Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Dynamic Time Signature Recognition, Tempo Inference, and Beat Tracking through the Metrogram Transform

James M. Cozens, Simon J. Godsill

[1]Signal Processing and Communications Laboratory, University of Cambridge Engineering Department, CB2 1PZ

Corresponding author: James M. Cozens (email: jmc257@cam.ac.uk).

JMC is funded by the EPSRC Doctoral Training Partnership (DTP)

ABSTRACT This paper proposes a probabilistic approach for extracting time-varying and irregular time signature information from polyphonic audio extracts, subsequently providing beat and bar line positions given inferred time signature divisions. This is achieved via dynamically evaluating the beat tempo as a function of time through finding an optimal compromise in beat and bar alignment in the time and tempo domains. Time signature divisions are determined based on a new representation, termed the Metrogram, that presents time-varying information regarding rhythmic and metric periodicities in the Tempogram. Our methodology is characterised by its ability to provide a distribution over metric interpretations, offering insights into the diverse ways music can be rhythmically perceived. Results indicate high-level accuracy for a variety of polyphonic extracts containing irregular, complex, irrational, and time-varying time signatures. Accuracy rivalling state-of-the-art methodologies is also reported in a beat tracking task performed on the standard Ballroom Dataset. The paper offers insights into the field of dynamic time signature recognition and beat tracking, offering a valuable and versatile resource for the analysis, composition, and performance of music.

INDEX TERMS Audio signal processing, Beat tracking, Dynamic time signature recognition, Metrogram transform, Music analysis, Polyphonic extracts, Rhythmic periodicity, Tempo inference, Time-varying time signatures, Transcription.

I. Introduction

The process of transcribing complex polyphonic performances to musical notation is a notoriously arduous task, requiring the ability to distinguish numerous instrumental lines, separable often only via remarkably subtle variances in timbre, frequency, and waveform characteristics. Automated music transcription is a field of great significance in the music and educational industry, especially for primarily improvised genres, such as jazz. In particular, one of the most significant and challenging tasks is that of time signature inference and beat tracking, especially in metrically ambiguous extracts, a common occurrence across genres [1]–[3]. Inaccurate beat tracking for performances with rubato (time-varying tempo) or multiple metric interpretations, regardless of the accuracy of note detections, will result in poor quality transcriptions given the misalignment of key rhythmic structures in the transcription [4], [5].

Several competing methods exist in the literature for tempo and time signature estimation from audio, primarily through the employment of Tempograms [21], [22] and similarity matrices [17], [20], respectively. Various probabilistic methods are employed to facilitate meter detection [19], [25], with Hidden Markov Models (HMMs) paired with bar pointer models, in particular, proving successful for beat and downbeat estimation tasks for genre-specific applications [6], [8]–[10], [12], [13], [15]. The advent of Deep Learning (DL) approaches has presented numerous genre-specific methods for beat and downbeat tracking, primarily through the employment of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Temporal Convolutional Networks (TCNs) [23], [24], [34]–[38]. HMM bar pointer models have also been successfully applied to irregular ("odd") time signature beat tracking tasks [6], [8], [13], [14], however, the approaches assume a constant metric division (often a priori). Likewise, for methods that tackle time-varying time signatures using HMMs [10], [16] or RNNS [24], [39], the models employ cascade formats, such that joint beat and bar tracking is performed sequentially

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

resulting in beat lengths independent of metric properties. The subjectivity of perceived pulse (beat) positions and lengths is addressed in literature through the employment of agents [5], [31], yet the process has not been generalised to joint time signature and beat tracking tasks. In general, the majority of algorithms assume either a constant tempo, beat length, or time signature division over time, providing methods for extracting each independently, and thus poor performance is reported for extracts with rubato and irregular beat and time signature changes [25]. Likewise, although high-level accuracy is reported for DL and HMM approaches for specific genres, poor performance is often reported on unseen genres [42] given that fundamentally, the models are constrained by the training data and genre-specific rhythmic and metric properties.

This paper instead proposes transforms in conjunction with novel probabilistic tracking algorithms, aimed specifically at extracting time-varying rhythmic and metric features from non-genre specific polyphonic extracts, such as providing metric interpretations, detecting time signature changes, rubato, and irregular beat length, which is often present in jazz and other improvisatory genres. As such, the paper evaluates the model on a custom dataset featuring a variety of metrically diverse extracts, with varying time signatures, metric modulations, and rubato, as well as on the Ballroom Dataset, in order to verify its capability in fixed time signature beat tracking against the state of the art.

II. Methodology

A. Overview

This paper proposes a posterior distribution that is maximised in order to optimally fit the model to the audio extract, with respect to a time series of tempo values and a phase offset, which together determine the relative positions of the bar and beat (tatum) times in the extract. The system iteratively samples the hyperparameter space of the model, optimising the posterior conditioned on the sampled hyperparameters with respect to the tempo values and phase offset until satisfactory convergence is achieved for each sampling iteration. The optimised posterior probabilities for the sampled hyperparameter configurations are compared and iterated until an appropriate global solution is found. The posterior distribution is constructed based upon a Note Onset Detection Function (NODF), a 2D Morlet Convolver, and a Fundamental Tempogram, which is used finally to generate the Metrogram, as described in the following sections. The analysis is presented in continuous time for simplicity, although of course, the practical implementation involves a time and frequency discretisation.

B. Note Onset Detection Function

As input to the Tempograms, a note onset detection function (NODF) is required. This paper proposes a two-dimensional NODF that predominantly exploits frequency domain information in order to distinguish more complex instrumental lines in polyphonic and polytimbral environments, which may have otherwise been hidden in the analytic envelope of the time domain.

The proposed NODF is based on the smoothed time-derivative of a spectrogram-like representation, using variable resolution wavelet basis functions in the analysis step:

$$f\left(\omega,\, t\right) = \left|x\left(t\right) \circledast W_{\omega}\left(t\right)\right| = \left|\int_{-\infty}^{\infty} x\left(\tau\right) W_{\omega}\left(t-\tau\right)\, d\tau\right|, \quad (1)$$

where $f\left(\omega,\, t\right)$ is the magnitude of the function evaluated at frequency ω, and time t. $x(t)$ is the continuous time audio signal (mono), and $W_{\omega}\left(t\right)$ is a wavelet function for frequency ω. Here we employ the Morlet wavelet, defined as a complex exponential with a Gaussian envelope [44]:

$$W_{\omega}\left(t\right) = \frac{1}{\sigma_{\omega}} \exp\left(-\frac{1}{2\sigma_{\omega}^{2}} t^{2}\right) \exp\left(i\omega t\right). \quad (2)$$

A wavelet-based approach is employed in our analysis of polyphonic music due to its ability to finely tune time-frequency resolutions at a semitone level, for example. A smoothed derivative approximation of $f\left(\omega,\, t\right)$ is then obtained by convolving, with respect to time, the first derivative of a Gaussian kernel, $g'_{\sigma_d}\left(t\right)$, with standard deviation σ_d. Subsequently, the output is summed over N harmonics of the frequency ω; the resulting value is limited to be above zero so that it is suited for note onset detection, rather than note release. As such, the method is analogous to the Harmonic Product Spectrum (HPS) [11], [18] used in the detection of fundamentals, however, our approach employs summations instead of multiplications of harmonics:

$$D\left(\omega,\, t\right) = \max\left\{0,\, \sum_{n=1}^{N} f\left(n\omega,\, t\right) \circledast g'_{\sigma_d}\left(t\right) \sqrt{\sigma_{\omega}^{2}+\sigma_{d}^{2}}\right\}, \quad (3)$$

where n is the harmonic index ($1 \leq n \leq N$). $\sqrt{\sigma_{\omega}^{2}+\sigma_{d}^{2}}$ is employed as a normalisation term. The inner function, $f\left(\omega,\, t\right) \circledast g'_{\sigma_d}\left(t\right)$, can be expanded as:

$$\begin{aligned} &= \int_{-\infty}^{\infty} \left|\int_{-\infty}^{\infty} x\left(t_1-\tau\right) \exp\left(-\frac{1}{2\sigma_{\omega}^{2}}\tau\left(\tau - 2i\omega\sigma_{\omega}^{2}\right)\right) d\tau\right| \\ &\quad \times \frac{\left(t_1-t\right)}{\sigma_{\omega}\sigma_{d}^{3}} \exp\left(-\frac{1}{2\sigma_{d}^{2}}\left(t-t_1\right)^{2}\right) dt_1. \end{aligned} \quad (4)$$

$D\left(\omega,\, t\right)$ is now evaluated for each of the 88 semitones $\omega\left(s\right) = A_0 2^{\frac{s-1}{12}}$, where $A_0 = 2\pi \times 27.5$, to yield the NODF. The parameter σ_{ω}, which determines the time-frequency resolution of the analysis, is chosen to respect the logarithmic spacing of the musical pitches, such that the frequency width is one-sixth of the distance between adjacent semitones. A factor of 6 is chosen as a reasonable compromise between frequency resolution and the potential inclusion of non-equally-tempered tones. The frequency resolution resulting from the wavelet analysis step is obtained from the Fourier Transform of the proposed wavelet:

$$\bar{W}_{\omega}\left(\phi\right) = \mathcal{F}\left\{W_{\omega}\left(t\right)\right\} = \sqrt{2\pi} \exp\left(-\frac{\left(\phi-\omega\right)^{2}\sigma_{\omega}^{2}}{2}\right). \quad (5)$$

which is of course a Gaussian in the frequency domain with a standard deviation equal to the inverse of the wavelet parameter σ_{ω}. Thus, σ_{ω} can be expressed as:

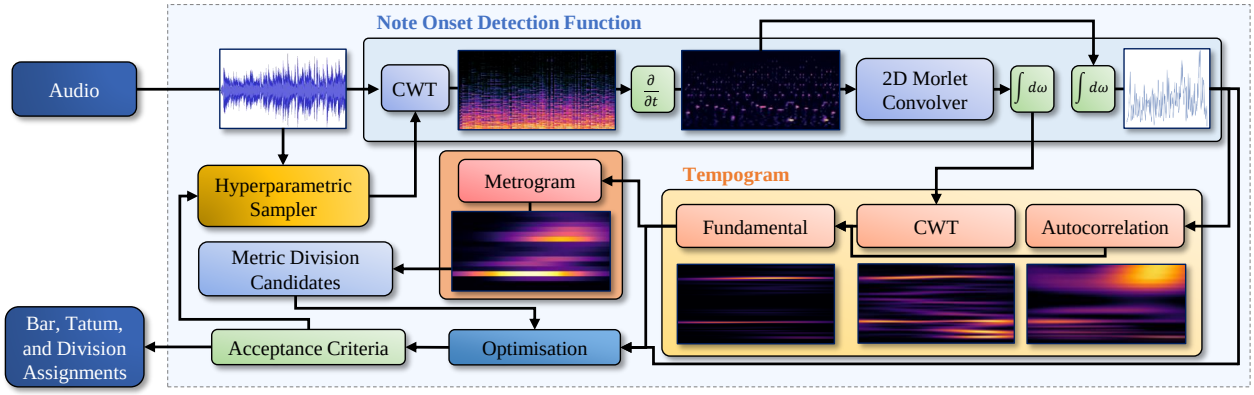


FIGURE 1: Simplified overview of the proposed architecture

$$\sigma_\omega(s) = \frac{12}{(\omega(s+1) - \omega(s-1))} = \frac{12}{A_0}\left(2^{\frac{s}{12}} - 2^{\frac{s-2}{12}}\right)^{-1}. \quad (6)$$

A one-dimensional NODF (required in Eq. (10)), is then obtained as a weighted sum of the two-dimensional form (Eq. (3)) across s, corresponding to the 88 semitones:

$$D(t) = \frac{1}{\sqrt{2\pi}} \sum_{s=1}^{88} D(\omega(s), t)\, \bar{V}(s) \sqrt{\sigma_\omega(s)^2 + \sigma_d^2}, \quad (7)$$

where $\bar{V}(s)$ is a (normalised) weighting function with respect to the semitone number, s. The following expression is proposed that prioritises lower frequencies with strength q, which is typically desirable given the strong dependence of beat on the bass notes:

$$\bar{V}(s) = \left(\frac{88-s}{88}\right)^q \cdot \left[\int_0^{88} \left(\frac{88-s}{88}\right)^q ds\right]^{-1} = \left(\frac{1+q}{88}\right) \cdot \left(\frac{88-s}{88}\right)^q.$$

C. Fundamental Tempogram

An alternative to the conventional Tempogram, a *Fundamental Tempogram* (so termed because of its ability to attenuate harmonics and sub-harmonics of the true tempo), is now proposed in order to extract metric information from the NODF. As a starting point, this could be attempted by performing further wavelet analysis of the one-dimensional NODF $D(t)$ (Eq. (7)), with frequency centred upon candidate beats-per-minute (bpm) ω_{bpm}:

$$R_{CWT}(\omega_{bpm}, t) = \left| D(t) \circledast W_{\omega_{bpm}}(t) \right|^2. \quad (8)$$

The wavelet function, $W_{\omega_{bpm}}(t)$, is defined in terms of the bpm (ω_{bpm}) and equivalent receptive field (standard deviation) $\sigma_{\omega_{bpm}}$, as follows:

$$W_{\omega_{bpm}}(t) = \frac{1}{\sigma_{\omega_{bpm}}} \exp\left(-\frac{1}{2\sigma_{\omega_{bpm}}^2} t^2\right) \exp\left(\frac{i\pi\omega_{bpm}}{60} t\right). \quad (9)$$

However, given that information regarding note positioning and spacing in the frequency domain is lost when evaluating $D(t)$ from $D(\omega(s), t)$, the following alternative expression employing the two-dimensional NODF, Eq. (3), is proposed:

$$R_{CWT}(\omega_{bpm}, t) = \int_0^\infty \left| D(\omega(s), t) \circledast W_{2D}(s, t) \right|^2 ds = \int_0^\infty \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} D(\omega(j), \tau)\, W_{2D}(s-j, t-\tau)\, d\tau\, dj \right|^2 ds,$$

where $W_{2D}(s, t)$ is a 2D Morlet wavelet defined as:

$$W_{2D}(s, t) = \frac{1}{\sigma_{\omega_s}\sigma_{\omega_{bpm}}} \exp\left(-\frac{1}{2}\left[\left(\frac{s}{\sigma_{\omega_s}}\right)^2 + \left(\frac{t}{\sigma_{\omega_{bpm}}}\right)^2\right] + \frac{i\pi\omega_{bpm}}{60} t\right).$$

Parameter σ_{ω_s} is the standard deviation of the wavelet in component s. The resulting function is equivalent to a standard Morlet function in the time domain, weighted by a Gaussian in the (semitone) frequency domain.

The purpose of increasing the dimensionality of the wavelet whilst evaluating the CWT Tempogram is to capture correlations between note positions at neighbouring tatum and bar lines. The s component of the wavelet effectively weights neighbouring note positions by a Gaussian with standard deviation σ_{ω_s} semitones, such that closer note values at neighbouring tatum and bar lines produce greater responses in the final CWT Tempogram; typically, $\sigma_{\omega_s} = 12$ is employed. Thus, periodicities in certain frequency ranges are extracted with greater accuracy. The relationship between the three-dimensional space generated by this process and the final CWT Tempogram can be observed in the animation available here.

As a further aid to determining beat tempo, an Autocorrelation Tempogram is also constructed. This takes the one-dimensional NODF $D(t)$ (Eq. (7)) as its input, autocorrelates it, and then smooths the result through convolution with a Gaussian function having the same standard deviation as the CWT Tempogram ($\sigma_{\omega_{bpm}}$):

$$R_{AC}(\omega_{bpm}, t) = D(t)\, D\left(t + \frac{60}{\omega_{bpm}}\right) \circledast \exp\left(-\frac{1}{2\sigma_{\omega_{bpm}}^2} t^2\right) = \int_{-\infty}^{\infty} D(\tau)\, D\left(\tau + \frac{60}{\omega_{bpm}}\right) \exp\left(-\frac{1}{2\sigma_{\omega_{bpm}}^2}(\tau - t)^2\right) d\tau. \quad (10)$$

As a consequence of their construction, R_{CWT} has harmonics associated with the rhythmic components, whilst

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2023.3344048

R_{AC} has sub-harmonics. Harmonic and tempo ambiguity in Tempograms have been encountered previously in the literature, with certain methods proposed, such as exploiting the combined properties of the Autocorrelation (R_{AC} and Fourier Tempograms (R_{DFT}) through multiplication [16], [27], [28] and performing octave removal [26]. However, as a consequence of the methods employed (such as DFT as opposed to CWT) to generate R_{AC} and R_{CWT} in [16], [27], [28], and the limitations of purely targeting octave removal in [26], fundamental "rhythmic" frequencies remain largely inseparable from their harmonics. This paper by contrast proposes combining the properties of the presented CWT and Autocorrelation Tempograms, employing the same receptive fields $\sigma_{\omega_{bpm}}$ in each case, via computing the geometric mean of the two arrays such that harmonics and sub-harmonics are attenuated and the resulting Tempogram normalised. Employing the notation $\lceil x \rceil^+ = \max\{0, x\}$, the final proposed Fundamental Tempogram is obtained as:

$$R_F\left(\omega_{bpm},\, t\right) = \sqrt{R_{CWT}\left(\omega_{bpm},\, t\right) \lceil R_{AC}\left(\omega_{bpm},\, t\right) \rceil^+}. \quad (11)$$

D. Metrogram Transform for Time Signature Recognition

As observed in the Tempogram plots (such as Figure 2), tempo trajectories present in the Tempogram provide insight into the type of time signature divisions represented in the music, specifically the ratios between the tempo lines. A few papers explore the concept of ratios in the Tempogram [7], [27], [29], in order to facilitate meter tracking and genre classification. However, in these papers, ratios are evaluated causally with respect to a specified tempo and for a limited discrete subset of metric divisions, and in [29], the dependency on time is lost given this fixed-tempo assumption. Here however we propose a transform that exploits the Tempogram properties to extract rhythmic ratios present in the music, independent of tempo information, thus enabling the time-varying metric characteristics to be evaluated in continuous form. To achieve this, the proposed transform, named the Metrogram, involves evaluating the multiplicative equivalent of the autocorrelation function in the frequency domain. The proposed Metrogram can be expressed as:

$$P\left(k,\, t\right) = \frac{1}{Z\left(k\right)} \int_{0^+}^{\infty} R_F\left(\omega_{bpm},\, t\right) R_F\left(k\omega_{bpm},\, t\right)\, d\omega_{bpm}, \quad (12)$$

where k is the rhythmic ratio to be evaluated, and $Z\left(k\right)$ is a constant factor with respect to k; typically $Z\left(k\right) \propto k^{-(p+1)}$ is suitable ($0.5 < p < 1.5$), given that higher time signature divisions are inherently more sensitive to rhythmic ambiguity, and therefore larger Metrogram ratios should be weighted accordingly. Note that the evaluation of $P\left(k,\, t\right)$ is not limited by integer values of k, and can thus be employed to detect polyrhythms in music, for example. To determine the primary time signature division as a function of time given the Metrogram, $\hat{k}\left(t\right) = \max_k\{P\left(k,\, t\right)\}$ can be employed.

The Fundamental Tempogram's equivalent receptive field is noteworthy; both the Autocorrelation and CWT Tempograms employ Gaussian window functions with a standard deviation of $\sigma_{\omega_{bpm}}$, defining this receptive field. A small receptive field might miss correlations between adjacent bar lines, whilst a sufficiently large $\sigma_{\omega_{bpm}}$ might not capture localised metric variation. A suitable range, influenced by the musical genre, has been found experimentally to lie between $1.5s < \sigma_{\omega_{bpm}} < 5s$, a result in line with previous works [13]. Figure 2 illustrates the receptive field's effect on the Fundamental Tempogram and the subsequent influence on the integer-valued Metrogram division assignments.

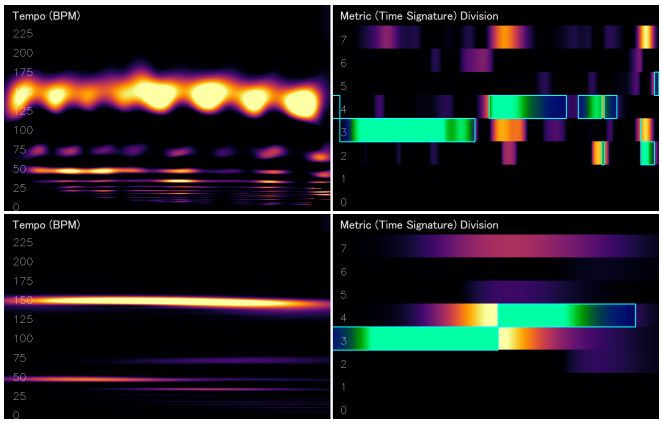


FIGURE 2: Fundamental Tempograms (left) with the corresponding Metrograms (right) for two receptive fields for a recording modulating from 3/4 to 4/4. The threshold division assignments over time are illustrated in blue in the Metrograms. Top: $\sigma_{\omega_{bpm}} = 0.5s$. Bottom: $\sigma_{\omega_{bpm}} = 3.0s$. X-axis = time. An animation showing the gradual transition between the extremes of the receptive field can be seen here

E. Tempo Inference and Beat Tracking

Having defined the various input components to the tempo and beat tracking functions, we now describe time-varying tempo inference through simultaneous optimisation with respect to 6 probabilistic objective functions. Specifically, the various inputs are discretised into arrays, denoted by the NODF, $D_{1:N}$, the Fundamental Tempogram, $R_{1:N}(\omega_{bpm})$, and the bar-aligned metric divisions extracted from the Metrogram, $\hat{k}_{1:N}$, forming the input dataset $x_{1:N} = \{D_{1:N},\, \hat{k}_{1:N},\, R_{1:N}\left(\omega\right)\}$. In [6], [8], [13], [14] Markov models were applied to meter and rhythm. Here a posterior distribution is proposed directly in terms of the bpm over time, $\lambda_{1:N}$, and the phase offset within a bar, ϕ:

$$p(\lambda_{1:N},\, \phi \,|\, x_{1:N}) \propto p(x_{1:N} \,|\, \lambda_{1:N},\, \phi)\, p\left(\phi,\, \lambda_{1:N}\right), \quad (13)$$

where:

$$p(x_{1:N} \,|\, \lambda_{1:N},\, \phi) \propto \prod_{n=1}^{N} p(x_n \,|\, \lambda_{1:N},\, \phi). \quad (14)$$

This paper proposes a likelihood that is constructed to account for all metric phenomena previously described, with no dependency on training data:

$$p(x_n \,|\, \lambda_{1:N},\, \phi) \propto \prod_{j=1}^{6} \mu_j, \quad (15)$$

where j, $1 \leq j \leq 6$, are the indices associated with the following proposed probabilistic objective functions:

$$\mu_1 = \sigma\left(C_0 R_n\left(\lambda_n\right) - b_0\right) \tag{16}$$

$$\mu_2 = \sigma\left(C_1 R_n\left(\lambda_n\right) R_n\left(\lfloor \lambda_n / \hat{k}_n \rceil\right) - b_1\right) \tag{17}$$

$$\mu_3 = \exp\left(-\frac{1}{2\sigma_{bar}^2}\left[D_n - \cos^{2r_1}\left(\frac{\phi}{\hat{k}_0} + \frac{\pi\Delta t}{60}\sum_{m=1}^{n}\frac{\lambda_m}{\hat{k}_m}\right)\right]^2\right) \tag{18}$$

$$\mu_4 = \exp\left(-\frac{1}{2\sigma_{tatum}^2}\left[D_n - \cos^{2r_2}\left(\phi + \frac{\pi\Delta t}{60}\sum_{m=1}^{n}\lambda_m\right)\right]^2\right) \tag{19}$$

$$\mu_5 = \sigma\left(C_2 D_n \cos^{2r_1}\left(\frac{\phi}{\hat{k}_0} + \frac{\pi\Delta t}{60}\sum_{m=1}^{n}\frac{\lambda_m}{\hat{k}_m}\right) - b_2\right) \tag{20}$$

$$\mu_6 = \sigma\left(C_3 D_n \cos^{2r_2}\left(\phi + \frac{\pi\Delta t}{60}\sum_{m=1}^{n}\lambda_m\right) - b_3\right). \tag{21}$$

Equation (16) encourages the maximisation of the fit with respect to the Fundamental Tempogram, and Eq. (17) maximises the fit of the tatum and division tempo trajectories with respect to the Tempogram array. Likewise, Eq. (18) and (20) maximise the fit of the bar lines with respect to the NODF, and Eq. (19) and (21) maximise the fit of the tatum lines. C_0, C_1, C_2, C_3 are scaling constants, and b_0, b_1, b_2, b_3 are biases. The function $\cos^{2r}\theta$ is employed to enforce convergence on the possible bar and tatum line alignments, with hyperparameters r_1, r_2 utilised such that alignments are encouraged with strength r with respect to the time domain NODF rendered previously. Note, previous methods that model time-varying time signature divisions employ cascade formats [10], [24], [39], such that joint beat and bar tracking is performed sequentially, as opposed to our proposed method that simultaneously optimises with respect to beat and bar allocations, ensuring beat length variations are appropriately modelled. The prior is then specified as:

$$p\left(\phi,\, \lambda_{1:N}\right) = p(\phi)\, p(\lambda_{1:N}) \tag{22}$$

$$\propto p(\lambda_N \,|\, \lambda_{1:N-1})\, p(\lambda_{N-1} \,|\, \lambda_{1:N-2}), \ldots, p(\lambda_2 \,|\, \lambda_1)\, p(\lambda_1) \tag{23}$$

$$= p(\lambda_1) \prod_{n=1}^{N-1} p(\lambda_{n+1} \,|\, \lambda_n) \propto \prod_{n=1}^{N-1} p(\lambda_{n+1} \,|\, \lambda_n). \tag{24}$$

Eq. (23) and (24) are possible given $p(\phi)$ and $p(\lambda_1)$ are assumed to be uniform in the ranges $0 \leq \phi \leq \max_n\{\hat{k}_n\}\pi$, $0 \leq \lambda_1 \leq \omega_{bpm,max}$, and due to the Markovian nature of the proposed prior on $\lambda_{1:N}$:

$$p\left(\lambda_{n+1} \,|\, \lambda_n\right) \propto \exp\left[-\frac{1}{2\left(\sigma_{\lambda_d}\Delta t\right)^2}\left(\lambda_{n+1} - \lambda_n\right)^2\right] \{\lambda_n \geq 0\}. \tag{25}$$

Thus, taking the partial derivative of the negative log of the posterior ($\mathcal{L}\left(\lambda_{1:N},\, \phi\right) := -\log p(\lambda_{1:N},\, \phi \,|\, x_{1:N})$) with respect to λ_n and ϕ results in Eq. (26) and (27) respectively.

The model parameters can be initialised through maximisation over the Metrogram inner terms in Eq. (12) given the determined $\hat{k}_n$ values (multiplied by $\hat{k}_n$ for the beat tempo):

$$\lambda_n = \hat{k}_n \max_{\omega_{bpm}}\left\{R_n\left(\omega_{bpm}\right) R_n\left(\hat{k}_n \omega_{bpm}\right)\right\}. \tag{28}$$

Other models typically employ the Viterbi algorithm for optimisation [6], [8]–[10], [12], [13] or Monte Carlo (MC) techniques [15]; for the structure of our continuous-parameter model an iterative Gradient Descent method [30] is appropriate, as follows:

$$\begin{bmatrix}\boldsymbol{\lambda}\\ \phi\end{bmatrix}_{e+1} = \begin{bmatrix}\boldsymbol{\lambda}\\ \phi\end{bmatrix}_{e} - \eta\begin{bmatrix}\frac{\partial}{\partial\boldsymbol{\lambda}}\\ \frac{\partial}{\partial\phi}\end{bmatrix}\mathcal{L}\left(\boldsymbol{\lambda}\right)\big|_e, \tag{29}$$

where e is the training epoch, and $\boldsymbol{\lambda} = \{\lambda_n\}_{n=1}^{N}$. In order to avoid becoming trapped in local minima during convergence, a stochastic step is added. For every epoch, with probability p_s ($p_s \approx 0.2$), the algorithm evaluates the posterior probability at a grid of ϕ values, $\phi = \phi_e + n_s\pi$, for $-\lceil \hat{k}_{max}/2 \rceil \leq n_s \leq \lceil \hat{k}_{max}/2 \rceil$ and integer n_s. $\hat{k}_{max}$ is the maximum division extracted from the whole extract and ϕ_e corresponds to ϕ computed during the current epoch. The algorithm then updates ϕ according to the maximum posterior probability (MAP estimate) across the specified range of n_s. This specific range of ϕ is chosen given that each beat corresponds to π phase, thus, this step effectively samples the neighbouring beats to ensure the correct beat-bar alignment position has been chosen.

Upon satisfactory convergence of the parameters $\boldsymbol{\lambda}$ and ϕ, the bar (right) and tatum (left) times in terms of the sampling index n can be determined by solving the following equations for n (with appropriate linear interpolation):

$$mod\left(\frac{\Delta t}{60}\sum_{n=1}^{N}\lambda_n + \frac{\phi}{\pi},\, 1\right) = 0, mod\left(\frac{\Delta t}{60}\sum_{n=1}^{N}\frac{\lambda_n}{\hat{k}_n} + \frac{\phi}{\pi\hat{k}_0},\, 1\right) = 0.$$

F. Monte Carlo Hyperparameter Sampling

Given the inherent subjectivity in time signature recognition [1], the proposed posterior exhibits multiple local maxima dependent on hyperparameter selections. This arises primarily due to polyrhythmic and polymetric elements, prevalent in genres like jazz [3]. For instance, a waltz in $3/4$ could be perceived as $6/8$, $12/8$, or $2/2$ due to a multi-layered rhythmic hierarchy (as observed in Figure 4). Hyperparameters include q, the weight in the 2D NODF, $\sigma_{\omega_{bpm}}$, the Tempogram receptive field, and σ_{ω_s}, the 2D wavelet standard deviation. To determine the most suitable maxima, a global posterior maximum with respect to the hyperparameters can be identified (MAP estimate) by employing Monte Carlo (MC) sampling of the model's hyperparameter space for J iterations. Subsequently, for each sampling iteration, the posterior is evaluated through gradient descent of the negative log posterior (as per Eq. (29)) conditioned on these hyperparameters. The extended posterior is thus given by:

$$\begin{aligned} &p(\lambda_{1:N},\, \phi,\, \sigma_{\omega_s},\, \sigma_{\omega_{bpm}},\, q \,|\, x_{1:N}) \\ &\propto p(\lambda_{1:N},\, \phi) p(x_{1:N} \,|\, \sigma_{\omega_s},\, \sigma_{\omega_{bpm}},\, q,\, \lambda_{1:N},\, \phi) \\ &\quad \times p(\sigma_{\omega_s}) p(\sigma_{\omega_{bpm}}) p(q). \end{aligned} \tag{30}$$

$$
\begin{aligned}
\frac{\partial}{\partial \lambda_n}\mathcal{L}\left(\lambda_{1:N},\,\phi\right) = &-\frac{C_0\frac{\partial}{\partial\lambda}R_n\left(\lambda_n\right)\exp\left[-C_0R_n\left(\lambda_n\right)+b_0\right]}{1+\exp\left[-C_0R_n\left(\lambda_n\right)+b_0\right]}\\
&-\frac{C_1\left[R_n\left(\lfloor\lambda_n/\hat{k}_n\rceil\right)\frac{\partial}{\partial\lambda}R_n\left(\lambda_n\right)+\frac{1}{\hat{k}_n}R_n\left(\lambda_n\right)\frac{\partial}{\partial\lambda}R_n\left(\lfloor\lambda_n/\hat{k}_n\rceil\right)\right]}{1+\exp\left[-C_1R_n\left(\lambda_n\right)R_n\left(\lfloor\lambda_n/\hat{k}_n\rceil\right)+b_1\right]}\\
&\times\exp\left[-C_1R_n\left(\lambda_n\right)R_F\left(\lfloor\lambda_n/\hat{k}_n\rceil,\,n\right)+b_1\right]\\
&+\frac{1}{\left(\sigma_{\lambda_d}\Delta t\right)^2}\left(2\lambda_n-\lambda_{n+1}-\lambda_{n-1}\right)\\
&+\frac{2}{\sigma_{bar}^2}\sum_{m=n}^{N}\left(D_m-\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)\right)r_1\cos^{2r_1-1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)\sin\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)\frac{\pi\Delta t}{60\hat{k}_n}\\
&+\frac{2}{\sigma_{beat}^2}\sum_{m=n}^{N}\left(D_m-\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)\right)r_2\cos^{2r_2-1}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)\sin\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)\frac{\pi\Delta t}{60}\\
&+\sum_{m=n}^{N}\frac{2C_2D_mr_1\cos^{2r_1-1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)\sin\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)\frac{\pi\Delta t}{60\hat{k}_n}}{1+\exp\left[-C_2D_m\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)+b_2\right]}\\
&\times\exp\left[-C_2D_m\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\frac{\lambda_j}{\hat{k}_j}\right)+b_2\right]\\
&+\sum_{m=n}^{N}\frac{2C_3D_mr_2\cos^{2r_2-1}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)\sin\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)\frac{\pi\Delta t}{60}}{1+\exp\left[-C_3D_m\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)+b_3\right]}\\
&\times\exp\left[-C_3D_m\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{m}\lambda_j\right)+b_3\right]
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \phi}\mathcal{L}\left(\lambda_{1:N},\,\phi\right) = &\frac{2}{\hat{k}_0\sigma_{bar}^2}\sum_{n=1}^{N}\left(D_n-\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)\right)r_1\cos^{2r_1-1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)\sin\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)\\
&+\frac{2}{\sigma_{beat}^2}\sum_{n=1}^{N}\left(D_n-\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)\right)r_2\cos^{2r_2-1}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)\sin\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)\\
&+\frac{1}{\hat{k}_0}\sum_{n=1}^{N}\frac{2C_2D_nr_1\cos^{2r_1-1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)\sin\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)}{1+\exp\left[-C_2D_n\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)+b_2\right]}\\
&\times\exp\left[-C_2D_n\cos^{2r_1}\left(\frac{\phi}{\hat{k}_0}+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\frac{\lambda_j}{\hat{k}_j}\right)+b_2\right]\\
&+\sum_{n=1}^{N}\frac{2C_3D_nr_2\cos^{2r_2-1}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)\sin\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)}{1+\exp\left[-C_3D_n\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)+b_3\right]}\\
&\times\exp\left[-C_3D_n\cos^{2r_2}\left(\phi+\frac{\pi\Delta t}{60}\sum_{j=1}^{n}\lambda_j\right)+b_3\right]
\end{aligned}
\tag{27}
$$

Each hyperparameter σ_{ω_s}, $\sigma_{\omega_{bpm}}$ and q is assigned a separate gamma distribution as its prior, $p(\theta|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$. Subsequently, the allocation results obtained from the sampling iteration that achieved the greatest posterior value are taken. The corresponding sampling iteration index, i, can thus be expressed as:

$$
i=\max_j\left\{p\left(\lambda_{1:N},\,\phi,\,\sigma_{\omega_s}^{(j)},\,\sigma_{\omega_{bpm}}^{(j)},\,q^{(j)}\,|\,x_{1:N}\right)\right\},\tag{31}
$$

where $\sigma_{\omega_s}^{(j)}$, $\sigma_{\omega_{bpm}}^{(j)}$, $q^{(j)}$ are the hyperparameters sampled from the priors during the jth sampling iteration; note that the posteriors here are computed with respect to the converged $\lambda_{1:N}$, ϕ values resulting from the optimisation step. A notable benefit of this proposed approach is the ability to provide the user with a variety of metric interpretations, corresponding to each sampling iteration, alongside the single highest probability assignment.

III. Results and Discussion

In order to evaluate the performance of our approach on extracts with both beat and time-signature variations, a metrically diverse custom dataset is presented. Although several other datasets have been presented with either irregular or time-varying time signatures [6], [8], [10], [13], [14], [16], [32], none are available for extracts that exhibit both simultaneously with rubato, as is the case in improvisatory

genres such as jazz. This custom dataset was recorded by the first author and features 43 extracts (primarily jazz piano) with primary time signature divisions of 2, 3, 4, 5, and 7 (35 extracts), as well as several metrically ambiguous extracts with dynamic (time-varying) time signature divisions and tempo (rubato). A custom-designed algorithm was developed for the labelling of ground truth assignments based upon quarter-speed playback and expert domain knowledge provided by the first author.

The following widely-used evaluation techniques [33] are employed: The F-measure, Cemgil score, P-score, and the Mean Squared Error (MSE). As input to the optimisation step, the terms $x_n = \{D_n, \hat{k}_n, R_n(\omega)\}$ are computed once every 1500 data samples at 44.1kHz. This compression factor, found through experimentation, provides a suitable compromise in temporal accuracy and computational efficiency, a result consistent with previous findings [6], [31]. The results for the 35 primary division extracts are shown in Table 1. One example extract's beat assignment results are plotted in Figure 3. Likewise, another example extract, Figure 5, is presented with its Fundamental Tempogram, Metrogram (with the division allocations highlighted in blue), and the final beat (grey vertical lines) and Bar (bright orange vertical lines) allocations superimposed over the CWT NODF intensities and audio waveform. Figure 4 additionally presents three metric interpretations provided by the algorithm for a metrically ambiguous orchestral extract. As depicted in Table 1, the algorithm displays high-level accuracy for both tatum and bar line assignments, with a tatum F-measure mean of 0.967 attesting to its strong performance in handling rubato, syncopation, and rhythmic irregularities. Every time signature division in the dataset was accurately inferred. Evaluation metrics largely remain consistent across the different time signature divisions (see Table 1). For the case of more subjective metric assignments, the algorithm was able to infer compatible interpretations for each, as observed in the example extracts presented in Figures 5 and 4.

Our model is designed to account for localised metric and rhythmic variations, such as time-varying time signature divisions and beat tempo, providing a set of probability-ranked metric interpretations, and is hence more flexible than the majority of other approaches in the literature. Nevertheless, comparisons with state-of-the-art approaches are considered for fixed-time signature recognition to specifically evaluate its fixed-division beat tracking capabilities. For this purpose, evaluation is performed on the Ballroom dataset (BDS) [43]; our algorithm's performance is shown in Table 2. Its overall performance is then compared with a number of state-of-the-art algorithms in Table 3. Overall, our results are competitive with the leading methods for fixed time-signature methods, surpassed only by some of the deep learning methods that employ 8-fold cross-validation. For context, 8-fold cross-evaluation involves training on a fraction (k-1)/k of the dataset and testing on the remaining 1/k. As such, while these models may exhibit high F-measures for specific datasets, there is potential for overfitting to particular genre-specific metric and rhythmic features, as evidenced when tested on unfamiliar datasets [42] (the BDS is primarily drum-tracked with minimal tempo variation). A distinctive feature of our approach is its zero-training necessity and versatility to general polyphonic audio extracts, facilitating broader applications in various rhythmic and metric domains, as exhibited by the performance in the custom dataset. Accordingly, an interesting consequence of the proposed method is that lower BDS performance figures from our method are predominantly attributed to the model providing localised and global alternate metric interpretations (such as hemiolas or polymetric features), especially in the Waltz category, a possibility that is not entertained by either the fixed or dynamic time-signature methods. Indeed, this is a challenge faced previously in the literature, with numerous studies finding the misalignment of objective beat tracking scores with subjective participant scores due to varying metric interpretations [1]–[4].

IV. Conclusion

This paper presents a probabilistic approach for the extraction of time-varying and irregular time signatures from polyphonic audio extracts, whilst also providing beat tracking estimates according to the inferred metric properties. Central to this approach is the Metrogram, a novel representation that captures time-varying information on rhythmic and metric periodicities within the Tempogram. A unique feature of our methodology is its ability to provide a distribution over metric interpretations through hyperparameter sampling of the posterior, offering insights into the diverse ways music can be rhythmically perceived. This aspect is particularly crucial for handling complex, irrational, and fluctuating time signatures commonly found in polyphonic extracts, especially in rhythmically and metrically diverse genres such as jazz. To demonstrate this unique feature, this paper presents a dataset consisting of irregular, irrational, and time-varying time signatures, with overall high level accuracy reported. This level of accuracy is attributable to the unique probabilistic approach to dynamically evaluating the optimal balance in beat and bar alignment across both time and tempo domains. Likewise, empirical evaluations of the algorithm's fixed time signature beat tracking capabilities are presented for the standard Ballroom Dataset, with accuracy rivalling state-of-the-art methodologies.

Looking ahead, the potential applications of this algorithm are extensive; its adaptability and capacity to interpret complex rhythmic and polymetric structures open new avenues for music composition, performance, and analysis. Ultimately, our algorithm balances adaptability and accuracy, demonstrating capability in handling complex rhythmic and metric structures and offering versatility across diverse polyphonic audio environments, paving the way for a deeper understanding and appreciation of the complexities inherent in music.

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2023.3344048

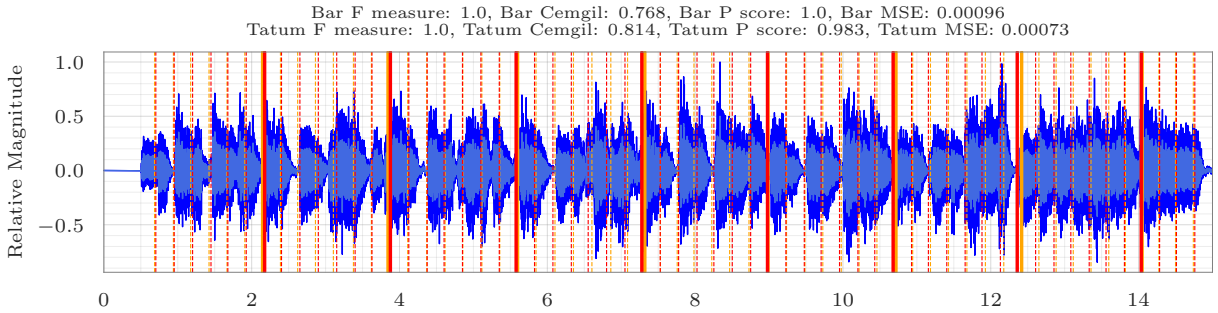
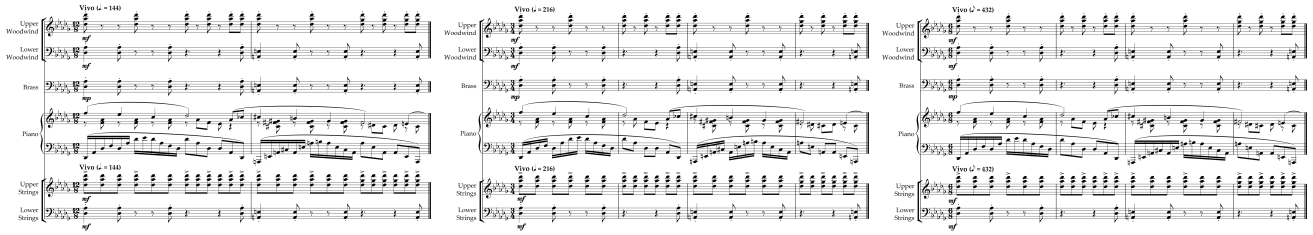
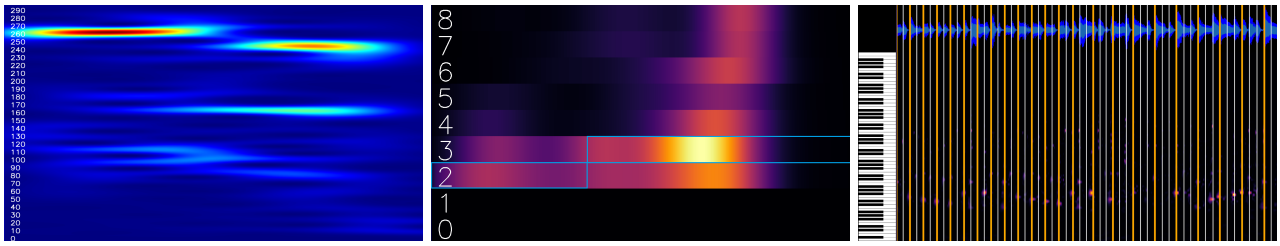


FIGURE 3: Predicted and Ground Truth Bar and Tatum alignment results superimposed over the audio waveform for an extract with a primary time signature division of 7. Red = Predicted; Orange = Ground Truth; x-axis = time (s)



(a) 12/8 (or 4/4) - Video available here (b) 3/4 - Video available here (c) 6/8 - Video available here

FIGURE 4: A selection of metric interpretations provided by the algorithm for a metrically ambiguous orchestral extract (Piano Concerto No.4 by the first Author). A reduced score is shown for convenience (full scores available in the links)



(a) F. Tempogram (y-axis=Tempo, bpm) (b) Metrogram (y-axis=Division) (c) Allocations (Bar=Orange, Beat=Grey)

FIGURE 5: Results for an extract which modulates from 2/4 to 3/4 (x-axis = time for 15s extract). Video available here

Division	F-meas. (B)	Cemgil (B)	P-score (B)	MSE (B)	F-meas. (T)	Cemgil (T)	P-score (T)	MSE (T)
2 or 4	0.962	0.773	0.997	0.00111	0.961	0.777	0.940	0.00368
3	0.954	0.723	1.000	0.00148	0.933	0.724	0.905	0.00170
5	1.000	0.835	1.000	0.00066	1.000	0.837	1.000	0.00066
7	1.000	0.781	1.000	0.00096	0.996	0.790	0.944	0.00090
Overall	0.972	0.774	0.999	0.00109	0.967	0.777	0.943	0.00238

TABLE 1: Evaluation Statistics for the Jazz Piano dataset (B = Bar, T = Tatum). Animated dataset results available here

Genre	No. of Extracts	F-measure
ChaChaCha	111	0.954
Jive	60	0.944
Quickstep	82	0.963
Rumba-American	7	0.907
Rumba-International	51	0.956
Rumba-Misc	40	0.878
Samba	86	0.920
Tango	86	0.894
VienneseWaltz	65	0.917
Waltz	110	0.817
Total	698	0.913

TABLE 2: F-measure scores for the Ballroom Dataset, broken down by genre for our method

Method	F-measure
Our Method	0.913
Krebs et al. [45]	0.855
Multi-Model + DBN [42]	0.910†
Zapata et al. [46]	0.767
Davies and Böck [36]	0.933†
BeatNet [39]	0.774⋆
Elowsson [37]	0.925†
Aubio [40]	0.567⋆
Spectral TCN [35]	0.956†
Böck et al. [38]	0.938†
IBT [41]	0.708⋆

TABLE 3: Comparison of F-measure scores: † refers to 8-fold cross-validated models, and ⋆ are online methods

REFERENCES

[1] È. Poudrier and B. Repp, "Can Musicians Track Two Different Beats Simultaneously?," *Music Perception: An Interdisciplinary Journal*, vol. 30, pp. 369–390, 2013

[2] E. Poudrier, "Multiple Temporalities: Speeds, Beat Cues, and Beat Tracking in Carter's Instrumental Music," in *Proceedings of the Joint Conference of the American Musicological Society (AMS), Society for Music Theory (SMT), and Society for Ethnomusicology (SEM)*, New Orleans, LA, Nov. 2012

[3] C.-Y. Chiu, M. Müller, M. E. P. Davies, A. W.-Y. Su, and Y.-H. Yang, "An Analysis Method for Metric-Level Switching in Beat Tracking," *IEEE Signal Processing Letters*, vol. 29, pp. 2153–2157, 2022

[4] M. E. P. Davies and S. Böck, "Evaluating the Evaluation Measures for Beat Tracking," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[5] M. A. Miguel, M. Sigman, and D. Fernandez Slezak, "From beat tracking to beat expectation: Cognitive-based beat tracking for capturing pulse clarity through time," *PLoS ONE*, vol. 15, no. 11, Art. no. e0242207, Nov. 2020.

[6] A. Holzapfel, F. Krebs, and A. Srinivasamurthy, "Tracking the 'odd': Meter inference in a culturally diverse music corpus," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 425-430.

[7] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007

[8] A. Srinivasamurthy, A. Holzapfel, A. T. Cemgil, and X. Serra, "Particle Filters for Efficient Meter Tracking with Dynamic Bayesian Networks," in *Proc. of the 16th Int. Society for Music Information Retrieval Conf. (ISMIR 2015)*, Malaga, Spain, 2015, pp. 197–203

[9] A. Holzapfel and T. Grill, "Bayesian meter tracking on learned signal representations," in *Proc. of the ISMIR-International Conference on Music Information Retrieval*, 2016, pp. 262–268

[10] N. Whiteley, A. Cemgil, and S. Godsill, "Bayesian modelling of temporal structure in musical audio," in *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, 2006

[11] L. Su, T.-Y. Chuang, and Y.-H. Yang, "Exploiting Frequency, Periodicity and Harmonicity Using Advanced Time-Frequency Concentration Techniques for Multipitch Estimation of Choir and Symphony," in *Proc. of ISMIR*, 2016.

[12] F. Krebs, S. Böck, and G. Widmer, "An Efficient State-Space Model for Joint Tempo and Meter Tracking," in *Proc. of ISMIR*, 2015, pp. 72-78

[13] A. Srinivasamurthy, A. Holzapfel, A. T. Cemgil and X. Serra, "A generalized Bayesian model for tracking long metrical cycles in acoustic music signals," 2016 IEEE ICASSP, Shanghai, China, 2016, pp. 76-80

[14] A. Srinivasamurthy, A. Holzapfel, and X. Serra, "In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music," *Journal of New Music Research*, vol. 43, no. 1, pp. 94–114, 2014

[15] A. T. Cemgil and B. Kappen, "Monte Carlo methods for tempo tracking and rhythm quantization," *Journal of Artificial Intelligence Research*, vol. 18, pp. 45–81, 2003.

[16] E. Quinton, K. O'Hanlon, S. Dixon, and M. Sandler, "Tracking metrical structure changes with sparse-NMF," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 41–45

[17] M. Gainza, "Automatic musical meter detection," in *2009 IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 329-332.

[18] R. M. Bittner et al., "Deep Salience Representations for F0 Estimation in Polyphonic Music," in *Proc. of ISMIR*, 2017.

[19] A. P. Klapuri, A. J. Eronen and J. T. Astola, "Analysis of the meter of acoustic musical signals," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 342-355, Jan. 2006.

[20] M. Gainza and E. Coyle, "Time signature detection by using a multi resolution audio similarity matrix," in *122nd Audio Engineering Society Convention*, Vienna, Austria, 2007.

[21] G. Percival and G. Tzanetakis, "Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1765-1776, Dec. 2014.

[22] A. Cemgil, A. T., B. Kappen, P. Desain, and H. Honing, "On Tempo Tracking: Tempogram Representation and Kalman Filtering," *Journal of New Music Research*, vol. 28, pp. 259-, 2001.

[23] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR-2011)*, Miami, FL, USA, 2011, pp. 669-674.

[24] M. Fuentes, B. Mcfee, H. C. Crayencour, S. Essid, and J. P. Bello, "A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning," in *ICASSP 2019-2019 IEEE ICASSP*, Brighton, UK, 2019, pp. 481-485.

[25] J. Abimbola, D. Kostrzewa, and P. Kasprowski, "Time Signature Detection: A Survey," *Sensors (Basel)*, vol. 21, no. 19, 6494, Sep. 2021.

[26] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A mid-level tempo representation for musicsignals," in *2010 IEEE ICASSP*, Dallas, TX, USA, 2010, pp. 5522-5525.

[27] P, Geoffroy. "Rhythm Classification Using Spectral Rhythm Patterns." ISMIR. 2005.

[28] P, Geoffroy. "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal." IEEE Transactions on Audio, Speech, and Language Processing 19.5 (2010): 1242-1252.

[29] M. Prockup, A. Ehmann, F. Gouyon, E. Schmidt, and Y. Kim, "Modeling Rhythm at Scale with the Music Genome Project," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 2015.

[30] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[31] S. Dixon, "Evaluation of the audio beat tracking system BeatRoot," *Journal of New Music Research*, vol. 36, no. 1, 2007.

[32] A. Holzapfel, M. E. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012

[33] M. Davies, N. Degara Quintela, and M. Plumbley, "Evaluation Methods for Musical Audio Beat Tracking Algorithms," 2009.

[34] G. Song and Z. Wang, "An Efficient Hidden Markov Model with Periodic Recurrent Neural Network Observer for Music Beat Tracking," Electronics, vol. 11, no. 24, pp. 4186, 2022

[35] Böck, S. and Davies, M. E. P., "Deconstruct, Analyse, Reconstruct: how to improve tempo, beat, and downbeat estimation," in *ISMIR*, 2020.

[36] M. E. P. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking," in *Proc. of the 27th European Signal Processing Conf.*, 2019.

[37] A. Elowsson, "Beat tracking with a cepstroid invariant neural network," in *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, 2016, pp. 351–357.

[38] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, 2016, pp. 255–261.

[39] M. Heydari, F. Cwitkowitz, and Z. Duan, "BeatNet: CRNN and Particle Filtering for Online Joint Beat Downbeat and Meter Tracking," arXiv preprint arXiv:2108.03576, 2021.

[40] P. M. Brossier, "Automatic annotation of musical audio for interactive applications," P. dissertation, Ed., Queen Marry University, London, UK, August 2006, pp. 58–102.

[41] J. L. Oliveira, F. Gouyon, L. G. Martins, , and L. P. Reis, "IBT: A real-time tempo and beat tracking system," in In Proc. of the 11th Intl. Conf. on Music Information Retrieval (ISMIR), 2014, pp. 291–296.

[42] S. Böck, F. Krebs, and G. Widmer, "A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles," in *International Society for Music Information Retrieval Conference*, 2014

[43] F. Gouyon et al., "An experimental comparison of audio tempo induction algorithms," in IEEE Trans. on Audio, Speech, and Lang. Processing, vol. 14, no. 5, pp. 1832-1844, Sept. 2006.

[44] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media," Geophysics, vol. 47, no. 2, pp. 203-221, 1982.

[45] F. Krebs, S. Böck, and G. Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *Proc. 14th Int. Soc. for Music Information Retrieval Conf. (ISMIR 2013)*, Curitiba, Brazil, Nov. 2013, pp. 227–232.

[46] J. R. Zapata, M. E. P. Davies, and E. Gómez, "Multi-feature beat tracking," *IEEE/ACM Trans. on Audio, Speech, and Lang. Processing*, vol. 22, no. 4, pp. 816–825, Apr. 2014.