TEMPLATE-BASED ESTIMATION OF TEMPO: USING UNSUPERVISED OR SUPERVISED LEARNING TO CREATE BETTER SPECTRAL TEMPLATES

Geoffroy Peeters, *

IRCAM - CNRS STMS,
Sound Analysis/Synthesis Team, Paris, France
peeters@ircam.fr

ABSTRACT

In this paper, we study tempo estimation using spectral templates coming from unsupervised or supervised learning given a database annotated into tempo. More precisely, we study the inclusion of these templates in our tempo estimation algorithm of [1]. For this, we consider as periodicity observation a 48-dimensions vector obtained by sampling the value of the amplitude of the DFT at tempo-related frequencies. We name it spectral template. A set of reference spectral templates is then learned in an unsupervised or supervised way from an annotated database. These reference spectral templates combined with all the possible tempo assumptions constitute the hidden states which we decode using a Viterbi algorithm. Experiments are then performed on the "ballroom dancer" test-set which allows concluding on improvement over state-of-the-art. In particular, we discuss the use of prior tempo probabilities. It should be noted however that these results are only indicative considering that the training and test-set are the same in this preliminary experiment.

1. INTRODUCTION

Given the importance of tempo information for a large number of Music Information Retrieval tasks (front-end for many beat-tracking algorithms, therefore for many downbeat-tracking, chord estimation, cover version detection or beat-synchronous algorithms, direct use of tempo information for performing search over music databases) and given the performance obtained by current algorithms (see [2] for an overview of recent results), tempo estimation still remains an important research field.

1.1. Related works

Tempo estimation algorithms can be first classified according to the analyzed materials: - symbolic data or - audio data. Algorithms based on audio analysis usually start by a front-end which either - plays the role of an "audio-to-symbolic" translator [3], [4], - or extracts frame-based audio features such as energy, energy variations, ... [5], [6], [7]. Depending on the kind of information provided by this front-end and the context of the application, a large variety of processes are used to track/estimate the tempo: - time interval histograms [8] [9], - periodicity measure (Fourier transform, auto-correlation function, narrowed-ACF, wavelets, comb filter-bank). The periodicity measure can be used - to estimate directly the tempo or - to serve as observation for the estimation of the whole metrical structure through (probabilistic) models [7] [4] [10]. Some authors propose the use of templates for tempo estimation - in the time/phase domain [10] [7] [11], - in the spectral domain [1]. We refer the reader to [12] for a detailed report on state of the art tempo estimation algorithms. The work which is the most closely related to our work is the one of [13]. In this, Eronen proposes a more detailed use of templates using a database of templates and a K-NN-regression to find the best tempo corresponding to a target template representing an unknown signal.

* This work was partly supported by the "Quaero" Program funded by Oseo French agency for innovation.

1.2. Motivating previous results

In this paper, we study the extension of our previous tempo-estimation algorithm [1] to the inclusion of spectral templates specific to each rhythm-class. The algorithm proposed in [1] relies on the simultaneous estimation of tempo and meter using a set of spectral templates representing all the possible combinations of tempo and meter. For this, three meters are considered: 22 (binary grouping of tactus into bar/ binary subdivision into tatum), 23 (binary/ ternary) and 32 (ternary/ binary). The spectral templates corresponding to the three meters have been manually drawn by observation. We named them Meter-Beat-Subdivision-Templates (MBSTs). The combination of tempo and meter are modeled as hidden states and a Viterbi algorithm is used to find the most-likely hidden states path over time given the DFT/ Frequency-Mapped-ACF periodicity observations.

Because the MBSTs only partially represent the various possible spectral templates for a given tempo and meter, we propose in this paper to use spectral templates obtained by unsupervised or supervised learning given a database annotated into tempo. This should allow to better represent the diversity of rhythm characteristics. For this, we rely on our recent results of [14] and [15].

In [14], we show that using spectral-templates derived directly from the DFT allows obtaining a high classification accuracy (88%) for music genres which are related to the rhythm characteristics of the music ("ballroom dancer" test-set). From [14], we can conclude that the characteristics of a rhythm are well described by spectral-templates derived from the DFT.

In [15], we propose a "copy and scale" method for obtaining directly the estimation of tempo, beat-positions and class of an unknown item. A K-NN algorithm (which uses a complex distance applied to a complex-spectral-template representation) is used to find the closest database-item to a given unknown target. The annotation of the closest item is then "copied and scaled (to the unknown target's tempo and position)" to obtain the estimation of tempo, beat-positions and class of the unknown target. In [15] we only test a limited set of tempo assumption (the ones corresponding to the potential octave errors of the algorithm of [1]). Also, only the subset of the K-NN database for which the item has an

initial tempo close to the target's tempo assumption is considered. This inherently creates a dependency between rhythm pattern and tempo, and applies a sort of prior tempo probability. We show that using this simple and direct method, tempo accuracy (excluding octave error) can be improved from 65.47% (using [1]) to 67.62%. When considering only the subset of items which have been correctly classified, it improves from 60.33% (using [1]) to 97.88%. From this, we can conclude that the characteristics of the rhythm provided by the spectral-templates allow improving tempo estimation; we also conclude that the characteristics of the rhythm are specific to the tempo, i.e. not all rhythm patterns can be found at a given tempo. However, no temporal continuity between frames are taken into account in the K-NN approach of [15].

1.3. Paper content and organization

In this study, we propose to replace the MBST approach of [1] by a Spectral-Template approach obtained either by using unsupervised or supervised learning on a database annotated into tempo. While in [1], the DFT/ Frequency-Mapped ACF was used as a periodicity measure, we use here a sampled and tempo-normalized amplitude DFT. We show in [14] [15] that the DFT provides better results for rhythm description (not for tempo estimation) than the DFT/ FM-ACF.

In part 2 we explain the proposed approach. We explain the sampled and tempo-normalized spectral templates representation (part 2.1), how the reference spectral templates m_j and tempo B_j are created using unsupervised or supervised learning (part 2.2), and the way we introduce them in the hidden states of our Viterbi decoding algorithm (part 2.3). We then perform an evaluation on the 698-tracks "ballroom dancer" test-set. We compare the use of our previous MBST, and the proposed unsupervised or supervised Reference Spectral-Templates. We also test the influence of the use of prior tempo observation on the results. We finally conclude in part 4 and gives direction for future works. We give an overview of the proposed approach in Figure 1.

2. TEMPO ESTIMATION USING UNSUPERVISED OR SUPERVISED TRAINED SPECTRAL TEMPLATES

2.1. Spectral Templates representation

In [14], we have proposed to represent the rhythmic content of a signal frame using sampled and tempo-normalized values of the amplitude spectrum of the local onset-energy function.

For a given audio item, we first extract an onset-energy-function representing at each time the likelihood of an onset. In this study, we have used the method proposed in [1] but any other methods can be used. We note $o(n)$ the corresponding function. It has a sampling rate of 172Hz. Around each time frame t_m, we compute the amplitude spectrum of $o(n)$ using a hamming analysis window of length 8s and a hop size of 0.5s. We note it $Y_o(f_k, t_m)$ where f_k represent the Fourier frequencies. Considering a tempo $b(t_m)$ at time t_m (expressed in Hz), we then sampled $Y_o(f_k, t_m)$ at the frequencies $f_k = b(t_m) \cdot f_l$ with $f_l = [\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, ...8]$. These frequencies represent the harmonic series corresponding to a 4/4 meter ($\frac{1}{4}b(t_m)$) and a 3/4 meter ($\frac{1}{3}b(t_m)$) up to the frequency $8b(t_m)$. We note it $Y_o(l, t_m, b(t_m))$, $l \in [1, 48]$. It is 48-dimensions vector. $Y_o(l, t_m, b(t_m))$ is then normalized by its maximum value over l. It should be noted that for two tracks with the same rhythm-pattern but with different tempi b, their spectral-template $Y_o(l, t_m, b)$ will be similar. This is because $Y_o(l, t_m, b)$ has been made tempo-independent. In the case of tempo estimation, b is an unknown variable. The goal is therefore to find b such that the corresponding vector $Y_o(l, t_m, b)$ looks close to one of the $j \in [1, J]$ prototype vectors $Y_j(l)$. The prototype vectors are trained using real data but annotated tempo $\hat{b}$. They represent therefore the average (over many o) shape of $Y_o(l, t_m, b = \hat{b})$ for various rhythms. Therefore, if $Y_o(l, t_m, b)$ is close to one of the $Y_j(l)$, it is likely to have a tempo b. We now study the estimation of the $Y_j(l)$.

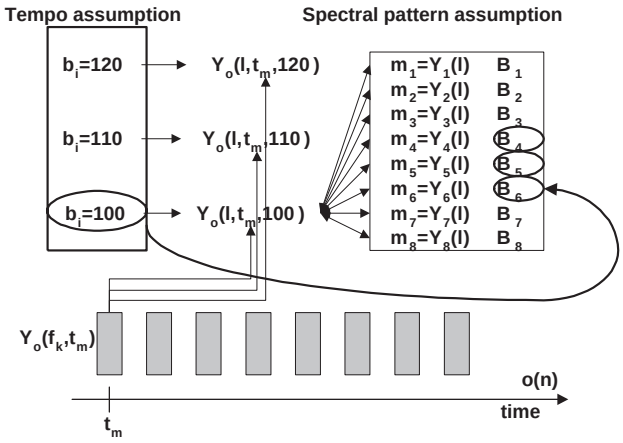

Figure 1: *The periodicity observation $Y_o(f_k, t_m)$ is extracted over time. At each frame, we estimate the likelihood that $Y_o(f_k, t_m)$ corresponds to the hidden state s_{ij} defined as tempo b_i and Reference Spectral-Template m_j. For each tempo assumption b_i (b_i=100 in the figure), we create from $Y_o(f_k, t_m)$ the sampled and tempo-normalized vector $Y_o(l, t_m, b_i)$ ($l \in [1, 48]$), named Spectral Template. $Y_o(l, t_m, b_i)$ is then compared to the set of Reference Spectral Templates $m_j = Y_j(l)$ ($j \in [1, J]$ with $J = 8$ in the present experiment) using a one-minus-cosine distance. This distance is considered as the likelihood to observe s_{ij} given the observation of $Y_o(f_k, t_m)$. Only the subset of m_j which initial tempo B_j is close to the tempo assumption b_i are considered ($b_i = 100$ and B_4, B_5, B_6 in the figure).*

2.2. Creating reference spectral templates

Given a test-set annotated into tempo $\hat{b}$, we extract the series of sampled and tempo-normalized vector $Y_o(l, t_m, \hat{b})$ for each track of the test-set and compute, for each track, its vector of mean-value-over-time $Y_o(l, \hat{b})$. From this set of vectors, we study two possibilities to create reference spectral templates.

Unsupervised learning: Given the whole set of files of the test-set, we apply a fuzzy k-means algorithm. For this, we consider for each track the vector obtained by concatenating the spectral template $Y_o(l, \hat{b})$ and the tempo $\hat{b}$ (48+1=49 dimensions). Doing this, we force the clustering to group tracks which are similar in spectral template $Y_o(l, \hat{b})$ but also in tempo $\hat{b}$. For a clustering using J clusters, the resulting centroids of the clusters provide the set of J reference spectral templates $m_j = Y_j(l)$ (we ommit the $\hat{b}$ variable in the function since the templates are tempo-independent) and their associated reference tempo noted B_j. For

the evaluation of part 3, we have used $J = 8$ clusters.

Supervised learning: Given a test-set annotated into J music genre classes, we compute for each class the mean value of the $Y_o(l, \hat{b})$ and tempo $\hat{b}$ for the tracks belonging to this class. The resulting mean-vectors provide the set of J reference spectral templates $m_j = Y_j(l)$ and their associated reference tempo noted B_j. Since our learning will be based on the "ballroom dancer" test-set which has 8 classes, $J = 8$ also in this case.

2.3. Introduction into hidden variables

We explain the introduction of the spectral templates in the Viterbi decoding algorithm of [1] using the same notation as in [1].

The dominant periodicities $Y_o(f_k, t_m)$ are estimated at each time t_m. $Y_o(f_k, t_m)$ does not only depend on the tempo but also on the characteristics of the rhythm. We therefore look for the temporal path of tempo and rhythm characteristics that best explain $Y_o(f_k, t_m)$. We define a hidden state as a specific combination of a tempo frequency b_i and a specific reference spectral-template m_j. The three probabilities of the Viterbi decoding are:

- the prior probability of each state: $p_{prior}(s_{ij}(t_0))$
- the transition probability between two states: $p_t(s_{ij}(t_{m+1})|s_{kl}(t_m))$
- the emission probability of the states: $p_{emi}(Y_o(f_k, t_m)|s_{ij}(t_m))$

In part 3, we will test two prior probabilities. The first is the one described in [1], it favors the detection of tempo around 120bpm (in the range 50-150bpm) but do not favor any reference spectral-templates in particular. It is modeled as a Gaussian pdf $p_{prior}(s_{ij}(t_0)) = p_{prior}(b_i(t_0)) = N_{\mu=120,\sigma=80}(b_i)$. The second is a uniform probability, i.e. it does not favor any tempo or reference spectral-templates in particular.

The transition probability is the one defined in [1], i.e. it favors tempo continuity over time and disadvantage reference spectral-template changes.

2.3.1. Emission probability in the case of MBST

In [1], the emission probabilities are computed using a score. For a specific tempo b_i and MBST m_j, we compute a score defined as a weighted sum of the values of $Y_o(f_k, t_m)$ at specific frequencies: $score_{i,j}(Y_o(f_k, t_m)) = \sum_{r=1}^{5} \alpha_{j,r} \cdot Y_o(f_k = \beta_r b_i, t_m)$, where $\underline{\beta}$ represents the various ratios of the considered frequency f_k to the tempo frequency b_i of the state s_{ij}: $\underline{\beta} = \left[\frac{1}{3}, \frac{1}{2}, 1, 1.5, 2, 3\right]$. These ratios correspond to significant frequency components for the triple meter, duple meter, tempo, "penalty", simple and compound meter. $\underline{\alpha}_j$ represents the weightings of each of these components. These weightings depend on the MBST m_j of the state s_{ij} and have been manually chosen to better discriminate the various MBSTs (see [1] for details).

2.3.2. Emission probability in the case of spectral-templates

In the case of the spectral-templates, the $score_{i,j}(Y_o(f_k, t_m))$ (the probability to observe $Y_o(f_k, t_m)$ given tempo b_i and reference spectral-templates m_j) is computed as follows.

For a given tempo assumption b_i, we first compute (using the method explained in part 2.1) the sampled and tempo-normalized spectral representation $Y_o(l, t_m, b_i)$ corresponding to this b_i. From the set of reference spectral-templates m_j, we then select the ones which have a reference tempo B_j "close" to the current tempo assumption b_i. The "closeness" is defined as $abs(log_2(\frac{b_i}{B_j})) < 0.3785$. The m_j which are not "close" receive a likelihood (score) of zero. For the subset of "close" m_j, we compute the one-minus-cosine distance between $Y_o(l, t_m, b_i)$ and the reference spectral templates $m_j = Y_j(l)$.

3. EVALUATION

3.1. Test-set

We perform the evaluation on the "ballroom dancer" test-set [16]. We have chosen this test-set since the music genres provided with it (ChaChaCha, Jive, Quickstep, Rumba, Samba, Tango, Viennese-Waltz and Slow-Waltz) are closely related to the rhythm characteristics of the tracks. It therefore facilitates the experiment when using reference spectral-templates obtained by supervised learning based on music genre classes. It should be noted that both spectral-templates (unsupervised or supervised) learning and evaluation of tempo estimation are performed on the same test-set. Results obtained should therefore only be considered as indicative preliminary results.

3.2. Evaluation scenario

For each track, we compare the annotated tempo and the estimated tempo obtained using • m_j= the 22/23/32 MBST, • m_j= the reference spectral-templates obtained by unsupervised learning (ST-Unsupervised), • m_j= the reference spectral-templates obtained by supervised learning (ST-Supervised).

We also compare • the use of a prior tempo probability favoring tempo detection around 120bpm, • the use of a uniform tempo probability which do not favor any tempo in particular.

3.3. Evaluation rules

To measure the performances of tempo estimation, we have used the two measures proposed by [17]: • Accuracy1: measures the number of tracks for which the estimated tempo is within a 4% Tolerance Window of the annotated tempo, • Accuracy2: within 4% of either 1/3, 1/2, 1, 2 or 3 the annotated tempo. Accuracy 2 therefore considers octave errors as correct.

3.4. Results and discussion

The global Accuracy 1/2 are indicated into Table 1. Detailed Accuracy 1/2 per class are given into Table 2. We also indicate into Table 2 the mean-over-class Accuracy 1/2 which differ from the global ones considering the non-equal distribution of the test-set.

Tempo estimation			Acc1	Acc2
MBST 22/23/32	Prior tempo 120	DFT	65,0%	89,4%
	No Prior	DFT	44,0%	87,1%
ST Unsupervised	Prior tempo 120	DFT	63,9%	90,7%
	No Prior	DFT	**72,9%**	**93,4%**
ST Supervised	Prior tempo 120	DFT	62,5%	89,1%
	No Prior	DFT	**75,2%**	**94,8%**

Table 1: *Tempo estimation in terms of Accuracy 1 and 2 using MBST, ST-Unsupervised or ST-Supervised on the "ballroom dancer" test-set..*

	Prior tempo 120						No Prior tempo					
	MBST 22/23/		ST Unsuperv		ST Supervise		MBST 22/23/		ST Unsuperv		ST Supervise	
	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2
ChaChaCha	99%	**100%**	**100%**	**100%**	97%	97%	5%	88%	86%	97%	85%	97%
Jive	**85%**	92%	27%	**100%**	22%	95%	78%	80%	83%	90%	82%	88%
Quickstep	6%	85%	0%	**100%**	0%	100%	84%	88%	**87%**	93%	**87%**	91%
Rumba	72%	94%	**87%**	92%	86%	91%	6%	94%	30%	98%	31%	**99%**
Samba	3%	66%	**95%**	**95%**	92%	92%	2%	80%	65%	78%	73%	86%
Tango	99%	99%	**100%**	**100%**	**100%**	**100%**	80%	99%	99%	**100%**	99%	**100%**
Viennese Waltz	**83%**	**92%**	0%	85%	0%	83%	91%	**92%**	82%	92%	91%	**97%**
Slow Waltz	**68%**	85%	60%	61%	60%	61%	45%	76%	64%	95%	67%	**96%**
Mean over class	65%	89%	59%	92%	57%	90%	49%	87%	74%	93%	**77%**	**94%**

Table 2: *Tempo estimation in terms of Accuracy 1 and 2 per class using MBST, ST-Unsupervised or ST-Supervised on the "ballroom dancer" test-set..*

Using prior tempo probabilities: The baseline results obtained with the MBST are Acc1=65% (Acc2=89.4%) [1]. Using the ST-Unsupervised leads to 63.9% (90.7%), i.e. a slightly better Accuracy 2 than MBST but a lower Accuracy 1. Using the ST-Supervised leads to 62.5% (89.1%), i.e. lower Accuracy 1 and 2 than MBST and ST-Unsupervised.

As can be seen in Table 2, ST-Unsupervised provides a perfect Accuracy2 for the classes ChaChaCha, Jive, Quickstep and Tango (100%). This explains the mean-over-class-Accuracy2 of 92% (global average Accuracy2 of 90.7%). However the Accuracy1 obtained for the classes Jive, Quickstep and Viennese-Waltz is very low. The ST-Unsupervised approach seems to suffer from important octave errors.

Using uniform tempo probabilities: Because we did not observe these octave errors in our KNN-approach of [15] and because this KNN-approach do not use any prior tempo probability[2] we redo the same experiment neglecting the prior tempo probability, i.e. we set the prior tempo probability to a uniform probability.

The MBST approach now leads to 44.0% (87.1%) while the ST-Unsupervised approach leads to **72.9%** (**93.4%**) and the ST-Supervised to **75.2%** (**94.8%**). Therefore, the use of prior information has an inverse influence on the MBST and ST-Unsupervised/ST-Supervised approaches: a prior tempo probability is beneficial for the MBST approach while a uniform tempo probability is beneficial for the ST-Unsupervised/ST-Supervised approaches.

4. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed the use of Spectral Templates obtained by unsupervised or supervised learning on a database annotated into tempo. The results obtained with the ST-Unsupervised and ST-Supervised approaches without prior tempo probability improve upon previously published results. The best results obtained during the ISMIR-2004 [17] tempo induction contest on this test-set were 63.2% (92%). The results obtained here are above: 72.9% (93.4%) with ST-Unsupervised and 75.2% (94.8%) with ST-Supervised. It should be noted however that the results presented here at not directly comparable to the one of [17] since we use a learning stage which use the characteristics of the test-set.

The performances obtained with the ST-Unsupervised approach, while lower than the ones obtained with the ST-Supervised, are very promising since ST-Unsupervised does not require a database labeled into classes (of music genre or rhythm genre). Any database annotated into tempo can therefore be used for the creation of the reference spectral templates m_j.

Future works will therefore concentrate on extending the set of m_j and on testing the validity of the proposed approach when the test-set has not been used for the creation of the reference spectral templates.

[1] It should be noted that the results indicated here are not directly comparable with the ones published in [1] for the same test-set. Indeed, while in [1] we have used the DFT/FM-ACF as periodicity observation, we use here the DFT instead in order to be able to use the spectral templates proposed in [14].

[2] It should be noted that prior tempo information is somehow indirectly encoded by the dependency between m_j and B_j and the fact that we only consider the m_j with a B_j close to the tempo assumption b_i.

5. REFERENCES

[1] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. Special Issue on Music Information Retrieval Based on Signal Processing, pp. Article ID 67215, 14 pages, 2007, doi:10.1155/2007/67215.

[2] MIREX, "Audio beat tracking contest," 2009.

[3] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. of IEEE ICASSP*, Phoenix, Arizona, USA, 1999, pp. 3089–3092.

[4] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[5] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, 1998.

[6] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. of ISMIR*, Paris, France, 2002, pp. 150–156.

[7] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[8] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, no. 1, pp. 39–58, 2001.

[9] F. Gouyon, P. Herrera, and P. Cano, "Pulse-dependent analyses of percussive music," in *Proc. of AES 22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002, pp. 396–401.

[10] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *J. Audio Eng. Soc.*, vol. 51, no. 4, pp. 226–233, 2003.

[11] M. Wright, W. Schloss, and G. Tzanetakis, "Analyzing afro-cuban rhythms using rotation-aware clave template matching with dynamic programming," in *Proc. of ISMIR*, Philadelphia, PA, USA, 2008, pp. 647–652.

[12] F. Gouyon and S. Dixon, "A review of rhythm description systems," *Computer Music Journal*, vol. 29, no. 1, pp. 34–54, 2005.

[13] A. Eronen and A. Klapuri, "Music tempo estimation with k-nn regression," *IEEE Trans on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 50–57, 2010.

[14] G. Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *submitted to IEEE. Trans. on Audio, Speech and Language Processing*, 2009.

[15] G. Peeters, "Copy and scale mehod for doing time-localized m.i.r. estimation: application to beat-tracking," in *ACM Multimedia 2010, MML 2010: 3rd International Workshop on Machine Learning and Music*, Firenze, Italy, 2010.

[16] Ballroom-Dancers.com, ," .

[17] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 5, pp. 1832–1844, 2006.