Predicting agreement and disagreement in the perception of tempo

Geoffroy Peeters and Ugo Marchand

STMS - IRCAM - CNRS - UPMC
geoffroy.peeters@ircam.fr, ugo.marchand@ircam.fr,
WWW home page: http://www.ircam.fr

Abstract. In the absence of a music score, tempo can only be defined in terms of its perception. Thus recent studies have focused on the estimation of perceptual tempo such as defined by listening experiments. So far, algorithms have been proposed to estimate the tempo when people agree on it. In this paper, we study the case when people disagree on the perception of tempo and propose an algorithm to predict this disagreement. For this, we hypothesize that the perception of tempo is correlated to a set of variations of various viewpoints on the audio content: energy, harmony, spectral-balance variations and short-term-similarity-rate. We hypothesize that when those variations are coherent a shared perception of tempo is favoured and when they are not, people may perceive different tempi. We then propose various statistical models to predict the agreement or disagreement in the perception of tempo from these audio features. Finally, we evaluate the models using a test-set resulting from the perceptual experiment performed at Last-FM in 2011.

Keywords: tempo estimation, perceptual tempo, tempo agreement, disagreement

1 Introduction

Tempo is one of the most predominant perceptual elements of music. For this reason, and given its use in numerous applications (search by tempo, beat-synchronous processing, beat-synchronous analysis, musicology ...) there has been and there are still many studies related to the estimation of tempo from an audio signal (see [8] for a good overview).

While tempo is a predominant elements, Moelants and McKinney [12] highlighted the fact that people can perceive different tempi for a single track. For this reason, recent studies have started focusing on the problem of estimating the "perceptual tempo". This is usually done for the subset of audio tracks for which people agree on the tempo. In this paper we start studying the case when people disagree.

1.1 Formalisation

We denote by a an audio track and by t its tempo. The task of tempo estimation can be expressed as finding the function f such that $f(a) = \hat{t} \simeq t$. Considering that different users, denoted by u, can perceive different tempi for the same audio track, the ideal model could be expressed as $f(a, u) = \hat{t}_u \simeq t_u$.

Previous research on the estimation of perceptual tempo (see part 1.2) consider mainly audio tracks t for which the perception of the tempo is shared. This can be expressed as $f(a, \forall u) = \hat{t}$. The prediction is therefore independent of the user u.

Before attempting to create the whole model $f(a, u) = \hat{t}_u \simeq t_u$, we concentrate here on predicting the audio tracks a for which the perception is not shared: $t_u \neq t_{u'}$ or $f(a, u) \neq f(a, u')$. We consider that the disagreement on tempo perception is due to

1. the preferences of the specific users,
2. the specific characteristics of the audio track; it may contain ambiguities in the rhythm or in the hierarchical organization of it.

In the current work we only focus on the second point. We therefore estimate a function $f(a)$ which indicates this ambiguity and allows predicting whether users will share the perception of tempo (Agreement) or not (Disagreement).

1.2 Related works

One of the first studies related to the perception of tempo and the sharing of its perception is the one of Moelants and McKinney [12]. This study presents and discusses the results of three experiments where subjects were asked to tap to the beat of musical excerpts. Experiments 1 and 2 lead to a unimodal perceived tempo distribution with resonant tempo centered on 128 bpm and 140 bpm respectively[1]. They therefore assume that a preferential tempo exists around 120 bpm and that "...pieces with a clear beat around 120 bpm are very likely to be perceived in this tempo by a large majority of the listeners.". An important assumption presented in this work is that "the relation between the predominant perceived tempi and the resonant tempo of the model could be used to predict the ambiguity of tempo across listeners (and vice versa) ...if a musical excerpt contains a metrical level whose tempo lies near the resonant tempo, the perceived tempo across listeners (i.e., perceived tempo distribution) is likely to be dominated by the tempo of that metrical level and be relatively unambiguous". In our work, this assumption will be used for the development of our first prediction model. In [12], the authors have choosen a resonant tempo interval within $[110 - 170]$ bpm. As we will see in our own experiment (see part 3), these values are specific to the test-set used. In [12], a model is then proposed to predict, from acoustic analyses, the musical excerpts that would deviate from the proposed resonance model.

Surprisingly few other studies have dealt with the problem of tempo agreement/ disagreement except the recent one of Zapata et al. [19] which uses mutual

[1] Experiment 3 is performed on musical excerpts specifically chosen for their extremely slow or fast tempo and leads to a bi-modal distribution with peaks around 50 and 200 bpm. Because of this, we do not consider the results of it here.

agreement of a committee of beat trackers to establish a threshold for perceptually acceptable beat tracking.

As opposed to studies on agreement/ disagreement, a larger set of studies exists for the estimation of "perceptual tempo" (the case when user agree), perceptual tempo classes or octave error correction.

Seyerlehner proposes in [17] an instance-based machine learning approach (KNN) to infer perceived tempo. For this, the rhythm content of each audio item is represented using either a Fluctuation Patterns or an Auto-correllation function. Two audio items are then compared using Pearson correlation coefficient between their representations. For an unknown item, the K most similar items are found and the most frequent tempo among the K is assigned to the unknown item.

Xiao proposes in [18] a system for correcting the octave errors of the tempo estimation provided by a dedicated algorithm. The idea is that the timbre of a song is correlated to its tempo. Hence, the content of audio files are represented using MFCCs only. An 8-component GMM is then used to model the joint MFCC and annotated tempo T_a distribution. For an unknown track, a first tempo estimation T_e is made and its MFCCs extracted. The likelihoods corresponding to the union of the MFCCs and either T_e, $T_e/3$, $T_e/2$ $\ldots$ is evaluated given the trained GMM. The largest likelihood gives the tempo to the track.

Chen proposes in [2] a method to correct automatically octave errors. The assumption used is that the perception of tempo is correlated to some moods ("aggressive" and "frantic" usually relates to "fast" tempo while "romantic" and "sentimental" relates to "slow" tempi). A system is first used to estimate automatically the mood of a given track. Four tempo categories are considered: "very slow", "somewhat slow", "somewhat fast" and "very fast". A SVM is then used to train four models corresponding to the tempi using the 101-moods feature vector as observation. Given the estimation of the tempo category, a set of rules is proposed to correct the estimation of tempo provided by an algorithm.

The work of Hockman [9] considers only a binary problem: "fast" and "slow" tempo classes. Using Last.fm A.P.I., artist and track names corresponding to the "fast" and "slow" tags have been selected. The corresponding audio signal is obtained using YouTube A.P.I. This leads to a test-set of 397 items. 80 different features related to the onset detection function, pitch, loudness and timbre are then extracted using jAudio. Among the various classifiers tested (KNN, SVM, C4.5, AdaBoost . . .), AdaBoost achieved the best performance.

Gkiokas [7] studies both the problem of continuous tempo estimation and tempo class estimation. The content of an audio signal is represented by a sophisticated feature vector. For this 8 energy bands are passed to a set of resonators. The output is summed-up by a set of filter-bank and DCT applied. Binary one-vs-one SVM classifier and SVM regression are then used to predict the tempo classes and continuous tempo. For the later, peak picking is used to refine the tempo estimation.

As opposed to previous studies, the work of Peeters et al. [15] is one of the few to study perceptual tempo estimation on real annotated perceptual tempo data (derived from the perceptual experiment performed at Last-FM in 2011). They propose four feature sets to describe the audio content and propose the use of GMM-Regression [3] to model the relationship between the audio features and the perceptual tempo.

1.3 Paper organization

The goal of the present study is to predict user Agreement or Disagreement on tempo perception using only the audio content.

For this, we first represent the content of an audio file by a set of cues that we assume are related to the perception of tempo: variation of energy, short-term-similarity, spectral balance variation and harmonic variation. We successfully validated these four functions in [15] for the estimation of perceptual tempo (in the case $f(a, \forall u) = \hat{t}$). We briefly summarized these functions in part 2.1.

In part 2.2, we then propose various prediction models to model the relationship between the audio content and the Agreement and Disagreement on tempo perception. The corresponding systems are summed up in Figure 1.

In part 3, we evaluate the performance of the various prediction models in a usual classification task into tempo Agreement and Disagreement using the Last-FM 2011 test-set.

Finally, in part 4, we conclude on the results and present our future works.

2 Prediction model for tempo Agreement and Disagreement

2.1 Audio features

We briefly summarized here the four audio feature sets used to represent the audio content. We refer the reader to [15] for more details.

Energy variation $d_{ener}(\lambda)$: The aim of this function is to highlight the presence of onsets in the signal by using the variation of the energy content inside several frequency bands. This function is usually denoted by “spectral flux” [10]. In [14] we proposed to compute it using the reassigned spectrogram [4]. The later allows obtaining a better separation between adjacent frequency bands and a better temporal localization. In the following we consider as observation, the autocorrelation of this function denoted by $d_{ener}(\lambda)$ where λ denotes “lags” in second.

Short-term event repetition $d_{sim}(\lambda)$: We make the assumption that the perception of tempo is related to the rate of the short-term repetitions of events (such as the repetition of events with same pitch or same timbre). In order to highlight these repetitions, we compute a Self-Similarity-Matrix [5] (SSM) and measure the rate of repetitions in it. In order to represent the various type of repetitions (pitch or timbre repetitions) we use the method we proposed in [13]. We then convert the SSM into a Lag-matrix [1] and sum its contributions over time to obtain the rate of repetitions for each lag. We denote this function by $d_{sim}(\lambda)$.

Spectral balance variation $d_{specbal}(\lambda)$*:* For music with drums, the balance between the energy content in high and low frequencies at a given time depends on the presence of the instruments: low > high if a kick is present, high > low when a snare is present. For a typical pop song in a 4/4 meter, we then observe over time a variation of this balance at half the tempo rate. This variation can therefore be used to infer the tempo. In [16] we propose to compute a spectral-balance function by computing the ratio between the energy content at high-frequency to the low-frequency one. We then compare the values of the balance function over a one bar duration to the typical template of a kick/snare/kick/snare profile. We consider as observation the autocorrelation of this function, which we denote by $d_{specbal}(\lambda)$.

Harmonic variation $d_{harmo}(\lambda)$*:* Popular music is often based on a succession of harmonically homogeneous segments named "chords". The rate of this succession is proportional to the tempo (often one or two chords per bar). Rather than estimating the chord succession, we estimate the rate at which segments of stable harmonic content vary. In [15] we proposed to represent this using Chroma variations over time. The variation is computed by convolving a Chroma Self-Similarity-Matrix with a novelty kernel [6] whose length represent the assumption of chord duration. The diagonal of the resulting convolved matrix is then considered as the harmonic variation. We consider as observation the autocorrelation of this function, which we denote by $d_{harmo}(\lambda)$.

Dimension reduction: The four feature sets are denoted by $d_i(\lambda)$ with $i \in \{ener, sim, specbal, harmo\}$ and where λ denotes the lags (expressed in seconds). In order to reduce the dimensionality of those, we apply a filter-bank over the lag-axis λ of each feature set. For this, we created 20 filters logarithmically spaced between 32 and 208bpm with a triangular shape. Each feature vector $d_i(\lambda)$ is then multiplied by this filter-bank leading to a 20-dim vector, denoted by $d_i(b)$ where $b \in [1, 20]$ denotes the number of the filter. To further reduce the dimensionality and de-correlate the various dimensions, we also tested the application of the Principal Component Analysis (PCA). We only keep the principal axes which explain more than 10% of the overall variance.

2.2 Prediction models

We propose here four prediction models to model the relation-ship between the audio feature sets (part 2.1) and the Agreement and Disagreement on tempo perception. The four prediction models are summed up in Figure 1.

2.2.1. Model MM (Ener and Sim): As mentioned in part 1.2, our first model is based on the assumption of Moelants and McKinney [12] that "if a musical excerpt contains a metrical level whose tempo lies near the resonant tempo, the perceived tempo across listeners is likely to be dominated by the tempo of that metrical level and be relatively unambiguous". In [12], a resonant tempo interval is defined as $[110 - 170]$ bpm. Our first prediction model hence looks if

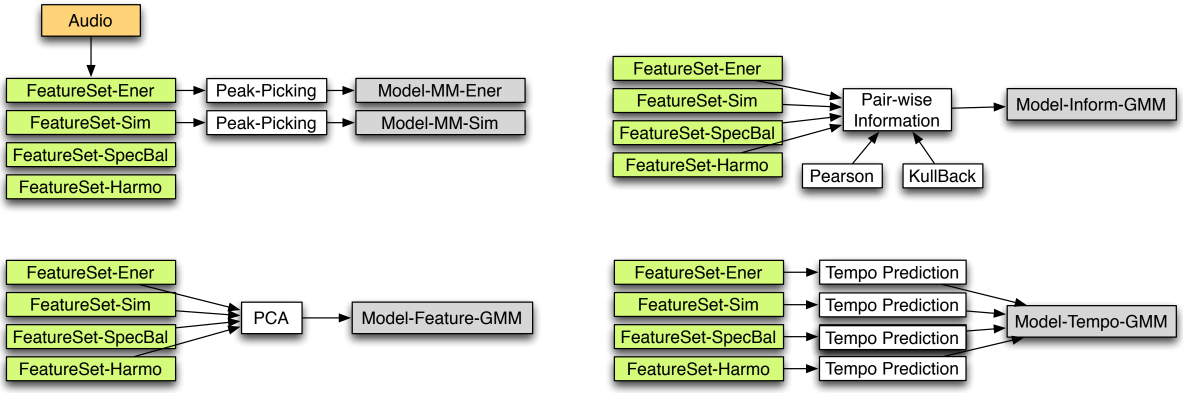

Fig. 1. Flowchart of the computation of the four prediction models

a major peak of a periodicity function exists within this interval. For this, we use as observations the audio feature functions in the frequency domain: $d_i(\omega)$ (i.e. using the DFT instead of the auto-correlation) and without dimensionality reduction. We then look if one of the two main peaks of each periodicity function $d_i(\omega)$ lies within the interval $[110-170]$ bpm. If this is the case, we predict an Agreement on tempo perception; if not, we predict Disagreement.

By experiment, we found that only the two audio feature $d_{ener}(\omega)$ and $d_{sim}(\omega)$ lead to good results. This leads to two different models: MM (ener) or MM (sim).

Illustration: We illustrate this in Figure 2 where we represent the function $d_{ener}(\omega)$, the detected peaks, the two major peaks, the $[110-170]$ bpm interval (green vertical lines) and the preferential 120 bpm tempo (red dotted vertical line). Since no major peaks exist within the resonant interval, this track will be assigned to the Disagreement class.

2.2.2. Model Feature-GMM: Our second model is our baseline model. In this, we estimate directly the Agreement and Disagreement classes using the audio features $d_i(b)$. In order to reduce the dimensionality we apply PCA to the four feature sets[2]. Using the reduced features, we then train a Gaussian Mixture Model (GMM) for the class Agreement (A) and Disagreement (D). By experimentation we found that the following configuration leads to the best results: 4-mixtures for each class with full-covariance matrices. The classification of an unknown track is then done by maximum-a posteriori estimation.

2.2.3. Model Inform-GMM (Pearson and KL): The feature sets $d_i(b)$ represent the periodicities of the audio signal using various view points i. We assume that

[2] As explained in part 2.1, we only keep the principal axes which explain more than 10% of the overall variance. This leads to a final vector of 34-dimensions instead of 4*20=80 dimensions.

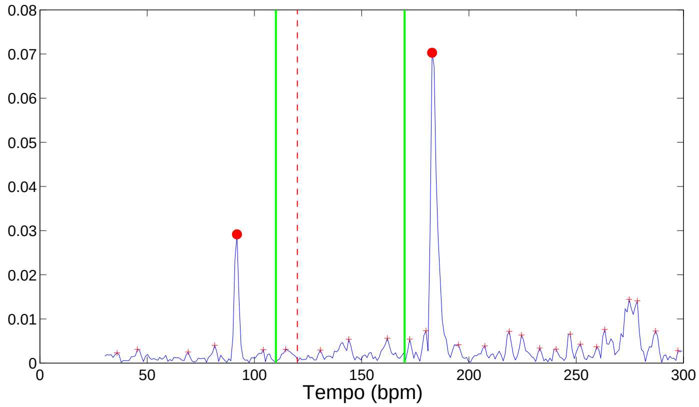

Fig. 2. Illustration of the Model MM (ener) based on Moelants and McKinney assumption [12].

if two vectors $\underline{d}_i$ and $\underline{d}_{i'}$ bring the same information on the periodicity of the audio signal, they will also do on the perception of tempo, hence favoring a shared (Agreement) tempo perception.

In our third model, we therefore predict A and D by measuring the information shared by the four feature sets. For each track, we create a 6-dim vector made of the information shared between each pair of feature vector $\underline{d}_i$: $\underline{C} = [c(\underline{d}_1, \underline{d}_2), c(\underline{d}_1, \underline{d}_3), c(\underline{d}_1, \underline{d}_4), c(\underline{d}_2, \underline{d}_3) \ldots]$. In order to measure the shared information, we will test for c the use of the Pearson correlation and of the symmetrized Kullback-Leibler divergence (KL) between $\underline{d}_i$ and $\underline{d}_{i'}$.

The resulting 6-dim vectors $\underline{C}$ are used to train a GMM (same configuration as before) for the class Agreement (A) and Disagreement (D). The classification of an unknown track is then done by maximum-a posteriori estimation.

Illustration: In Figure 3, we illustrate the correlation between the four feature sets for a track belonging to the Agreement class (left) and to the Disagreement class (right)[3]. As can be seen on the left (Agreement), the positions of the peaks of the ener, sim and specbal functions are correlated to each other's. We assume that this correlation will favour a shared perception of tempo. On the right part (Disagreement), the positions of the peaks are less correlated. In particular the sim function has a one-fourth periodicity compared to the ener function, the specbal a half periodicity. We assume that this will handicap a shared perception of tempo.

2.2.4. Model Tempo-GMM: Our last prediction model is also based on measuring the agreement between the various view points i. But instead of predicting this

[3] It should be noted that for easiness of understanding we represent in Figure 3 the features $d_i(\lambda)$ while the $\underline{C}$ is computed on $d_i(b)$.

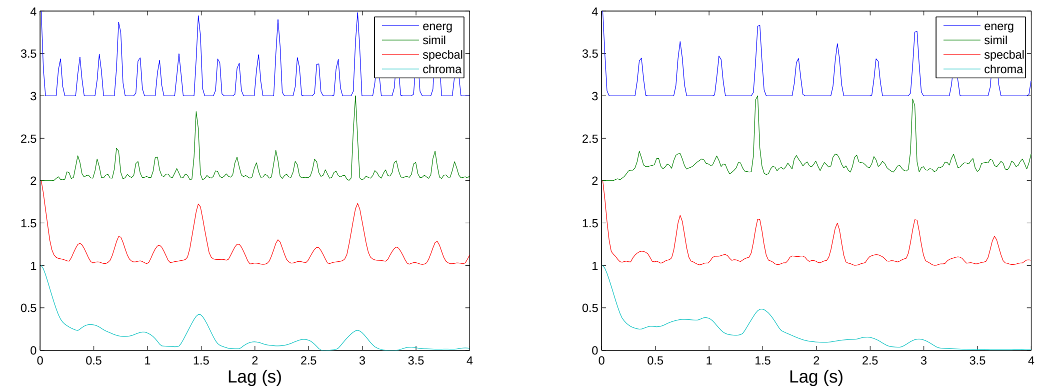

Fig. 3. [Left part] from top-to-bottom ener, sim, specbal and harmo functions for a track belonging to the Agreement class; [right part] same for the Disagreement class.

agreement directly from the audio features (as above), we measure the agreement between the tempo estimation obtained using the audio features independently.

For this, we first create a tempo estimation algorithm for each feature sets: $\hat{t}_i = f(d_i(\lambda))$. Each of these tempo estimation is made using our previous GMM-Regression methods as described in [15]. Each track a is then represented by a 4-dim feature vector where each dimension represent the prediction of tempo using a specific feature set: $[\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{specbal}, \hat{t}_{harmo}]$. The resulting 4-dim vectors are used to train the final GMM (same configuration as before) for the class Agreement (A) and Disagreement (D). The classification of an unknown track is then done by maximum-a posteriori estimation.

3 Experiment

We evaluate here the four models presented in part 2.2 to predict automatically the Agreement or Disagreement on tempo perception using only the audio content.

3.1 Test-Set

In the experiment performed at Last-FM in 2011 [11], users were asked to listen to audio extracts, qualify them into 3 perceptual tempo classes and quantify their tempo (in bpm). We denote by $t_{a,u}$ the quantified tempo provided by user u for track a. Although not explicit in the paper [11], we consider here that the audio extracts have constant tempo over time and that the annotations have been made accordingly. The raw results of this experiment are kindly provided by Last-FM. The global test-set of the experiment is made up of 4006 items but not all items were annotated by all annotators. Considering the fact that these

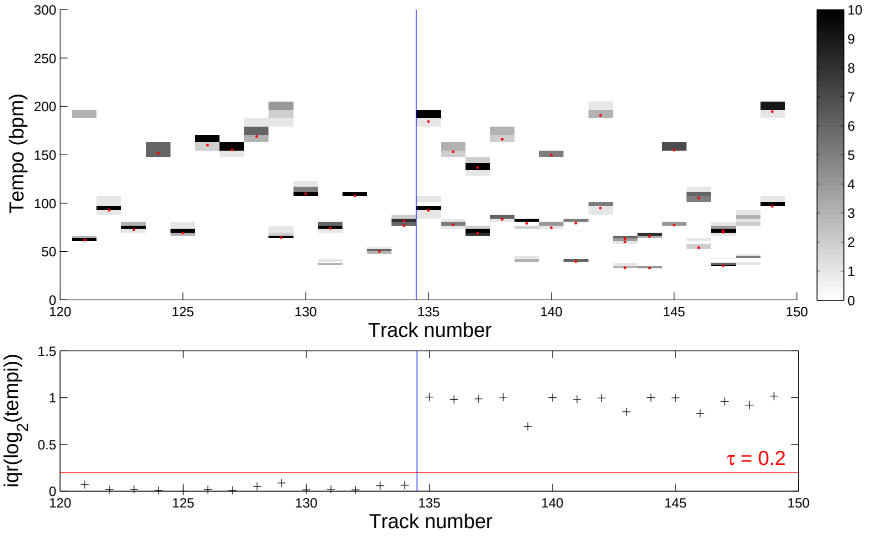

Fig. 4. [Top part] For each track a we represent the various annotated tempi $t_{a,u}$ in the form of a histogram. [Bottom part] For each track a, we represent the computed IQR_a. We superimposed to it the threshold τ that allows deciding on the assignment of the track to the Agreement (left tracks) or Disagreement (right part).

annotations have been obtained using a crowd-sourcing approach, and therefore that some of these annotations may be unreliable, we only consider the subset of items a for which at least 10 different annotations u are available. This leads to a subset of 249 items.

For copyright reason, the Last-FM test-set is distributed without the audio tracks. For each item, we used the 7-Digital API in order to access a 30s audio extract from which audio features has been extracted. This has been done querying the API using the provided artist, album and title names.We have listened to all audio extracts to confirm the assumption that their tempi are constant over time

Assigning a track to the Agreement or Disagreement class: We assign each audio track a to one of the two classes Agreement (A) or Disagreement (D) based on the spread of the tempo annotations $t_{a,u}$ for this track. This spread is computed using the Inter-Quartile-Range (IQR)[4] of the annotations expressed in log-scale[5]:

[4] The IQR is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles. It is considered more robust to the presence of outliers than the standard deviation.

[5] The log-scale is used to take into account the logarithmic character of tempo. In log-scale, the intervals $[80 - 85]$ bpm and $[160 - 170]$ bpm are equivalent.

$\mathrm{IQR}_a\,(\log_2(t_{a,u}))$. The assignment of a track a to one the two classes is based on the comparison of IQR_a to a threshold τ. If $\mathrm{IQR}_a < \tau$, Agreement is assigned to track a, if $\mathrm{IQR}_a \leq \tau$, Disagreement is assigned. By experimentation we found $\tau = 0.2$ to be a reliable value. This process leads to a balanced distribution of the test-set over classes: #(A)=134, #(D)=115.

Illustration: In Figure 4 we represent the histogram of the tempi $t_{a,u}$ annotated for each track a and the corresponding IQR_a derived from those.

3.2 Experimental protocol

Each experiment has been done using a five-fold cross-validation, i.e. models are trained using 4 folds and evaluated using the remaining one. Each fold is tested in turn. Results are presented as mean value over the five-folds. When GMM are used, in order to reduce the sensitivity on the initialization of the GMM-EM algorithm, we tested 1000 random initializations.

In the following, we present the results of the two-classes categorization problem (A and D) in terms of class-Recall[6] (i.e. the Recall of each class) and in terms of mean-Recall, i.e. mean of the class-Recalls[7].

3.3 Results

Results are presented in Table 1. For comparison, a random classifier for a two-class problem would lead to a Recall of 50%. As can be seen, only the models MM (Sim), Inform-GMM (KL) and Tempo-GMM lead to results above a random classifier. The best results (mean Recall of 70%) are obtained with the model Tempo-GMM (predicting the Agreement/Disagreement using four individual tempo predictions). This model largely exceeds the other models.

Table 1. Results of classification into Agreement and Disagreement using five-fold cross-validation for the various prediction models presented in part 2.2.

Model	Recall(A)	Recall(D)	Mean Recall
MM (Ener)	62.69 %	42.61 %	52.65%
MM (Sim)	56.71 %	58.26 %	57.49%
Feature-GMM	55.21 %	45.22 %	50.22%
Inform-GMM (Pearson)	51.51 %	49.57 %	50.54%
Inform-GMM (KL)	61.17 %	50.43 %	55.80%
Tempo-GMM	**73.73%**	**66.52%**	**70.10%**

[6] $\mathrm{Recall} = \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

[7] As opposed to Precision, the Recall is not sensitive on class distribution hence the mean-over-class-Recall is preferred over the F-Measure.

Discussion on the results obtained with the model MM: The model MM is derived from Moelants and McKinney experiment assuming a preferential tempo around 120 bpm. Considering the bad results obtained in our experiment with this model, we would like to check the preferential tempo assumption. For this, we computed the histogram of all annotated tempi for the tracks of our test-set. This is represented in Figure 5. As can be seen, the distribution differs from the one obtained in experiments 1 and 2 from [12]. In our case, the distribution is bimodal with two predominant peaks around 87 and 175 bpm. This difference may be due to the different test-sets, experimental protocol and users. The resonant model that best fits our distribution has a frequency of 80 bpm (instead of 120 bpm in [12]). We therefore redid our experiment changing the preferential tempo interval in our prediction model to $[60 - 100]$ bpm (instead of $[110 - 170]$ bpm in [12]). However this didn't change the results in a positive way: mean-Recall(MM-Ener)=50.39%, mean-Recall(MM-Sim)=42.49%.

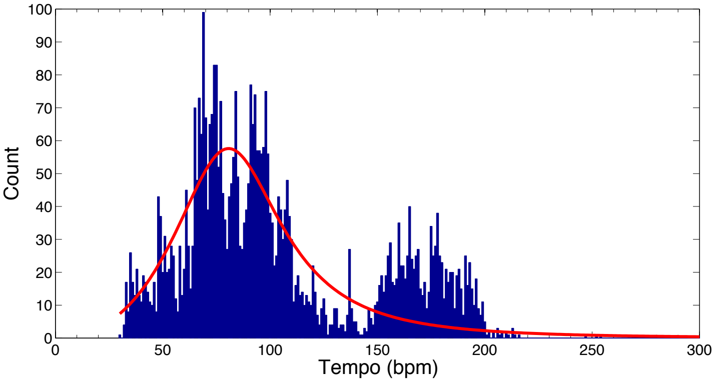

Fig. 5. Histogram of tempi annotation for the tracks of the Last-FM test-set. We superimposed to it the resonant model as proposed by Moelants and McKinney [12] with a frequency of 80 bpm.

Detailed results for the model Tempo-GMM: In Table 2, we present the detailed results in the case of the Model-Tempo-GMM. Those indicate that the class Agreement is more easily recognized than the class Disagreement. In order to have a better insight into the model, we represent in Figure 6 the relationship between the four estimated tempi $\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{specbal}, \hat{t}_{harmo}$ for data belonging to the classes Agreement (red plus sign) and Disagreement (blue crosses). As can be seen, the estimated tempi for the class Agreement are more correlated (closer to the main diagonal) than the ones for the class Disagreement (distribution mainly outside the main diagonal). This validates our assumption that the sharing of the perception of tempo may be related to the agreement between the various acoustical cues.

Table 2. Confusion matrix between class Agreement and Disagreement for Model Tempo-GMM. Results are presented in terms of number of items (not in percent).

	$\hat{T}(A)$	$\hat{N}$ (D)
T (A)	98.8	35.2
N (D)	38.5	76.5

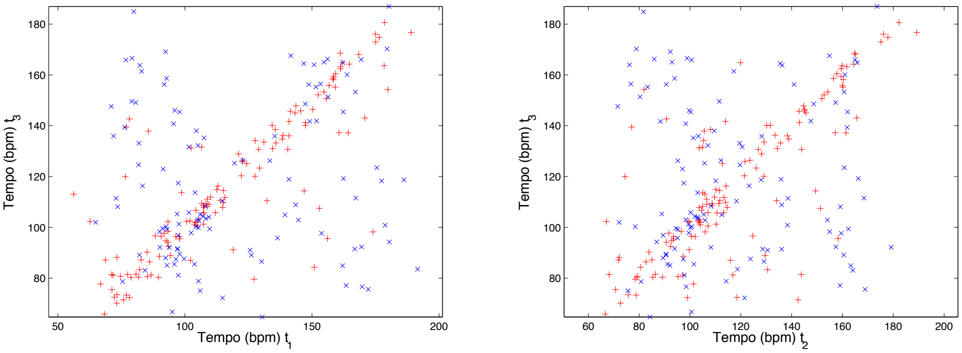

Fig. 6. Each panel represents the relationship between the estimated tempo for [left part] $t_1 = \hat{t}_{ener}/t_3 = \hat{t}_{specbal}$, [right part] $t_2 = \hat{t}_{sim}/t_3 = \hat{t}_{specbal}$. Red plus signs represent data belonging to the Agreement class, blue crosses to the Disagreement class.

4 Conclusion

In this paper, we studied the prediction of agreement and disagreement on tempo perception using only the audio content. For this we proposed four audio feature sets representing the variation of energy, harmony, spectral-balance and the short-term-similarity-rate. We considered the prediction of agreement and disagreement as a two classes problem. We then proposed four statistical models to represent the relationship between the audio feature and the two classes.

The first model is based on Moelants and McKinney [12] assumption that agreement is partly due to the presence of a main periodicity close to the user preferential tempo of 120 bpm. With our test-set (derived from the Last-FM 2011 test-set) we didn't find such a preferential tempo but rather two preferential tempi around 87 and 175 bpm. The prediction model we created using [12] assumption reached a just-above-random mean-Recall of 57% (using the sim function).

The second model predict the two classes directly from the audio features using GMMs. It performed the same as a random two-class classifier.

The third and fourth model use the *agreement* of the various acoustical cues provided by the audio features to predict tempo Agreement or Disagreement. The third model uses information redundancy between the audio feature sets (using either Pearson correlation or symmetrized Kullback-Leibler divergence) and models those using GMM. It reached a just-above-random mean-Recall of 55% (with the symmetrized Kullback-Leibler divergence).

The fourth model uses the four feature sets independently to predict four independent tempi. GMMs is then use to model those four tempi. The corresponding model leads to a 70% mean-Recall. Detailed results showed that for the class Agreement the four estimated tempi are more correlated to each other's than for the class Disagreement. This somehow validates our assumption that the sharing of tempo perception (Agreement) is facilitated by the coherence of the acoustical cues.

Future works will now concentrate on introducing the user variable u in order to create the whole model $f(a, u) = \hat{t}_u$. However, this will require accessing data annotated by the same users u for the same tracks a.

Acknowledgements

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

References

1. M. Bartsch and G. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proc. of IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)*, pages 15–18, New Paltz, NY, USA, 2001.
2. C.W. Chen, M. Cremer, K. Lee, P. DiMaria, and H.H. Wu. Improving perceived tempo estimation by statistical modeling of higher level musical descriptors. In *Proc. of the 126th AES Convention*, Munich, Germany, 2009.
3. T. En-Najjary, O. Rosec, and T. Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *Proc. of Eurospeech*, Geneva, Switzerland, 2003.
4. P. Flandrin. *Time-Frequency/Time-Scale Analysis.* Academic Press, San Diego, California, 1999.
5. Jonathan Foote. Visualizing music and audio using self-similarity. In *Proc. of ACM Multimedia*, pages 77–80, Orlando, Florida, USA, 1999.
6. Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, pages 452–455, New York City, NY, USA, 2000.
7. Aggelos Gkiokas, Vassilis Katsouros, and George Carayannis. Reducing tempo octave errors by periodicity vector coding and svm learning. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, 2012.
8. F. Gouyon, Anssi Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(5):1832–1844, 2006.
9. Jason Hockman and Ichiro Fujinaga. Fast vs slow: Learning tempo octaves from user data. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Utrecht, The Netherlands, 2010.
10. J. Laroche. Efficient tempo and beat tracking in audio recordings. *JAES (Journal of the Audio Engineering Society)*, 51(4):226–233, 2003.

11. Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.
12. Dirk Moelants and Martin F. McKinney. Tempo perception and musical content: What makes a piece slow, fast, or temporally ambiguous? In *Proc. of ICMPC (International Conference of Music Perception and Cognition)*. Northwestern University, Evanston, Illinois (Chicago,USA), 2004.
13. Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Vienna, Austria, 2007.
14. Geoffroy Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Applied Signal Processing*, 2007(1):158–158, 2007. doi:10.1155/2007/67215.
15. Geoffroy Peeters and Joachim Flocon-Cholet. Perceptual tempo estimation using gmm regression. In *Proc. of ACM Multimedia/ MIRUM (Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies)*, Nara, Japan, November 2012.
16. Geoffroy Peeters and Hélène Papadopoulos. Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(6):1754–1769, August 2011.
17. Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer. From rhythm patterns to perceived tempo. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Vienna, Austria, 2007.
18. Linxing Xiao, Aibo Tian, Wen Li, and Jie Zhou. Using a stastic model to capture the association between timbre and perceived tempo. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Philadelphia, PA, USA, 2008.
19. José R Zapata, André Holzapfel, Matthew EP Davies, Joao L Oliveira, and Fabien Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, 2012.