Towards a Class-Based Representation of Perceptual Tempo for Music Retrieval

Ching-Wei Chen
Media Technology Lab
Gracenote, Inc.
Emeryville, CA, USA
cwchen@gracenote.com

Kyogu Lee
Graduate School of
Convergence Science *
Seoul National University
Seoul, Korea
kglee@snu.ac.kr
* work done while at Gracenote

Ho-Hsiang Wu
Music and Audio Research Lab *
New York University
New York, NY, USA
hhw230@nyu.edu
* work done while at Gracenote

Abstract— **Tempo is a common criterion by which humans describe and categorize music, and this has spawned a large amount of research in the field of automatic tempo estimation. Most tempo estimation systems focus mainly on detecting the temporal repetition and periodicity present within a signal, and represent tempo as a count of beats-per-minute (BPM). However, in real-world music retrieval applications such as music navigation and playlist generation, a rough perceptual representation of tempo may be more appropriate than a BPM representation. In this paper, the problem of tempo estimation is presented as a statistical classification problem. Four Perceptual Tempo Classes are defined which correspond to rough semantic terms that average users may use to describe tempo. Statistical models of each class are built using low-level audio features. Experimental results show that the Perceptual Tempo Class representation outperforms several conventional BPM-based tempo estimation systems when applied to the tasks of music navigation and playlist generation.**

Keywords-music information retrieval, music classification, tempo estimation, perceptual tempo, music navigation, playlist generation, music similarity

I. INTRODUCTION

As an attribute for describing and categorizing music, tempo is one of the most basic and intuitive to humans. However, the word "tempo" is broadly defined: it sometimes refers to the rate of periodic repetition in an audio signal, usually given as a count of beats-per-minute (BPM). Tempo may also be used to express a feeling or mood, such as the Classical tempo mark "Allegro" which means "quick and lively", or "Largo" which means "slow and stately". People may also associate a tempo with an activity or experience, for example, jogging is fast, reading is slow, excitement is fast, and relaxation is slow.

Automatic tempo estimation techniques are useful in many music information retrieval applications, such as music transcription, auto-accompaniment, and beat matching. In these applications, a BPM-based representation of tempo is necessary to identify the exact rate of repetition of audio beat events.

However, there are other applications of music information retrieval such as music navigation, and automatic playlist generation, wherein a BPM representation may be unnecessary, or even inappropriate. Tempo-based music navigation allows a user to organize and browse through a collection of music based on the tempo of each song. Tempo-based playlisting allows a user to create a grouping or sequence of songs that match the perceptual tempo of a specified seed song, or that fit a particular activity or situation that they may associate with a particular tempo; for example to create a "fast playlist" for jogging, or a "slow playlist" for relaxing. Users of these applications are more likely to select songs using descriptive terms such as "slow", or "fast", rather than a particular BPM value, or even a BPM range. Therefore, in these applications, the objective is simply to identify and group songs that a listener considers to share roughly similar tempo qualities, rather than to extract a precise measure of the rates of repetition present in the recording.

While automatic tempo estimation methods exist that can accurately detect tempo in the form of a BPM repetition rate, there are many challenges to using these BPM estimates to describe the perceptual tempo qualities of a song. A good example of these challenges is illustrated by considering the following two songs: "New Favorite" by Alison Krauss and Union Station, and "She Is Beautiful" by Andrew W.K.. "New Favorite" has sparse instrumentation consisting of acoustic guitars and a gentle female voice. While the vocal melody is sung at a very slow pace, there is a prominent muted guitar being strummed at a rate of roughly 162 BPM. "She Is Beautiful" has a much fuller arrangement, including distorted electric guitar and a full drum kit, and features an aggressive male vocal lead. The drums beats are very prominent, and have a repetition rate of roughly 166 BPM. When these songs are analyzed by 4 commercially available tempo estimation systems (beaTunes[1] , The Echo Nest API[2] , MARSYAS[3] , and MixMeister BPM Analyzer[4]), the tempo estimated by 3 out of the 4 systems is between 162-166 BPM for both songs. Considering these BPM estimates alone, one may be led to believe that the tempos of these two songs are very similar. However, when 5 test subjects were asked to describe each song as either "Slow" or "Fast", "New Favorite" was unanimously described as "Slow", while "She Is Beautiful" was unanimously described as "Fast". Clearly

[1] *http://www.beatunes.com*
[2] *http://developer.echonest.com*
[3] *http://marsyas.sness.net*
[4] *http://www.mixmeister.com/bpmanalyzer*

human perception of tempo involves more factors than are captured in a BPM rate.

In this paper we introduce the Perceptual Tempo Class (PTC), a class-based representation of perceptual tempo, distinct from the traditional BPM representation, which can be used to categorize and group songs that a listener perceives to share a similar tempo quality. The proposed system uses a Gaussian mixture model classifier to categorize audio signals into these Perceptual Tempo Classes based on low-level audio features. Tempo-based music navigation and playlist generation tasks are simulated to compare the results of our system against several commercially available tempo estimation systems.

II. Related Work

A. Tempo Estimation

There has been much research dedicated to automated tempo estimation (see [1] and [2] for surveys). One approach to tempo estimation from acoustic signals (such as [3]) works by first detecting significant onset events in the frequency domain, and then analyzing those events to find the onset interval that best represents the tempo of the song. Other approaches do not explicitly detect onset events, but rather directly analyze the envelope of the audio signal in the spectral or sub-band domain. Different approaches then use a resonator filterbank [4], an autocorrelation function [5], or self-similarity analysis [6] to estimate the prominence or strength of repetitions in the envelope signal, over a range of repetition rates, or lag times.

At the 2004 International Conference on Music Information Retrieval (ISMIR), a large-scale evaluation of several tempo estimation systems was conducted, comparing the estimated tempos of submitted systems against a ground truth dataset of expert annotated tempos. The ground truth tempo was given as a single BPM value, derived from the ``foot-tapping rate'' of the annotator, and was intended to represent the most salient tempo within the music, as perceived by a listener. Accordingly, the output of each tempo estimation algorithm was a single BPM estimate.

However, as noted in the results and discussion of this evaluation [7], the most common error in tempo estimation algorithms is a doubling or halving of the BPM, or octave errors. This is understandable as most popular music is polyphonic and polyrhythmic, and features several instruments playing notes at rates that are often integrally-related multiples of each other. Other integer ratio errors (4/3, 2/3, and to a lesser extent, 1/3, 3/2, and 3/1) are also common, and can be attributed to misidentification of the "down" beats, as well as triple and other compound meters. These common errors mean that even a small error in the tempo estimation algorithm (e.g. detecting only alternate beats) may result in a large error in the estimated BPM (e.g. doubling or halving).

B. Perceptual Tempo Estimation

In recent years, the concept of perceptual tempo has garnered more attention [2][8]. In the Music Information Retrieval Evaluation eXchange (MIREX) evaluation of tempo extraction in 2005 [9] and 2006, the ground truth was updated to include the two most salient perceptual BPMs, along with a weighting factor indicating the relative perceptual significance of each BPM. The performance metric was also updated to assign partial scores for correctly identifying one of the two annotated BPMs, and also considered the relative weighting factor between the two estimates.

These changes correctly acknowledged that in most popular music there is no single "correct answer" for tempo detection, but rather that more than one repetition rate may be regarded as an accurate description of the tempo of a song. The updated performance metric also accounted for the known prevalence of integer ratio errors, which allowed tempo estimation algorithms to achieve higher accuracy ratings. However, while the evaluated tempo estimation algorithms did incorporate this concept of multiple BPM estimates and weighting factor, they still did not directly address the underlying question of what causes a listener to perceive one piece of music as being "faster" or "slower" than another.

It is only more recent work that has demonstrated the effect that mood [10], timbre [11], and rhythmic pattern [12] have on identifying the most perceptually salient BPM rate in a piece of music. These methods typically begin by obtaining several likely BPM hypotheses using traditional periodicity-based tempo estimation techniques, and then apply statistical classification of other non-periodicity-based features to select the most perceptually salient BPM out of those hypotheses.

These recent methods are certainly promising approaches to modeling human perception of tempo. However, given that the objective of tempo-based music navigation and playlist generation is simply to group music into perceptual categories such as "Slow" or "Fast", their use of a BPM representation of tempo is unnecessary, and in some ways an answer to the wrong question.

III. Proposed System

We define four Perceptual Tempo Classes as shown in Table I, where each class is defined using descriptive and semantic terms, without referring to a BPM value or range. These classes are intended to correspond to descriptive terms that a listener might use in a tempo-based music navigation or playlist generation application.

The proposed system is built around a statistical model of

TABLE I. Proposed Perceptual Tempo Classes

Perceptual Tempo Class	Name	Description
PTC1	Very Slow	Unambiguously and extremely slow
PTC2	Somewhat Slow	Slightly ambiguous tempo, but relatively slow
PTC3	Somewhat Fast	Slightly ambiguous tempo, but relatively fast
PTC4	Very Fast	Unambiguously and extremely fast

four perceptual tempo classes derived from frame-based low-level audio features. We use four standard audio features widely used in music signal processing applications, including Mel-Frequency Cepstral Coefficients (MFCCs), Audio Spectral Envelope (ASE), Spectral Flatness Measure (SFM), and Chromagram. These features correspond to the timbre, noisiness, and harmonic qualities of the audio. To contrast our method with traditional approaches to tempo estimation, we intentionally avoid the use of periodicity-based features.

All audio signals are obtained in PCM format, down-sampled to 11025 Hz and down-mixed to a mono signal. We then extract the aforementioned audio features, using a 30-ms analysis window with no overlap. The raw feature frames are then combined into groups of 4 frames (120-ms), with a 50% group overlap. The feature vector is then composed of the mean and variance of each feature component over each 4-frame group. The final feature vector consists of 128 elements.

Linear discriminant analysis (LDA) is used to find the eigenvectors of the training set feature vector space that best separate the tempo classes from each other. The 50 eigenvectors with the highest eigenvalues are then used to transform feature vectors to a lower dimensional space before training and classification.

A Gaussian mixture model (GMM) of each tempo class is trained using two mixture components for each class. For an unseen input feature, the predicted class is selected by computing the log-likelihoods from all four GMMs, then selecting the one with the highest likelihood, as shown in (1),

$$c^* = \arg\max_c \sum_{i=1}^{N} \log(p(f_i \mid m_c)), \qquad (1)$$

where c* is the predicted class, N is the number of feature frames in a given signal, f_i is the i^{th} feature frame, and m_c is the GMM model for class c.

IV. Experiments and Results

A. Dataset and Ground Truth

Most publicly available audio datasets have ground truth tempo obtained by experts "tapping" the most salient perceived tempo, and recording this tapping rate in BPM. Since categorizing music into perceptual tempo classes is quite different from the concept of "tapping" a tempo, we have chosen to produce our own dataset and ground truth for this evaluation. The entire dataset consists of 1162 audio tracks from a variety of music genres shown in Table II. Although a majority of the songs in the dataset belong to the Rock and Pop genres, this is still a reasonably diverse distribution of genres given the broad definition of Pop and Rock, as well as the distribution of genres in the music collection of a typical user. In order to limit the effects of varying tempos across the duration of a song, as well as to reduce listening and processing times, 30-second clips of the audio tracks were used throughout the training and testing stages.

TABLE II. Distribution of Genres in Dataset

Genre	Number of tracks
Alternative	45
Blues	23
Classical	58
Country & Folk	63
Dance & House	19
Easy Listening	31
Electronica	58
Indie Rock	102
Jazz	74
Latin	20
Metal	29
New Age	37
Pop	156
Punk	55
R & B	73
Rap	31
Reggae	21
Rock	223
Soundtrack	12
Traditional	32

Of these 1162 clips, 422 clips from all genres are set aside as the training set for perceptual tempo classification. To annotate these 422 clips, 5 subjects were asked to listen to each audio clip and then label it with a number between 1 and 4, corresponding to the 4 perceptual tempo classes described in Table I. The ground truth perceptual tempo class for each clip was then taken as the rounded average of the ratings of the 5 subjects.

Table III shows the distribution of the aggregated tempo class annotations in the training set. The variance in annotations for each tempo class represents the degree of disagreement among the 5 annotators. As might be expected, the variances for tempo classes 1 and 4 are lower than those

TABLE III. Distribution of Tempo Class Annotations in the Ground Truth

Perceptual Tempo Class	Number of Tracks	Variance
PTC1	74	0.17
PTC2	131	0.27
PTC3	155	0.24
PTC4	62	0.19

for tempo classes 2 and 3, since the extreme tempo classes are by definition less ambiguous. However, the fact that the average variance for all 4 tempo classes is almost a quarter of a full tempo class is a reminder that even categorizing songs into one of only 4 perceptual tempo classes is a subjective task.

B. Classification Results

After training the tempo classifier as described in Section III, the perceptual tempo class of each of the songs in the training set was extracted using 4-fold cross-validation. The track-level results of the classifier are shown in Table IV.

The accuracy of the proposed tempo classification system, as given by the average F-score of all of the 4 tempo classes, is 63%. It is also notable that most incorrect predictions are within 1 tempo class of the ground truth. The average absolute error between the predicted tempo class and the true tempo class for the proposed tempo classifier is 0.39, or less than half of a full tempo class. Thus a "very slow" song is rarely mistaken as "fast" (either class 3 or 4), while a "very fast" song is rarely mistaken as "slow" (either class 1 or 2). This average error is also comparable to, though slightly higher than the variances in the ground truth annotations shown in Table III.

C. Tempo-based Music Navigation

The objective of the tempo-based music navigation task is to categorize the tempo of each track in the training set into the correct perceptual tempo class as annotated in the ground truth. The same 4 commercially available tempo estimation systems cited in Section I were evaluated: beaTunes, The Echo Nest API, MARSYAS, and MixMeister BPM Analyzer, abbreviated henceforth as BT, EN, MARS, and MM, respectively. Tempo estimates were extracted from the 422 clips in the training set using each of the 4 tempo estimation systems. The BT and MM systems provide a single BPM estimate, while the EN system provides a single BPM estimate and an associated confidence value, and the MARS system outputs two BPM estimates and respective confidence values.

Mapping the output of these 4 tempo estimation systems into the 4 perceptual tempo classes is not straightforward, since by definition the perceptual tempo classes do not refer to any particular BPM or BPM range. However, since a BPM value is the only measure of tempo available, the only possible approach is to define a set of BPM ranges that map into each of the perceptual tempo classes, similar to how the tempo marks in classical music are usually associated with a BPM range. For the first experiment we use a fixed set of heuristically determined BPM ranges roughly corresponding to each perceptual tempo class (< 70 BPM, 71-110 BPM, 111-150 BPM, and > 150 BPM). A second experiment is run using BPM ranges determined adaptively based on the distribution of estimated BPMs. Since the ground truth tells us the number of songs from the training set that belong to each perceptual tempo class (see Table III), we map the tracks with the 74 lowest estimated BPMs to PTC1, the tracks with the next 131 higher estimated BPMs to PTC2, and so on until the number of songs in each estimated tempo class is equal to the distribution in the ground truth.

TABLE IV. CONFUSION MATRIX OF PROPOSED PERCEPTUAL TEMPO CLASSIFIER

		Predicted Perceptual Tempo Class			
		PTC1	***PTC2***	***PTC3***	***PTC4***
Ground truth	***PTC1***	45	29	0	0
	PTC2	18	84	29	0
	PTC3	3	30	98	24
	PTC4	2	2	18	40

TABLE V. CONFUSION MATRIX OF MARS SYSTEM USING FIXED BPM RANGES

		Estimated BPM Ranges			
		"Very Slow" < 70 BPM	***"Somewhat Slow" 71-110 BPM***	***"Somewhat Fast" 111-150 BPM***	***"Very Fast" > 150 BPM***
Ground truth	***PTC1***	13	23	13	25
	PTC2	35	34	28	34
	PTC3	27	36	52	40
	PTC4	5	18	15	24

Table V shows the confusion matrix between the ground truth perceptual tempo classes and the mapped tempo class estimated by the MARS system using the fixed BPM range approach. The confusion matrices of the other tempo estimation algorithms show similar trends, and are not shown here.

Looking at Table V, we see a significantly lower precision and recall than the proposed tempo classification system (Table IV), as well as a much higher incidence of 2- and 3-class prediction errors, meaning that more "very slow" songs are mistaken for "fast" songs, and more "very fast" songs are mistaken for "slow" songs. Results using the adaptive approach for mapping BPMs to perceptual tempo classes remain roughly the same, and the confusion matrices are not shown here.

The average F-score accuracies across all 4 tempo classes and the average class error for all systems, using both the fixed and adaptive BPM mapping approaches, are summarized in Table VI. The proposed tempo classification system achieves about 1.7 times the accuracy and less than half the average class error of the best BPM-based tempo estimation system.

D. Tempo-based Playlist Generation

To simulate a playlist generation task, 20 seed songs were selected from outside the entire 1162 song dataset. The 740 unlabeled test set songs were used as the pool from which playlist candidates are selected. To evaluate systems on each perceptual tempo class, the 20 seed tracks were composed of 5 tracks from each of the 4 perceptual tempo

classes, as annotated by the same 5 subjects who labeled the training set.

The tempos of each of the 740 songs in the test set and the 20 seed songs were then estimated using the same 4 tempo estimation systems under evaluation in the previous experiment. For each seed song, a tempo distance was computed between the seed song's tempo and each of the songs in the test set. This tempo distance was calculated as the absolute difference in BPM between the two estimated tempos. The 10 songs with the smallest distance from the seed song were chosen as the resulting playlist.

To test the proposed tempo classification system, the perceptual tempo class of each of the 740 songs in the test set and the 20 seed songs was predicted using the classifier described in Section 3. For each seed song, a playlist was generated by randomly selecting 10 songs from the test set that were classified with the same tempo class as the seed song.

As an experimental control (abbreviated as CT), 20 playlists were generated by randomly selecting 10 songs from the entire test set for each seed song.

Given 20 seed songs and 6 algorithms, a total of 120 playlists were generated, each containing the seed song along with 10 candidate songs.

The playlists were presented by a test operator to 3 test subjects in a single-blind manner. The test subjects were first asked to listen to the seed song of each playlist. They then listened to each of the candidates in the playlist, and were asked to mark songs they felt did not fit with the perceptual tempo of the seed song. Finally, the results from all test subjects were collated, and the rejection rate for each playlist was used as a metric for comparing the performances of the evaluated tempo estimation algorithm.

Fig. 1 shows the mean and standard deviation of the number of rejected songs from the playlists generated by the 6 algorithms used in evaluation. The first four sets of bars show the rejection rates for each perceptual tempo class and the last set shows the rejection rates averaged across all classes.

TABLE VI. RESULTS OF TEMPO-BASED MUSIC NAVIGATION TASK

Tempo Estimation System	Average Accuracy	Average Class Error
BT – Fixed	20%	1.17
BT – Adaptive	24%	1.03
EN – Fixed	38%	0.82
EN – Adaptive	37%	0.88
MARS – Fixed	29%	1.07
MARS – Adaptive	25%	1.00
MM – Fixed	27%	0.81
MM – Adaptive	27%	0.98
Proposed System	63%	0.39

It may be seen from Fig. 1 that the proposed system PP consistently outperforms all other systems. Taking all 120 playlists into consideration, we see that the subjects on average rejected about 3 songs out of 10 from the playlists generated by the proposed system (PP). Random generation of playlists (CT) resulted in about 6.5 rejections on average, which was the highest among all six systems. The other 4 evaluated systems produced roughly 5-6 rejected songs in a playlist on average, where EN performed best and MM yielded the worst results.

V. DISCUSSION AND FUTURE WORK

The results in Fig. 1 show clear differences in performance for the different perceptual tempo classes. The 4 BPM-based algorithms all perform their worst (have the

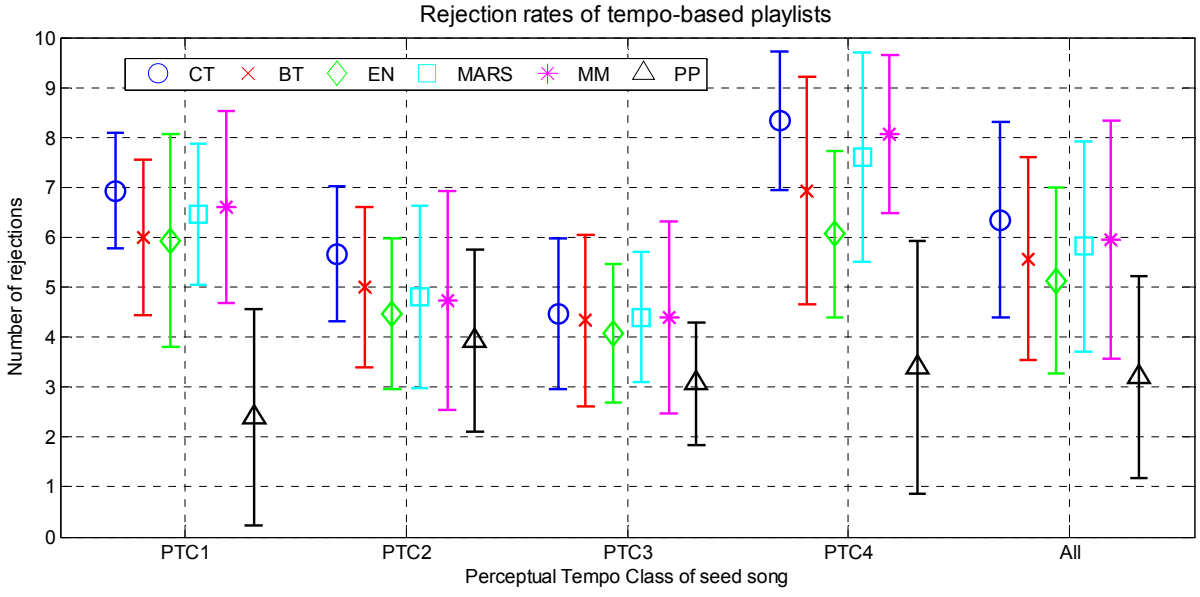

Figure 1. Mean and standard deviation of the rejection rates from six systems, for each perceptual tempo class, and for the average over all classes. CT = Control, BT = beaTunes, EN = Echo Nest, MARS = Marsyas, MM = MixMeister, and PP = Proposed System.

highest rejection rates) in PTC1 and PTC4. This may be explained by the fact that BPM-based systems are prone to octave errors, which causes the systems to mistake "very slow" songs for a "fast" song, and vice versa, which is highly objectionable to the listener. The fact that the Control system also exhibits the highest rejection rates in PTC1 and PTC4 may be explained simply by the relatively fewer numbers of "Very Slow" and "Very Fast" songs in the test set (as is true of the training set, as shown in Table III). Meanwhile, the proposed tempo classification system manages to achieve much better performances than the other systems in these same perceptual tempo classes. This suggests that the frame-level audio features used in the statistical classifier do capture aspects of the signal that humans use to distinguish between "Very Slow" and "Very Fast" music. It also suggests that traditional periodicity-based features may not be as important for making these same distinctions. In future work it would be valuable to experiment with different audio features and attributes, such as loudness, percussiveness, and genre, as inputs to the classifier.

Another observation from Fig. 1 is the relatively large variances in the rejection rates in all perceptual tempo classes and across all tempo estimation algorithms. We believe the high variance is due to a relatively small number of seed songs and test subjects. It would be valuable in future work to repeat the experiments using a larger number of seeds and test subjects, which may reduce these variances.

We hope to provide a means for sharing the dataset used in this paper with the research community in the near future. In the meantime, information about the datasets used in this paper may be obtained on request by contacting the authors.

VI. Conclusion

In this paper, we presented the perceptual tempo estimation problem as a classification problem, based on the hypothesis that a class-based representation of perceived tempo is more appropriate than the BPM representation in certain music information retrieval applications such as music navigation and playlist generation. Experimental results show that the proposed system consistently outperforms traditional BPM-based methods in both of these tasks.

References

[1] F. Gouyon, S. Dixon: "A review of automatic rhythm description system," Computer Music Journal, vol. 29, no. 1, pp.34-54, 2005.

[2] B.Y. Chua, G. Ku: "Determination of Perceptual Tempo of Music", Lecture Notes in Computer Science, vol. 3310, pp. 61-70, Springer-Verlag , 2005.

[3] M. Goto, Y. Muraoka: "Music Understanding At The Beat Level – Real Time Beat Tracking For Audio Signals", Working Notes of the IJCAI-95 Workshop on Computational Auditory Scene Analysis, pp. 68-75, 1995.

[4] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals", Journal of the Acoustical Society of America, vol. 103, no. 1, pp. 588-601, 1998.

[5] G. Tzanetakis, G. Essl, P. Cook, "Automatic Musical Genre Classification Of Audio Signals", Proceedings of the International Symposium on Music Information Retrieval, 2001.

[6] J. Foote, S. Uchihashi, "The Beat Spectrum: A New Approach To Rhythm Analysis", Proceedings of the International Conference on Multimedia and Expo, 2001.

[7] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano: "An Experimental Comparison of Audio Tempo Induction Algorithms", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1832-1844, 2006.

[8] D. Moelants, M. McKinney: "Tempo Perception and Musical Content: What Makes a Piece Fast, Slow, or Temporally Ambiguous?", Proceedings of the 8th International Conference on Music Perception & Cognition, 2004.

[9] F. Gouyon, S. Dixon: "Influence of Input Features in Perceptual Tempo Induction", 2nd Annual Music Information Retrieval eXchange (MIREX), 2005.

[10] C.W. Chen, M. Cremer, K. Lee, P. DiMaria, H.H. Wu: "Improving Perceived Tempo Estimation By Statistical Modeling of Higher Level Musical Descriptors", Proceedings of the 126th Audio Engineering Society Convention, 2009.

[11] L. Xiao, A. Tian, W. Li, J. Zhou: "Using a Statistic Model to Capture the Association between Timbre and Perceived Tempo", Proceedings of the International Conference on Music Information Retrieval, 2008.

[12] K. Seyerlehner, G. Widmer, D. Schnitzer: "From Rhythm Patterns to Perceived Tempo", Proceedings of the International Conference on Music Information Retrieval, 2007.