IMPROVED HIDDEN MARKOV MODEL PARTIAL TRACKING THROUGH TIME-FREQUENCY ANALYSIS

Corey Kereliuk[*]
SPCL, Music Technology
Schulich School of Music
Montréal, Canada
corey.kereliuk@mail.mcgill.ca

Philippe Depalle[*]
SPCL, Music Technology
Schulich School of Music
Montréal, Canada
depalle@music.mcgill.ca

ABSTRACT

In this article we propose a modification to the combinatorial hidden Markov model developed in [1] for tracking partial frequency trajectories. We employ the Wigner-Ville distribution and Hough transform in order to (re)estimate the frequency and chirp rate of partials in each analysis frame. We estimate the initial phase and amplitude of each partial by minimizing the squared error in the time-domain. We then formulate a new scoring criterion for the hidden Markov model which makes the tracker more robust for non-stationary and noisy signals. We achieve good performance tracking crossing linear chirps and crossing FM signals in white noise as well as real instrument recordings.

1. INTRODUCTION

Additive models for sound synthesis are popular due to their potential for high quality synthesis and their flexibility with respect to sound transformations and control. The additive model is given as:

$$x(t) = \Re\left(\sum_{l=1}^{L(t)} a_l(t) e^{j\phi_l(t)}\right) \tag{1}$$

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

where $a_l(t)$, $\omega_l(t)$, and $\phi_l(0)$ are the amplitude, frequency and initial phase of the l^{th} partial, respectively. Typically, these parameters are evaluated for every $t = nH/F_s$ where n is the sample number, F_s is the sampling frequency and H is the hop size. The model parameters are undersampled and will need to be interpolated in order to calculate the signal. Before we can perform this interpolation we must first organize the parameter estimates into trajectories (ie: assign each parameter to a trajectory, l, at every time frame). This process is referred to as peak continuation or partial tracking. In this paper we adopt the latter terminology.

Many different strategies and algorithms have been developed for partial tracking over the years. McAulay and Quatieri (MQ) developed one of the first partial tracking algorithms in the context of speech coding [2]. Their method uses a simple metric designed to minimize local frequency differences between analysis frames. The MQ method ignores the fact that some peaks may be spurious and uses a quasi-stationary signal assumption. The MQ method was modified in [3] to allow partial trajectories to 'sleep' and in [4] for use with a reassigned bandwidth enhanced model. New strategies based on linear prediction coding (LPC) have been presented in [5] and [6]. The LPC method uses past samples in each trajectory to predict the best match in the current frame and can interpolate missing peaks. In [7] an adaptive method is presented which uses B-splines to estimate the parameters of the additive model. The authors in [1] developed a hidden Markov model (HMM) for partial tracking which optimizes the partial trajectories jointly across an analysis window. This method considers spurious peaks, and performs well in a number of difficult tracking situations.

In this paper we describe several improvements to the HMM in [1] that make it even more suitable for non-stationary and noisy signal analysis. We describe how the Wigner-Ville distribution can be used to estimate the frequency and chirp rate of spectral peaks, and then illustrate the potential of this technique for detecting crossing frequency tracks in the presence of noise. We also describe how to estimate the amplitude and initial phase of detected peaks. In the second part of this paper we describe our HMM scoring criterion, and provide sample results produced by our system.

The rest of this paper is organized into the following sections. In section 2 we give an overview of our partial tracking system. In section 3 we explain the methodology we used to estimate spectral parameters, and in section 4 we describe the HMM partial tracking. In section 5 we show examples which demonstrate the efficacy of our technique.

2. OVERVIEW

The block diagram in figure 1 shows the basic elements of our additive analysis/synthesis system. As illustrated the system can be roughly divided into three stages: preprocessing, parameter estimation, and synthesis.

The intent of the preprocessing stage is to mitigate the effect of interference terms due to the quadratic nature of the Wigner-Ville distribution (discussed in section 3.1).

The short-time spectrum is computed by windowing the input signal and applying the fast Fourier transform (FFT). The local maxima are then extracted from the FFT and used to control a bank of linear phase, finite impulse response band-pass filters. Linear phase filters are used so that the initial phase can be recovered at a later stage. Each band-pass filter is centered on a FFT peak, and cut-off frequencies are taken midway between adjacent peaks.

Ideally, the output from each band-pass filter would be a mono-component signal, although this is not absolutely required since our system is capable of estimating the parameters of low order

[*] Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)

multicomponent signals. In section 3 we show how the Wigner-Ville distribution and Hough transform can be used to estimate the parameters of each signal produced by the preprocessing stage.

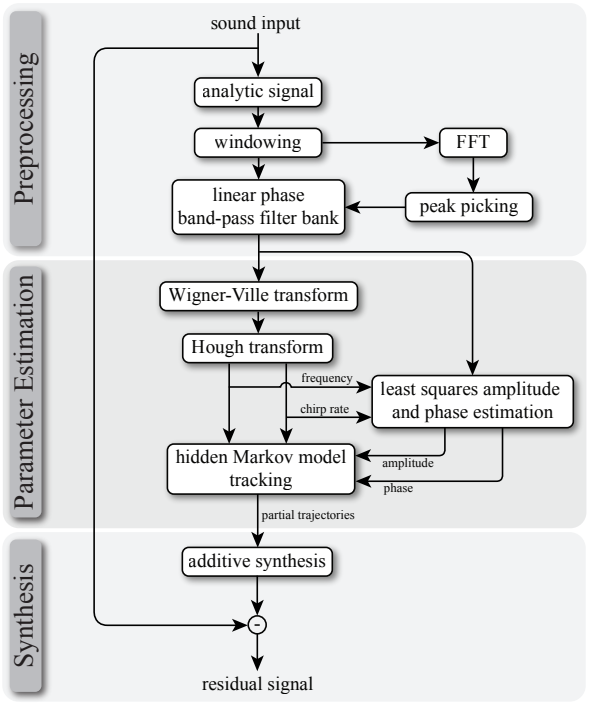

Figure 1: *Block diagram of proposed system.*

3. PARAMETER ESTIMATION

3.1. The Wigner-Ville Distribution

The Wigner-Ville distribution (WVD) was first described in [8], in the context of quantum thermodynamics and then again in [9], in the context of signal analysis. The WVD is a member of Cohen's class of bilinear time-frequency distributions [10] which includes the often used spectrogram, and many other time-frequency distributions used in the audio community [11][12].

We are motivated to use the WVD because it exhibits a superior time-frequency resolution to the spectrogram (in fact, it can be shown that the spectrogram is a smoothed version of the WVD). The equation for the WVD is given as [13]:

$$X_{WVD}(t,\omega) = \int_{-\infty}^{\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j\omega\tau}d\tau \quad (3)$$

If x is real, its analytic associate is typically used in order to remove negative frequencies. Additionally, the analytic associate prevents aliasing from negative frequencies in the discrete WVD (the Nyquist frequency is 4x the highest frequency in the discrete WVD). It is informative to examine the WVD of a complex linear chirp. A complex linear chirp is defined as:

$$x(t) = ae^{j\Phi(t)} \quad (4)$$

$$\Phi(t) = \phi_0 + \omega_0 t + \pi\alpha t^2 \quad (5)$$

where a is the amplitude, ϕ_0 is the initial phase, ω_0 is the frequency at time zero, and α is the chirp rate. The chirp has the following instantaneous frequency (IF) law:

$$\Phi'(t) = \frac{d\Phi}{dt} = \omega_0 + 2\pi\alpha t \quad (6)$$

The WVD of the chirp is:

$$X_{WVD}(t,\omega) = \int_{-\infty}^{\infty} a^2 e^{j(\Phi(t+\tau/2)-\Phi(t-\tau/2))}e^{-j\omega\tau}d\tau \quad (7)$$

$$= a^2\int_{-\infty}^{\infty} e^{-j(\omega-\omega_o-2\pi\alpha t)\tau}d\tau \quad (8)$$

$$= 2\pi a^2\delta(\omega-\omega_o-2\pi\alpha t) \quad (9)$$

This expression is non-zero when $\omega = \omega_0 + 2\pi\alpha t$, and thus the WVD forms a ridge in the time-frequency plane equal to the IF law of the chirp. For this reason the WVD is well suited to the analysis of first order FM signals.

A well known problem with the WVD is the occurrence of inner and outer interference terms which tend to obfuscate its interpretation. Outer interference terms occur in the WVD of multicomponent signals due to cross terms in the quadratic expansion of the signal. Figure 2 illustrates cross terms between two linear chirps. Inner interference terms result from non-linear modulations of the IF-law and may appear in monocomponent signals such as the FM signal in figure 3.

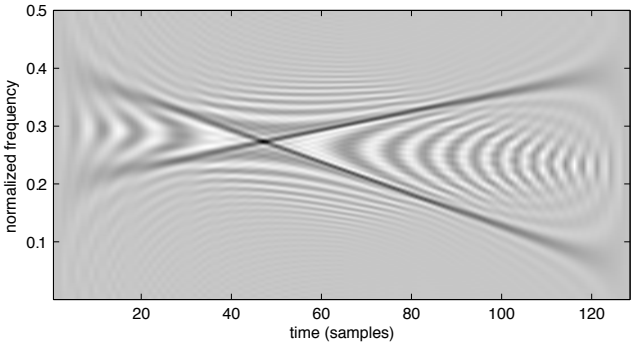

Figure 2: *WVD of crossing linear chirps. Outer interference (cross) terms clearly visible.*

If we restrict our analysis window such that the windowed signal has a near linear IF law we can reduce the effect of inner interference terms. Likewise, if we use a bank of bandpass filters (as in figure 1) we can largely eliminate the effect of outer interference terms from out-of-band partials. In the sequel we demonstrate how the Hough transform can be used to estimate the parameters of linear FM signals even when there are crossing chirps in the filter band.

3.2. The Hough Transform

The Hough transform (HT) is an image processing tool used to find lines and other complex patterns in images [14]. The HT exploits the point-line duality in order to map image pixels to a 2D slope-intercept parameter space. We can apply the HT to the WVD in

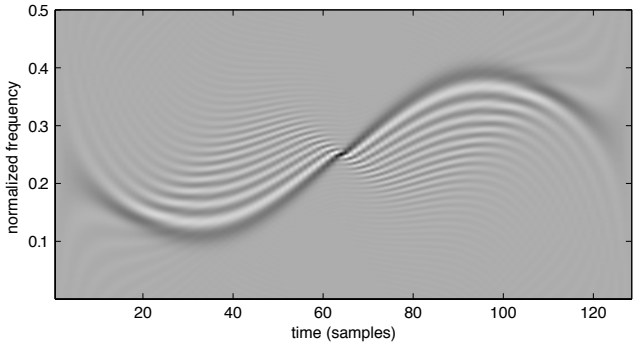

Figure 3: *WVD of monocomponent signal with sinusoidal IF law. Inner interference terms clearly visible.*

order to search for straight lines (frequency ridges) in the time-frequency plane. The HT of the WVD is an integration over all straight lines in the time-frequency plane:

$$X_{WH}(\omega_0, \alpha) = \int X_{WVD}(t, \omega_0 + 2\pi\alpha t)dt \qquad (10)$$

Peaks in the HT give the initial frequency ω_o, and chirp rate α, of ridges in the time-frequency plane. It has been shown that the outer interference terms of the WVD are amplitude modulated and zero mean so that their energy contribution is reduced via the integration in equation 10 [15]. The HT of the WVD of two crossing linear chirps is shown in figure 4. At SNR levels greater than 2dB estimates from the HT approach the Cramer-Rao bounds [15].

Using the HT in conjunction with the WVD allows us to detect multiple overlapping chirps which is an advantage over other first order FM estimators such as [16][17]. As described previously, we limit the number of partials in the HT by using a bank of linear phase band-pass filters. This is because the number of outer interference terms grows at a rate of $L(L-1)$, where L is the number of partials in the WVD. Clearly the outer interference terms will become unwieldy if the number of partials is not limited. Thus we use band-pass filters to reduce the number of partials in each analysis.

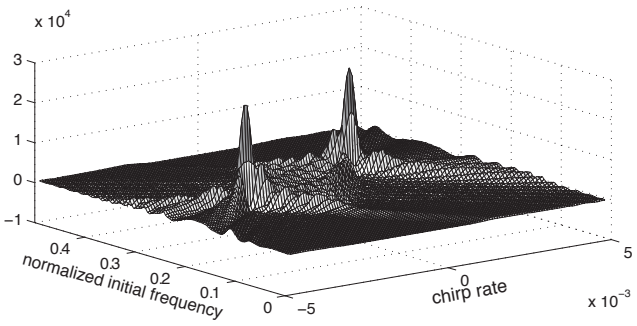

Figure 4: *Hough transform of WVD of crossing linear chirps.*

3.3. Initial Phase and Amplitude Estimation

It is not possible to estimate the initial phase using the WVD because it is an energy distribution. In order to estimate the initial phase and amplitude we use a least squares error estimate in the time domain. This is done by minimizing the following matrix equation:

$$\mathbf{x} - \begin{bmatrix} \hat{\mathbf{x}}_1 & \hat{\mathbf{x}}_2 & \cdots & \hat{\mathbf{x}}_N \end{bmatrix} \begin{bmatrix} a_1 e^{j\phi_1} \\ a_2 e^{j\phi_2} \\ \vdots \\ a_N e^{j\phi_N} \end{bmatrix} \qquad (11)$$

where $\mathbf{x}$ is a column vector containing time domain samples from the original signal, $\hat{\mathbf{x}}_\mathbf{i}$ is a column vector containing time domain samples from the i^{th} chirp estimate, and $a_i e^{j\phi_i}$ is the amplitude and initial phase of the i^{th} chirp to be estimated.

The least squares technique allows us to estimate the amplitude and initial phase for crossing chirps, which would be difficult using the short time Fourier transform (STFT). Figure 5 shows the phase error from two crossing constant amplitude FM modulated partials. The solid line shows the error in the STFT phase estimate, and the dashed line shows the error in the least squares phase estimate.

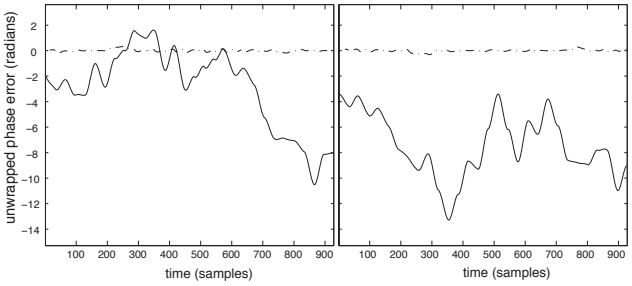

Figure 5: *Phase error for two crossing constant amplitude FM modulated partials. Partial 1 (left). Partial 2 (right). The STFT phase error is shown using a solid line, and the least squares phase error is shown using a dashed line.*

4. HMM PARTIAL TRACKING

Hidden Markov models are used to describe processes which emit observable/measurable symbols that occur jointly with a set of underlying hidden states [18]. The partial tracking problem can be formulated as an HMM if we consider spectral peaks as the observable symbols emitted from a set of underlying partial trajectories.

Using the same notation and definition from [1], the elements of the HMM are:

- h_k is the number of spectral peaks at time k.
- $I_k(j)$ is the trajectory assigned to peak j at time k. For useful trajectories $I_k(j) > 0$. $I_k(j) = 0$ is reserved for spurious trajectories.
- $S_k = (\mathbf{I_{k-1}}, \mathbf{I_k})$ is the hidden state at time k (the set of partial trajectories connecting peaks at frame $k-1$ to the peaks at frame k).
- $\omega_k(j)$, $\alpha_k(j)$, $a_k(j)$ are the frequency, chirp rate, and amplitude of the j^{th} peak at time k. Notice that in the work presented here the chirp rate is explicitly measured, whereas

in [1] the chirp rate was deduced as a frequency difference between consecutive analysis frames.

- $\theta_k(j, r, t)$ is the *matching criterion* between peaks j, r and t at times k, $k-1$, and $k-2$, respectively. The matching criterion is used to develop an analytical expression for the state transition probabilities in the HMM. The principal difference between our HMM and the one developed in [1] is our definition of the matching criterion.

In this model the probability of observing a set of spectral peaks either zero or one, and thus the HMM is purely combinatorial. The fact that some peaks may be due to noise/noisy measurements is taken into account when defining the state transition probabilities.

4.1. State Transition Probabilities

The matching criterion assigns a score to every three point path defined by the peaks j, r, and t in frames k, $k-1$, and $k-2$, respectively ($T = H/F_s$ is the time between analysis frames):

$$\theta_k(j,r,t) = \begin{cases} e^{-\frac{\Delta\omega_k(j,r)^2 + \Delta\omega_k(r,t)^2}{\sigma_\omega^2}} e^{-\frac{\Delta a_k(j,r,t)^2}{\sigma_a^2}} & \text{if } I_k(j) > 0 \\ 1 - (1-\mu)e^{-\frac{\Delta\omega_k(j,r)^2 + \Delta\omega_k(r,t)^2}{\sigma_\omega^2}} \cdot & \\ \qquad e^{-\frac{\Delta a_k(j,r,t)^2}{\sigma_a^2}} & \text{if } I_k(j) = 0 \end{cases} \tag{12}$$

where:

$$\Delta\omega_k(j,r) = \left[\omega_{k-1}(r) + 2\pi\alpha_{k-1}(r)\frac{T}{2}\right] - \left[\omega_k(j) - 2\pi\alpha_k(j)\frac{T}{2}\right] \tag{13}$$

and:

$$\Delta a_k(j,r,t) = [a_k(j) - a_{k-1}(r)] - [a_{k-1}(r) - a_{k-2}(t)] \tag{14}$$

When evaluating the matching criterion we consider each peak as either a useful peak or spurious peak. We must enumerate every possible combination of useful and spurious paths in order to capture the underlying trajectory. Equation 13 evaluates the inter-frame frequency error based on the estimated chirp rate (figure 6 depicts this equation). Equation 14 records the difference in amplitude change between frames. Small values of $\Delta\omega_k$ and Δa_k will lead to high useful scores (low spurious scores) in the matching criterion. In other words the matching criterion promotes the continuity of frequency and amplitude trajectories, and penalizes discontinuities. The parameters σ_ω, σ_a, μ are used to control the sensitivity of the matching criterion.

In [1] the matching criterion was also designed to preserve the continuity of frequency slopes, however, with no explicit chirp rate estimate their criterion was maladjusted in certain tracking situations. For example consider the set of peaks shown in figure 7. The peaks in the highlighted path have a very high continuity according to the criterion in [1]. Our new criterion, which benefits from the chirp rate estimate, would reject this path as spurious since the chirp rate estimate leads to a discontinuous frequency trajectory.

Given the matching criterion in equation 12 we define the state transition score as:

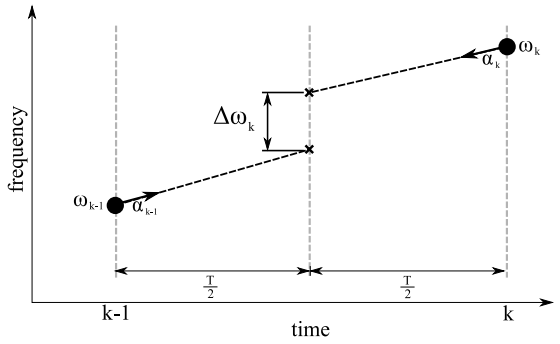

Figure 6: *Illustration of frequency scoring from equation 13.*

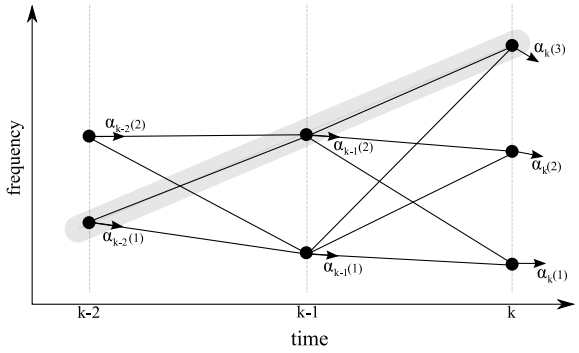

Figure 7: *Spectral peaks at three analysis frames. Solid lines indicate all possible trajectories.*

$$G(S_{k-1}, S_k) = \prod_{j=1}^{h_k} \theta_k(j,r,t) \tag{15}$$

where r and t are chosen such that trajectories are matched across states: $I_{k-2}(t) = I_{k-1}(r) = I_k(j)$. $\mathbf{G}$ is a state transition matrix, which can be normalized to make the state transitions scores into true probabilities. Since our HMM is not intended to be generative (our application is decoding) we do not need to normalize our state transition matrix. The optimal path through the trellis of spectral peaks is then decoded by applying the Viterbi algorithm [18].

4.2. High Level Considerations

We use the same high-level procedure to detect partial birth/death as was used in [1]. The Viterbi decoding is performed on a window of several analysis frames, and this window slides along the temporal axis one frame at a time. The birth/death of partials is detected by searching for appearing/disappearing partials from frame to frame.

4.3. Computational Cost/Implementation Details

The computational tractability of the HMM is strongly dependant on the number of peaks in each analysis frame. If h_k is the number of peaks in the current frame, then there are $N_k = h_k \cdot h_{k-1} \cdot h_{k-2}$ paths that can be drawn between the peaks in frames $k-2$, $k-1$, and k. For these N_k paths we must consider all cases (ie: that there are 0 useful trajectories and h_k spurious trajectories, 1 useful

trajectory and $h_k - 1$ spurious trajectories, ..., h_k useful trajectories and 0 spurious trajectories). The number of states that must be computed for a single frame are:

$$\sum_{p=0}^{h_k} \frac{N_k!}{p!(N_k - p)!} \tag{16}$$

Clearly, the number of states grows exponentially with the number of peaks detected in each analysis frame. In order to make the HMM computationally tractable we have employed a number of strategies. First, we disallow trajectories that have large frequency deviations. Second, we partition the frequency domain into a number of overlapping windows. This reduces N_k and h_k in each window, and significantly reduces the number of combinations computed in 16. In our implementation we use a variable window size and frequency overlap factor of 50 % and then join overlapping trajectories into single trajectories after the Viterbi algorithm runs.

5. RESULTS

Figure 8 shows tracking results for two crossing chirps in a short burst of white gaussian noise. The signal is well modeled as evidenced by the lack of chirp signals in the residual.

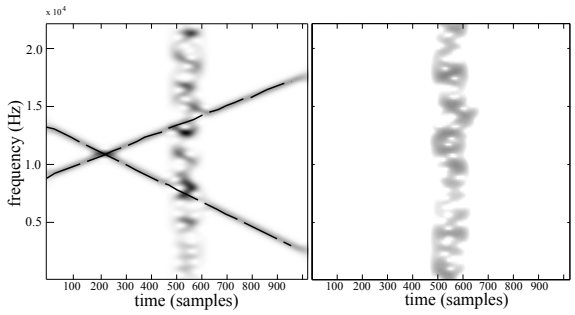

Figure 8: *Spectrogram of crossing chirps with white gaussian noise burst (SNR -10 dB). Detected partial tracks superimposed in dashed black lines (left). Residual spectrogram (right).*

We are able to track even highly non-stationary signals such as crossing FM modulated signals embedded in white gaussian noise (see figure 9).

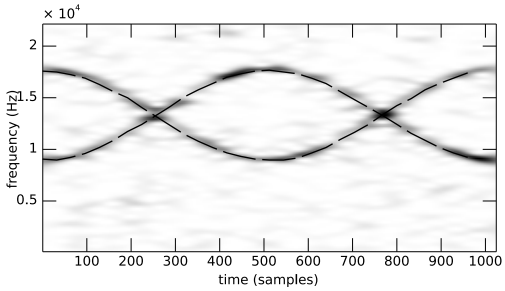

Figure 9: *Spectrogram of crossing FM signals in white gaussian noise (SNR 2 dB). Detected partial tracks superimposed in dashed black lines.*

Figure 10 compares the tracking performance of our HMM with the one from [1]. Notice how our system is able to track fast modulations, whereas the tracker from [1] has trouble distinguishing between partials at key frames.

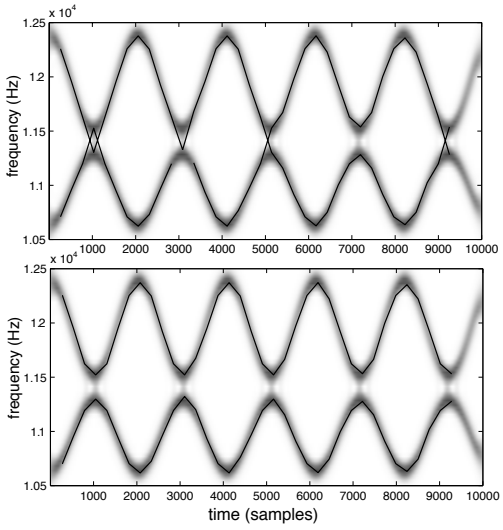

Figure 10: *Tracking performance of the HMM from [1] (top) vs. the system presented in this paper (bottom).*

In the following examples we use the reconstruction signal to noise ratio (R-SNR) to help quantify our results. The R-SNR is defined as:

$$\text{R-SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1}(x(n) - \hat{x}(n))^2}\right) \tag{17}$$

where $x(n)$ is the original signal, and $\hat{x}(n)$ is the estimated signal from the additive model. The R-SNR is a useful measure if the residual signal energy is primarily due to analysis errors (and not noise). Figure 11 shows the tracking results for an upward glissando on a violin. The R-SNR of the glissando is 39.5 dB.

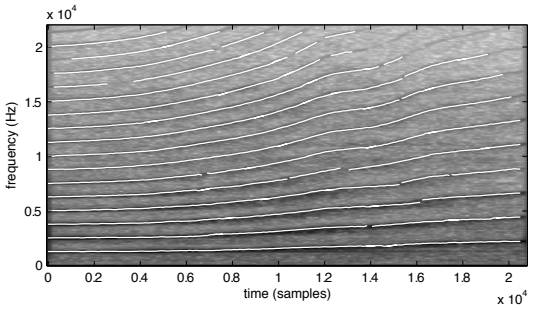

Figure 11: *Spectrogram of upward glissando on a violin. Detected partials superimposed in white. 39.5 dB R-SNR.*

Figure 12 shows the tracking results for a vocal falsetto with strong vibrato. The R-SNR for this signal is 60.7 dB.

Figure 13 shows overlapping upward and downward glissandi on a violin. We are able to detect many of the crossing partials in this difficult example. The R-SNR of this signal is 10.2 dB.

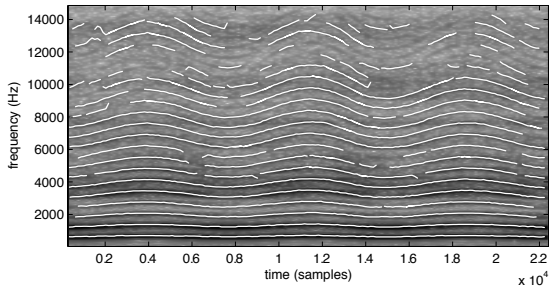

Figure 12: *Spectrogram of vocal falsetto with strong vibrato. Detected partials superimposed in white. 60.7 dB R-SNR.*

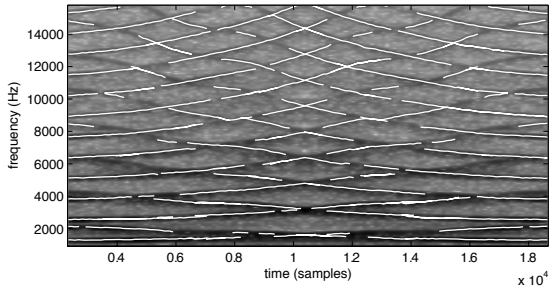

Figure 13: *Spectrogram of overlapping upward and downward glissandi on a violin. Detected partials superimposed in white. 10.2 dB R-SNR.*

6. CONCLUSIONS AND FUTURE WORK

In this paper we have outlined the major elements in an HMM-based partial tracker for additive synthesis. We have demonstrated how the Wigner-Ville and Hough transforms can be used to estimate the parameters of a first order FM model, and shown how these estimates can improve the matching criterion for HMM-based partial tracking. We have devised a number of strategies to make the HMM computationally tractable, and have implemented the complete system in Matlab. We have achieved good tracking results for synthetic sounds and monophonic instrument recordings. At present we are working to improve the management of crossing partials in polyphonic instrument recordings. We are also experimenting with linear prediction in order to interpolate/join closely spaced trajectories.

7. ACKNOWLEDGEMENTS

This research is supported by a grant from NSERC (Natural Sciences and Engineering Research Council of Canada).

8. REFERENCES

[1] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden Markov models," *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 242–245, 1993.

[2] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, July 1986.

[3] X. Serra and J. Smith III, "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.

[4] K. Fitz and L. Haken, "Bandwidth enhanced sinusoidal modeling in lemur," *Proceedings of the International Computer Music Conference (ICMC)*, pp. 154–157, 1995.

[5] M. Lagrange, S. Marchand, M. Raspaud, and J.B. Rault, "Enhanced partial tracking using linear prediction," *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 141–146, 2003.

[6] M. Lagrange, S. Marchand, and Rault, "Tracking partials for the sinusoidal modeling of polyphonic sounds," *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 229–232, 2005.

[7] A. Röbel, "Adaptive additive modeling with continuous parameter trajectories," *IEEE transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1440–1453, 2006.

[8] E. Wigner, "On the quantum theory for thermodynamic equilibrium," *Physical Review*, vol. 40, pp. 749–759, 1932.

[9] J. Ville, "Theorie et applications de la notion de signal analytique," *Cables et Transmission*, vol. 2, no. 1, pp. 61–74, 1948.

[10] L. Cohen, "Time-frequency distributions - A review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[11] T. Lysaght and J. Timoney, "Timbre morphing using the modal distribution," *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 191–194, 2002.

[12] J.J. Wells and D.T. Murphy, "Real-time partial tracking in an augmented additive synthesis system," *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 93–96, 2002.

[13] T. Claasen and W.F.G. Mecklenbrauker, "The Wigner distribution - A tool for time-frequency signal analysis. I. continuous time signals," *Philips Jl Research*, vol. 35, pp. 217–250, 1980.

[14] P.V. Hough, "Methods and means to recognize complex patterns," *U.S. Patent 3.069.654*, 1962.

[15] S. Barbarossa, "Analysis of multicomponent LFM signals by a combined Wigner-Hough transform," *IEEE Transactions on Signal Processing*, vol. 43, no. 6, pp. 1511–1515, 1995.

[16] M. Abe and J.O. Smith III, "AM/FM rate estimation for time-varying sinusoidal modeling," *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2005.

[17] M. Betser, P. Collen, G. Richard, and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 505–517, 2008.

[18] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.