TEMPO ESTIMATION ON MULTIPLE METRICAL LEVELS WITH SEQUENCY FLUX AND MULTIRESOLUTION ANALYSIS

Savvas KAZAZIS (savvas.kazazis@mail.mcgill.ca) [1,2]

[1] *School of Music Studies*, **Aristotle University of Thessaloniki**, Thessaloniki Greece
[2] *Department of Informatics and Telecommunications*, **National and Kapodistrian University of Athens**, Athens Greece

ABSTRACT

We present a method that accurately estimates multiple tempi simultaneously by acting on different time-scales of the signal. We introduce the term sequency flux, which refers to an audio feature derived from the Walsh-Hadamard transform, as an alternative to the widely used spectral flux. The signal derived from sequency flux is further decomposed using the Empirical Mode Decomposition (EMD) and multiple tempo hypotheses are formed by evaluating the autocorrelation function on the intrinsic mode functions (IMFs) over a range of analysis window lengths. An inference method is then proposed that filters out weak hypotheses. The inference lies on measuring the inter-onset intervals (IOIs) found within each IMF, and on how strongly these correspond to the respective tempo hypotheses. The method is extremely simple to implement yet robust, because it does not rely on complex heuristics, the use of arbitrary and signal dependent thresholds, or learning.

1. INTRODUCTION

Tempo perception and beat induction are fundamental musical traits that allow one to enjoy and appreciate music. It is because of these traits that one can dance to the music, or participate in collaborative music-making activities. As Honing nicely puts it [1]: "Without it no music". It is therefore no surprise that topics related to tempo estimation and beat tracking have received much attention in the music information retrieval community [2]. Most algorithms aim to estimate at least one of the three following metrical levels: the tatum (or temporal atom; [3]), the tactus (or beat) and the measure (or bar). These levels relate to the detection of musical events operating on different time scales. The tatum refers to the shortest durational value of events (e.g., a sixty-fourth note). The tactus is the most prominent level and refers to the tapping-rate that a listener is most likely to tap to while listening to music. Tempo is usually derived from the tactus level. The measure level relates to the rate of harmonic changes, or to the period of rhythmic and melodic patterns.

In the fields of computational ethnomusicology and especially in performance studies, one may be interested in analyzing the morphological structures of musical events that occur on metrical levels other than the tatum, tactus, and measure. Stylistic conventions (related to a particular genre, for example) and aesthetic principles of performers and composers may appear both in macro and micro time scales, other than the measure or tatum levels [4]. However, the vast majority of algorithms aim at estimating just one metrical level, which is often related to tactus (i.e., the tempo). Although one may assume that tempi that are multiples or submultiples of the basic tempo are equally valid, this assumption may not always agree with listeners' perceptions (Section 2). This is because actual events may not be present (or cannot be implied) on hypothetical metrical levels. As Goto mentions in [5], "The principal reason that beat tracking is intrinsically difficult is that it is the problem of inferring an original beat structure that is not expressed explicitly." In addition, systems which detect just one metrical level, may have limited success when applied to musical pieces composed with polyrhythms.

In this work, we propose a method that aims to estimate all the characteristic time scales in which musical events occur. This is accomplished through a decomposition method (i.e., the Empirical Mode Decomposition; Section 3), and the formation of multiple tempo hypotheses that are derived from the resulting components of the decomposition. We are particularly interested in testing whether these components actually carry tempo information on metrical levels that match listeners' perceptions. In the following, we first present related work, and some algorithms that were evaluated in the MIREX campaigns along with results obtained from two of the most popular deep learning approaches (Section 2). In Section 3, we present the core elements of the proposed method, and in Section 4 the method itself along with evaluation results on a limited dataset from the MIREX campaign, which served as a case study. In Section 5 we summarize our conclusions and point to some future directions.

2. RELATED WORK & CHALLENGES

2.1 Onset Detection & Spectral Flux

Most tempo estimation methods start by first applying an onset detection function to the signal in order to estimate the positions of note onsets. The onset detection function can be energy-based or spectral-based. Energy based methods usually consist of the following steps. The audio signal is first rectified (e.g., by squaring it) and then

smoothed by a low-pass filter in order to estimate its (power) amplitude envelope. Sudden energy changes in the amplitude envelope are then captured through differentiation. For onset detection, it is usually assumed that only the energy increments observed in the differentiated amplitude envelope are relevant to note onsets and therefore, the signal is half-wave rectified in a subsequent step. The last step involves a peak-picking strategy (e.g., [6, 7]) in order to eliminate spurious peaks, which may not correspond to actual note onsets. Energy-based methods generally provide acceptable results when applied to musical pieces that exhibit strong percussive onsets. However, the results are relatively poor when these methods are applied to musics which mainly consist of soft onsets (e.g., orchestral pieces, string quartets, chorales), or to dense musical passages in which notes occur asynchronously and occupy different frequency bands.

Spectral-based methods [8] offer an improvement over energy-based methods because the differentiation occurs within particular frequency bands of the audio signal, and therefore minimize the possibility of an onset to go unnoticed due to energy masking. The first step in spectral-based methods is the computation of *spectral-flux*, which refers to the computation of the short-time Fourier Transform (STFT) followed by differentiation (across time) over each frequency band. Alternatively, some methods [9] group the frequency bins of the STFT into several bands (e.g., according to Mel-scale) and differentiation take place between those bands instead of the individual frequency bins. Many methods also use the power instead of the magnitude of frequency bands or some other form of logarithmic compression [10] prior to differentiation. The differences are then half-wave rectified and summed across all frequency bands. The last step usually involves some peak-picking strategy as in the energy-based methods.

Dixon [11] compared several different methods related to the computation of spectral flux for the purpose of onset detection, but found that the magnitude-based spectral flux (as described above) performs in par with other more complicated methods (e.g., methods that also take phase into account). Energy fluctuations in a particular spectral band (e.g., tremolo) may lead to the detection of erroneous onsets. Although thresholding through a peak-picking strategy may improve the results, fluctuations in frequency (e.g., vibrato) also cause "blurring" of the peaks in the final onset detection function. A solution to this problem was proposed by Böck et al. [12] through an onset detection algorithm with vibrato and tremolo suppression ("Superflux"), which reduces the number of false-positive onset detections for musical pieces that exhibit strong vibrato. However, it should be noted that the reduction of false positive onsets through some form of suppression require the use of carefully chosen thresholds and that the optimal values of which may vary across musical genres, or instrumentation.

2.2 Periodicity Analysis & Tempo Estimation

Tempo estimation usually relies on finding the dominant periodicity of the onset detection function. One of the most commonly used methods in detecting such periodicities is the autocorrelation function (e.g., [13, 14]). Other methods include the Fourier Transform [10] and comb filterbanks [9, 15] or a bank of resonators applied to chroma features derived from the harmonic and percussive components of the signal [16]. Peeters [17] explores the fact that octave uncertainties present in Fourier-based methods and autocorrelation-based methods occur in inverse domains (i.e., in frequency domain and time-lag-domain, respectively) and proposes a combination of the two. Another group of methods detect dominant periodicities directly from the inter-onset intervals (IOIs) found in the onset detection function through histogramming [18], or some clustering scheme [19]. Many tempo estimation systems are also equipped with an induction algorithm [19] in order to make more accurate predictions given a set of tempi candidates, which possibly correspond to different metrical levels (e.g., tatum, tactus). Common algorithms used for this task include the use of Hidden Markov Models (HMMs; e.g., [9,17]) and dynamic programming (e.g., [16, 20]).

In relation to modelling multiple hierarchical metrical levels, one of the first systems was proposed by Goto [5]. The system potentially recognizes hierarchical beat structures consisting of the quarter note level, the half-note level, and the measure-level. The latter is computed under the assumption that the time signature of the musical track is 4/4. In relation to quarter note level (i.e., the beat level) the system assumes that the tempo range is between 61 BPM and 185 BPM for drum tracks, or between 61 BPM and 120 BPM for music without drums. Klappuri et al. [9] proposed a method that simultaneously estimates the metrical levels of tatum, tactus (i.e., the beat), and measure. The HMM is designed to take into account temporal dependencies between successive estimates of each metrical level, and imposes explicit dependencies between the periods and phases of the tatum, tactus and measure levels. Lartillot et al. [21] present a method that tracks hierarchical metrical structures at various levels expressed as a set of detected periodicities in pairwise harmonic relations. This method requires no training (compared to the method presented in [9], for example) and relies on signal processing (e.g., by using methods similar to spectral flux and autocorrelation) as well as heuristics-based peak tracking.

Deep learning approaches to tempo estimation include the models of [22] and [23]. In addition to being widely cited and having publicly available open-source implementations, these methods are also considered to be some of the most accurate in relation to particular datasets. Table 1 provides the evaluation results of these two approaches, and the rest of the methods mentioned above. These results stem from the MIREX tempo estimation campaigns in which the task was to estimate two tempi correctly along with their perceptual salience (see also Section 4).

From this table we can see that the deep learning approaches perform better than the other methods in estimating at least one tempo correctly (although perhaps in some cases with questionable statistical significance). However, their accuracy in estimating correctly a second tempo is

Reference	[22]	[23]	[16]
One tempo correct	0.99	0.98	0.94
Both tempi correct	0.69	0.66	0.62
Reference	[9]	[20]	[21]
One tempo correct	0.94	0.93	0.92
Both tempi correct	0.61	0.46	0.57

Table 1. Evaluation results of some algorithms submitted to the MIREX campaigns over the years. (Results from [21].)

relatively low. Interestingly, methods designed for modelling multiple hierarchical levels perform less well in this task compared to other generic methods. These results are rather discouraging in relation to the relatively high 8% tolerance level, which was used in the MIREX campaigns, instead of the widely accepted 4% tolerance level and according to which Accuracy 1 (*Acc1*) is measured.

Another popular and relatively accurate deep learning approach for simultaneously tracking tempo, beat, and downbeat is proposed by Böck et al. [24]. While this method estimates a single tempo (based on datasets with tempo and/or beat tracking data), an additional tempo hypothesis could potentially be formed through the estimation of downbeats. However, downbeats do not always directly relate to the measure level. Table 2 lists the evaluation results of this method on tempo estimation, alongside some previously mentioned methods on specific datasets, as reported in [24]. Comparing these results to those for secondary tempo estimation in Table 1, it can be inferred that nearly perfect Accuracy 2 (*Acc2*) scores rarely reflect listeners' perceptions of tempo.

Reference	[24]	[23]	[16]	[13]
Lowest *Acc1*	0.830	0.769	0.651	0.506
Highest *Acc1*	0.870	0.821	0.725	0.733
Lowest *Acc2*	0.950	0.926	0.922	0.924
Highest *Acc2*	0.990	0.976	0.979	0.972

Table 2. Evaluation results of the method presented in [24] compared to other algorithms. "Lowest *Acc*" and "Highest *Acc*" represent the minimum and maximum accuracy scores observed across the datasets used in the study. (Results from [24].)

3. SEQUENCY FLUX & THE EMPIRICAL MODE DECOMPOSITION

3.1 The Walsh-Hadamard Transform & Sequency Flux

It is well known that narrow-band signals that possess harmonic structures such as voiced speech, or sustained and pitched instrument sounds, can be well represented by the FFT through a small number of harmonic complex exponentials. Abruptly changing signals, such as percussive onsets, unvoiced speech, or complex musical passages, are often random-like, "noisy", and require a large number of Fourier coefficients in order to be accurately represented, and therefore may not be characterized efficiently. Such broadband signals can be characterized more efficiently by a broadband class of basis functions such as the Walsh functions [25]. The Walsh-Hadamard transform is analogous to the Fourier transform, but instead of sinusoids it consists of a set of periodic and aperiodic rectangular waves that take only two amplitude values of +1 and -1. These waves are characterized by their sequency (as opposed to frequency in the Fourier transform), which denotes the number of zero-crossings the function makes per unit time.

Hadamard matrices of order $M = 2^k$ can be generated recursively from the following relationship [26]:

$$\mathbf{H}(k+1) = \begin{bmatrix} \mathbf{H}(k) & \mathbf{H}(k) \\ \mathbf{H}(k) & -\mathbf{H}(k) \end{bmatrix}, \quad k = 0, 1, 2, ..., \quad (1)$$

with $H(0) = 1$.The Walsh-Hadamard transform of a real valued signal $\mathbf{x}$ of length M can then be computed from:

$$\mathbf{X} = \mathbf{x} \cdot \mathbf{H}^{(M)}, \quad (2)$$

where the coefficients of $\mathbf{X}$ correspond to the weights of the Walsh functions.

The Walsh-Hadamard transform can be applied sequentially to (possibly overlapping) frames of x in a similar way to the short-term Fourier Transform (STFT) with the only minor restriction being that the frame size should be a power of two. If we let $X(n, s)$ represent the *s*th sequency of the *n*th frame, then the computation of sequency flux ($SeqFlux$) becomes identical to the computation of (magnitude-based) spectral flux:

$$SeqFlux[n] = \sum_{s=1}^{S} |X(n, s)| - |X(n-1, s)| \quad (3)$$

3.2 The Empirical Mode Decomposition (EMD)

The Empirical Mode Decomposition (EMD) [27] is a nonlinear and adaptive method for analyzing non-stationary signals and data. With this method, the signal is represented as a sum of zero-mean amplitude and frequency modulated (AM/FM) components. In the context of EMD, these components are called Intrinsic Mode Functions (IMFs) and (in theory) must satisfy the following two conditions: (i) the number of extrema (i.e., the local maxima and local minima of the signal) and the number of zero crossings must be equal, or differ at most by one; and (ii) the envelopes defined by local maxima and local minima must be symmetric, or in other words, the mean value of the upper and lower envelopes must be zero.

EMD is based on an iterative process that successively removes the fine structure of the signal through some form of adaptive and signal-dependent time variant filtering. The IMFs are computed according to the following steps [28]:

1. Estimate all the maxima and minima of the signal $x(t)$

2. Construct the upper ($u(t)$) and lower ($l(t)$) envelopes of the signal by interpolating the maxima and minima, respectively
3. Compute the mean envelope $m(t) = (u(t)+l(t))/2$
4. Extract the detail $d(t) = x(t) - m(t)$
5. Iterate from step (1) to (4) on $d(t)$ until it can be considered zero-mean (and therefore an IMF)
6. Set $d(t)$ to be an IMF, $i(t) = d(t)$
7. Repeat from the beginning on the residual, $r(t) = x(t) - i(t)$ as the data
8. Stop the process if $r(t)$ remains approximately constant, or is monotonic

The procedure of repeated iterations that occur in step 5 is known as *sifting*. The algorithm will eventually converge, because the number of extrema is decreased when progressing from one residual to the next.

To the best of authors' knowledge, there are only a few approaches for tempo estimation using the EMD algorithm [29, 30]. In [29], Pikrakis et al. used EMD to segment music recordings into regions that exhibit similar rhythmic characteristics, and to analyze the diagonals of the self-similarity matrix of those regions in order to extract the tempo. The method presented in [30] is conceptually similar to the one we present here, and therefore merits some discussion. This method computes IMFs of the input signal, and estimates their corresponding amplitude envelopes using low-pass filtering and half-wave rectification. The periodicities of each IMF are estimated through the autocorrelation function (ACF). Tempo induction is made in two steps. In the first step, the IMFs that do not exhibit harmonically related peaks in the respective ACF are removed. In the second step, the induction algorithm searches for one-to-one (peak) correspondences and harmonic relations (in terms of time-lags) between the ACF peaks of a particular IMF in relation to subsequent ones. If no match is found, the induction method fails to evaluate the tempo. Although this method is conceptually similar to the one presented in this paper, it is fundamentally different to the one proposed in this work and which we detail in the next section.

4. MULTIRESOLUTION TEMPO ESTIMATION

4.1 Formation of Multiple Tempo Hypotheses

The method starts with a pre-processing stage that includes the following steps: conversion to mono by selecting the channel with the highest root mean square amplitude; resampling to 22.5 kHz; and removal of the global DC offset. After this stage, the Walsh-Hadamard transform is computed using a window length of 512 samples and a hop-size of 256 samples, which corresponds to about 11.6 ms. The sequency flux is then computed according to Equation (3). We deliberately avoid the use of half-wave rectification (HWR) on sequency flux, because we assume that offsets also play a role in tempo perception [1]. To achieve a better resolution, the frame rate of sequency flux gets upsampled to 200 Hz. The next stage is the computation of EMD in order to derive the IMFs of sequency flux. For the computation of EMD the algorithm of [31] is used with the default settings provided by the authors.

The periodicities of sequency flux along with the periodicities of each IMF are estimated using a windowed-autocorrelation function. The (normalized) autocorrelation function is computed using a set of window lengths ranging from 2 s to 12 s with 1 s increments and a hop-size of 5 ms. The dominant periodicities of each frame per window length are estimated by finding the global maxima of the ACFs within the range of 30 to 600 BPM. Another set of periodicities is computed using the peaks of the ACFs that occur after the first minima, which in some cases may coincide with the global maxima.

A first set of tempo hypotheses for each IMF and the original sequency flux, and for each window length, is formed using the median values of the dominant periodicities (although it is acknowledged that the mode in some cases could give a more reliable estimate). Hypotheses that do not agree with the medians of the second set of periodicities are removed from this set. This "filtering" stage is done to ensure that the estimations are relatively stable, because if the ACF is noisy it may not exhibit a clear peak structure.

A second set of tempo hypotheses is formed using the medians of the second set of detected periodicities. Only tempi that are lower than the minimum tempo of the first set of hypotheses are kept. This "filtering" stage is specifically done to utilize the periodicities found in higher IMFs (i.e., lower tempi): in these IMFs the ACF decays slowly to zero and therefore, the peak of the ACF that occurs after the first minimum is a more robust estimator of the dominant tempo than the global maximum.

4.2 Induction

Depending on the complexity of the signal, some of the hypotheses formed according to the description given above might be wrong, including hypotheses derived from the original sequency flux. This could be due to various reasons such as: signals having a weak periodic structure and the inability of the ACF to detect periodicities within these structures; inaccurate IMFs generated from EMD due to over-iteration [28]; or even the choice of the summary statistic that is used to estimate the global tempo, which in this case is the median.

In the induction stage, we seek rational related tempi. For each hypothesis and from its respective waveform (i.e., sequency flux, or an IMF), we compute the total number of occurrences of the IOI that matches that hypothesis. The IOIs are rounded to the nearest BPM value of the ACF's resolution. This computation considers all possible interval pairs, not just consecutive ones. In the next step, we check whether integer multiples of the hypothesis that has the maximum number of IOI occurrences exist. If that is the case, that hypothesis along with its multiples is considered as valid.

If no multiples are found, we use the hypotheses derived

from sequency flux. As a reminder, there are as many hypotheses as the number of different window lengths, which are used in the computation of windowed ACF. From these hypotheses we choose the one that has the fastest tempo. Multiples and submultiples of this hypothesis (which may also be found in IMFs) are also considered to be valid. The decision to use the fastest tempo estimated from sequency flux, is based on the rationale that one has to first estimate the tatum before estimating the beat, or similarly, to first estimate the beat before estimating the measure.

4.3 Results

The proposed method was evaluated on twenty music excerpts provided by the 2006 Music Information Retrieval Evaluation eXchange campaign (MIREX06) for the Audio Tempo Extraction Competition from 2006 to 2018. The excerpts have a 30-seconds duration and are of constant tempo. The dataset includes a wide range of tempi, musical styles, genres, instrumentation and tempo estimations (derived from tapping-data) made by 40 listeners. It's also worth noting that some excerpts contain high levels of background noise, making them useful for testing algorithmic robustness in such scenarios. The ground-truth data consist of two tempi with their respective salience, because not all listeners were tapping on the same metrical level (e.g., the second tempo could be two, or three times faster than the first one). The algorithms were evaluated according to their ability to track both tempi correctly, and a *P*-score which relates the two estimations with their respective salience levels.

In this work, we are interested in evaluating: the accuracy of multiple tempo hypotheses formed in subsection 4.1; and whether the induction method presented in subsection 4.2 actually filters out the weak hypotheses while retaining the correct tempo. We also used a 4% tolerance level (*Acc1* value) on the estimations instead of the 8% used in the competition. The evaluation showed that the proposed method generates hypotheses that include at least one of the two tempi in 100% of the cases, and hypotheses that include both tempi in 85% of the cases. The induction algorithm preserved 95% of the correct tempo hypotheses by incorrectly rejecting one correct tempo, which indicates that the induction step could be improved.

Although these excerpts were meant to be used as a "training" dataset for the algorithms submitted to the MIREX campaigns, in this work this dataset was used as a case study. This allowed us to: (i) test whether there is any particular advantage of using a set of different window lengths when computing the ACF instead of just one; (ii) inspect the relative contribution of each IMF and sequency flux to the formation of correct tempo hypotheses; and (iii) identify pitfalls related to the induction process. Fig. 1 shows the number of occurrences of sequency flux and each IMF in the set of correct tempo hypotheses per excerpt. Notably, for some excerpts, the correct tempo was estimated exclusively from the IMFs and not from sequency flux. This is also the case for the slow tempi of the dataset, for which the correct tempo could only be estimated from the high IMFs.

5. CONCLUSIONS

We presented a multiresolution method that is capable of estimating multiple tempi from a detection function. In this work, the detection function was computed using the Walsh-Hadamard transform and it was therefore termed *sequency flux*, as opposed to the widely used spectral flux. The method acts on different time scales of the signal because it detects periodicities using a set of different window-lengths for the computation of the windowed-ACF, and because the EMD decomposes the sequency flux into signals that exhibit different characteristic time scales. In this way, the dominant periodicities of sequency flux that correspond to different time scales are emphasized and most importantly, these periodicities can be directly and more clearly observed from the IOIs of each IMF. The method is simple to implement and does not rely on empirically set thresholds or values that depend on the complexity of the signal (e.g., genre, instrumentation).

In general, the method generates some tempo hypotheses that could be considered as incorrect, and which are not harmonically related. This is the major reason why an induction stage was employed. However, the existence of metrical levels that are not harmonically related may indicate harmonic changes or rhythmic structures that occur *within* a "harmonic" time scale, but further work is needed to verify this hypothesis. In addition to the generation of non-harmonically related tempo hypotheses, in some cases, the method failed to estimate both tempi correctly, which indicates that some tempi cannot be directly derived from sequency flux or the IMFs. These results could be due to various reasons, some of which are already mentioned in Section 4. Some reasons that are specifically related to the EMD include the possible over-iteration, which may have occurred in the sifting stage, and the so-called "mode mixing" problem. Sifting the signal too much may have generated artificial modes that do not actually appear in the data [28]. The mode mixing problem refers to the presence of similar time scales in different modes, which may have distorted the median periodicity values found from these IMFs. This issue could potentially be addressed by using variants of the EMD such as the EEMD [32] or the CEEMD [33]. However, in this work the EMD was preferred over other methods because of its simplicity.

Clearly, a more sophisticated and robust induction algorithm is needed for this method to work on musical pieces with polyrhythms. However, the simple induction algorithm presented here achieves 95% accurate results on this case study but obviously, the method needs to be evaluated on larger datasets before making any claims related to its overall accuracy. Additionally, this method is currently not suitable for real-time applications due to its computational complexity. While the Fast Walsh-Hadamard transform is computationally more efficient than the Fast Fourier Transform (since it utilizes only real number additions and subtractions) the proposed method is still slow for real-time applications. This is because the ACF computation with multiple window lengths, the chosen EMD algorithm, and the current induction process, all contribute to the overall complexity. Nonetheless, the method finds

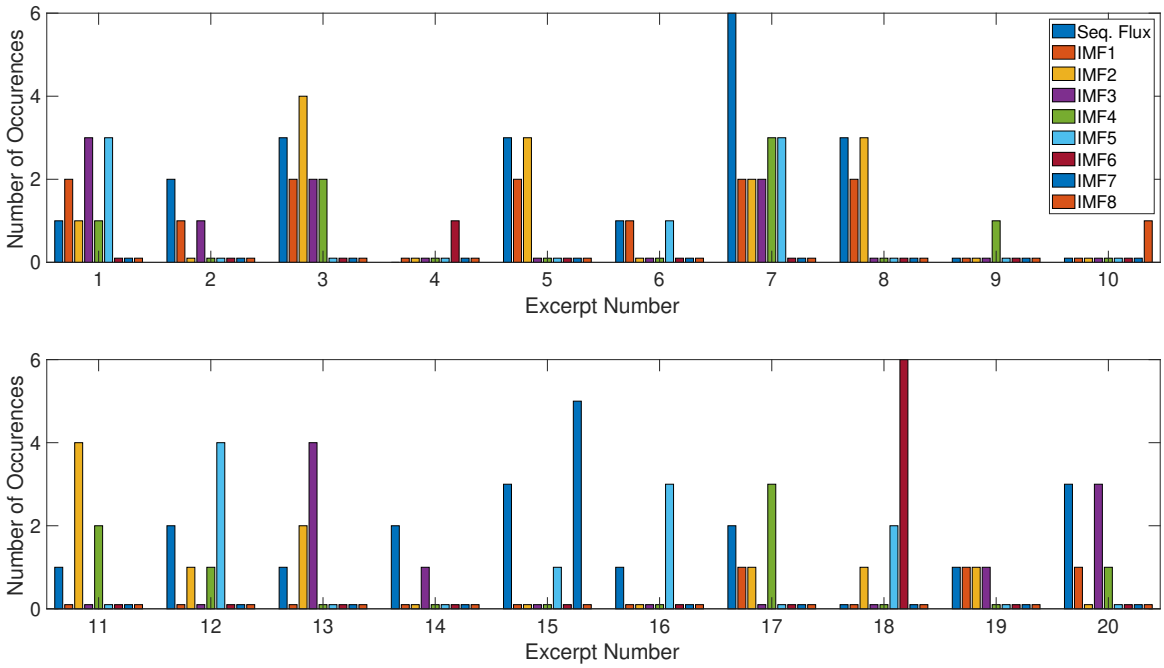

Figure 1. Number of occurrences of each IMF and sequency flux in the set of correct tempo hypotheses per excerpt.

many (offline) applications in areas such as: computational musicology (e.g., studying the interaction between rhythmic components in music); audio-to-MIDI transcription systems (e.g., capturing multiple rhythmic layers in polyphonic music); content-based music retrieval (e.g., searching based on rhythmic similarities); and automatic accompaniment generation (e.g., creating accompaniments that better follow a melody's rhythmic structure).

In the future, we plan to adapt this method for pieces with non-constant tempi and explore its beat tracking capability on different metrical levels. In conclusion, according to the results of this case study, our initial hypothesis is justified: the IMFs do carry tempo information on multiple hierarchical metrical levels. Hopefully, this information will be used to improve existing approaches related to meter induction in general, and more specifically to compound meter induction.

Acknowledgments

This research was co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call Research - Create - Innovate (MIS 5149205).

6. REFERENCES

[1] H. Honing, "Without it no music: Beat induction as a fundamental musical trait," in *Annals of the New York Academy of Sciences*, 2012, pp. 85–91.

[2] H. Schreiber, J. Urbano, and M. Müller, "Music tempo estimation: Are we done yet?" in *Transactions of the Int. Society for Music Information Retrieval*, 2020.

[3] J. A. Bilmes, "Techniques to foster drum machine expressivity," in *Proc. Int. Computer Music Conference*, 1993.

[4] G. Kokkonis, N. Ordoulidis, G. Evagelou, and S. Kazazis, "Analysing tempo stability in Greek rebetiko music. the case of Vasilis Tsitsanis's repertoire," in *Digital Technologies Applied to Music Research. Methodologies, Projects and Challenges*, 2024.

[5] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," in *Journal of New Music Research*, 2002.

[6] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research," in *Proc. Int. Society for Music Information Retrieval Conference*, 2016.

[7] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proc. Int. Society for Music Information Retrieval Conference*, 2012.

[8] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," in *IEEE Signal Processing Letters*, vol. 11, 2004, pp. 553–556.

[9] A. Klapuri, A. J. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," in *IEEE Trans-*

actions on Audio, Speech, and Language Processing, vol. 14, 2006, pp. 342–355.

[10] P. Grosche and M. Müller, "Computing predominant local periodicity information in music recordings," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 33–36.

[11] S. Dixon, "Onset detection revisited," in *Proc. Int. Conference on Digital Audio Effects*, 2006.

[12] S. Böck and G. Widmer, "Local group delay based vibrato and tremolo suppression for onset detection," in *Proc. Int. Society for Music Information Retrieval Conference*, 2013.

[13] G. Percival and G. Tzanetakis, "Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, 2014, pp. 1765–1776.

[14] S. Dixon, E. Pampalk, and G. Widmer, "Classification of dance music by periodicity patterns," in *Proc. Int. Society for Music Information Retrieval Conference*, 2003.

[15] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals." in *Journal of the Acoustical Society of America*, vol. 103, 1998, pp. 588–601.

[16] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 421–424.

[17] G. Peeters, "Template-based estimation of time-varying tempo," in *Advances in Signal Processing*, 2007, pp. 1–14.

[18] F. Gouyon, P. Herrera, and P. Cano, "Pulse-dependent analyses of percussive music," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. IV–4174–IV–4174.

[19] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," in *Journal of New Music Research*, vol. 30, 2001, pp. 39 – 58.

[20] D. P. W. Ellis, "Beat tracking by dynamic programming," in *Journal of New Music Research*, vol. 36, 2007, pp. 51 – 60.

[21] O. Lartillot and D. Grandjean, "Tempo and metrical analysis by tracking multiple metrical levels using autocorrelation," in *Applied Sciences*, 2019.

[22] S. Böck, F. Krebs, and G. Widmer, "Accurate tempo estimation based on recurrent neural networks and resonating comb filters," in *Proc. Int. Society for Music Information Retrieval Conference*, 2015.

[23] H. Schreiber and M. Müller, "A single-step approach to musical tempo estimation using a convolutional neural network," in *Proc.Int. Society for Music Information Retrieval Conference*, 2018.

[24] S. Böck and M. E. P. Davies, "Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation," in *Proc. Int. Society for Music Information Retrieval Conference*, 2020.

[25] A. Spanias, "A hybrid transform method for analysis/synthesis of speech," in *Proc. IEEE Global Telecommunications Conference GLOBECOM '91: Countdown to the New Millennium. Conference Record*, vol. 2, 1991, pp. 719–724.

[26] D. S. Stoffer, "Walsh-fourier analysis and its statistical applications," in *Journal of the American Statistical Association*, vol. 86, 1991, pp. 461–479.

[27] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C.-C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," in *Journal of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, 1998, pp. 903 – 995.

[28] G. Wang, X. Chen, F.-L. Qiao, Z. Wu, and N. Huang, "On intrinsic mode function." in *Advances in Adaptive Data Analysis*, vol. 2, 2010, pp. 277–293.

[29] A. Pikrakis and S. Theodoridis, "An application of empirical mode decomposition on tempo induction from music recordings," in *Proc. Int. Society for Music Information Retrieval Conference*, 2007.

[30] K. Trohidis and L. J. Hadjileontiadis, "Tempo induction from music recordings using ensemble empirical mode decomposition analysis," in *Computer Music Journal*, vol. 35, 2011, pp. 83–97.

[31] G. Rilling, P. Flandrin, and P. Gonçalves, "On empirical mode decomposition and its algorithms," in *In Proc. of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 2003.

[32] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," in *Advances in data science and adaptive analysis*, vol. 1, 2009, pp. 1–41.

[33] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4144–4147.