Beyond the beat: Modeling metric structure in music and performance

Stefan T. Tomic[a)]
Center for Mind and Brain, University of California, Davis, California 95618

Petr Janata[2,3,b)]
[2]*Center for Mind and Brain, University of California, Davis, California, 95618*
[3]*and Department of Psychology, University of California, Davis, California, 95618*

(Received 14 February 2008; revised 30 September 2008; accepted 5 October 2008)

Current models for capturing metric structure of recordings of music are concerned primarily with the task of tempo and beat estimation. Even though these models have the potential for extracting other metric and rhythmic information, this potential has not been realized. In this paper, a model for describing the general metric structure of audio signals and behavioral data is presented. This model employs reson filters, rather than the comb filters used in earlier models. The oscillatory nature of reson filters is investigated, as they may be better suited for extracting multiple metric levels in the onset patterns of acoustic signals. The model is tested with several types of sequences of Dirac impulses as inputs, in order to investigate the model's sensitivity to timing variations and accent structure. The model's responses to natural stimuli are illustrated, both for excerpts of recorded music from a large database utilized by tempo-estimation models, and sequences of taps from a bimanual tapping task. Finally, the relationship of the model to several other beat-finding and rhythm models is discussed, and several applications and extensions for the model are suggested.
 [DOI: 10.1121/1.3006382]

PACS number(s): 43.75.Zz, 43.75.Cd, 43.75.Xz, 43.75.St [DD] Pages: 4024–4041

I. INTRODUCTION

Describing the metric and rhythmic structure of a piece of music is a complex endeavor, in part, because the concepts of rhythm and meter encompass many types of timing features and relationships that are found in a piece of music, such as the tempo or the accent structure (London, 2004). It is generally accepted, however, that a description of music's rhythmic properties must take into account metrical structure, tempo, and timing (Gouyon and Dixon, 2005). Nevertheless, most current rhythm models that analyze acoustic signals focus primarily on finding the beat and tempo. The process of beat extraction is relevant both to cognitive psychologists interested in the mechanisms of sensorimotor synchronization with music (Large, 2000; Repp, 2005; Snyder and Krumhansl, 2001), as well as to modelers interested in automatic music classification and retrieval (Gouyon *et al.*, 2004; Tzanetakis and Cook, 2002). Not surprisingly, a model's success is evaluated usually by comparing the model's output with the judgments of experts who have annotated a musical excerpt or score with markings that correspond to events at the metric level that they view as the tactus. Despite their utility, beat-finding models that are optimized to mirror the judgments of experts are solving a narrow perceptual problem that is somewhat restricted to musical selections in which an unambiguous beat is discernable. When hearing a piece of music, not all listeners may agree on which metric level corresponds to the tactus, and when asked to tap along with the music, they may produce taps at different metric levels. It is therefore of interest to develop quantitative methods that describe the relative prominence and time-varying properties of the different metric levels in a piece of music.

In this paper, our objective is to present a means of modeling the timing characteristics in a signal that allow one to ask questions about the presence and relative prominence of different metric levels and rhythmic patterns. We accommodate both raw audio signals and recorded behavior from tapping experiments, thus providing a common framework for characterizing both the sensory and motor aspects of sensorimotor synchronization tasks. The initial useful stage of the model produces a continuous output rather than an output of a discrete symbolic nature. The model output is of a similar nature to the work of Todd (1994), which produces a continuous output to represent rhythmic structure. We also provide a demonstration of how one may convert the continuous model output to a discrete symbolic representation that can be interpreted as a classification of the prevalent meter and a means of representing competing metric hypotheses. The framework we use can be seen as a variation on two models that were developed primarily for the purpose of beat estimation (Scheirer *et al.*, 1998; Klapuri, Eronen, & Astola, 2006). We first describe some of the salient features of these two models and our adaptations to them.

Scheirer's model (Scheirer, 1998) incorporates a preprocessing stage that separates the original audio signal into separate frequency bands. This effectively models the simultaneous rhythmic tracking of different streams in the fre-

a)Electronic mail: sttomic@ucdavis.edu
b)Author to whom correspondence should be addressed. Electronic mail: pjanata@ucdavis.edu

quency domain. We deviate slightly from this initial processing stage by utilizing the Auditory Peripheral Module from the IPEM Toolbox (Leman *et al.*, 2001). This module models the information stream from the outer ear to the auditory nerve. The process separates the signal into 40 bands on a critical band scale and easily substitutes for the filterbank in the first stage of Scheirer's model. After performing onset detection on each of the bands, Scheirer (1998) employed a filterbank of comb filters for each frequency subband. The filterbank outputs are then summed across frequency subbands. His model finds the comb filter with the highest energy output, which is used as an estimate of the tempo of the piece. Although Scheirer (1998) recognized that less prominent peaks are present in the comb filter outputs, he did not investigate the use of these peaks for identifying other metric levels present in the onset pattern.

The model by Klapuri *et al.* (2006) expands on Scheirer's model by improving the various processing stages, providing fine-tuning parameters and implementing a hidden Markov model which determines the frequencies and onset locations at the metric levels corresponding to the tempo, tactus, and measure. This model is currently regarded as one of the most accurate models of tempo and beat induction (Gouyon *et al.*, 2006). Although Klapuri *et al.* (2006) advanced the beat-finding ideas of Scheirer's model by taking into account the information across the three metric levels, one area that they do not investigate is the identification of other metrical hierarchies such as polyrhythms.

Our most significant departure from these two models is our use of reson filters instead of comb filters for the analysis of the periodicities present in the onset patterns. As described in more detail below, we find reson filters to be better suited to the demands of tracking the relative energy at multiple metric levels. The final stage of our model produces a surface plot that illustrates the metric levels of an excerpt as they vary over time. We refer to this plot as an *average periodicity surface* (APS). From the APS, we derive a *mean periodicity profile* (MPP), which depicts the mean metric hierarchy over the course of the excerpt. Next, we describe and illustrate in detail the processing stages of our model, and demonstrate how our model can be used to examine the metric and rhythmic properties of simulated signals, music, and bimanual taps produced in behavioral tasks.

II. METHODS

A. Model overview

Our model was written in MATLAB (The MathWorks, Inc., Natick, MA). Figure 1 schematizes the processing steps. First, we pass the audio signal through the Auditory Peripheral Module of the IPEM Toolbox. The output of the module is an auditory nerve image (ANI), which represents the auditory information stream along the VIIIth cranial nerve and results in 40 channels of data. Next, we extract the amplitude envelopes by performing a root mean square (rms) calculation on the output of each channel. Onset patterns are then extracted by a difference calculation and half-wave rectification. We then sum every eight adjacent channels to produce five bands. Each of the five bands is passed through a bank

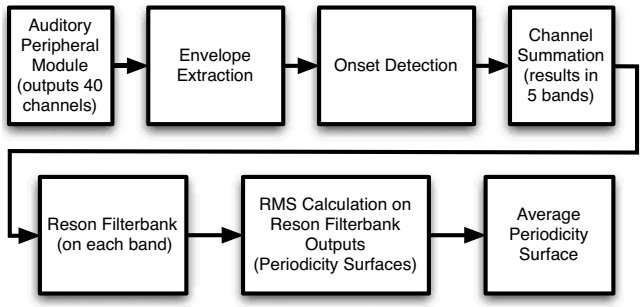

FIG. 1. Overview of the model's processing stages.

of reson filters. A rms calculation is performed on each band to facilitate the identification of the relative amplitudes of each filter. We refer to the surface plot of the rms of each band as a *periodicity surface*. The five periodicity surfaces are then averaged to produce an APS.

B. Auditory peripheral module

The Auditory Peripheral Module of the IPEM Toolbox is based on a model devised by Van Immerseel and Martens (1992). The first processing stage involves filtering the audio signal with a second-order low pass filter with a resonance frequency of 4 kHz. This filter approximates the frequency response of the ear. The low pass filtered audio signal is passed through a series of 40 bandpass filters. The resulting bands are referred to as channels. Each channel is then passed through a hair cell model, which is a forward-driven gain controlled amplifier that incorporates half-wave rectification and dynamic range compression. Each channel is subsequently low pass filtered at 1250 Hz. The resulting 40 channels simulate neuronal bandpass properties and firing rate encoding of the primary auditory nerve. The output of the model is referred to as an ANI. Each channel of the ANI has a bandwidth of one critical band and is spaced half a critical band from its neighboring channels. The 40 ANI channel center frequencies range between 141 and 8877 Hz. Figure 2(a) illustrates the outputs of channels 4, 10, and 30, with center frequencies of 252 Hz, 507 Hz, and 3266 Hz, respectively, for a musical excerpt (a sample excerpt of "Goodies" by Ciara featuring Petey Pablo, as downloaded from the Apple Music Store). The firing rate encoding of each channel is sampled at 2205 Hz. The incorporation of the Auditory Peripheral Module means that further processing steps in our rhythm model operate on a model of the auditory nerve rather than on bands of the audio signal itself. This model easily substitutes for the bank of bandpass filters employed by Scheirer (1998). Klapuri *et al.* (2006) employed 36 bands on a critical band scale, so the number and spacing of our filters more closely fit their configuration, although their filters range between 50 Hz and 20 kHz, while our range is narrower.

C. Envelope extraction

The ANI channels are smoothed with a rms calculation. We use a frame size of 50 ms and a frame step of one sample. This provides an estimation of the envelope of each channel. The rms signals are then downsampled to 100 Hz

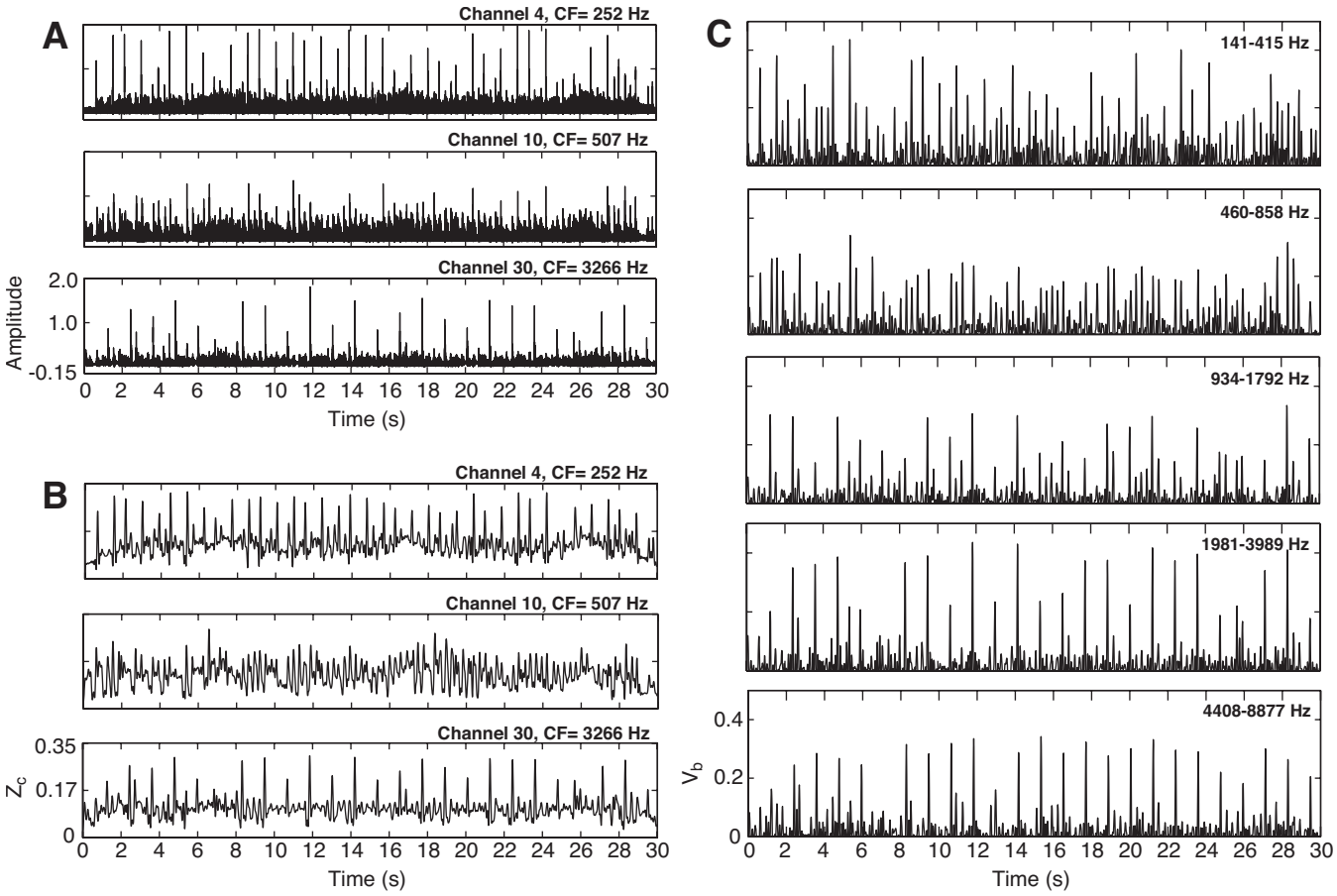

FIG. 2. Signal plots corresponding to (a) ANI channels 4, 10, and 30 with center frequencies 252, 507, and 3266 Hz. (b) The same ANI channels after performing a rms, downsampling, and lowpass filtering. The calculation of $z_c(n)$ is discussed in Sec. II C. (c) The resulting five bands after onset detection and summation of every eight adjacent channels. Equation (2) is used to calculate $v_b(n)$.

and low pass filtered by a 12th order Butterworth filter with a low pass frequency of 10 Hz for further smoothing. The downsampled and low pass filtered rms outputs of the same three channels are illustrated in Fig. 2(b).

D. Onset detection

We use $z_c(n)$ to represent the signal of each channel after performing the rms, downsampling, and low pass filtering calculations. First-order differencing and half-wave rectification are applied to $z_c(n)$. Half-wave rectification simply removes all negative values by setting them to zero.

$$z_c'(n) = \max(0, z_c(n) - z_c(n-1)). \tag{1}$$

The output $z_c'(n)$ provides an estimate of the onset pattern of the signal, reflecting both the rates and magnitudes of the onsets.

E. Channel summation

Sets of adjacent bands (channels) are next combined into single bands. This summation across the bands reduces some of the redundancy that is visibly present across the bands and provides an estimate of the temporal properties of broader spectral regions. In addition, the subsequent computational overhead is reduced considerably. Klapuri *et al.* (2006) summed the outputs from every nine adjacent bands, resulting in four final bands. We sum every eight adjacent channels, resulting in five bands (Fig. 2(c)). We chose the greatest divisor of the original 40 bands that produced periodicity surfaces (see Sec. II G) with noticeably different characteristics. This provided the best reduction of redundancy across the bands while still maintaining the different characteristics between frequency ranges.

$$v_b(n) = \sum_{c=(b-1)m_0+1}^{bm_0} z_c'(n), \quad b = 1, \ldots, b_0. \tag{2}$$

Note that this equation is identical to the one used by Klapuri *et al.* (2006), except that we exchange the b and c subscripts. In our model, we use c to denote ANI channel index and b to denote the summed band index. m_0 is the number of adjacent channels to sum and the calculation results in $b_0 = c_0/m_0$ bands.

F. Reson filterbanks

The signal from each band is passed through a bank of resonators, following the models of Scheirer (1998) and Klapuri *et al.* (2006). The bank of resonators for each band provides an analysis of the periodicities present in the onset patterns of that band. The models by Scheirer (1998) and Klapuri *et al.* (2006) employ comb filters as resonators. We

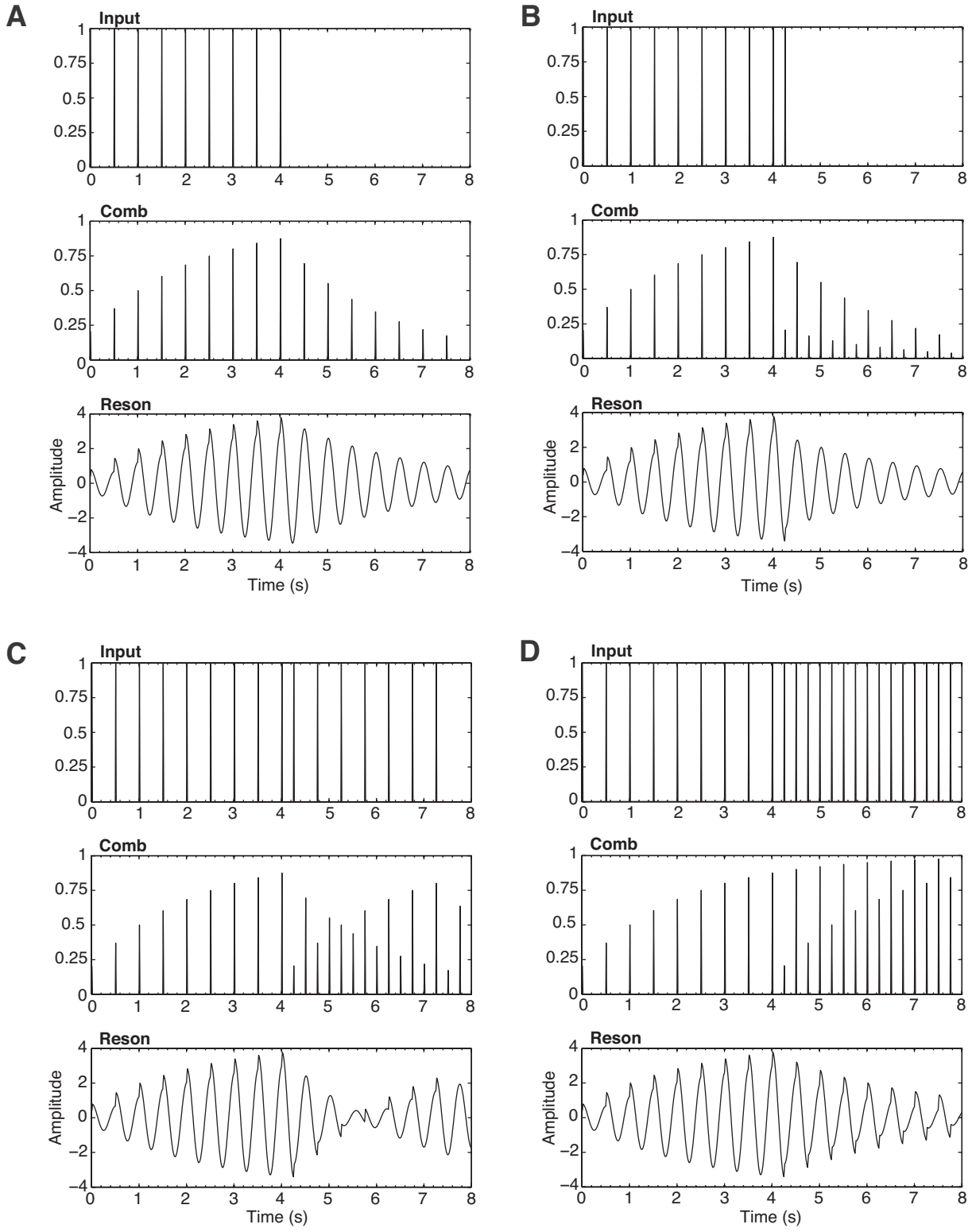

FIG. 3. Comparison of comb filter and reson filter responses. The reson filter has a resonant frequency of 2 Hz, and the comb filter's lowest resonance is at 2 Hz. Both filters are passed four different series of Dirac impulses for comparison: (a) a series at 2 Hz ending at 4 s, (b) a series at 2 Hz, followed by a single impulse at 4.25 s which is antiphase with both of the filter outputs, (c) a series at 2 Hz, phase shifted at 4.25 s, and (d) a series at 2 Hz which immediately doubles in frequency to 4 Hz after 4 s.

found that comb filters did not adjust quickly enough to changes in the onset patterns of the input signal, however; whereas *reson* filters, a type of damped oscillator (Steiglitz, 1996), adjusted much more quickly to changes in the onset patterns.

Figure 3 illustrates the responses of a comb filter and a reson filter, both tuned to a frequency of 2 Hz, to a variety of input signals. Figure 3(a) illustrates the responses of the comb filter and reson filter when they are passed an isochronous series of Dirac impulses at 2 Hz for 4 s, followed by 4

s of silence. The responses of both filters peak at 4 s as expected, and then decay. However, when an antiphase onset at 4.25 s follows the same series of isochronous pulses in the input (Fig. 3(b)), the amplitude envelope of the comb filter output is the same as if it was not passed the antiphase onset. The reson filter output in Fig. 3(b), however, is attenuated slightly after 4.25 s as compared with the reson filter output of Fig. 3(a). Figure 3(c) illustrates how each of the filters adjusts to a phase shift in the isochronous sequence. Both filters go through a phase adjustment period between 5.25 and 6.25 s. The comb filter output, however, still contains remnants of the original phase after 6.25 s, while the effects of the original phase are not as apparent in the reson filter after 6.25 s. Figure 3(d) illustrates how the two filters react to a doubling of the isochronous sequence after 4 s. The comb filter outputs a second set of impulses that are antiphase with the original series, but the overall amplitude envelope of the comb filter is not affected by a doubling of the frequency of the isochronous series. In contrast, the energy in the reson filter (which is now effectively tuned to a period that is twice as long as the period in the pulse train) decays after 4 s at approximately the same rate as if silence had been introduced as in Fig. 3(a).

The phase perturbations in Figs. 3(b) and 3(c) and the doubling of the frequency of the isochronous sequence in Fig. 3(d) actually increase the overall energy of the comb filter response by introducing additional antiphase impulses, while causing an attenuation of the overall energy of the reson filter response. The reason why a reson filter is attenuated when presented with antiphase impulses is that the output signal of the filter assumes both positive and negative values, whereas the comb filter outputs only positive values. Thus, the amount of constructive or destructive interference imposed by an incoming signal on the output of a reson filter depends on the phase angle at which it arrives. This characteristic made the overall responses of the reson filter adapt more quickly to changes in input signal timing. In addition, it prevented the reson filter from responding to input signals arriving at twice the frequency to which the filter was tuned.

A reson filter is often implemented as a two-pole filter. Reson filters tuned to low frequencies exhibit high rolloffs below the resonance frequency. In order to circumvent this issue, we use a version of the reson filter called *reson_z* (Steiglitz, 1996). Reson_z introduces two zeros to the filter at $z=1$ and $z=-1$, resulting in a two-pole two-zero filter. The addition of two zeros solves this problem by improving the low frequency rolloff (for low frequency filters) or improving the high frequency rolloff (for high frequency filters). The transfer function for the reson_z filter is

$$H(z) = \frac{g(1 - z^{-2})}{1 - 2R\cos\ \theta z^{-1} + R^2 z^{-2}}, \tag{3}$$

where g is a gain factor (see section on reson filter gain below) and R is the radius of the poles. The value of R affects the sharpness of the reson filter response and is approximately related to the half-power bandwidth by

$$R \approx 1 - B/2, \tag{4}$$

where B is the half-power bandwidth. The value for R is calculated from the desired half-power bandwidth for each filter. Our model employs a series of 99 reson filters between 0.25 and 10 Hz. The number of filters is determined by the methods used to calibrate the filters (described below).

1. Reson filter decay

Scheirer (1998) and Klapuri *et al.* (2006) tuned their resonators with an equivalent *half-energy* time, where half-energy refers to the time it takes the output of each resonator to reach half amplitude. Our resonators are tuned with a decay rate based on the number of periods of the filters. High frequency resonators in our model exhibit far too much energy when employing the same half-energy time. This is evident when a series of quick notes, such as triplets or quintuplets, cause a high frequency filter to ring. The high frequency filter, with an equivalent decay time as the other filters, exhibits more energy than a lower frequency filter that represents the tactus frequency, since it is being reinforced at zero phase for each of these onsets. These rapid onsets, however, are perceived as subdivisions of the beat and should not raise the energy of high frequency filters above the energy of the filter that represents the tactus. We determined the decay rate of the filters by measuring the filter's *Q-factor*, which is the ratio of a filter's center frequency f_c to its bandwidth B.

$$Q = \frac{f_c}{B}. \tag{5}$$

Q-factor describes a filter's sharpness. It also describes the number of periods that a resonator's impulse response decays by a factor of $e^{-\pi}$, which is approximately −27 dB (Smith, 2007). We use a constant Q-factor of 13 for all the filters, which means that the filters all decay by approximately −27 dB after 13 periods. Since the Q-factor is constant across all the resonators, this also means that resonators with higher resonance frequencies also have proportionally wider bandwidths.

2. Reson filter spacing

There are at least two primary issues to address when choosing an appropriate spacing of resonators. First, the frequency domain must be adequately sampled. Second, the rolloffs of the magnitude responses of the filters should not meet too far below the peak magnitudes. If the magnitude responses do meet far below the peak magnitude responses, then certain frequencies are not captured by the filters as well as others.

Since the filters have a constant Q-factor, their bandwidths cannot be changed. The two issues mentioned above are adequately addressed by spacing the filters at a point above their half-power bandwidth. The half-power bandwidth is defined as the width of the magnitude response of the filter at its half-power points ($1/\sqrt{2}$ times the peak magnitude or −3 dB from the peak response). To produce reson filter responses that overlap above their half-power points, we space the filters so that their resonance frequencies are

half of a half-power bandwidth apart. The peak magnitude response of a reson filter can differ slightly from the pole angle of the filter, so an adjustment is made to approximate the true resonance frequency of each filter (Steiglitz, 1994). This results in 99 total filters with resonance frequencies from 0.25 to 10 Hz. There are 66 filters with resonance frequencies below 3 Hz. For the purpose of testing the model's responses to various input types, we found this to be a sufficient sampling of periodicities commonly found at the tactus level.

3. Reson filter gain

In some cases, it is possible for low frequency resonators, particularly below 1 Hz, to exhibit more energy than is appropriate if they are to be regarded as a model of perceptual salience. This issue arises when a series of rapid onsets is fed into a low frequency resonator. Multiple rapid onsets that all occur near the peak (zero phase) of the resonator will increase the overall output of the resonator more than a single onset would. Although we have yet to perform systematic calibrations of the model to match the perceived salience at different periodicities (see Sec. IV), we accommodate the need for such adjustments by introducing a gain factor g [see Eq. (3)] to the reson filter, which varies with the resonance frequencies of the filters. We vary the gain across each resonator bank between 0.1 and 0.4. A beta probability density function with parameters $\alpha=2$ and $\beta=5$ is used to vary the gain across resonators. This results in a gain that varies across the resonators in a semiparabolic shape, starting with a gain of 0.1 at 0.25 Hz, rising to a maximum gain of 0.4 at 2.23 Hz, and dropping back to a gain of 0.1 at 10 Hz. A lower gain for filters with resonances at higher frequencies also improves the resonator bank output slightly by reducing the prominence of high frequency filter outputs.

G. rms calculation on reson filters

In order to provide a clearer picture of the relative amplitudes of the resonator outputs, we perform a rms calculation over a sliding time frame at each sample point. The rms frame size we use is 2 s. Each rms frame begins 2 s before the current sample point and ends at the current sample point (i.e., each sample point is at the end of each rms frame). For time points prior to 2 s into the signal, a frame size equal to the number of samples up to the current sample point is used. The use of a frame size that gradually increases to 2 s at the beginning of the signal produces a ramping effect of the rms plot. The rms calculation smoothes the resonator outputs, facilitating the identification of relative energy levels of the resonators over time. Figure 4 illustrates the resonator outputs from each band (left panels) and the rms outputs (right panels). We refer to each of the rms plots as a *periodicity surface*. Note that the periodicity surface allows for a rapid appraisal of the metric relationships, if any, between the periodicities that are present in the signal of that band at any given moment in time.

H. Average periodicity surface

The rms outputs for each band are then averaged to produce a composite rms plot (bottom center of Fig. 4). We refer to this surface plot as the APS of a musical excerpt. The APS is then averaged in time over the whole excerpt to create a MPP (Fig. 4, bottom right panel). This graph allows one to identify the prominent periodicities in the excerpt. Further investigation of this plot allows one to infer various metric relationships and properties of the periodicities that are present. The peak widths may indicate the relative accuracy of a given frequency over the course of the excerpt. Ratios of peak frequencies suggest the meter of the excerpt. For example, in the MPP in Fig. 4, the peaks at filter frequencies of 0.84, 1.71, and 3.38 Hz present a 1:2:4 set of ratios indicating duple meter. However, more complex metric relationships also appear to be present as represented by the small peak at 2.50 Hz that exists in a 3:1 ratio with the peak at 0.84 Hz.

I. Audio signal inputs versus other inputs

The model described above can, in principle, be used to identify the periodicities present in other data streams. Of particular interest to us are the rhythmic properties of listeners' movements associated with hearing a piece of music. Such movements might include finger taps, key presses, or even the more complicated movements of multiple limbs. In order to compute periodicity surfaces for the latter types of data, it is necessary to adapt the model in order to accommodate heterogeneous signal types. While an audio signal arising from a performance of a musical piece can be analyzed by the model with no need for modification, other signals, such as MIDI data obtained from a drum pad or keyboard, require special treatment. With MIDI data, it is not appropriate to process the data through portions of the model concerned with simulating the auditory nerve firing rate and onset pattern extraction. A straightforward solution for working with MIDI data of this sort is to convert each *note on* message into a Dirac impulse. The signal produced in this fashion is then passed directly into a single resonator bank in order to produce periodicity surfaces and MPPs. Currently, for bimanual tasks, we do not differentiate between left and right hand taps. The amplitude of each Dirac impulse is linearly scaled by calculating the ratio of the note's MIDI velocity to the maximum velocity. We also mask out double strikes for each hand by removing onsets that follow an intertap interval below a certain threshold (50 ms). Since we employ only one resonator bank for Dirac impulse inputs, we refer to a surface plot of the rms of the single resonator bank output as a periodicity surface instead of an APS. A MPP is then calculated directly from the periodicity surface.

III. RESULTS

We evaluated the behavior of our model in several ways. First, we compared the performance of our model with the performance of existing models of tempo and beat estimation. Second, we inspected the behavior of our model with clearly defined input types: multiple series of isochronous Dirac impulses that varied in timing and in amplitude. Third,

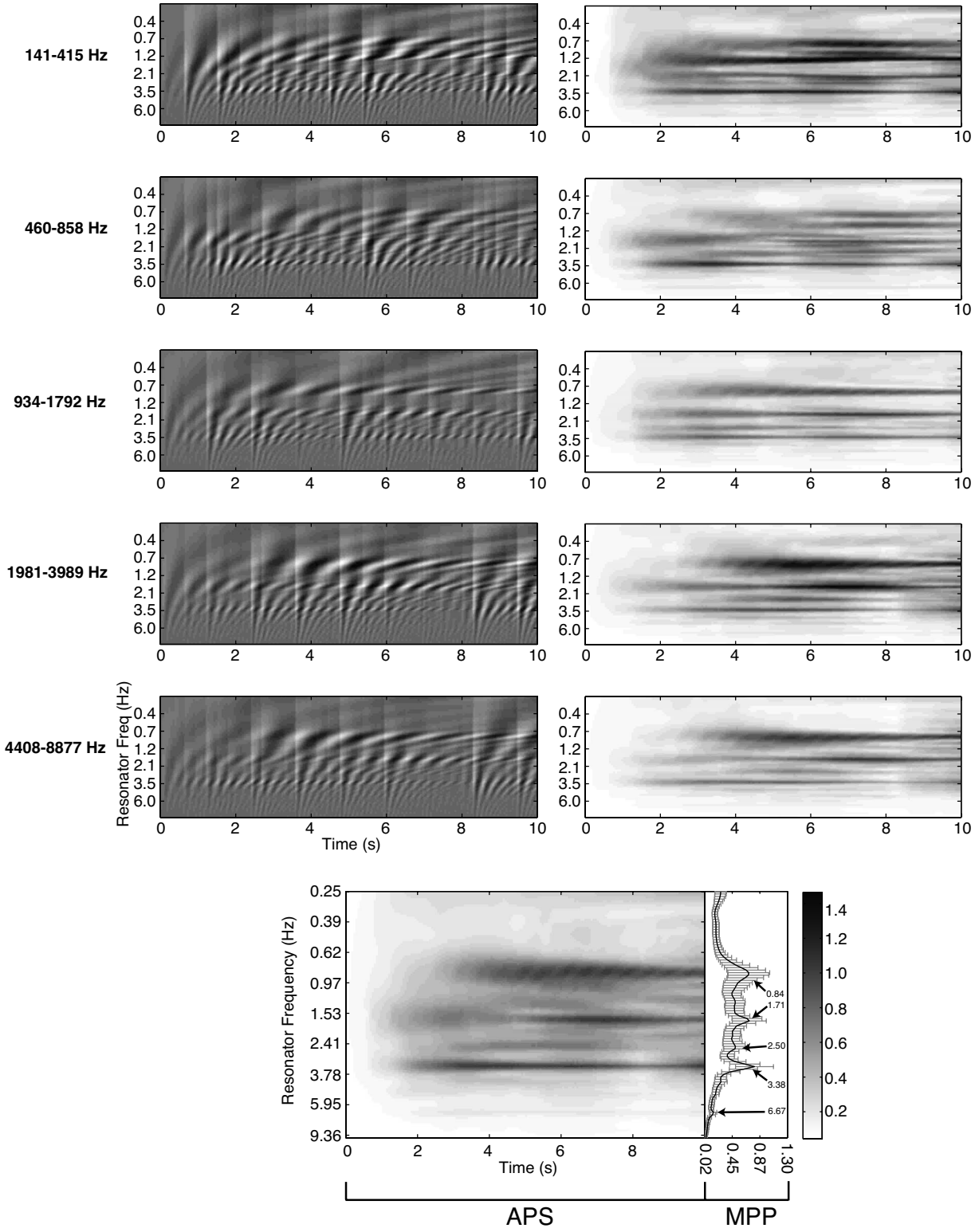

FIG. 4. Surface plots of the resonator outputs from each of the five bands are shown on the upper left. Grayscale values range from negative (black) to positive (white) values. Surface plots of the rms of the resonator outputs, referred to as periodicity surfaces, of the same five bands are shown on the upper right. Grayscale values range from zero (white) to positive (black). The APS, which resulted from averaging the five periodicity surfaces, is shown on the bottom surface plot. The MPP is on the right adjacent side of the APS. The black line of the MPP plots the mean of the APS over time, while the gray bars plot the standard deviation of the APS over time.

we tested the model with musical stimuli and behavioral response data (MIDI recordings of tapping data). Finally, in order to illustrate that the model output may be converted to a symbolic representation, we developed a technique for classifying the meter of simple accent patterns and tested the meter classification technique for accuracy.

A. Model accuracy

In order to verify that our model produces measurements with accuracy levels that are on par with other recent rhythm models, we processed a large set of audio clips that were used in the tempo induction contest conducted by Gouyon,

et al. (2006). Eleven algorithms were compared in the contest article and the results for a 12th algorithm were reported on the contest website.[1] In the contest, 3199 audio clips were used, comprised of three different data sets. Since the first data set (DJ loops) was copyrighted and not readily available, we processed our model with the second and third data sets (1163 of the 3199 audio clips used in the contest). The second data set consisted of 30 s excerpts of ballroom dancing music.[2] The third data set used in the contest consisted of 20 s song excerpts, apparently hand picked by the authors, from several different musical genres. As a ground truth tempo, the contest organizers asked a professional musician to place beat marks on each of the song excerpts. The tempo was computed as the mean of the annotated interbeat intervals. The annotated tempi were available to us from the contest website, so we were able to test our model with the same criteria as the contest algorithms.

There were two measurements used to determine the level of accuracy of the algorithms in the contest. Both measurements involved tempo estimation, even though many algorithms also provided other information such as beat locations. The first measurement (accuracy 1) calculated the percentage of tempo estimations that were correct for each algorithm within a 4% error margin. The second measurement (accuracy 2) calculated the percentage of tempo estimations that were correct within a 4% error margin for each algorithm, considering an estimation at half, double, three times, or one-third of the annotated tempo as correct.

The calculation of accuracies 1 and 2 for our model were 35.00% and 78.42%, respectively. The contest website reported a range of 25.22%–67.29% (mean 37.53% and median 34.02%) for accuracy 1 and a range of 50.73%–85.01% (mean 69.10% and median 71.69%) for accuracy 2 on all three data sets. Although we tested our model on only two of the three data sets, comprising approximately one-third of the total number of audio excerpts, it is reasonable to expect our model to perform at comparable accuracy levels on the complete data set. If our model performed with these scores in the contest using the complete data set, it would have ranked 7th of 13 for accuracy 1 and 4th of 13 for accuracy 2.

B. Timing variability in isochronous pulse sequences

We examined the response of our model to an isochronous series of Dirac impulses at 1 Hz without any timing deviations (Fig. 5(a)). As expected, the MPP exhibited the highest energy at the frequency of the impulses. Peaks were also observed for resonators at or near integer multiples (harmonics) of the peak frequency. The reason why a perfectly isochronous input yields energy at higher harmonics of the fundamental frequency is that each impulse is perfectly phase aligned with the higher harmonics. As the harmonic number increases, the impulses skip a number of periods of the filter, explaining the reduced energy level at each higher harmonic. An important distinction between reson filters and comb filters is that reson filters will not exhibit high energy at subharmonics since reson filters have a cancellation effect when presented with inputs at or near their anti-phase state.

We predicted that the energy in the higher harmonics would be sensitive to timing deviations from a strict isochronous series because any given deviation represents a larger phase perturbation at higher frequencies. Gaussian distributions have been used in other models to approximate timing deviations of human performance (Cemgil *et al.*, 2000; Gouyon *et al.*, 2002), so we utilized this distribution for introducing timing deviations. We varied the timing of the pulses by a Gaussian distribution function with a mean $\mu = 0$ and a standard deviation variable σ. Timing deviations produced by nonintentional motor noise tend to deviate between 10 to 100 ms (Desain and Honing, 1993), so we used standard deviations of 20, 40, and 80 ms to provide details on how our model reacts to timing deviations within this range. Figures 5(b)–5(d) illustrate the resonator outputs, periodicity surfaces, and MPPs produced from a series of impulses drawn from Gaussian distributions with standard deviations of 20 ms, 40 ms, and 80 ms, respectively. The most noticeable effect of the timing deviations introduced in Figs. 5(b) and 5(c) as compared with Fig. 5(a) was a reduction of harmonic peak amplitude, particularly above the third harmonic. The upper harmonics in the periodicity surface and MPP dissipated completely in the 80 ms condition.

Although not surprising that the filters corresponding to the higher harmonics would experience the variability in the input train as greater phase variability than would filters for lower harmonics, thereby leading to more damping of the filters' responses, the results of these simulations affirmed basic properties of the model and illustrated that periodicity surfaces were very stable across a significant range of timing deviations. The largest peak in the MPP, which corresponded to the frequency of the impulses, was very prominent, even when the standard deviation in the timing was as large as 80 ms. Since it is critical for rhythm models to be sensitive to timing (Honing, 2001), it was fortunate to affirm that timing variability is represented in our model.

C. Accent structure

In order to make the claim that our model can represent certain defining aspects of meter, we employed accented isochronous sequences as inputs to a single resonator bank. Figure 6(a) illustrates the resonator bank output, periodicity surface, and MPP produced by a series of Dirac impulses at 1 Hz with no variation, in a similar fashion to the series in Fig. 5(a) except that the amplitudes of the impulses were at half of the maximum amplitude. We used impulses at maximum amplitude and impulses at half-maximum amplitude for accented and unaccented onsets.

When every other impulse was accented (Fig. 6(b)), a peak emerged at 0.5 Hz and higher harmonics of 0.5 Hz. The MPP produced when accenting every third impulse (Fig. 6(c)) likewise contained peaks corresponding to the frequency of the accents, at 0.34 Hz, and harmonics of this frequency. It is important to note that the closest frequency to 0.33 Hz and corresponding harmonics were captured, and that if we employed a finer sampling of the frequency spectrum, peaks would occur closer to 0.33 Hz and integer multiples of this frequency.

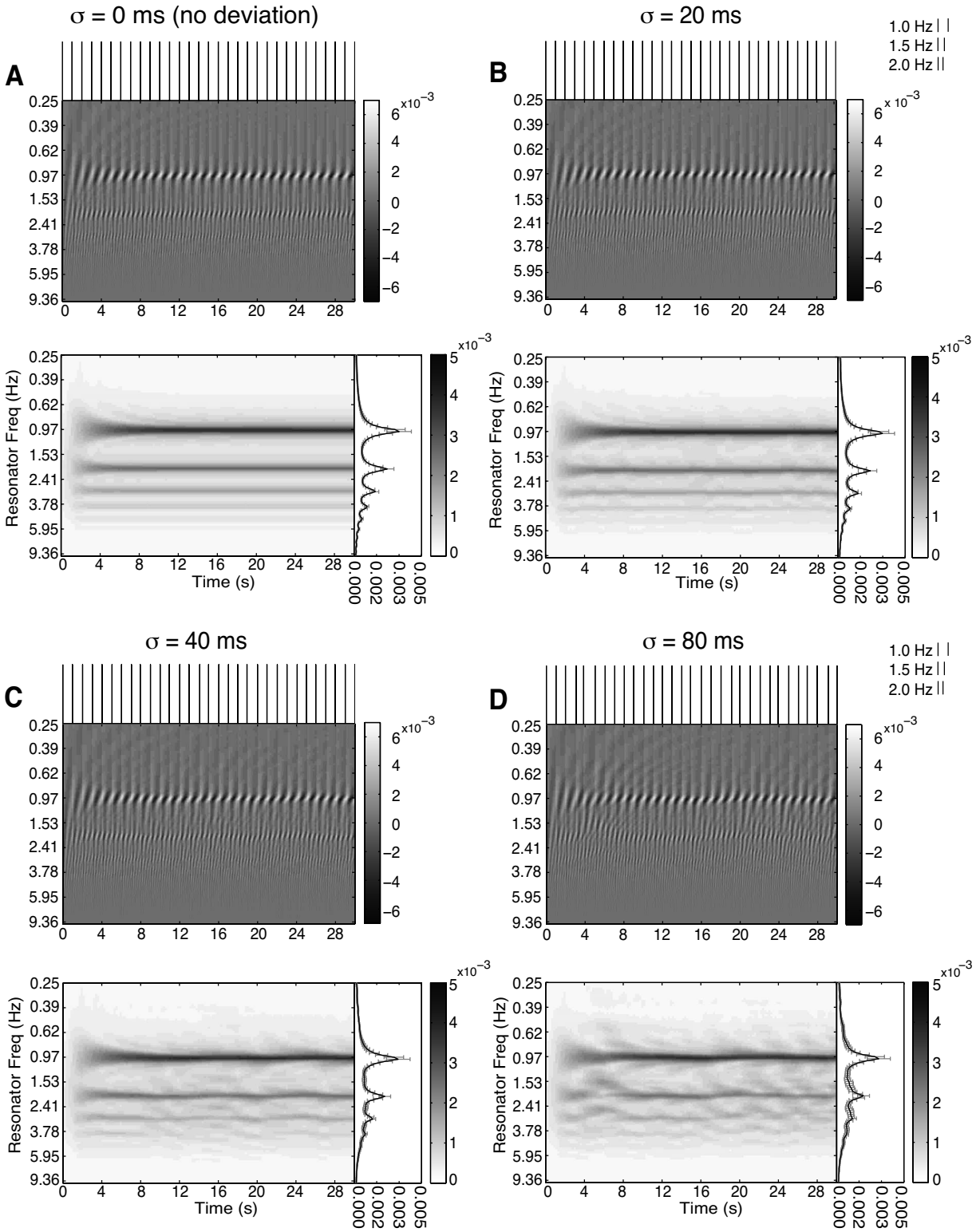

FIG. 5. The four sets of resonator outputs, periodicity surfaces, and MPPs illustrate how a resonator bank behaves when Gaussian noise is added to the timing of Dirac impulses in an otherwise isochronous series. (a) No deviation (σ=0 ms). Deviations in onset times are drawn from Gaussian distributions with (b) σ=20 ms, (c) σ=40 ms, and (d) σ=80 ms.

Figure 6(d) illustrates the resonator bank output, periodicity surface, and MPP from a series of impulses with accents alternating every second and third beats. The alternating accent pattern on Fig. 6(d) produced a MPP with small peaks at 0.41 and 0.60 Hz. These were harmonics of a peak that would have emerged at 0.2 Hz if the resonators spanned to this frequency. Harmonics at 0.8 and 1.2 Hz were not apparent probably because they fused with the wide peak at 1.0 Hz. This MPP represented a 1:5 metric ratio present in the accent structure. This illustrated how the model captures

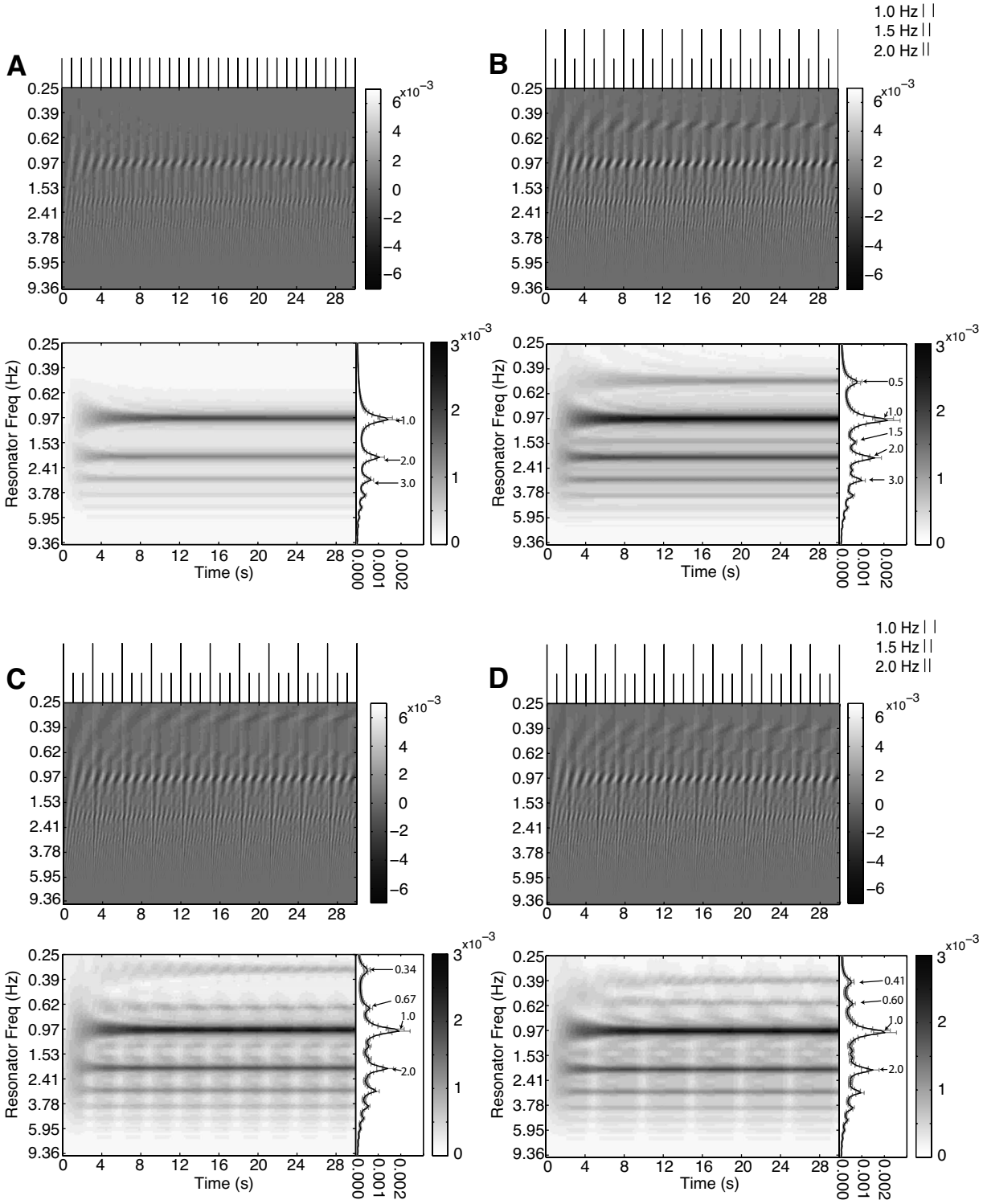

FIG. 6. The four sets of resonator outputs, periodicity surfaces, and MPPs illustrate how a resonator bank behaves when presented with a series of Dirac impulses that has an accent structure. Nonaccented impulses are assigned half the maximum amplitude, while accented impulses are assigned the maximum amplitude. (a) The input isochronous series has no accented impulses. (b) Every other impulse is accented. (c) Every third impulse is accented. (d) Accents are alternated every second and third impulses.

groupings of accent structures. One way to view the accent structure is that the first two accents are each followed by an accent every five onsets and are simply phase shifted from one another by two onsets.

The three accent patterns (Figs. 6(b)–6(d)) illustrated how accent structure is represented in periodicity surfaces and MPPs. The peak at 1 Hz and corresponding harmonics remained prominent. These peaks reflected the overall frequency of the impulses and were also reinforced as harmonics of the accent pattern frequencies in all three patterns. The results confirmed that accents, which often help to define metric structure in music, produce additional peaks and therefore contribute to the interpretation of metric levels present in a musical excerpt.

In order to test whether reson filters provide an advantage over comb filters in representing accent structure, we passed the same accent patterns through comb filters. Since there is a trend for the comb filterbank power response to increase as the filter delay increases (Klapuri *et al.*, 2006), it was difficult to see the effects of Dirac impulses on comb filter outputs. By convolving the same four series of Dirac impulses with a Gaussian distribution function, the peaks in the comb filterbanks were much clearer. We used a Gaussian distribution function with a full-width at half-maximum measurement of approximately 200 ms. We also passed this same convolved accent pattern through a bank of reson filters. Since the gain function that we normally use across the reson filterbank was tuned specifically for those filters, it was not appropriate for the comb filterbank. Therefore, we used a constant gain of 0.4 across the reson filterbank to provide a better comparison with the comb filterbank output. The comb filterbank was tuned to Scheirer's specifications, with a half-energy time of 1.5 s across all of the filters.

Figures 7(a)–7(d) illustrate the periodicity surfaces produced by a reson filterbank (top plots in each panel) and a comb filterbank (bottom plots in each panel). As evident in the figures, the reson filterbank produces peaks corresponding to the overall frequency of the spikes at 1 Hz and peaks corresponding to the accent frequencies in a similar fashion to Figs. 6(a)–6(d). Since we did not use an overall weighting function, a trend of the reson filters to produce higher amplitudes at lower frequencies was evident, but the peaks were still very clear. The periodicity surfaces produced from the comb filterbank, on the other hand, did not significantly change when presented with different accent patterns. The reason why the comb filterbank did not illustrate the frequency of the accent patterns could be that comb filters tuned to frequencies at subharmonics (frequencies that evenly divide the fundamental frequency) of the input also exhibit high energy. Therefore, accent patterns, which usually evenly divide an overall beat structure, are probably masked by these subharmonics. One aspect that is readily noticeable on the periodicity surfaces produced by comb filters in Figs. 7(c) and 7(d) is a series of light vertical bands. These are artifacts of the windowing technique used in calculating the rms for the periodicity surfaces. The same artifacts are also somewhat noticeable in the periodicity surfaces produced by the reson filters. These artifacts were not noticeable in other periodicity surfaces we produced with reson filters. The artifacts could possibly be avoided by using a leaky integrator for the rms, for which values further in the past are scaled progressively to lower values.

D. Percussion patterns

We constructed a MAX/MSP (Puckette and Zicarelli, 2006) patch for defining and playing MIDI percussion patterns. The MAX/MSP patch was executed on a G5 Macintosh, and the MIDI notes were rendered by an E-Mu Proteus 2000 synthesizer. The analog audio output from the synthesizer was sent back to the G5 Macintosh and recorded by a second MAX/MSP patch. We passed a series of 30 different percussion patterns through the model. Two of the patterns are discussed here. Both loops had a tempo of 90 bpm, where bpm indicates the number of quarter notes per minute rather than the perceived tempo. The first loop (Fig. 8(a)) was rendered with the sound of a single high hat (Proteus preset 68, bank 3, note value 68) and iterated 12 times resulting in a segment that was 32 s long. The APS of the excerpt exhibited a prominent peak at 0.75 Hz with many smaller peaks at greater frequencies.

The second loop decomposed the same note pattern into two instrument sounds, a "snap" (Proteus preset 68, bank 3, note value 79) and a high hat. Some of the notes from the score in Fig. 8(a) were played by only one of the two instruments while other notes were simultaneously played by both instruments to produce the score in Fig. 8(b). The two instruments were chosen for their differing timbres, so that we could inspect how differential activation of the five spectral bands affected the APS. This loop also iterated 12 times and the excerpt was approximately 32 s long. The most remarkable difference in the APS and MPP of this excerpt from the single instrument case was that the peak at 1.53 Hz was more prominent. The peak at 0.75 Hz was also very prominent but slightly lower than the peak at 1.53 Hz.

The APSs of the percussion patterns illustrated our model's sensitivity to timbral content. From inspecting the score in Fig. 8(b), it may have been possible for simultaneous onsets of both instruments to effectively produce accents on the periodicity surface of the same frequency band. However, a subsequent inspection of the periodicity surfaces from each band (data not shown) revealed that each of the instruments largely affected the periodicity surfaces of different bands. Therefore, the resulting APS in Fig. 8(b) resulted from separate periodicity analyses of the instruments which were subsequently integrated in the APS. This result suggests that it may be possible to use the model to investigate the effects of shifts in selective attention to timbre on rhythm perception (see Sec. IV).

E. Musical stimuli

Figures 8(c) and 8(d) illustrate the APSs and MPPs of two musical stimuli that we recently used during an experiment in which participants were instructed to tap isochronously and bimanually on a MIDI drum pad to the beat of the music (Janata *et al.*, 2007). Since multiple peaks are readily apparent in the APSs and MPPs of the stimuli, where each of the peaks indicates a prominent metric level in the music, it is of interest to inspect the frequencies and ratios of the peaks and speculate on how these relate to the content of the musical excerpts.

Herbie Hancock's "Maiden Voyage" produced a MPP with many peaks. The highest peak was at 1.78 Hz. Many other peaks were present, however, at 0.51, 0.78, 1.01, 1.27, 1.53, 2.32, and 4.08 Hz. With the exception of the peak at 2.32 Hz, which approximated a 4:3 ratio with the highest peak, these frequencies did not appear to approximate any simple ratios. This suggested that the excerpt contained complex rhythms. The piece is characterized by a prominent slow

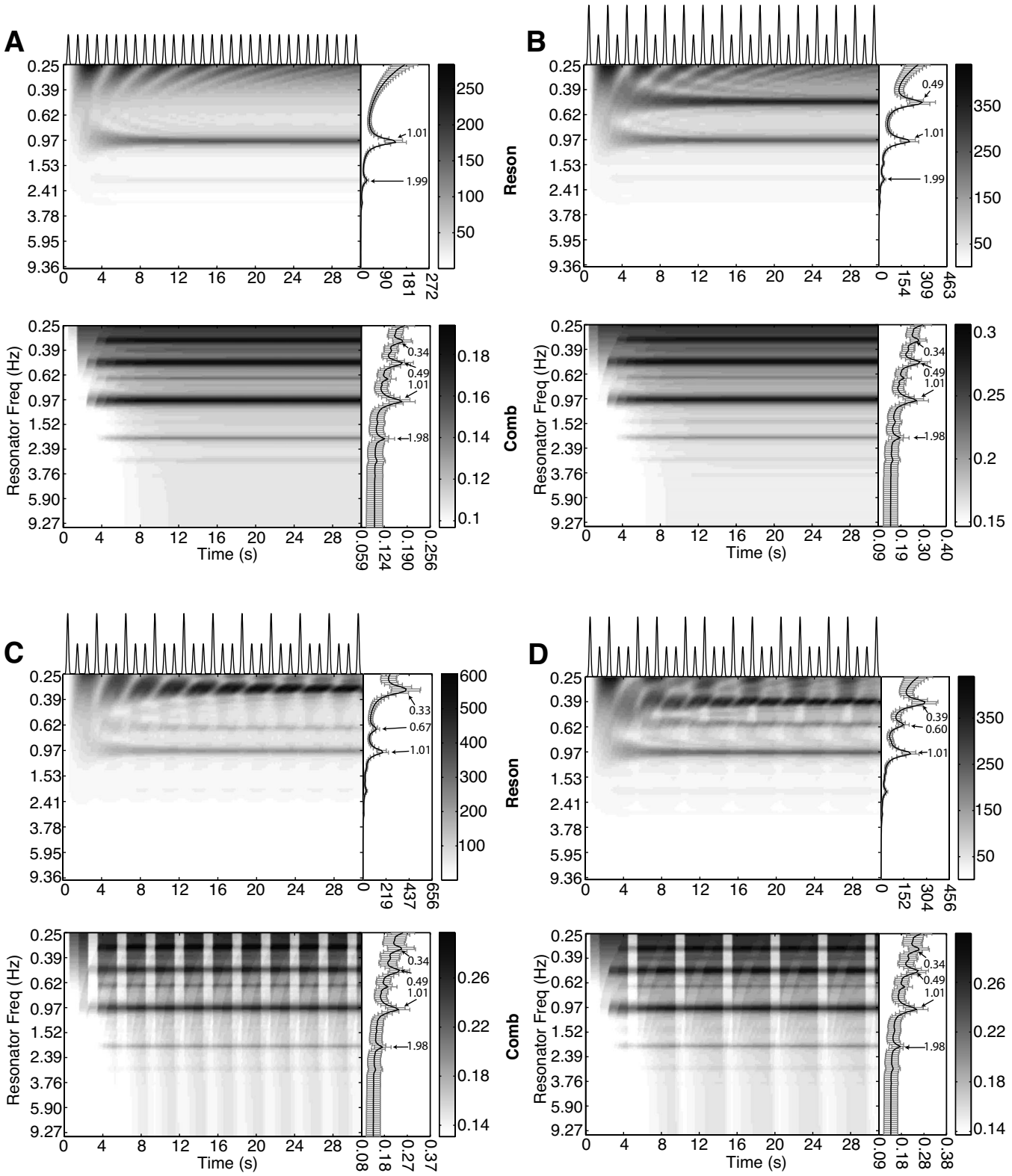

FIG. 7. The same accent patterns of Dirac impulses in Fig. 6 were convolved with a Gaussian distribution function and used as inputs for comparing periodicity surfaces produced by reson filters [top plots in (a)–(d)] with periodicity surfaces produced by comb filters [bottom plots in (a)–(d)]. A constant gain for the reson filters was used for a more equivalent comparison. The comb filters were tuned according to Scheirer (1998). The reson filters clearly illustrate the frequencies present in the accent patterns while the peaks in the periodicity surfaces produced by comb filters do not significantly change with the introduction of accents.

syncopated piano part, with less prominent but complex bass and drum parts, and an intermittent slow melody played by brass instruments.

By contrast, "It's a Wrap (Bye, Bye)" by "FH1 (Funky Hobo 1)" produced a MPP that provided a clear illustration of a metric hierarchy. The highest peak was at 1.53 Hz. The second highest peak was at 0.78 Hz and the third highest peak was at 3.02 Hz, which approximated ratios of 1:2 and

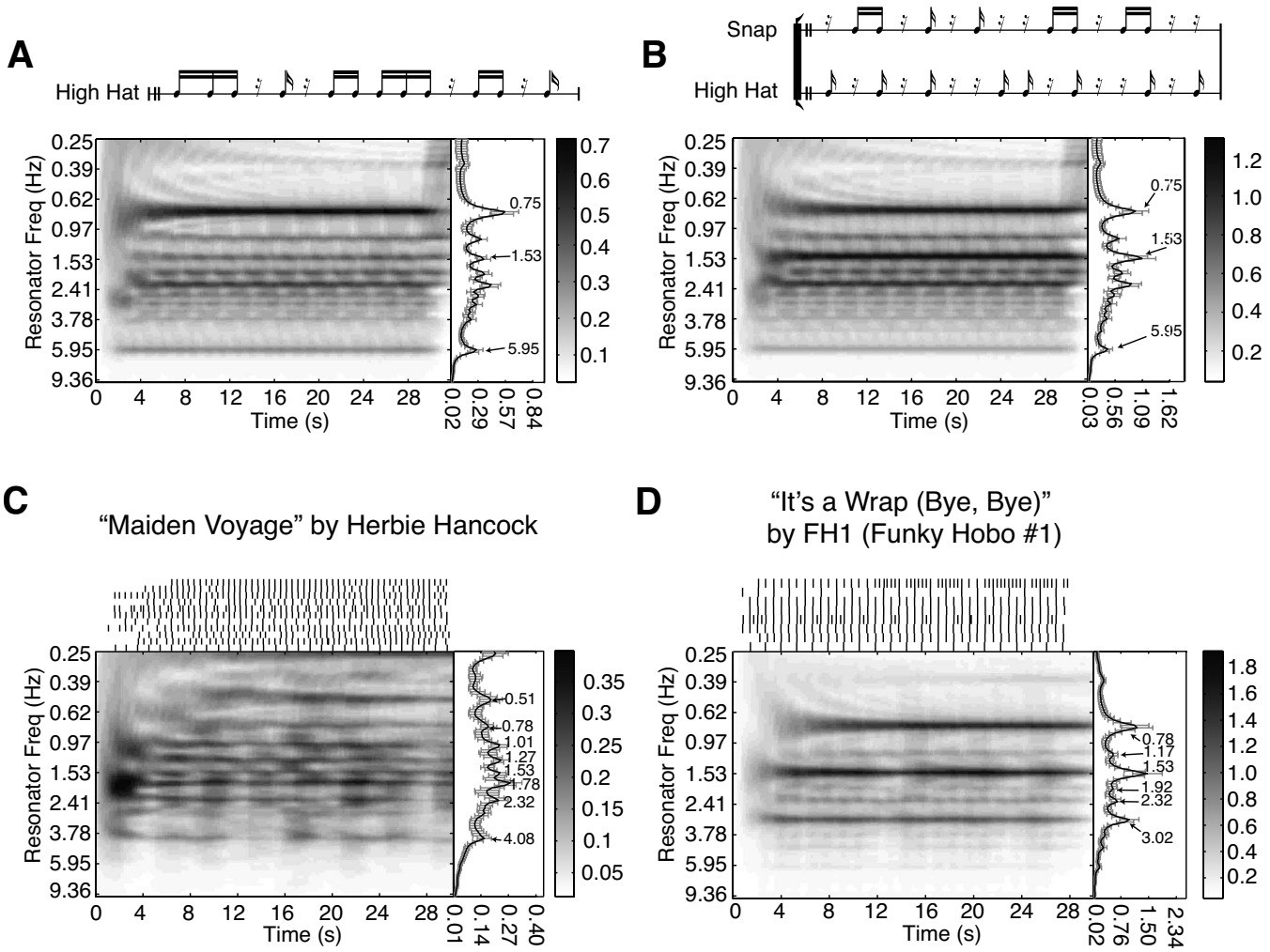

FIG. 8. APSs and MPPs produced from stimuli. (a) A pattern played by a single high hat. The score above the plot notates a 2.67 s long pattern that is iterated 12 times, resulting in a 32.0 s excerpt. The excerpt is used as an input to the model to produce the APS and MPP. (b) A stimulus with two instruments: snap and high hat for which every note from the excerpt in (a) is played by one or both of the instruments. The duration of the score is also 2.67 s long and iterated 12 times. [(c) and (d)] APSs and MPPs of excerpts from commercial musical recordings retrieved from the iTunes store. Each row in the raster plots above each APS corresponds to a different participant, and shows the times at which he/she tapped when instructed to tap along with the perceived beat in the music.

2:1 to the peak frequency. A prominent electronic "clap" sound played at approximately 47 bpm, resulting in the peak at 0.78 Hz. A bass guitar played at twice the tempo of the clap, likely contributing to the most prominent peak at 1.53 Hz. The peak at 3.02 Hz may have simply emerged as a harmonic of the peak at 1.53 Hz. The low amplitude peaks at 1.17, 1.92, and 2.32 Hz approximated 3:4, 5:4, and 3:2 ratios with the highest peak, respectively. Although these small peaks suggested subtle rhythmic arrangements, the piece has little rhythmic variation, which explains the three strong peaks at very simple ratios. The APS of the piece by FH1 illustrated that it changed very little over time, while the APS of the Herbie Hancock piece illustrated a more dynamic piece. The MPP of the excerpt by FH1 also contained prominent peaks at simpler ratios than the MPP of the Herbie Hancock excerpt. The contrast between these two excerpts illustrated how greater dynamics over time in the APS and peak frequencies at higher integer ratios in both the APS and MPP might indicate a greater rhythmic complexity.

The rows of rasters above the APSs in Figs. 8(c) and 8(d) show the tap onsets produced during individual trials. Each row corresponds to a different participant. It is evident from the rasters that participants chose different metric levels as the tactus of the music, even for a piece in which the metric hierarchy is clear. Although this example is limited to two musical excerpts, it illustrates the importance of considering individual variability and models that depict multiple metric relationships in work that seeks to understand how individuals synchronize with music. Although we have not devised a method for determining the metric level participants that are attending to, it may be possible to make comparisons using APSs and MPPs of tapping responses corresponding to these stimuli. In Sec. III F, we illustrate how the tapping data may be processed.

F. Tapping data

We processed two tapping trials from the experiment described above in Sec. III E. The excerpt playing during both trials was the same excerpt of Maiden Voyage by Herbie Hancock. The first trial (Fig. 9(a)) was a bimanual isochronous tapping task. The second trial (Fig. 9(b)) was a

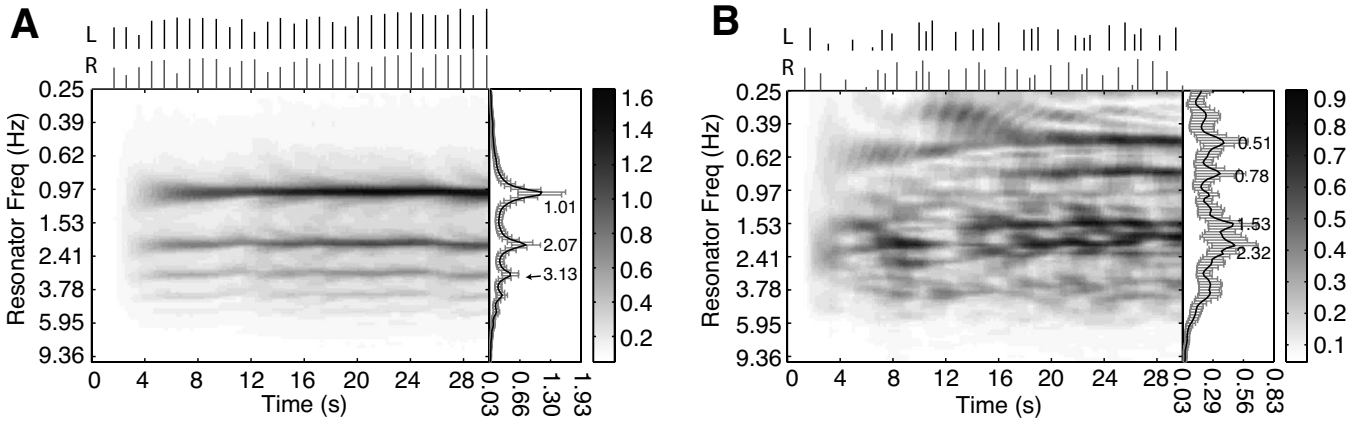

FIG. 9. Periodicity surfaces and MPPs of response tapping data recorded as MIDI. The MIDI note ons are converted to Dirac impulses that are subsequently amplitude scaled by MIDI velocity, and then passed to a single resonator bank. The vertical lines at the top plot the amplitude scaled Dirac impulses produced by the left and right hand taps. The trials shown are from (a) an isochronous tapping task and (b) a free-form tapping task. The stimulus being presented during both trials was the musical excerpt for which the APS and MPP were illustrated in Fig. 8(c).

free-form tapping trial, in which the participant was instructed to tap along with the music as he/she saw fit. The lower left and lower right sections of the drum pad on which the participant tapped produced unique MIDI notes, allowing us to distinguish between left hand and right hand taps. Each note on event from the MIDI data was converted to a Dirac impulse, scaled by the velocity value of the note on event. The series of impulses was passed through a single resonator bank, using the method described in Sec. II I.

The MPP of the isochronous tapping task illustrated a clear peak at 1.01 Hz, estimating the approximate tempo of the participant's tapping. Harmonics of the highest peak appear as expected. The free-form tapping task in Fig. 9(b) illustrates that the participant was capturing a few different metric levels evident in the APS and MPP of the stimulus. All of the peak frequencies in the MPP produced by this tapping pattern were also present in the MPP produced by the stimulus shown in Fig. 8(c). The APSs and MPPs of tapping responses illustrate that it is possible to capture important metric properties of tapping data, and that these data can be compared with the APSs and MPPs of stimuli.

G. Determining the meter of musical excerpts

In order to use our model for analyses of stimuli and behavioral data, a conversion of the model output to a symbolic representation may be necessary. Since we wish to eventually use our model for multiple rhythmic measurements (see Sec. IV), the development of a comprehensive set of symbolic conversions would be beyond the scope of this paper. In order to illustrate one possible way in which the model can be converted into a symbolic output, we developed a simple technique that serves primarily as a proof-of-concept. As we have shown, the frequencies of the peaks in MPPs generally relate to one another by rational numbers, and the frequency ratios are closely linked to the metric patterns present in the inputs. We leveraged this property of MPPs in order to produce a simple automatic meter classifier that operated on inputs with varying accent patterns.

In a study conducted by Palmer and Krumhansl (1990), theoretical predictions of note occurrence frequencies on the quarter note and eighth note positions within a measure for various meters (2/4, 3/4, 4/4, and 6/8) were illustrated. We

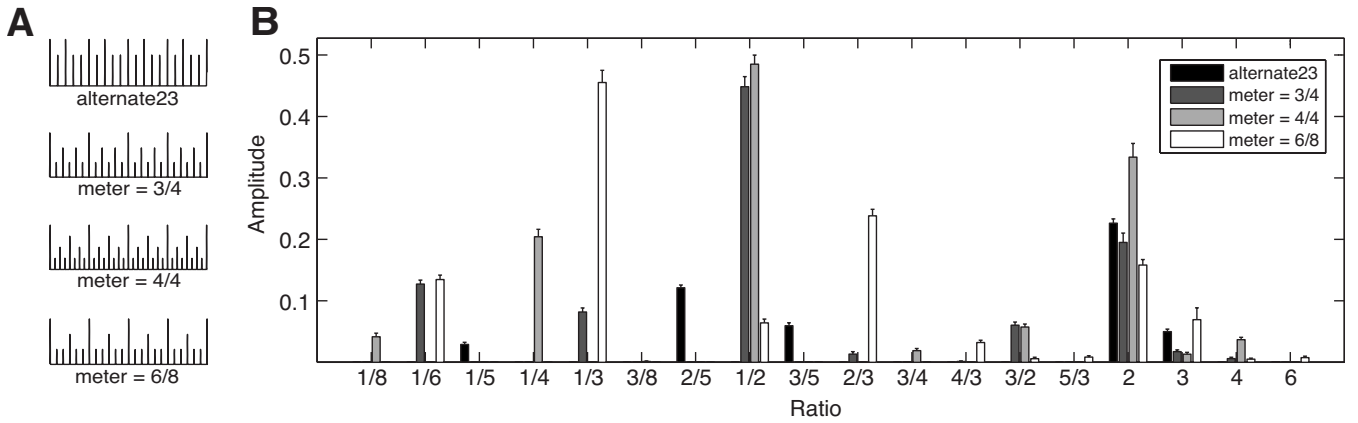

FIG. 10. (a) Accent patterns that were used as exemplars for four different metric classes. Four measures of each pattern are illustrated here. The amplitudes of the accented onsets (except the highest onset) and the tempi were randomized, while maintaining the accent hierarchy to produce 100 exemplars for each class. (b) Mean peak height values of the MPPs of the four sets of exemplars and their standard errors. The peaks were binned by the approximate frequency ratio relative to the frequency of the peak with the highest amplitude of each MPP.

used the note occurrence predictions to produce accent pattern exemplars for 3/4, 4/4, and 6/8 (Fig. 10(a)). Since the pattern illustrated for 2/4 was the same as for 4/4, we excluded 2/4 from our metric class list. As a fourth metric pattern, we used the same alternated two and three beat accent pattern illustrated in Fig. 6(d).

We varied the amplitudes of the onsets (except the highest one which stayed at maximum amplitude) by ±10% of the maximum amplitude. This allowed us to create a wide variety of patterns while still visibly maintaining the hierarchy of the accents. Across exemplars, tempo was manipulated by varying measure durations between 1.67 to 3.33 s. 100 exemplars (patterns) for each of the four classes were produced by randomizing the accent amplitudes and tempi with a uniform distribution within the ranges mentioned. Each pattern was repeated as many times as was necessary to produce 30 s excerpts.

After producing exemplars for each metric class and processing them through our model, we used a simple automatic peak finding algorithm to identify the frequencies and heights of the peaks in the MPPs. The peak finder locates points at which the slope of the MPP switches from positive to negative. Heights of the peaks were calculated from the higher of the left or right base of the peak (where the slope changes from negative to positive). Peaks with height values that were less than 5% of the amplitude range of the MPP were ignored. We then approximated the closest ratios that the peak frequencies represented in relation to one another, using the peak with the highest amplitude as a reference point with a ratio=1. For example, if the highest peak was at 3.0 Hz, it would be assigned a ratio=1, while peaks at 1.5 and 0.75 Hz would be assigned ratios of 1/2 and 1/4, respectively. We compiled a list of commonly found simple ratios in MPPs (1/8, 1/6, 1/5, 1/4, 1/3, 3/8, 2/5, 1/2, 3/5, 5/8, 2/3, 3/4, 4/5, 5/6, 7/8, 5/4, 4/3, 3/2, 5/3, 5/2) and integer multiples from 1–8. We then found the closest ratio or multiple that expressed the frequency of each peak in relation to the frequency of the highest peak. The frequency ratios of the peaks of all of the exemplars were successfully approximated within an error margin of 5%. The mean amplitudes of the peaks of the exemplars for each class, binned by ratio, are illustrated in Fig. 10(b).

In order to produce a pattern classifier for recognizing the similar peak relationships exhibited within each class, we implemented a feedforward neural network with backpropogation using MATLAB's Neural Network Toolbox. The default transfer function and learning rate associated with a network initialized by the newff() function were used. An input layer was fully interconnected with an output layer, but no hidden layer was used. All of the binned peak height values, with the exception of the peak at ratio=1, were used as inputs into the network. The MPPs of most exemplars had very prominent highest peaks, resulting in a strong preference for the value at ratio=1 in the weight matrix of the network when it was included. We were primarily interested in identifying other frequency ratios, such as 1/6, 1/4, 1/3, and 1/2 in the MPPs, since different metric classes were characterized by the presence or the absence of simple frequency ratios or multiples of the peaks. The output layer consisted of four units, thus representing each of the classes from which exemplars were drawn.

80% of the exemplars were selected randomly from each class in order to train the network. After training, the remaining exemplars were used to test the performance of the network. Each test exemplar was presented to the network and the identity of the winning output unit was compared to the identity of the class from which the exemplar was drawn. In this manner, we determined the proportion of exemplars that were identified correctly. Since the performance of the network is dependent on the subsets of exemplars chosen for training and testing, we reselected our training and testing sets, and reinitialized, retrained, and retested the network over 100 trial runs. We then calculated the mean performance of the network over all of the trials. The network performed very well at automatically classifying our test exemplars. The neural network correctly classified on average 92.9% (13.6% standard deviation) of the test exemplars. For 69 of the 100 trials, the network classified 100% of the test exemplars correctly.

Since the training and testing sets were very simple accent patterns, and a limited set of metric classes was chosen, the performance of this network is not indicative of how successful the technique would be with musical recordings or tapping performance data. However, the high success rate of the neural network on this simple data set indicated that conversions to symbolic representations are possible with our model and that a similar technique could be used for identifying patterns of peak ratios and amplitudes in recorded music and tapping patterns.

IV. DISCUSSION

In this paper, we described a model for performing periodicity analyses of audio stimuli and MIDI performance data. The model is based on banks of resonator filters. In contrast to similar models in which the primary objective is to identify the underlying beat in a piece of music (Scheirer, 1998; Klapuri *et al.* 2006), the objective of our model is to depict and describe the metric relationships that are present over the course of a piece of music or behavioral responses, such as tapping, to that music.

By inspecting the responses of our model to different types of input, we have an understanding of some of the model's basic response properties. The model regularly produces harmonics of the estimated tempo. We illustrated that upper harmonics are more sensitive to timing deviations than lower harmonics using isochronous series of Dirac impulses. More importantly for the analysis of metric structure, we showed that the model was sensitive to the accent structures present in the signal. The model was also sensitive to timbral changes when employing percussion sequences with different instrumentations. The periodicity surfaces and MPPs of the trials from the tapping experiment, as well as those of the music, illustrated various properties such as meter, rhythmic changes, and rhythmic complexity in both input types.

With a more informed understanding of our model's responses to various input types, we will now discuss how our

model relates to other models of meter and rhythm and suggest possible extensions and applications for our model.

A. Relationship of our model to other models of meter and rhythm

Models of metric and rhythmic structures have been developed to work on different types of input data and for different analytic purposes. Some models utilize symbolic data, e.g., musical scores, quantized interonset intervals, or MIDI data, whereas others operate directly on an audio signal. Analysis objectives include categorization for purposes of studying perception (Desain and Honing, 2003), complexity estimation (Shmulevich and Povel, 2000), beat finding (Scheirer, 1998; Klapuri *et al.* 2006), automatic music transcription (Cemgil *et al.*, 2000; Klapuri, 2004), automated genre classification (Dixon *et al.*, 2003; Tzanetakis and Cook, 2002), or combinations of these objectives. Below, we discuss the relationship of some of these models to our model and suggest how our model could be extended to accommodate these various objectives.

Two models that operate on symbolic data are those of Desain and Honing (2003) and Shmulevich and Povel (2000). Desain and Honing (2003) investigated the rhythmic categorization using groups of four onsets, defining three IOIs. They mapped the IOI triplets to a triangular surface on which each location represented a particular rhythmic pattern. This representation is useful for depicting the borders of perceived rhythmic categories because it is easy to show the areas of timing deviations around particular points in this space within which the percept of the rhythmic pattern does not change. Our cursory investigation of our model's responses to simple accent structures and tapping patterns shows that different rhythmic patterns manifest themselves as different distributions of peaks in the periodicity surfaces and MPPs, and that timing variability influences the presence of higher harmonics. Because our representational scheme is not limited to the relationships among groups of four events, it may be possible to use our model to examine perceived and produced rhythmic category structure in extended sequences.

Related to the issue of perceived rhythmic categories is perceived rhythmic complexity. While rhythmic complexity has been studied using IOI patterns and a heuristic for modeling the relative salience of measures with different event distributions (Shmulevich and Povel, 2000), APSs or MPPs might provide an alternative substrate for complexity calculations. For instance, one could imagine calculating the entropy of a MPP, with a greater number of peaks resulting in a larger entropy value. Alternatively, a distribution of the integer ratios of the peaks present in a MPP could serve as a basis for a complexity calculation. A preponderance of simple ratios, e.g., 2:1 and 3:1, in such a distribution would likely be associated with pieces of music that are perceived as less complex.

One obvious drawback to employing algorithms dependent on IOIs for rhythmic categorization or rhythmic complexity analysis is that onset timing information must either be readily available or extracted using an onset detection algorithm. A second drawback to using IOIs for this purpose is that the periodicities present in a grouping structure are not reflected in the distribution of IOIs. A simple histogram only reflects durations between onsets and does not reflect durations of groups of onsets. Therefore, methods employing IOI histograms must use various algorithms to extract these other durations. Our model readily captures the periodicities of pattern groups, as evidenced in the alternated two to three accent pattern (Fig. 6(d)), so therefore may provide a convenient mechanism for rhythmic categorization or rhythmic complexity measures.

Our model falls into a class of models that operate on nonsymbolic audio data. We provided a modification of the Scheirer model by incorporating the Auditory Peripheral Module of the IPEM Toolbox, inspection of periodicities other than the estimated tactus, and reson filters for frequency analysis. The incorporation of the Auditory Peripheral Module provided our model with filter approximations of the outer ear and cochlea, as well as an approximation of the mechanical to neural transduction of cochlear hair cells. This was a slight departure from the Scheirer model since this stage also effectively decomposes the signal into separate frequency bands. While models for extracting beat and meter are very useful for the purposes of automatic music transcription, the ability to extract other metric levels could be very useful to researchers of music perception. Scheirer (1998) only tracked the tempo and beat locations of frequencies between 1 and 3 Hz, and was not concerned with inspecting competing filters for extracting multiple metric levels. Klapuri *et al.* (2006) made significant steps toward tracking frequencies other than the tactus, namely, the tatum and measure. Our model illustrates that it is possible to extract other metric levels by directly inspecting the relative levels of the filter outputs.

Our most significant departure from the models of Scheirer (1998) and Klapuri *et al.* (2006), however, is our use of reson filters instead of comb filters. Reson filters have the property that they cycle through positive and negative values in response to an input (Fig. 3). Thus, a stimulus arriving out of phase with the oscillation established within a reson filter will reduce the energy in the filter depending on the magnitude of the arriving stimulus and its phase angle. It is this property that allows our model to represent the accent structure in an isochronous sequence of beats. When all of the beats are of equal magnitude, the output of filters at subharmonics of the period is suppressed due to the out-of-phase inputs. However, when an accent structure is imposed, the weak beats do not provide enough energy to entirely suppress the output of the filter that is tuned to the period of the strong beats.

Our model is by no means the first to employ oscillator-based filters for frequency analysis. Large (2000) used a bank of oscillators for periodicity analysis, though his model utilized a network of Hopf oscillators rather than reson filters. In terms of the preprocessing steps, such as the use of an auditory periphery model, envelope estimation, and an onset extraction process, the models are quite similar. A major difference between Large's model and our own, however, is that in his network active oscillators inhibit other oscillators. His model employs an inhibition matrix that causes the model to

favor harmonic and subharmonic multiples of 2 and 3. Large (2000) argued that the inhibitory nature of the oscillators in his model facilitates the representation of a large number of metrical patterns. In our model, this inhibitory property is not necessary, as multiple metric levels, and even competing metric organizations, e.g., triple meter in the presence of strong duple meter, are represented without it.

One aspect of our approach is to consider the overall periodicity surface rather than focusing on those metric levels that are related directly to the tactus level. The richer complement of information in the periodicity surfaces could assist in pattern classification efforts in areas such as genre classification. Such an approach has been illustrated by Tzanetakis and Cook (2002), who used autocorrelation to detect periodicities in the signal and derive a beat histogram. The bins in their beat histogram represent the prominence of tempi within the range of 40–200 bpm. They use the beat histogram for comparing and contrasting rhythmic properties of different musical genres. APSs from our model could be employed similarly, where our MPPs or the peak height values binned by frequency ratios (Fig. 10(b)) could be regarded as similar to their periodicity histograms.

B. Further modifications and possible applications

Intuitively, one may be inclined to consider the highest peak in the APS (for audio inputs) or in the periodicity surface (for MIDI inputs) to represent the tempo of the excerpt. Following this logic, other metric levels corresponding to subdivisions of the beat, syncopations, and polyrhythms would be represented by lower amplitude peaks, and measures would be represented somewhere in this hierarchy. While it is apparent that tempo, syncopations, polyrhythms, and measures are represented somewhere in the peaks of APSs and MPPs, we have yet to develop an automated algorithm for identifying and categorizing local rhythmic patterns. Since developing a beat-finding or rhythm transcription algorithm was not our primary goal, we left the details of such an implementation open. Our longer-term objective is to develop a model that we can use to investigate the perceptual salience of different metric levels and rhythmic groupings, along the lines of Large and Palmer (2002).

When analyzing the metric properties of stimuli and responses using our model, it is important to bear in mind that peaks in the APSs and MPPs reflect some combination of the actual periodicities in the input signal and higher harmonics of these periodicities. While it was possible to identify which peaks were harmonics in periodicity surfaces produced by simple patterns of Dirac impulses, this was much more difficult to do with audio inputs. If one wishes to devise an automatic extraction of the metric levels present in stimuli and responses, a process for compensating for the harmonics must be devised. One possibility would be to create templates that estimate the harmonics that peaks of varying widths would produce. One could then subtract the second and higher harmonics of the templates from periodicity surfaces and MPPs to estimate the metric levels present in the excerpts. We showed that the amplitude of the higher harmonics reflects the timing variability at the fundamental frequency. Thus, these correction templates could, in principle, be used to estimate the degree of timing variability.

The discrepancy between the APSs in Figs. 8(a) and 8(b), the first produced by a single high hat and the second produced by a high hat and snap, indicated that timbre had a strong influence over the final output of the model. Because periodicity surfaces are calculated for each of several spectral bands prior to averaging into an APS, each periodicity surface will reflect the temporal patterning in the parts played by the instruments whose energy falls in that spectral band. One can therefore foresee the calculation of APSs in which certain bands are weighted more strongly than others. Such APSs could be useful for calculating periodicity surfaces that predict what the salient periodicities should be if a listener is attending selectively to a particular spectral region. We are fully aware that there is not a straightforward mapping between instruments and spectral regions. A more accurate model of metric and rhythmic structures in polyphonic music should perhaps calculate periodicity surfaces for individual auditory object streams or combinations of object streams. Such a model would require additional auditory scene analysis stages and is beyond the scope of our current efforts.

Just as multiple periodicity surfaces are combined for auditory inputs, our model can, in principle, be used to combine multiple sources of performance data. For instance, separate periodicity surfaces could be calculated and combined for the left and right hands of tapping trials. The periodicity surfaces would look different if we calculated periodicity surfaces of the left and right hands separately, as we did for separate frequency bands of audio recordings, and then combined them into an APS. The many ways in which multiple input streams (both audio and performance) can be combined either prior to or following periodicity surface calculation raises questions about the most appropriate ways to combine them if one is to model cognitive or sensorimotor integration processes. One intuition is that a heuristic for combining the streams in the model should reflect mental grouping processes that underlie perception and behavior. In other words, if a snare and high-hat are perceived to create an aggregate rhythm, then a periodicity surface should be calculated on their combined input. Similarly, if tapping patterns performed by the left and right hands are perceived as unitary rather than as a combination of independent processes, it would be more appropriate to combine the information from both hands prior to periodicity surface calculation as we have done in this paper. A thorough treatment of these issues is necessarily left to future research.

V. CONCLUSION

Our model drew upon previous work in tempo and beat induction. However, instead of focusing on the estimation of tempi and beat locations, we investigated a method of describing multiple metric levels from musical audio files and MIDI tapping data. We investigated various behaviors of our model, such as sensitivity to timing and accent structure, the production of harmonic peaks, and sensitivity to timbre. This model may be useful for future studies that require the mea-

surement of periodic structure and the relationships between multiple metric levels in stimuli and response data collected from music cognition experiments. Although it is our hope that the model can be used to predict the perceptual salience of metric levels and to compare models of stimuli with performance data, we have not formally tested the ability of the model to do so. Nonetheless, we believe that the model provides an extensible framework for investigating a broad range of psychological phenomena and music analytical issues. The code for our model is publicly available for download.[3]

ACKNOWLEDGMENTS

We would like to thank Peter Keller and Bradley Vines for their comments and suggestions. We are also indebted to Joe Saavedra, Peter Keller, and Philip Front for developing the MAX/MSP patch for constructing the percussion patterns used in this study and to Rawi Nanakul for the collection of tapping data. This work was supported in part by a Templeton Advanced Research Program grant from the Metanexus Institute to P.J.

[1]http://mtg.upf.edu/ismir2004/contest/tempoContest
[2]Provided by BallroomDancers.com and labeled by this group with a tempo value.
[3]http://atonal.ucdavis.edu/projects/musical_spaces/rhythm/btb

Cemgil, A. T., Desain, P., and Kappen, B. (**2000**). "Rhythm quantization for transcription," Comput. Music J. **24**, 60–76.
Cemgil, A. T., Kappen, B., Desain, P., and Honing, H. (**2000**). "On tempo tracking: Tempogram representation and Kalman filtering," J. New Music Res. **29**, 259–273.
Desain, P., and Honing, H. (**1993**). "Tempo curves considered harmful," Contemp. Music Rev. **7**(2), 123–138.
Desain, P., and Honing, H. (**2003**). "The formation of rhythmic categories and metric priming," Perception **32**, 341–365.
Dixon, S., Pampalk, E., and Widmer, G. (**2003**). "Classification of dance music by periodicity patterns," in Proceedings of the Fourth International Conference of Music Information Retrieval, Baltimore, Maryland.
Gouyon, F., and Dixon, S. (**2005**). "A review of automatic rhythm description systems," Comput. Music J. **29**, 34–54.
Gouyon, F., Dixon, S., Pampalk, E., and Widmer, G. (**2004**)."Evaluating rhythmic descriptors for musical genre classification," in Proceedings of the AES 25th International Conference, London, UK.
Gouyon, F., Herrera, P., and Cano, P. (**2002**). "Pulse-dependent analyses of percussive music," in Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland.
Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., and Cano, P. (**2006**). "An experimental comparison of audio tempo induction algorithms," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1832–1844.
Honing, H. (**2001**). "From time to time: The representation of timing and tempo," Comput. Music J. **25**, 50–61.
Janata, P., Tomic, S. T., and Haberman, J. (**2007**). "Getting in the groove while tapping," in Poster presented at the Society of Music Perception, Montreal, Canada.
Klapuri, A. P. (**2004**). "Automatic music transcription as we know it today," J. New Music Res. **33**, 269–282.
Klapuri, A. P., Eronen, A. J., and Astola, J. T. (**2006**). "Analysis of the meter of acoustic musical signals," IEEE Trans. Audio, Speech, Lang. Process. **14**, 342–355.
Large, E. W. (**2000**). "On synchronizing movements to music," Hum. Mov. Sci. **19**, 527–566.
Large, E. W., and Palmer, C. (**2002**). "Perceiving temporal regularity in music," Cogn. Sci. **26**, 1–37.
Leman, M., Lesaffre, M., and Tanghe, K. (**2001**). Computer code IPEM Toolbox, Ghent University, Ghent, Belgium.
London, J. (**2004**). *Hearing in Time: Psychological Aspects of Musical Meter* (Oxford University Press, New York).
Palmer, C., and Krumhansl, C. (**1990**). "Mental representations of meter," J. Exp. Psychol. **16**, 728–741.
Puckette, M., and Zicarelli, D. (**2006**). Computer code MAX/MSP, IRCAM, Paris, France/Cycling'74, San Francisco, CA.
Repp, B. H. (**2005**). "Sensorimotor synchronization: A review of the tapping literature," Psychon. Bull. Rev. **12**, 969–992.
Scheirer, E. D. (**1998**). "Tempo and beat analysis of acoustic musical signals," J. Acoust. Soc. Am. **103**, 588–601.
Shmulevich, I., and Povel, D. J. (**2000**). "Measures of temporal pattern complexity," J. New Music Res. **29**, 61–69.
Smith, J. O. (**2007**). "Decay time is Q periods," *Introduction to Digital Filters with Audio Applications* (Standford University, Stanford, CA); http://ccrma.stanford.edu/~jos/filters/Decay_Time_Q_Periods.html (Last viewed October 2007).
Snyder, J., and Krumhansl, C. L. (**2001**). "Tapping to ragtime: Cues to pulse finding," Music Percept. **18**, 455–489.
Steiglitz, K. (**1994**). "A note on constant-gain digital resonators," Comput. Music J. **18**, 8–10.
Steiglitz, K. (**1996**). *A DSP Primer : With Applications to Digital Audio and Computer Music* (Addison-Wesley, Menlo Park, CA).
Todd, N. P. M. (**1994**). "The auditory primal sketch: A multi-scale model of rhythmic grouping," J. New Music Res. **23**, 25–70.
Tzanetakis, G., and Cook, P. (**2002**). "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process. **10**, 293–302.
Van Immerseel, L. M., and Martens, J. P. (**1992**). "Pitch and voiced unvoiced determination with an auditory model," J. Acoust. Soc. Am. **91**, 3511–3526.