a larger value of DD than noise and sinusoids. Accordingly the DD will achieve a very good discrimination of sidelobes. While the worst case signal appears to allow fairly good distinction between sinusoids and noise the modulation dependency of the distribution center does not allow very strict placement of the classification thresholds such that the DD achieves approximately similar discrimination between sinusoids and noise as the FCD.

The NBD descriptor can be viewed as a measure of the noise energy in the neighborhood of a sinusoidal spectral peak. Its performance can be explained in terms of the relation between the peak bandwidth and the total peak region L. The theoretical investigation of the NBD is very complicated even for the relatively simple case of constant amplitude chirps. The experimental investigation of the NBD distributions for modulated noise free sinusoidal peaks and

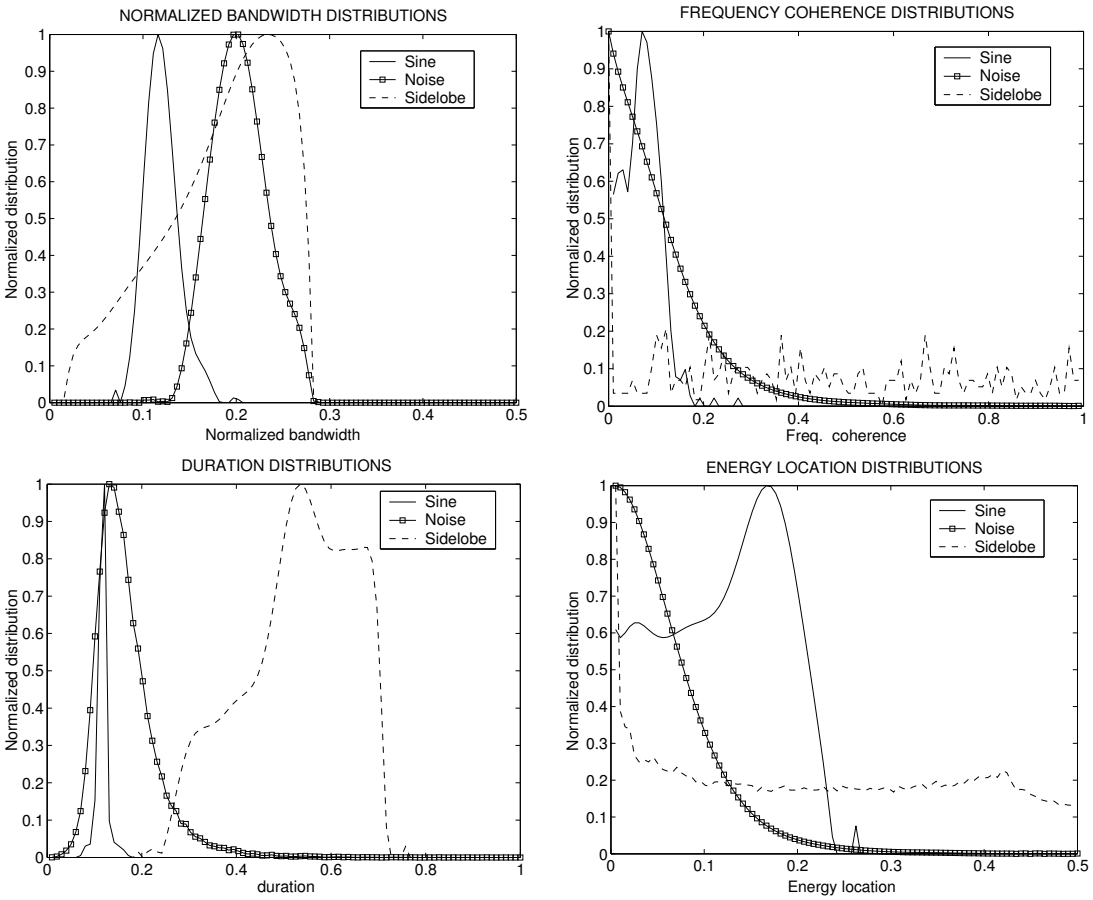

Figure 1: Distributions for peak descriptors used.

for noise peaks has shown that these distributions do not overlap at all making them a very good candidate for sinusoidal and noise classification. With increasing noise level in the sinusoidal signal the tail of the sinusoidal *NBD* distribution is moving right and overlaps slightly with the *NBD* noise distribution. To characterize the robustness of the descriptor with respect to noise we also have investigated into the dependency between classification errors for stationary sinusoids in noise as a function of the SNR. We used the maximum value of the sinusoidal *NBD* descriptor as classification threshold (0% classification errors for the sinusoidal peaks) and did allow 5% classifications errors for the noise peaks. The error rates are achieved for an SNR that keeps the noise floor -15dB below the sinusoidal peak. Due to amplitude and frequency modulation in the worst case scenario studied here the overlap is slightly larger but remains small compared to the overlap obtained for all the other descriptors. For sidelobe classification the *NBD* will only achieve low performance.

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed descriptors a preliminary binary decision tree for the peak classification has been established as follows: in the first level a sinusoidal and non-sinusoidal classification is performed. Then in the second level the non-sinusoidal peaks are classified into sidelobes and noise. The thresholds for both levels of classification have been obtained by means of analyzing the distributions shown in fig. 1. Because we have been interested to achieve nearly perfect sinusoidal detection we have set the *NBD* for the threshold classification such that 10% of noise peaks are misclassified. To exclude the sidelobes below the *NBD* threshold we require as second condition that the *DD* lies below a threshold such that no sidelobe peak is classified as sinusoid. The selected thresholds are listed in table (1). For our worst case signal we achieve less than 1% misclassification of sinusoidal peaks. Because sidelobe and sinusoidal *DD* distribution do hardly overlap the *DD* threshold need not be adapted to the signal at hand. The adaptable parameter for the first level of the decision tree is the *NBD* threshold. This threshold can be simply determined as a function of the noise classification error. Because the noise distribution does not change with the spectral envelop of the noise it can be rapidly created for a given window size and type and the *NBD* threshold can be automatically selected according to the noise classification error requested by a user. The thresholds shown in table (1) for the second level of the classification scheme have been selected according to fig. 1 such that each threshold achieves approximately similar classification error when distinguishing between noise and sidelobe peaks. The thresholds depend only weakly on the signal and can be kept constant for most applications.

The selected thresholds have been used to classify a number of artificial and real audio signals. Due to space constraints, we will present only one result of the algorithm applied to a real audio signal. The signal is a flute signal with vibrato taken from the Iowa University Database. We use this example to compare the proposed classification method to the correlation method mentioned in the introduction. In order to make the comparison meaningful, we have adjusted the thresholds for the correlation method such that for the worst case scenario signal it achieves the same percentage of sinusoidal peaks correctly classified.

In the top part of fig. 2 the spectrogram of the original signal is shown. Below it the classified spectrograms for both methods are

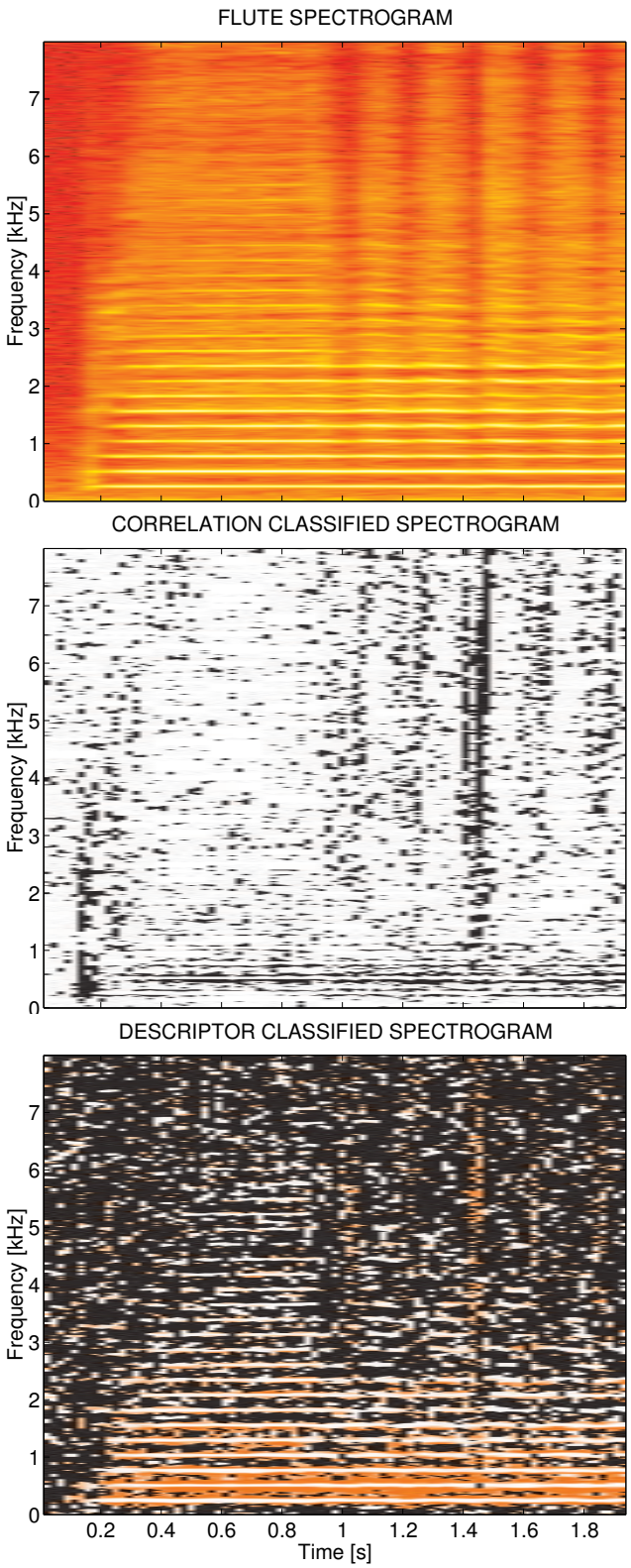

Figure 2: Flute vibrato signal: spectrogram (top), peaks classified by correlation (center), peaks classified by new descriptors (bottom). In the classified spectra the bins of all peaks are colored indicating the classification results as follows: white=sinusoid, black=noise, gray/orange=sidelobe.

sinusoid/non-sinusoid:	$NBD \leq 0.17$ & $DD \leq 0.18$
sidelobe/noise:	$DD \geq 0.28 \| FCD \geq 0.35 \| ELD \geq 0.25$

Table 1: thresholds for sinusoid/nonsinusoid detection in level 1 and for sidelobe/noise classification in level 2 of the binary decision tree.

drawn. The advantage of our approach (bottom) is evident. The bad performance of the correlation method can be explained by means of the distribution of the correlation descriptor for sinusoidal and noise peaks. To reliably detect peaks related to non stationary sinusoids the threshold for the correlation based descriptor has to be extended that much that nearly all noise peaks are considered sinusoids. Refined investigation showed that the results of the proposed method are always superior or equal to the correlation-based approach.

4. CONCLUSIONS

In this paper we have presented new descriptors for the classification of spectral peaks and have described preliminary results comparing the new classification method with a correlation-based approach. We have shown that the proposed descriptors achieve significantly better classification than the correlation-based descriptor if the signal contains only non-stationary sinusoids. The thresholds can be automatically adapted as a function of the desired noise classification error. Further investigation will be concerned with the use of the descriptors to obtain higher level features as for example voiced/unvoiced time frequency sections and partial tracks.

5. ACKNOWLEDGEMENTS

The first author of the paper would like to gratefully acknowledge the financial support of the Gobierno de Navrarra, Spain.

REFERENCES

[1] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003, pp. 344–349.

[2] X. Rodet, "Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models," in *Proc IEEE Time-Frequency and Time-Scale Workshop 97, (TFTS'97)*, 1997, p. ??

[3] S.W. Hainsworth, M.D. Macleod, and P.J. Wolfe, "Analysis of reassigned spectrograms for musical transcription," in *Proc. DAFX98 (Digital Audio Effects Workshop)*, 1998.

[4] P. Depalle, Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden Markov models," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1993, vol. I, pp. 242–245.

[5] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[6] L. Cohen, *Time-frequency analysis*, Signal Processing Series. Prentice Hall, 1995.