

A Frequency-Dependent Bilinear Time-Frequency Distribution for Improved Event Detection

Presented at the 1997 International Computer Music Conference, Thessaloniki, Greece

Andrew Sterian
asterian@umich.edu
<http://www-personal.umich.edu/~asterian>

Gregory H. Wakefield
ghw@eecs.umich.edu
<http://www.eecs.umich.edu/~ghw>

The University of Michigan
Department of Electrical Engineering and Computer Science
Ann Arbor, MI 48109-2122 USA

Abstract

The modal time-frequency distribution (TFD), a member of Cohen's bilinear class of distributions, has recently been applied to high resolution musical analysis and automated transcription. The smoothing by the modal TFD attenuates undesirable crossterms along the surface but also reduces resolution, and this smoothing is an effect applied uniformly over all time and frequency. We propose a frequency-dependent computation of the modal TFD which allows localized, frequency-varying smoothing. In addition, the model enables algorithms that adapt frequency-dependent tradeoffs based upon the data (in onset detection, for example) or upon prior knowledge of frequency content (Western 12-tone tuning, for example). This frequency-dependent model mirrors the multi-resolution properties of the constant-Q spectrogram and wavelet decomposition.

1 Introduction

Automated music transcription systems usually begin with a front-end stage that computes a time-frequency representation of the audio signal. Subsequent derivations of musical parameters, such as onset times and pitch, are estimated from this time-frequency image. An accurate, high-resolution time-frequency image leads to improved parameter estimation, and ultimately, improved transcription. For this reason, Cohen's class of bilinear time-frequency distributions [2], and more specifically the modal TFD [4] have been introduced as methods for generating high-resolution time-frequency images from musical audio signals.

With respect to the more conventional spectrogram and wavelet approaches, the modal TFD is superior in resolution but presents several challenges. In this paper, we consider extensions of the modal TFD to several forms of frequency-dependent processing.

2 Motivation and Background

Beginning with a time-frequency image, we want to estimate musically salient parameters: onset times, offset times, partial frequencies, and so on. Unfortunately, inherent estimation tradeoffs often conflict for different parameters. For example, good time resolution is required to accurately estimate onset times, but good frequency resolution improves pitch estimates. If we can

improve the resolution of the front-end stage, we can perhaps simplify subsequent processing. For this reason, computing a time-frequency representation in a frequency-dependent manner may be desirable. As an example, the constant-Q spectrogram uses varying time-support windows to effect a frequency-dependent tradeoff in time-frequency resolution. A similar frequency-dependent tradeoff is one of the attractive features of wavelet transform approaches.

An adaptive approach to detecting and tracking the partials of an instrument source also benefits from frequency-dependent processing since the parameters of the computation can be adapted independently in each frequency region. Finally, a frequency-dependent adaptive kernel can be used to compute a bilinear time-frequency distribution with smoothing parameters that are locally adapted in a frequency band rather than applied to the global image.

The latter case is relevant to musical transcription. The Wigner time-frequency distribution offers high resolution in time and frequency but requires some form of smoothing in order to suppress false indications of signal energy (commonly referred to as crossterms). Excessive smoothing can negate the gains in resolution; we want to perform the minimum amount of smoothing while still attenuating crossterms. Ideally, we would like to perform varying amounts of smoothing in localized regions of the time-frequency plane depending upon how much smoothing is required in each region.

We consider the modal time-frequency distribution as the front-end processor for transcription. This TFD was specifically designed for the analysis of signals that

¹ This work was supported by an NSERC scholarship to the first author and by the Office of the President of the University of Michigan through the MusEn Project

are well modeled by a sum of sinusoids [4]. Here, we present a brief introduction to the modal TFD as it forms the basis for our new frequency-dependent computation. Cohen's class of bilinear TFD's can be written as² [2]:

$$C(t, \omega; \phi) = \frac{1}{4\pi^2} \int \int \int s^*(u - \tau/2) s(u + \tau/2) \times \phi(\theta, \tau) e^{-j\theta t - j\tau\omega + j\theta u} du d\tau d\theta \quad (1)$$

The input signal is $s(t)$ and the function $\phi(\theta, \tau)$ is called the kernel, described here in the (θ, τ) ambiguity domain. (The (θ, τ) ambiguity domain is related to the (t, ω) time-frequency domain by a two-dimensional Fourier transform (see Figure 1).) As developed by Cohen, the kernel completely determines the properties of the time-frequency representation.

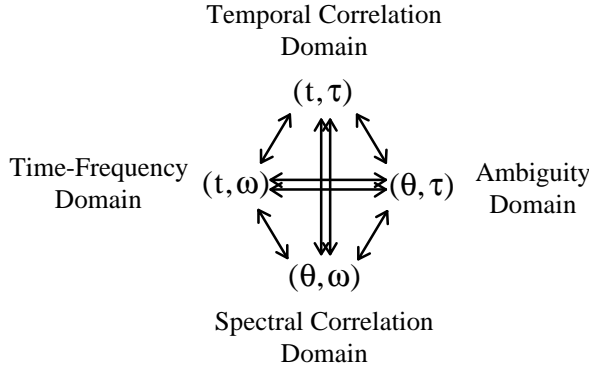


Figure 1: TFD's have equivalent representations in four domains. The domains are related by Fourier transforms (single arrows) or double Fourier transforms (double arrows).

The modal TFD is characterized by the modal kernel [4], which is given by:

$$\phi_{MK}(\theta, \tau) = h_F(\tau) H_T(\theta) \quad (2)$$

where $H_T(\theta)$ is a low-pass filter in the Fourier domain (with corresponding time-domain impulse response $h_T(t)$). The function $h_F(\tau)$ is a time-domain window function that truncates the infinite summation in (1) to allow for realizable implementations. It is the low-pass filter $H_T(\theta)$ that is of interest, however, as it effects the temporal smoothing necessary to suppress cross-terms. Substituting the modal kernel into (1) we have:

$$C_{MK}(t, \omega) = \mathbf{F}_{\tau \rightarrow \omega} \left\{ h_F(\tau) R(t, \tau; h_T(t)) \right\} \quad (3)$$

where

$$R(t, \tau; h_T(t)) = \int h_T(t - u) s(u + \frac{\tau}{2}) s^*(u - \frac{\tau}{2}) du \quad (4)$$

is the time-smoothed local autocorrelation function. The notation $\mathbf{F}_{\tau \rightarrow \omega}$ indicates the Fourier transform from the τ -dimension to the ω -dimension. We have dropped the scaling by $4\pi^2$ and, for brevity, will similarly drop all scaling factors in the sequel.

For suppression of crossterms, the low-pass cutoff frequency of $H_T(\theta)$ (or its Fourier dual, $h_T(t)$ as in (4) above) must be chosen to be smaller than the smallest frequency separation of components in the sum-of-sinusoids signal. The cutoff frequency should, however, be as large as possible in order to preserve temporal detail. For a single musical note of known pitch, the cutoff frequency is usually chosen to be slightly less than the fundamental frequency. The modal kernel has been used in this fashion for high-resolution analysis of piano notes [3].

When the input signal is polyphonic, however, the cutoff frequency must be chosen arbitrarily as the minimum separation of partials in frequency is an unknown quantity. Even if we had knowledge regarding the tuning used in the music, we could only infer a minimum partial frequency *spacing as a function of frequency*. For example, if we knew that the music was based on the Western 12-tone scale, then in the frequency region of 115 Hz we could expect partials at 110 Hz (A2) and 116.5 Hz (A#2). The cutoff frequency in this region could be chosen to be 6 Hz. In the neighborhood of 900 Hz we could expect partials at 880 Hz (A5) and 932.3 Hz (A#5). The cutoff frequency in this region would be 52 Hz and we would preserve more temporal detail due to a smaller degree of smoothing.

Unfortunately, we cannot effect a varying degree of smoothing since the cutoff frequency enters as a parameter in the computation of (4) which occurs prior to the Fourier transform in (3). We could compute several TFD's, each with a different cutoff frequency, and then combine the results (as proposed in [4] for the constant-Q modal TFD), but this approach is computationally demanding and unnecessary as we shall see.

3 Derivation

The modal TFD in (3) is computed by forming the smoothed autocorrelation in the temporal correlation domain and then taking the Fourier transform to enter the time-frequency domain (see Figure 1). An alternative approach is to form the smoothed autocorrelation in the spectral correlation domain and then compute a Fourier transform to once again end up in the time-frequency domain. We show that this approach gives us the desired frequency-dependent computation for the modal TFD.

We can rewrite equation (1) in an equivalent form that introduces computation in the spectral correlation domain [1]:

² Unless otherwise noted, all integrals are definite integrals over the entire real line.

$$C(t, \omega; \phi) = \int \int \Phi(\theta, \omega - \omega') S(\omega' + \theta/2) \times S^*(\omega' - \theta/2) e^{j\theta t} d\omega' d\theta \quad (5)$$

Now, $S(\omega)$ is the Fourier transform of the input signal $s(t)$, and $\Phi(\theta, \omega)$ is the kernel function expressed in the spectral correlation domain. For the modal kernel, we have:

$$\Phi_{MK}(\theta, \omega) = H_T(\theta) H_F(\omega) \quad (6)$$

In the temporal correlation domain computation of the modal TFD we required $h_F(\tau)$ to be a window function of finite support in order to limit the support of the smoothed correlation function $R(t, \tau; h_T(t))$. In the spectral correlation domain, $H_T(\theta)$ is band-limited by design and effectively limits the support of the integration in (5). $H_F(\omega)$ (the Fourier transform of $h_F(\tau)$) is not needed for realizability and we can remove it from consideration. Since $H_F(\omega)$ effects smoothing along the frequency dimension and we want to preserve as much frequency resolution as possible, we set $H_F(\omega) = \delta(\omega)$. That is, we choose $H_F(\omega)$ so that there is no smoothing in the frequency dimension. With this substitution, the modal TFD computed in the spectral correlation domain becomes:

$$C_{MK}(t, \omega) = \int H_T(\theta) S(\omega + \frac{\theta}{2}) S^*(\omega - \frac{\theta}{2}) e^{j\theta t} d\theta \quad (7)$$

Notice that frequency, ω , is now a parameter in the computation of (7), not a Fourier variable. This means that for each value of ω , we can use a different low-pass filter $H_T(\theta)$. This freedom forms the basis of our frequency-dependent computation.

We can simplify (7) somewhat if we can assume that $s(t)$ is real valued and that $H_T(\theta)$ is real and symmetric. Finally, if we assume that $H_T(\theta)$ is an "ideal" low-pass filter with cutoff frequency θ_c then we can write the modal TFD as:

$$C_{MK}(t, \omega) = \int_0^{\theta_c} \text{Re} \left\{ S(\omega - \frac{\theta}{2}) S(-\omega - \frac{\theta}{2}) e^{-j\theta t} \right\} d\theta \quad (8)$$

Allowing θ_c to vary with frequency ω implements frequency-dependent smoothing.

In the discrete implementation of the above, the "ideal" low-pass filter is actually designed using the frequency sampling method so that samples of $H_T(\theta)$ are 1 at computation points below θ_c and 0 elsewhere. This leads to a low-pass filter with some ripple in both the passband and stopband which have a negligible effect on the final outcome. (In addition, θ_c is usually small

compared to the sampling rate in a discrete implementation so that the reduced range of the integral means that the cost of computing the discrete approximation of the integral is greatly reduced.)

A difficulty with the above formulation involves the Fourier transform $S(\omega)$. For long signals, this computation may become impractical and some form of overlapped windowing is necessary. The choice of window length and degree of overlap is subject to tradeoffs in frequency resolution, temporal localization, and bias. We do not discuss these issues further in this paper except to mention that different approximations to $S(\omega)$ can be used at a given time to implement these tradeoffs in a frequency-dependent manner.

4 Applications

In Figure 2 we show a comparison between the spectrogram and the modal kernel computed in the spectral correlation domain. The source material consists of a B3-C4 piano chord played 5 times in rapid succession within a 800 ms span. The diagrams on the left side of the figure show frequency variation as a function of time in the neighborhood of 247 Hz and 262 Hz, corresponding to the fundamentals of B3 and C4, respectively. These graphs were obtained by picking peaks at the frequencies of interest. The entire 800 ms record is shown. The diagrams on the right side of the figure show amplitude variation as a function of time for the fundamental of C4 at 262 Hz. The amplitude scale is logarithmic, and once again, the entire 800 ms record is shown. The 5 note-onsets are visible in the amplitude variation graphs.

Graphs (a), (b), and (c) show results for the spectrogram computed with a Hamming window of duration 120 ms, 140 ms, and 160 ms, respectively. The classic time-frequency tradeoff is visible in these graphs. As the window becomes longer, the resolution in frequency improves but the amplitude profile is smoothed out, making the onsets more difficult to discern. The crossovers visible in the frequency variation graphs (the left side) of (a) and (b) are due to insufficient frequency resolution; the peaks come so close together that at certain times they merge and cease to be resolvable. Only at a window duration of 160 ms does the spectrogram clearly resolve the peaks for the entire duration of the 5 notes. With this window duration, however, the 3rd and 5th note onsets are nearly completely obscured.

The graph in (d) shows the frequency-dependent modal kernel computation for the same source material. Both high frequency and high time resolution are achieved simultaneously (note how the 3rd and 5th onsets are more clearly resolved).

Figure 3 shows one potential benefit of the

frequency-dependent modal TFD. This figure shows amplitude profiles of the same 5-chord sequence at 1044 Hz, the 5th partial of the C4 note. The profiles have been intentionally separated by 2 dB for clarity. The higher curve represents the usual modal TFD, while the lower curve shows the frequency-dependent computation. Note how in the lower curve the 3rd and 5th note onsets have a larger slope, and the separation between notes is generally better. This improvement is due to the lower smoothing at higher frequencies. In this case, the assumption that the source material conformed to a Western 12-tone scale was used to set the smoothing frequency θ_c based upon expected differences in partial frequencies.

5 Conclusions

We have presented a frequency-dependent implementation of the modal TFD as a front end processor for automated transcription. This new approach enables computationally efficient processing strategies that vary with frequency. As an example, frequency-dependent smoothing was shown to preserve more time resolution at higher frequencies, thereby reducing the uncertainty in onset detection.

References

[1] Loughlin, P. J., J.W. Pitton, and L.E. Atlas 1993. "Bilinear Time-Frequency Representations: New Insights and Properties," IEEE Trans. Sig. Proc., Volume 41, Number 2, pp.750-767.

[2] Cohen, L. 1995. *Time-Frequency Analysis*, New Jersey: Prentice-Hall, Inc.

[3] Guevara, R.C.L. 1997. *Modal Distribution Analysis and Sum of Sinusoid Synthesis of Piano Tones*, Ph.D. Dissertation, University of Michigan.

[4] Pielemeier, W.J., G.H. Wakefield, and M.H. Simoni 1996, "Time-Frequency Analysis of Musical Signals," Proc. IEEE, Volume 84, Number 9, pp.1216-1230.

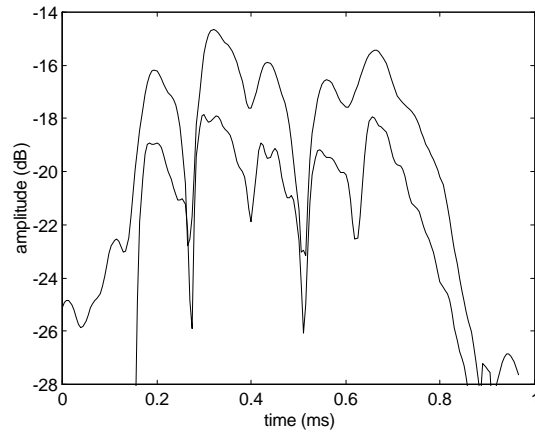


Figure 3: B3-C4 piano chord amplitude profile at 1044 Hz. The higher curve is derived from the standard modal TFD computation while the lower curve comes from the frequency-dependent modal TFD. Note the sharper transition at the 3rd and 5th onsets and the general improvement in note separation.

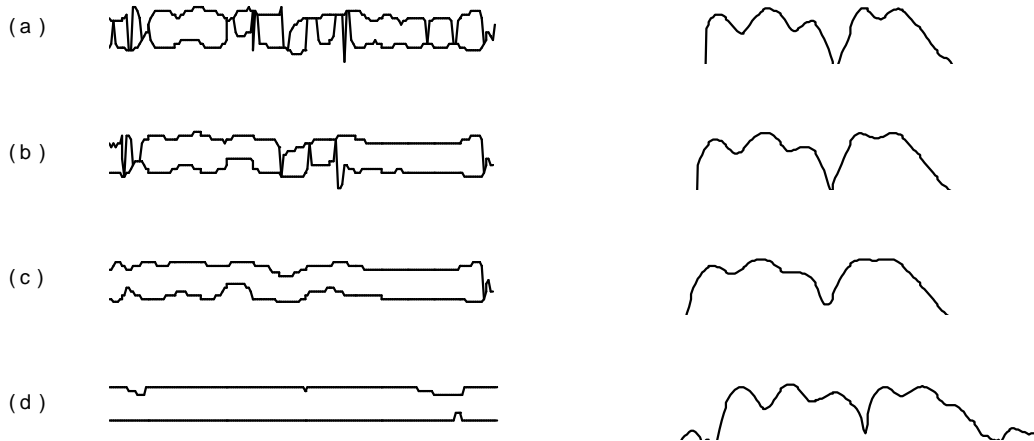


Figure 2: Spectrogram and modal TFD of fundamentals for 5 successive B3-C4 piano chords. Parts (a), (b), and (c) show the spectrogram using Hamming windows of length 120 ms, 140 ms, and 160 ms, respectively. Part (d) shows the modal TFD analysis. The left side shows frequency-vs.-time graphs near 247 Hz and 262 Hz (i.e., the fundamentals of B3 and C4). The right side shows log-amplitude-vs.-time graphs at 262 Hz. The modal TFD analysis results in good simultaneous time and frequency resolution, while the spectrogram analysis sacrifices one for the other. Both time axes extend over 800 ms.

