

A model-based approach to partial tracking for musical transcription

Presented at the 1998 SPIE Annual Meeting, San Diego, California

Andrew Sterian and Gregory H. Wakefield^a

Department of Electrical Engineering and Computer Science, The University of Michigan

ABSTRACT

We present a new method for musical partial tracking in the context of musical transcription using a time-frequency Kalman filter structure. The filter is based upon a model for the evolution of a partial in amplitude and frequency. The parameters of this model are determined from a statistical analysis of partial behavior across a wide range of pitch from four brass instruments. Statistics are computed independently for the partial attributes of frequency and log-power first differences. We present observed power spectral density shapes, total powers, and histograms, as well as least-squares approximations to these. We demonstrate that a Kalman filter tracker using this partial model is capable of tracking partials in music. We discuss how the filter structure naturally provides quality-of-fit information about the data for use in further processing and how this information can be used to perform partial track initiation and termination within a common framework. We propose that a model-based approach to partial tracking is preferable to existing approaches which generally use heuristic rules or birth/death notions over a small time neighborhood. The advantages include better performance in the presence of cluttered data (e.g., multi-voice material) and simplified tracking over missed observations.

Keywords: musical transcription, Kalman filter, music analysis

1. INTRODUCTION

The individual notes of many musical instruments can be described by an additive synthesis model comprising harmonically-spaced partials of time-varying amplitude but nearly time-independent frequency. In a time-frequency representation, these partials appear as well-defined horizontal structures when viewing the three-dimensional surface from above. For this reason, existing approaches to musical transcription generally form time-frequency images from audio data to reveal such horizontal structure. Subsequent integration tasks, such as the formation of notes from partials, rely upon the information conveyed by these structures, but relatively little systematic attention has been given as to how best to extract partial features from the time-frequency data. Heuristic rules or birth/death notions are commonly used to connect peaks that have been obtained from a peak picking analysis in a small time neighborhood. While such approaches are successful in handling single isolated notes, they generally break down when extended to multiple notes played simultaneously. For example, the presence of two partials closely spaced in frequency can lead to a beating effect in the time-frequency plane (an example is shown in Section 5). Without the expected horizontal structure, heuristic partial-formation algorithms may have a difficult time unraveling the interference. Similarly, the problem of missed observations (i.e., peak picking fails to correctly identify one or more peaks) can lead to a fragmented partial, thereby requiring more heuristics to determine when two or more partial segments “look like” they are just fragments of a greater whole and should be joined.

The partial formation step is just one processing stage of a complete transcription system, to be followed by a note formation stage which integrates the partial data into higher-level note constructs. Higher-level stages can be employed to recognize chords or rhythmic structures. In all of these stages, the quality of the data available at the input limits the efficacy of the algorithms. Whereas most approaches to transcription have focused on the higher-level stages, we choose instead to focus on the quality of data at the earliest stage which impacts all stages of processing.

In this paper we present an approach to partial formation for a limited class of musical instruments and signals that draws upon a mathematical model of a partial’s time-varying behavior. This model suggests the application of Kalman filtering to peak observations in the time-frequency plane to identify and extract partial-like structures without concern for the fine details of the partials. This fits in well with our broader goals in transcription, which are the identification of the primary attributes of notes, i.e., frequency and onset/offset time, rather than the more (mathematically) subtle aspects such as timbre and articulation. In addition, the Kalman filter not only produces estimates of partial amplitude and frequency but a quality-of-fit measure

^a Further author information —

A.S.: asterian@umich.edu; <http://www-personal.umich.edu/~asterian>

G.H.W.: ghw@eecs.umich.edu; <http://www.eecs.umich.edu/~ghw>

as well whereby we can judge how well our observations match the predicted models. This measure may be of use in the subsequent information integration stages of musical transcription.

We consider the following class of musical signals:

1. Harmonic (or nearly-harmonic) musical instruments. This instrument class is broad enough to admit a wide range of instruments, yet allows us to decompose each note into a sum of narrowband partials (the additive synthesis model) whose frequencies are integer (or close to integer) multiples of the fundamental frequency. Inharmonic instruments such as bells, drums, etc. are excluded from our signal class.
2. Time-invariant notes. Only notes played without vibrato, glissando, or other frequency or amplitude modulations are considered. The treatment of these colorations is left for future work.
3. The Brass Family. In this paper we consider four common instruments from the brass family: trumpet, french horn, trombone, and tuba. Our approach uses this restricted class to minimize modeling variability while still establishing “proof of concept”. Future work will evaluate the algorithm for data obtained from instruments outside the brass family and refine the model parameters as necessary based on the acoustic attributes of the other instruments.

Many existing approaches to partial formation for transcription¹⁻⁵ use intuitive notions of connecting peaks from one time step to another based upon nearness in amplitude and frequency. Some partial formation methods^{6,7} look beyond the immediate time neighborhood (using linear prediction, for example) to achieve some degree of filtering. Problems in partial formation (such as attributing the right peak to the right partial and handling partial splitting) have been approached using intuitive rule-based ideas⁸. The use of Kalman filtering for partial formation was inspired by a similar problem in radar tracking⁹: the formation of target tracks from discrete observations of a target’s position. Kalman filtering over the time-frequency surface has also been used to interpret meteorological data¹⁰. The statistics of musical partial behavior have been studied previously¹¹ (using an approach similar to the one in this paper) for the purposes of improving computer synthesis of music. Our statistical analysis is more specific to our partial tracking model and also considers a much larger set of partials.

The remainder of this paper is structured as follows. Section 2 presents the partial behavior model and the Kalman filter framework. Section 3 describes how the parameters of the model were obtained from statistical analyses of recordings of isolated notes played on brass instruments. In Section 4 we complete the Kalman filter implementation and describe the details of the partial tracker. In Section 5 we demonstrate the results of applying the new tracker to a musical recording. We also demonstrate how the tracker can resolve the problems of beating due to neighboring partial interference and missed observations.

2. A PARTIAL EVOLUTION MODEL

In this section we present a brief introduction to the decomposition of a musical note as a sum of partials and then present a discrete-system model for a partial’s evolution. Since our goal is to apply Kalman filtering to the partial’s model, we construct the model as a first-order recursion with additive driving noise.

Many musical instruments generate notes that are well-modeled as a time-varying sum of sinusoids¹². The waveform of a single note can be written as:

$$s(t) = \sum_{n=0}^{N-1} A_n(t) \text{Re}\{e^{j\theta_n(t)}\} \quad (1)$$

where:

$$\theta_n(t) = \int_0^t (\omega_n + \phi_n(\tau)) d\tau \quad (2)$$

The function $A_n(t)$ is the partial’s instantaneous amplitude and $\theta_n(t)$ is its instantaneous phase. The derivative of the latter is the instantaneous frequency and it comprises ω_n , the partial’s nominal frequency, and $\phi_n(t)$ which describes the partial’s instantaneous frequency variation. The most salient partials of such instruments are harmonically related so that $\omega_n \approx n\omega_0$, where the approximate equality indicates that a certain degree of inharmonicity may be present. The instantaneous frequency variation $\phi_n(t)$ encompasses both aleatoric timbral variations as well as performance effects such as vibrato and glissando. We

do not consider vibrato and glissando in this paper hence we consider the function $\phi_n(t)$ to be a driving noise process with no deterministic components.

2.1 Partial amplitude evolution

We propose a piecewise-linear fit to the log-power function $P(t) = 20 \log_{10} A(t)$ as a model of partial amplitude evolution for the purposes of musical transcription. This model is a good fit to instruments with nearly-constant partial power in steady state, such as brass and woodwind instruments playing notes without amplitude modulation as shown below in Figure 1. Additionally, many struck-string or plucked-string instruments, such as piano and harp, have exponentially decaying partial amplitudes, which results in linearly decaying values in log-power.

In discrete time, we assume a temporal sampling interval T_s on the time-frequency plane and write:

$$\begin{aligned} p_k &= p_{k-1} + v_k \\ v_k &= v_{k-1} + u_k \end{aligned} \quad (3)$$

where p_k is the log-power at discrete time step k , v_k is the log-power “velocity” (i.e., first-order time difference), and u_k is an unspecified driving noise process. We should note that the discrete log-power velocity is an approximation to the true (i.e., continuous time) derivative of the log-power function in the sense that the normalized DTFT of v_k , $|V(e^{j\omega})| = |2 \sin(\omega/2)|$ is an approximation to the ideal derivative response $|H(e^{j\omega})| = |\omega|$ over the interval $0 \leq \omega \leq \pi$. The sampling interval T_s can be reduced should we want to improve the quality of this approximation.

2.2 Partial frequency evolution

As mentioned above, we expect our partial frequencies to be very nearly constant, with only small variations due to the timbral characteristics of the instrument. Gross variations such as vibrato or glissando are not considered in our present model. A zero-order (i.e., constant) model is therefore appropriate:

$$f_k = f_{k-1} + w_k \quad (4)$$

where f_k is the partial frequency and w_k is an unspecified driving noise process.

Our models for partial amplitude and frequency suggest the system state vector:

$$\mathbf{x}_k = [p_k \ v_k \ f_k]^T \quad (5)$$

We augment \mathbf{x}_k with additional states in Section 4.

2.3 Observation vector

Estimators applied to the time-frequency surface give us estimates of a partial’s instantaneous amplitude and frequency at discrete points in time. Our observation vector is:

$$\mathbf{y}_k = [\hat{p}_k \ \hat{f}_k]^T \quad (6)$$

where \hat{p}_k and \hat{f}_k are the instantaneous log-power and frequency, respectively, at discrete time step k .

3. DRIVING PROCESS MODELS

In this section we are concerned with the driving processes u_k and w_k as defined in Section 2. Our approach is to characterize statistically the nature and degree of amplitude and frequency variation from partials extracted from instrument notes. We are initially concerned with the steady-state section of partials. Our instrument recordings were drawn from the McGill University Master Samples (MUMS)¹³. The recordings comprise isolated notes from four brass instruments played over the range of each instrument. Table 1 shows the instruments used in our analysis along with the corresponding MUMS volume and track numbers, the range of notes in each instrument, and the total number of partials contributed towards our analysis. Because of the large number of partials in our analysis set, we sought to perform the analysis using automated tools and without human intervention. Below we describe our analysis methodology.

MUMS Instrument Name	Volume	Track Number(s)	Note Range	Contributed Partials
B-flat trumpet, hard attack	7	14-19	E3-C#6	599
french horn	2	19	D2-D5	631
tenor trombone	2	22	E2-D#5	770
tuba	2	25	C2-G4	476

TABLE 1. The list of instruments used to perform the partial behavior analysis. The volume and track numbers refer to the MUMS indexing system. The number of partials from each instrument that contributed to the analysis is shown in the last column.

3.1 Time-frequency analysis of notes

For each note of an instrument, the first step was the computation of a time-frequency representation (TFR) of the audio waveform. We used the modal kernel TFR¹⁴, a member of Cohen’s class of bilinear time-frequency representations. This TFR was specifically designed for musical analysis and is well suited to our analysis task as it provides high-resolution estimates of instantaneous amplitude and frequency (as compared to more conventional tools such as the spectrogram).

The TFR’s were computed with a temporal sampling rate of $T_s = 2.5$ ms. The number of sampling points in the positive frequency dimension was chosen to be 8192 to yield a bin-to-bin spacing of 2.69 Hz at a 44.1 kHz sampling rate for the MUMS recordings. A maximum of 40 partials from each note was considered. The frequency lag window of the TFR was chosen to be a Chebyshev window with 100 dB of sidelobe attenuation.

3.2 Partial tracking

Once a note’s TFR was computed, a peak estimation and partial tracking algorithm was applied to the first 40 partials. This would appear to be a circular problem as it is the very task of partial tracking that originally concerns us. However, as these are isolated, single-voice recordings of known fundamental frequency, partial tracking can be performed fairly easily.

Prior to the actual partial tracking, our algorithm identified salient peaks. Each time slice of the TFR was divided into 40 regions centered upon the expected frequency of each partial. The first partial region was centered at the fundamental frequency, the second at twice the fundamental frequency, and so on. Within each region, the highest local maximum was retained. Other local maxima within each region were retained if their heights exceeded a distance, termed the *clutter threshold*, below the height of the highest local maximum. In our analysis, the clutter threshold was set at 5 dB. If no local maximum was present above an absolute minimum threshold (set to -80 dB) then no peak was retained for this region and this time slice.

After each peak was identified from a local maximum, its amplitude and frequency estimates were refined using a centroid calculation technique¹⁴. This process improves the accuracy of the estimates and has been shown to be superior to the more common parabolic interpolation. The centroid calculation region was allowed to extend away from the local maximum until the TFR value became non-positive or non-decreasing. An equal number of frequency bins were used on the left and the right sides of the local maximum. The number of frequency bins actually used in the centroid calculation was retained as a measure of quality for the peak.

The analysis proceeded in this fashion, one time slice after another, until the entire TFR was processed. At this point, partial tracking proceeded by connecting peaks across adjacent time slices within each one of the 40 frequency regions. First, the *global* maximum peak across all time slices in a given frequency region was identified. Partial tracking proceeded first forwards, then backwards from this global maximum. We will describe the tracking process in the forward direction only; the backwards tracking process is identical.

The principle behind connecting peaks into partial tracks was based upon nearest-neighbor frequency matching. That is, for a partial track fragment extending from time n to $n+k$, the peak at time $n+k+1$ closest in frequency to the peak at time $n+k$ was added to the track. The addition of this peak augmented the track so that it extended from time n to $n+k+1$. The tracking process continued at time $n+k+2$, and so on. A time slice with no peaks in the frequency region ended the partial track fragment.

A new partial track fragment was started when a time slice was again encountered with a valid peak in the frequency region. As with the global peak, the highest-power peak in the region was chosen to start the partial track if more than one peak was present in the region.

3.3 Data extraction

To compute the statistics of partial behavior we wanted to isolate those sections of each partial track that were estimated accurately. The number of points used in the centroid calculation of peak power and frequency was taken as a measure of accuracy. Similarly, the number of peaks within the clutter threshold of the highest peak indicated the expected amount of inter-peak interference. In our work we set the minimum number of points in the centroid calculation to 5 and the maximum number of clutter peaks to 0. These thresholds were determined empirically as good indicators of accurate data. Finally, track segments shorter than 100 ms were rejected.

An example of the results of this extraction process is shown in Figure 1. This figure shows the partial track in log-power and frequency as well as the two indicator measures (number of points in the centroid calculation and number of clutter peaks). The sections of the partial track that meet the extraction criteria are indicated in the figure by horizontal I-bars.

3.4 Statistical analysis

The extracted sections from each partial track were subjected to a spectral analysis in log-power differences (or “velocity”) and in frequency variation. For log-power velocity, each extracted partial segment was passed through a first-difference operator to obtain the log-power velocity v_k , defined in Section 2. Welch’s averaged periodogram method¹⁵ was used to estimate the power spectral density (PSD) for the entire partial. The log-power velocity was divided into 100 ms segments (with 50 ms of overlap between segments), each segment was windowed with a Blackman window, and then the magnitude squared of the DFT was computed over the segment. All DFT squared magnitudes were averaged to yield a single PSD estimate for the partial. We call this the *log-power velocity* PSD.

The same estimation process was applied to the partial frequencies, with two differences. First, no first-difference operation was involved. Secondly, the mean frequency over the entire partial was subtracted from all frequency values prior to spectral estimation since we are only interested in the variations in frequency around the mean. We call this spectral analysis the *frequency variation* PSD.

In addition to the PSDs, histograms of log-power velocities and frequency values were computed as estimates of the PDFs. For log-power velocity, the histogram was computed using 201 bins over the range [-2,2]. The frequency variation histogram used the same number of bins over the range [-40,40] Hz. These ranges were determined empirically after an initial analysis of the data allowed us to observe the usual ranges for variations over all instruments. Some outliers did exist outside of these ranges.

If the partial track did not contain at least 10 segments of valid data, the partial was deemed to be insufficiently significant and was excluded from all further analysis.

In collating the data, the frequency range from 0 Hz to 22050 Hz (the Nyquist frequency) was divided into 80 frequency regions of 275.6 Hz width. Each partial from an instrument was assigned to one of these frequency regions based upon the partial’s average frequency. The associated PSDs and histograms from these partials were then combined to form a single representative PSD and histogram for the given instrument and frequency region.

The averaging of PSDs was actually performed after each PSD was normalized to unity power, in order to distinguish between the spectral shape of the PSD and its overall power. We found this distinction

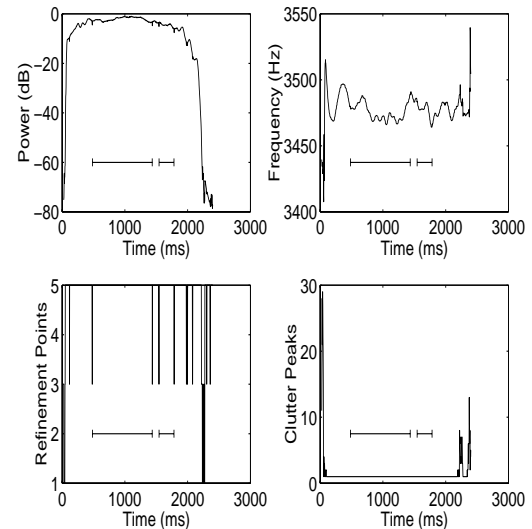


FIGURE 1. Some views of the extraction analysis for the 12th partial of a D4 note from the “B-flat trumpet, hard attack” instrument (MUMS Volume 7, Track 16, Index 1). The top-left graph shows the log-power, the top-right graph shows the frequency, the bottom-left graph shows the number of points used in the centroid calculation of peak power and frequency, and the bottom-right graph shows the number of clutter points for each peak. The horizontal I-bars indicate the regions of the partial that meet the data extraction criteria detailed in the text.

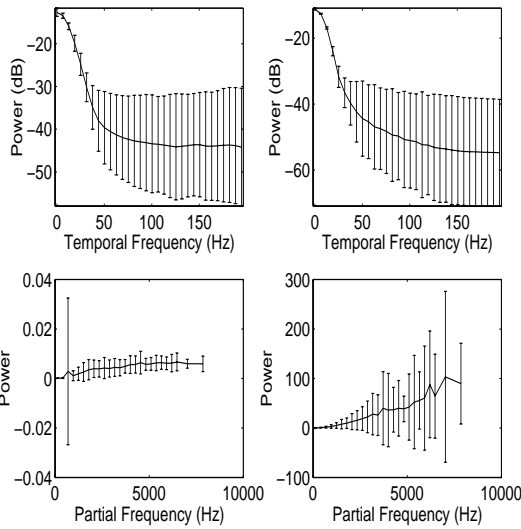


FIGURE 2. Some results of averaging across all instruments. The left side of the figure pertains to log-power velocity and the right side to frequency variation. The top row shows average normalized PSDs in the partial frequency neighborhood of 413 Hz. The bottom row shows total PSD power as a function of partial frequency. All error bars indicate one standard deviation. Only those frequency bins with at least 10 partials are shown.

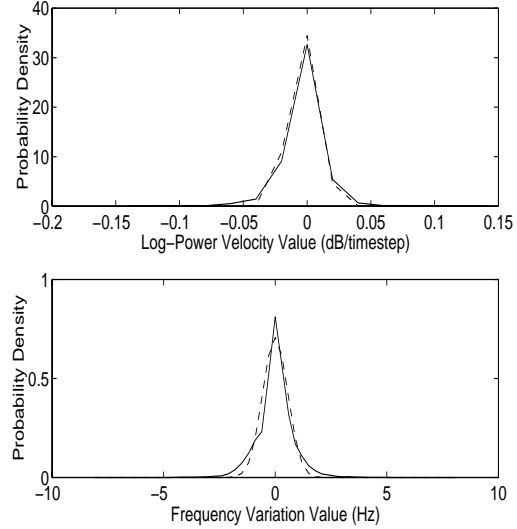


FIGURE 3. Empirical distribution functions across all instruments for log-power velocity values (top figure) and frequency variation values (bottom figure). This distribution represents the frequency bin centered on 413 Hz. The remaining frequency bins have similar results in terms of distribution shape but with varying standard deviations (see Figure 4). The best least-squares fit to a Gaussian distribution is shown with dashed lines.

useful as it removed much of the variance of the data from the spectral shape. This change allows us to represent the driving noise process in each frequency region by averages in two independent features (spectral shape and total power) rather than a single average PSD.

3.5 Results

Our analysis is summarized as follows:

1. Both the log-power velocity PSDs and the frequency variation PSDs are low-pass. Moreover, the shape of the low-pass spectrum is similar for all instruments and all partial frequencies. This is especially true for the frequency variation spectrum. Figure 2 shows the results of averaging across all instruments analyzed (see Table 1 for a list of instruments). The top-left graph shows the average normalized log-power velocity PSD in the partial frequency region centered on 413 Hz (as an example). The top-right graph shows the average normalized frequency variation PSD in the same frequency region. PSDs in the remaining frequency bins were all similar in appearance. Below we describe these PSDs parametrically.
2. The power in the log-power-velocity and frequency-variation processes increases fairly linearly with partial frequency. The bottom graphs of Figure 2 show the average trend of PSD power with partial frequency for log-power velocity (bottom-left graph) and frequency variation (bottom-right graph). Only those frequency regions comprising a minimum of 10 partials have been included.
3. The histograms increase in range and variance at higher frequency regions. The all-instrument histograms for log-power velocity and frequency deviation are shown in Figure 3 for the partial-frequency bin centered on 413 Hz. The remaining frequency bins have similar results in terms of histogram shape but increasing range and variance with increasing frequency (shown in Figure 4). The best least-squares fit to a Gaussian distribution is shown in Figure 3 using dashed lines. The empirical distributions appear more exponential than Gaussian (although the heavy tails of the distributions prevent a good match to the exponential distribution). Since our Kalman filter model treats the driving noise processes as Gaussian, we expect our tracking results to be slightly suboptimal.

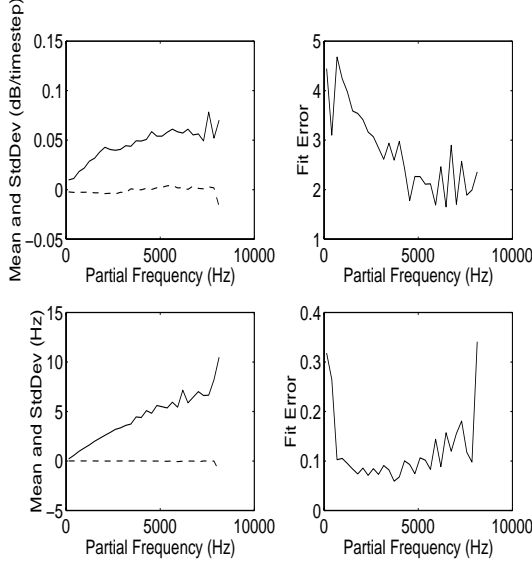


FIGURE 4. Graphs of best-least-squares fit of a Gaussian distribution to the empirical PDFs. The top left graph shows the standard deviation (solid line) and mean (dashed line) of the best-least-squares fit Gaussian distribution to the empirical log-power velocity PDFs. The top right graph shows the fit error. The bottom row shows the same quantities for the frequency deviation PDFs. Figure 3 presents these results for a single frequency bin (the second bin, centered at 413 Hz).

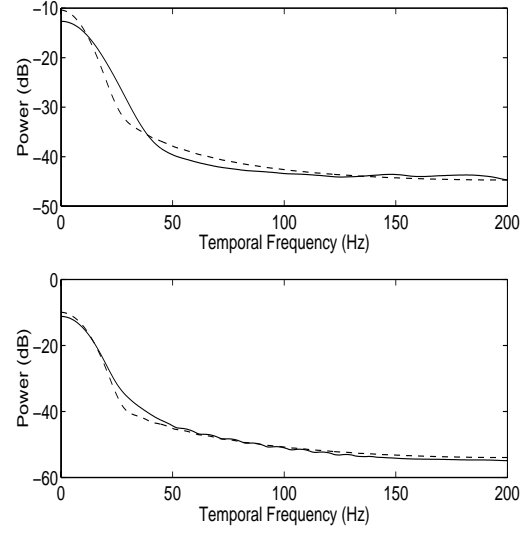


FIGURE 5. Average PSDs and least-squares fit PSDs for log-power velocity (top figure) and frequency variation (bottom figure). These PSDs represent the frequency region centered on 413 Hz. The solid lines represent the observed data and the dashed lines are the least-squares fits described by Equation 8 (with $\eta = 0$ for the frequency variation PSD). The parameters of these fits for all frequency regions is shown in Figure 6.

3.6 Parameterization

The partial model presented in Section 2 and our subsequent partial tracking algorithms require parametric forms for the preceding results. For the average PSDs, their low-pass shapes suggest that we fit the data with a single-pole rational transfer function:

$$H(z) = \frac{G}{1 - \alpha z^{-1}} \quad (7)$$

Our fitting experiments suggested that the above function was a good fit to the frequency variation PSD data at all frequencies, but for the log-power velocity PSDs, the spectrum rolled off too quickly. Adding a white noise “floor” to the model improved the fit to the log-power velocity PSDs. Our spectral model, then, is the sum of the single-pole low-pass spectrum and a white spectrum:

$$|H(e^{j\omega})|^2 = \frac{G^2}{(1 + \alpha^2) - 2\alpha \cos \omega} + |W(e^{j\omega})|^2 \quad (8)$$

where $|W(e^{j\omega})|^2 = \eta$ is the white noise process.

To fit this spectrum to our observed data, we performed a numerical least-squares minimization over the variables G , α , and η using the average PSD in each frequency region as the target of the minimization. Separate minimizations were performed for the log-power velocity PSDs and frequency variation PSDs. The goal was to fit the shape of these PSDs as closely as possible. Prior to computing a least-squares error measure between a particular instantiation of Equation 8 and the observed PSD, the putative spectrum $H(e^{j\omega})$ was convolved with the spectrum of the Blackman window used to compute the average PSDs. This convolution was performed to account for the spectral smearing present in the PSDs computed from the data.

The minimization results are shown in Figure 6 as graphs of pole radius α and white noise power η as a function of partial frequency. A typical fit of the spectrum in Equation 8 to the observed data is shown in Figure 5. The two graphs show the aver-

age PSDs (solid line) and fit PSDs (dashed line) on the same axes. The upper graph is for the log-power velocity PSD and the lower is for the frequency deviation PSD. These graphs are for the second partial frequency bin in the region of 413 Hz.

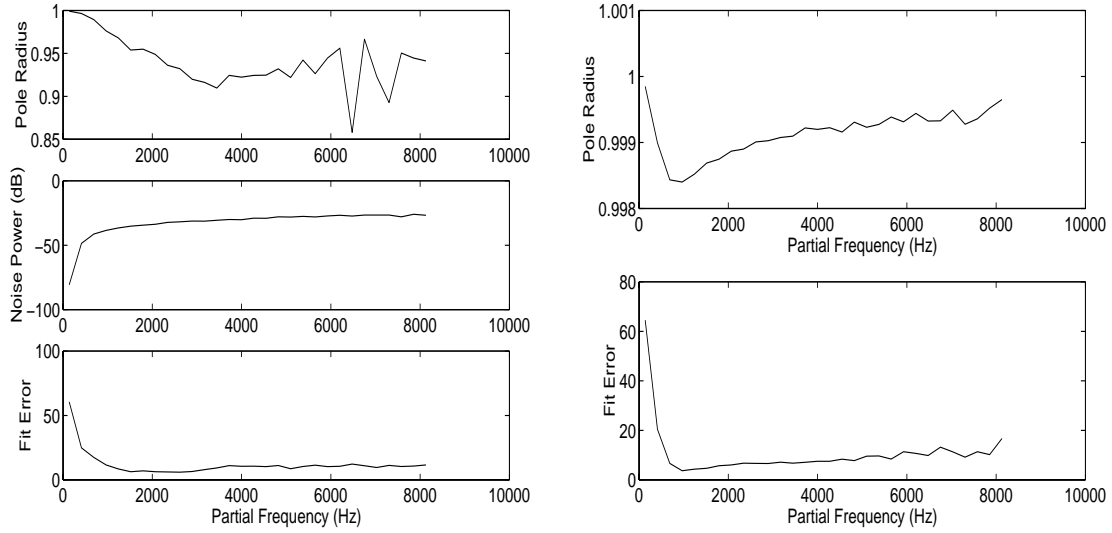


FIGURE 6. These graphs show the results from fitting the process model described by Equation 8 to the all-instrument average PSDs. The leftmost three graphs show results for the log-power velocity PSDs and the rightmost two graphs are for the frequency variation PSDs. The noise power for the frequency variation model was always 0 hence no graph is shown.

4. THE KALMAN TRACKER

4.1 Kalman filter implementation

Our Kalman tracker uses the following system description¹⁶:

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{v}_k\end{aligned}\quad (9)$$

where \mathbf{x}_k and \mathbf{y}_k are the system state and observation vectors (introduced in Section 2) at discrete time indices k . The \mathbf{A} , \mathbf{B} , and \mathbf{C} matrices arise from the partial evolution model described in the previous section. Finally, \mathbf{u}_k and \mathbf{v}_k are uncorrelated zero-mean Gaussian white noise processes with covariance matrices \mathbf{Q} and \mathbf{R} , respectively. For simplicity, we let $\mathbf{Q} = \mathbf{I}$ and represent non-unity noise powers in the \mathbf{B} matrix.

Beginning with the state vector presented in Equation 5, we augment it with additional state variables w_k and z_k to represent the colored driving noise processes:

$$\mathbf{x}_k = [p_k \ v_k \ w_k \ f_k \ z_k]^T \quad (10)$$

The remaining system matrices are:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & \alpha_p & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & \alpha_f \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & G_n \\ G_p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & G_f & 0 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_f^2 \end{bmatrix} \quad (11)$$

The α_p and α_f parameters are the pole radii for the log-power velocity and frequency variation PSD models, respectively (see Equation 7). Similarly, G_p and G_f are the gains that effect the expected overall driving noise power. $G_n = \sqrt{\eta}$ is the ampli-

tude scaling required for the white noise floor in the log-power velocity PSD model (i.e., the power η in the $W(e^{j\omega})$ process from Equation 8). The quantities σ_p^2 and σ_f^2 are the variances of the observation noise process \mathbf{v}_k . We do not have a theoretical method for determining these quantities, and are indeed more inclined to consider these adjustable parameters of the tracking algorithm rather than descriptors of “observation noise” (since we are assuming that our audio recordings are essentially noise-free). Values near 1.0 for both quantities seem to work well in practice. We note that our Kalman tracker uses a different set of state matrices (hence different Kalman filters) for each frequency band, corresponding to the data described in Section 3.

4.2 Track initiation

From the TFR of the audio waveform, a peak detector identifies local maxima in each time slice and obtains power and frequency estimates in the same fashion as the peak detector described in Section 2. Peaks that are not associated with existing tracks are used to initiate new tracks.

4.3 Track continuation

After a track has been initiated, each new time slice can continue the track if any peak falls within the *acceptance region* (or *gate*, as used in the radar tracking literature) of the track. This acceptance region depends upon a measure of distance between the peak’s power and frequency and the predicted power and frequency, as generated by the Kalman tracker. This distance function is⁹:

$$d_k^2 = \mathbf{e}_k^T (\mathbf{C}\mathbf{P}_k\mathbf{C}^T + \mathbf{R})^{-1} \mathbf{e}_k \quad (12)$$

where $\mathbf{e}_k = \mathbf{y}_k - \mathbf{C}\mathbf{x}_k$ is the error between the current actual observation and the predicted observation. The quantity $\mathbf{C}\mathbf{P}_k\mathbf{C}^T + \mathbf{R}$ is the covariance matrix of the error \mathbf{e}_k and \mathbf{P}_k is the expected covariance matrix of the system state \mathbf{x}_k (\mathbf{P}_k is obtained from the update equations of the Kalman filter). Thus, d_k^2 is the norm of the error normalized by the expected variance. This quantity is a reasonable measure of how “close” the observation is to where we expect it to be, given our partial evolution model.

To determine if a given peak observation should be associated with an existing track, we perform the threshold test $d_k^2 \leq \gamma$ on the error norm. This relation defines the acceptance region for each track. If the test is satisfied, the given peak observation is used to update the current track via the Kalman update equations. If a given peak does not fall within the acceptance region of any track, it is used to initiate a new track.

To determine an appropriate value for γ we assume that the observation errors are independent and Gaussian distributed. Then, d_k^2 is a sum-of-squares of two normalized Gaussian variables, hence d_k^2 has a Chi-square distribution with two degrees of freedom:

$$f_{d^2}(x) = \frac{1}{2}e^{-\frac{x}{2}}, \quad x \geq 0 \quad (13)$$

For a desired detection probability P_D (i.e., the probability that a peak observation that should be associated with a track does fall within the track’s acceptance region) we can compute the required value of γ :

$$P_D = \int_0^{\gamma} f_{d^2}(x) dx = 1 - e^{-\frac{\gamma}{2}} \quad (14)$$

Setting P_D to 0.99, for example, gives us $\gamma = 9.21$.

4.4 Track termination

A track that has not been updated with an observation for over 30 ms of consecutive time slices is terminated. This *ad hoc* criterion seems to work reasonably in practice but there exists ample room for a more careful treatment of this issue.

Similarly, we implement a minimum-duration rule for filtering out spurious tracks: all tracks shorter than 100 ms are deleted. Since our track initiation algorithm is highly optimistic — all peaks not assigned to existing tracks initiate new tracks — we need a method to eliminate the “noise” tracks initiated by the spurious peaks present in the time-frequency representation. This *ad hoc* criterion limits the maximum note density to 10 note events per second (assuming all notes are sustained for the full 100 ms duration). We believe this is a reasonable limitation for the moment, until the issue of spurious track deletion can be more closely investigated. In fact, we believe that this type of filtering is best left for the subsequent processing stages in the

transcription process. At the note formation level, for example, the lack of corroborative partials for a given spurious partial could lead to track deletion.

4.5 Quality-of-track information

The quantity d_k^2 described above can also be used as a quality-of-track measure for use by subsequent processing stages in the transcription process. A partial with consistently high values of d_k^2 suggests that the underlying observations are noisy (or are otherwise a poor fit to our partial model) so that the information conveyed by the track's existence should be given less weight than other tracks.

Another use for the d_k^2 measure is in making decisions about track initiation and termination at the note formation stage rather than at the partial formation stage. Using an artificially large value of γ in the tracker will incorporate peak observations from beyond the steady-state section of the track, i.e. the onset and the offset sections. The note formation algorithm can then use the value of d_k^2 to distinguish between the onset, steady-state, and offset sections of the track. More importantly, the d_k^2 values from multiple tracks can be correlated to further aid in onset and offset detection, especially for the difficult transcription problem of detecting repeated notes.

4.6 Kalman smoothing

Following the termination of a track, a backwards-time Kalman filter is applied to the data in order to effect fixed-interval smoothing¹⁶. The purpose of this step is to improve the accuracy of the data (more specifically, to reduce the expected covariance of the state estimates) by using both causal and anti-causal data to estimate the system states.

5. EXAMPLES

5.1 Brass duet

We presented an extract from a musical passage with known score played on trumpet and french horn to the tracker. The resulting tracks were then separated into two groups depending upon whether a track could be interpreted as belonging to a note sounding at the current time. Figure 7 shows the musical passage as well as the results of the tracking. The graph on the left shows the “accepted” partials while the graph on the right shows the remaining partials. For clarity, only the partials up to 2000 Hz are shown, but even this restricted range illustrates that a substantial amount of partial material has been extracted. The audio material contained some reverberation leading to some gaps in the filtered partials. The proper handling of these gaps is left for subsequent stages in transcription, where the partial information is integrated into higher-level note constructs. Also, some of the rejected partials in Figure 7 (near the beginning of the right graph) are actually still-sounding components of previous notes in the recording.

5.2 Closely spaced partials

To demonstrate the effectiveness of Kalman filtering in the presence of degraded data, we combined two of the MUMS B-flat trumpet notes (E3 and F#3) into a single audio waveform and presented it to the Kalman tracker. The fundamentals of the two notes (165 Hz and 185 Hz) were close enough in frequency to cause interference in the time-frequency distribution. An overhead view of the TFR peak-picking analysis in the region of both fundamentals is shown in the left graph of Figure 8. This graph shows the three most energetic peaks for each time slice in the frequency region. The peaks are connected with straight lines to the peaks of the adjacent time slices using a nearest-frequency criterion. Note how the upper fundamental (near 185 Hz) comprises a peak observation at nearly every time slice, but the lower fundamental (near 165 Hz) has several missing peaks due to the beating effect between the two partials.

The graph on the right of Figure 8 shows the results of applying the Kalman tracker to this data. The missing observations of the lower fundamental have been filtered and the underlying partials have been extracted by the Kalman tracker. We presented this particular example to highlight how our model-based approach can form partials without special *ad hoc* rules regarding how missing observations should be treated or how disjoint tracks should be connected.

6. CONCLUSIONS

We have presented a new approach to partial tracking for musical transcription based upon Kalman filtering in the time-frequency domain. The filter parameters were determined from a statistical analysis of musical partials from brass instrument recordings. We have demonstrated the ability of the Kalman tracker to extract partial material from music and to form partials from cluttered data. The new tracker provides quality-of-fit information for tracks as an aid in the subsequent integration of partials into notes.

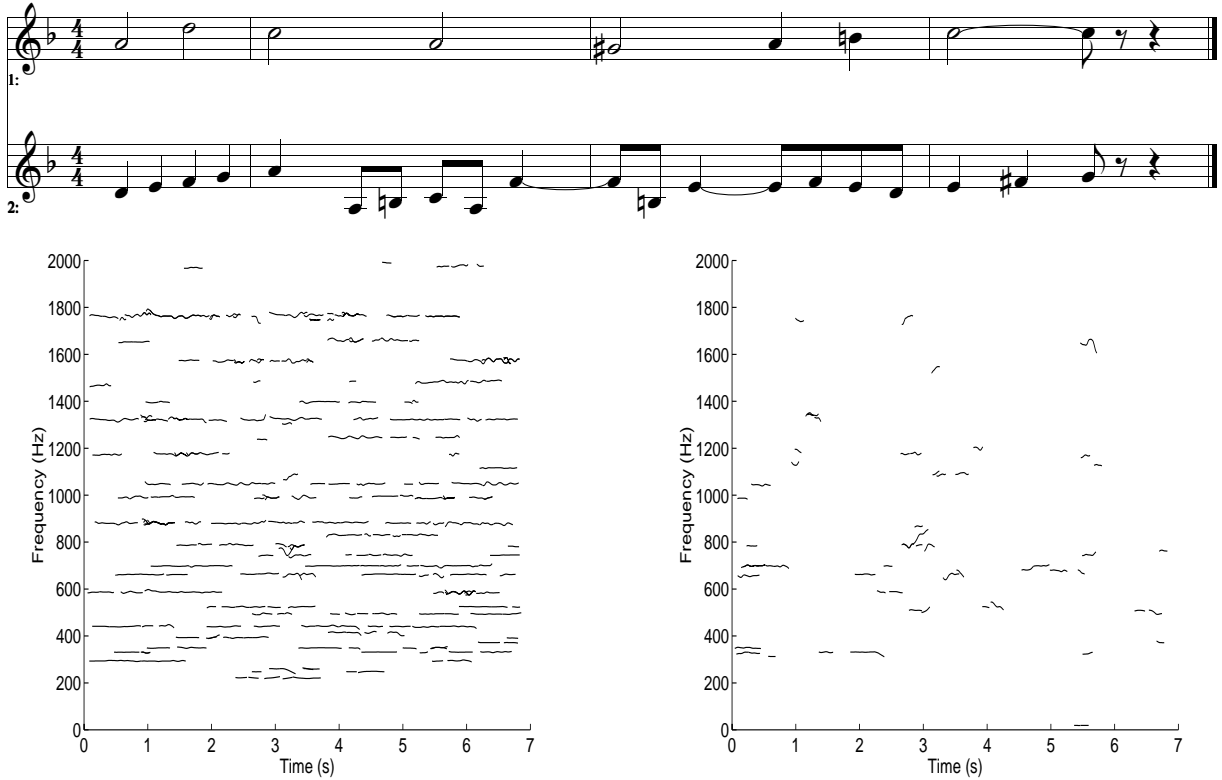


FIGURE 7. The staff at the top of the figure shows a score excerpt from “The Art of the Fugue” by J.S. Bach (Contrapunctus I). The top line was played on trumpet and the bottom line on french horn. This excerpt was presented to the Kalman tracker. The results of partial formation are shown in the two graphs above. The graph on the left shows an overhead view of the partials that could be attributable to notes in the source material. The graph on the right shows the “spurious” partials that bear no harmonic relationship to the notes sounding at the time. The separation of partials was performed manually.

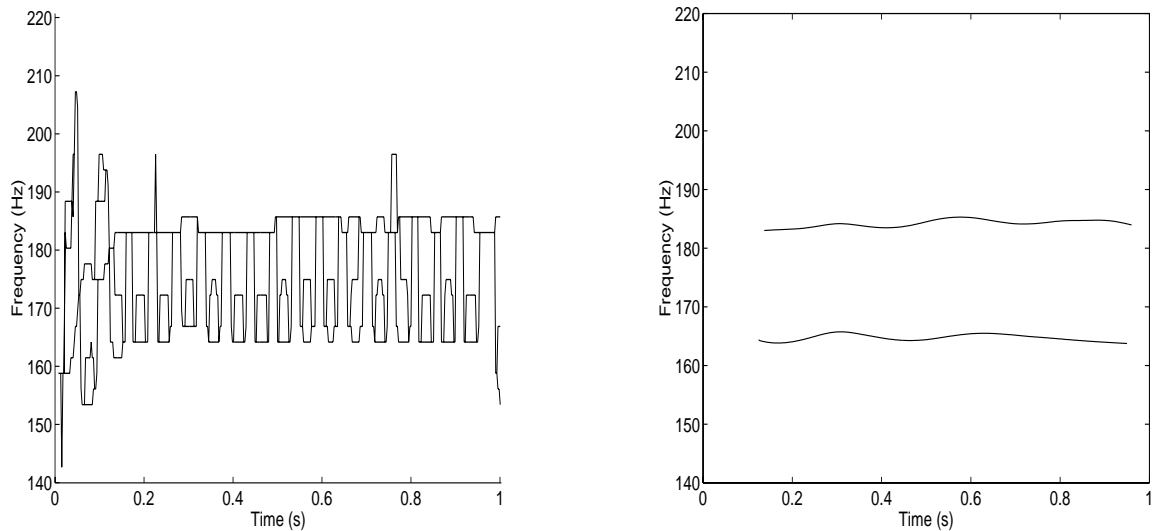


FIGURE 8. The left graph shows an overhead view of the TFR of a two-note chord formed by summing two notes (E3 and F#3) from the MUMS B-flat trumpet recordings. The frequency region shown is in the neighborhood of both fundamentals. The graph shows the location of the three most energetic peaks in the frequency region at each time slice. Peaks are present at the F#3 fundamental frequency at every time slice but note that the E3 fundamental has some missing peaks due to the beating effect between the two partials. This effect is due to the close separation of frequencies relative to the resolution limits of the TFR. The right graph shows the partials extracted from the data by the Kalman tracker.

7. ACKNOWLEDGMENTS

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.

8. REFERENCES

1. G.J. Brown and M. Cooke, "Perceptual grouping of musical sounds: a computational model," *Journal of New Music Research*, **23**, pp. 107-132, 1994.
2. J.C. Brown and B. Zhang, "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," *J. Acoust. Soc. Am.*, **89**(5), pp. 2346-2354, 1991.
3. R.C. Maher, *An approach for the separation of voices in composite musical signals*, Ph.D. Thesis, Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, 1989.
4. M. Piszczalski, *A computational model of music transcription*, Ph.D. Thesis, Department of Computer Science and Engineering, University of Michigan, Ann Arbor, 1986.
5. K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism," *Proc. Intl. Joint Conf. on Artificial Intelligence* **1**, pp. 158-164, 1995.
6. S. Foster, W.A. Schloss, and A.J. Rockmore, "Toward an intelligent editor of digital audio: signal processing methods," *Computer Music Journal*, **6**(1), pp. 42-51, 1982.
7. T. Nakatani, H.G. Okuno, and T. Kawabata, "Residue-driven architecture for computational auditory scene analysis," *Proc. Intl. Joint Conf. on Artificial Intelligence* **1**, pp. 165-172, 1995.
8. D.K. Mellinger, *Event formation and separation in musical sound*, Ph.D. Thesis, Department of Computer Science, Stanford University, Stanford, 1992.
9. S.S. Blackman, *Multiple-target tracking with radar applications*, Artech House, Norwood, Massachusetts, 1986.
10. W. Roguet, N. Martin, and A. Chehikian, "Tracking of frequency in a time-frequency representation," *Proc. IEEE Sig. Proc. Intl. Symposium on Time-Frequency and Time-Scale Analysis*, Paris, France, 1996.
11. S. Ando and K. Yamaguchi, "Statistical study of spectral parameters in musical instrument tones," *J. Acoust. Soc. Am.*, **94**(1), pp. 37-45, 1993.
12. A.H. Benade, *Fundamentals of musical acoustics*, Oxford University Press, New York, 1976.
13. F. Opolko and J. Wapnick, McGill University Master Samples, 11 CD-ROM set, Faculty of Music, McGill University, Montreal, Canada.
14. W.J. Pielemeier, G.H. Wakefield, and M.H. Simoni, "Time-frequency analysis of musical signals," *Proc. IEEE*, **84**(9), pp. 1216-1230, 1996.
15. S.M. Kay, *Modern Spectral Estimation*, Prentice-Hall, New Jersey, 1988.
16. C.K. Chui and G. Chen, *Kalman filtering with real-time applications*, Springer-Verlag, New York, 1991.