Predicting Agreement and Disagreement in the Perception of Tempo

Geoffroy Peeters, Ugo Marchand

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 by its perception by users. Thus recent studies have focused on the estimation of perceptual tempo defined by listening experiments. So far, algorithms have only 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 suppose 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 several 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 element 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 [9] for a good overview).

While tempo is a predominant element, Moelants and McKinney [14] 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" and perceptual tempo classes (such as "slow", "moderate" or "fast"). This is usually done for the subset of audio tracks for which people agree on the tempo. In this paper we study the case where people disagree.

1.1 Formalisation

We denote by a an audio track and by t_a its tempo. The task of tempo estimation can be expressed as finding the function f such that $f(a) = T_a \simeq t_a$. Considering

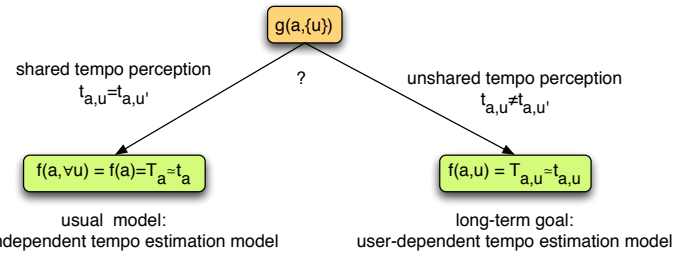


Fig. 1. $g(a, u)$ is a function that predicts tempo agreement and disagreement. Based on this prediction a user-independent or a user-dependent tempo estimation model is used.

that different users, denoted by u, can perceive different tempi for the same audio track, the ideal model can be expressed as $f(a, u) = T_{a,u} \simeq t_{a,u}$.

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

Our long-term goal is to create a user-dependent tempo prediction model $f(a, u) = T_{a,u} \simeq t_{a,u}$. As a first step toward this model, we study in this paper the prediction of the audio tracks a for which the perception is shared ($t_{a,u} = t_{a,u'}$) and for which it is not ($t_{a,u} \neq t_{a,u'}$). For this, we look for a function $g(a, \{u\})$ which can predict this shared perception for a given audio track a and a given set of user $\{u\}$ (see Figure 1). We consider that this disagreement of tempo perception is due to

1. the preferences of the specific users (which may be due to the users themselves or to the listening conditions such as the listening environment),
2. the specific characteristics of the audio track; it may contain ambiguities in its rhythm or in its hierarchical organization.

In this work we only focus on the second point. We therefore estimate a function $g(a)$ which indicates if an ambiguity exists and which can therefore be used to predict whether users will share the perception of tempo (agreement) or not (disagreement).

1.2 Related works

Studies on tempo agreement/disagreement estimation. One of the first studies related to the perception of tempo and the sharing of its perception is the one of Moelants and McKinney [14]. 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 a 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 [14], the authors have chosen a resonant tempo interval within $[110 - 170]$ bpm. During our own experiment (see part 3), we found that these values are specific to the test-set used. In [14], Moelants proposes a model to predict, from acoustic analyses, the musical excerpts that would deviate from the proposed resonance model.

Surprisingly no other studies have dealt with the problem of tempo agreement/ disagreement except the recent one of Zapata et al. [22] which uses mutual agreement of a committee of beat trackers to establish a threshold for perceptually acceptable beat tracking.

In the opposite, studies in the case of tempo agreement ($t_{a,u} = t_{a,u'}$) are numerous. In this case, the model simplifies to $f(a, \forall u) = T$ and aims at estimating "perceptual tempo", "perceptual tempo" classes or octave error correction.

Studies on "perceptual tempo" estimation. Seyerlehner [19] proposes 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.

Chua [3] distinguishes perceptual tempo from score tempo (annotated on the score) and foot-tapping tempo (which is centered around 80-100 bpm). He proposes an Improved Perceptual Tempo Estimator to determine automatically the perceptual tempo This IPTE determines the perceptual tempo (with frequency sub-band analysis, amplitude envelope autocorrelation then peak-picking) on 10 seconds-length segment, along with a likelyhood measure. The perceptual tempo is the tempo of the segment with the highest likelihood. On a test-set of 50 manually annotated musical excerpts, he evaluates his IPTE. The model failed for only 2 items.

[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 the specificities of these musical excerpts, we do not consider the results of it here.

Studies on "perceptual tempo" classes estimation. Hockman [10] considers only two classes: "fast" and "slow" tempo classes. Using Last.fm A.P.I., artists and tracks which have been assigned "fast" and "slow" tags are selected. The corresponding audio signal are then obtained using YouTube A.P.I. This leads to a test-set of 397 items. 80 different audio 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 [8] studies both the problem of continuous tempo estimation and tempo class estimation. The content of an audio signal is represented by a sophisticated set of audio features. 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 Support Vector Machine (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.

Studies on octave error correction. Chen [2] proposes a method to automatically correct octave errors. The assumption used is that the perception of tempo is correlated to the "mood" ("aggressive" and "frantic" mood usually relates to "fast" tempo while "romantic" and "sentimental" mood 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.

Xiao [21] proposes a system to correct the octave errors of the tempo estimation provided by a dedicated algorithm. The idea is that the timbre of a track is correlated to its tempo. To represent the timbre of an audio track, he uses the MFCCs. 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_a is made and its MFCCs extracted. The likelihoods corresponding to the union of the MFCCs and either T_a, $T_a/3$, $T_a/2$ ... is evaluated given the trained GMM. The largest likelihood gives the tempo of the track.

Studies that uses real annotated perceptual tempo. As opposed to previous studies, only the following work with real annotated perceptual tempo data.

McKinney [13] proposes to model the perceptual tempi assigned by the various users to a track by a histogram (instead of the single value used in previous studies). This histogram is derived from user tappings along 24 10-sec music excerpts. He then studies the automatic estimation of these histograms using 3 methods : resonator filter-bank, autocorrelation and IOI Histogram. All three

methods performs reasonably well on 24 tracks of 8 different genres. The methods usually find the first and the second largest peaks correctly, while having a lot of unwanted peaks.

Peeters et al. [17] studies the estimation of perceptual tempo using real annotated perceptual tempo data derived from the Last-FM 2011 experiment [12]. From these data, he only selects the subset of tracks for which tempo perception is shared among users ($t_{a,u} = t_{a,u'}$). He then proposes four feature sets to describe the audio content and proposes the use of GMM-Regression [4] to model the relationship between the audio features and the perceptual tempo.

1.3 Paper organization

The goal of this paper is to study the prediction of the agreement or disagreement among users on tempo perception using only the audio content. We try to predict this agreement/ disagreement using the function $g(a)$ (see Part 1.1 and Figure 1).

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 [17] for the estimation of perceptual tempo (in the case $t_{a,u} = t_{a,u'}$). We briefly summarize these functions in part 2.1.

In part 2.2, we then propose various prediction models $g(a)$ to model the relationship between the audio content and the agreement or disagreement on tempo perception. The corresponding systems are summed up in Figure 2.

In part 3, we evaluate the performance of the various prediction models in a usual classification task into tempo Agreement and tempo 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 $g(a)$ for the prediction of tempo agreement and disagreement

2.1 Audio features

We briefly summarize here the four audio feature sets used to represent the audio content. We refer the reader to [17] 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" [11]. In [15] we proposed to compute it using the reassigned spectrogram [5]. 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 [6] (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 [16]. 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 [18] 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 [17] 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 [7] 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.

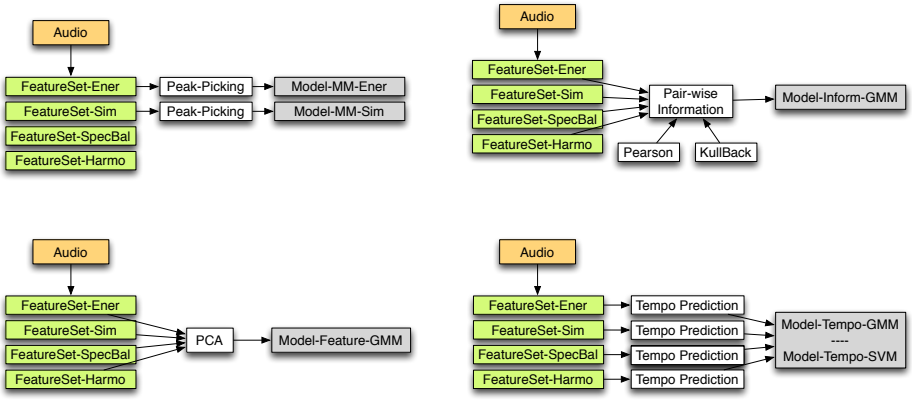


Fig. 2. Flowchart of the computation of the four prediction models

2.2 Prediction models

We propose here four prediction models to represent 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 2.

A. Model MM (Ener and Sim). As mentioned in part 1.2, our first model is based on the assumption of Moelants and McKinney [14] 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 [14], a resonant tempo interval is defined as $[110-170]$ bpm. Our first prediction model hence looks if 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 a disagreement.

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

Illustration: We illustrate this in Figure 3 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.

B. Model Feature-GMM. Our second model is our baseline model. In this, we estimate directly the agreement and disagreement classes using the audio

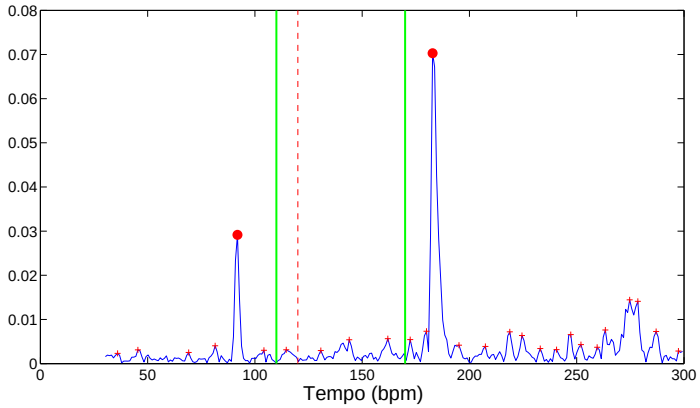


Fig. 3. Illustration of the Model MM (ener) based on Moelants and McKinney preferential tempo assumption [14].

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 ($\mathcal{A}$) and another for the class disagreement ($\mathcal{D}$). 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.

C. 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 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 $\mathcal{A}$ and $\mathcal{D}$ 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 the use 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 ($\mathcal{A}$) and disagreement ($\mathcal{D}$). The classification of an unknown track is then done by maximum-a posteriori estimation.

Illustration: In Figure 4, we illustrate the correlation between the four feature sets for a track belonging to the agreement class (left) and to the disagreement

[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.

(a) $\mathcal{A}$greement (b) $\mathcal{D}$isagreement

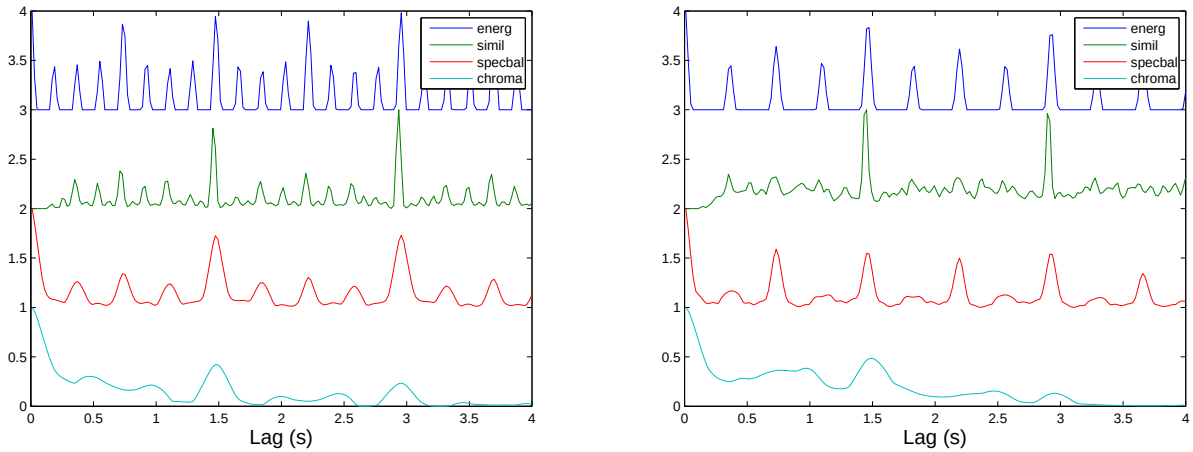


Fig. 4. [Left part] from top-to-bottom ener, sim, specbal and harmo functions for a track belonging to the agreement class; [right part] same for the disagreement class.

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.

D. Model Tempo-GMM and Model-Tempo-SVM. Our last prediction model is also based on measuring the agreement between the various view points i. But instead of predicting this 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: $T_i = f(d_i(\lambda))$. Each of these tempo estimation is made using our previous GMM-Regression methods as described in [17]. 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: $[T_{ener}, T_{sim}, T_{specbal}, T_{harmo}]$.

The resulting 4-dim vectors are used to train the final statistical model. For this, we compare two approaches:

[3] It should be noted that for easiness of understanding we represent in Figure 4 the features $d_i(\lambda)$ while the $\underline{C}$ is computed on $d_i(b)$.

– training a GMM (same configuration as before) for the class agreement ($\mathcal{A}$) and disagreement ($\mathcal{D}$); then use maximum-a posteriori estimation,
– training a binary Support Vector Machine (SVM) (we used a RBF kernel with $\gamma = 0.001$ and $C = 1.59$) to discriminate between the classes agreement ($\mathcal{A}$) and disagreement ($\mathcal{D}$).

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 [12], 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 [12], 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 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 ($\mathcal{A}$) or disagreement ($\mathcal{D}$) 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]: $\mathrm{IQR}_a\left(\log_2(t_{a,u})\right)$. 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 $\mathrm{IQR}_a \geq \tau$, disagreement is assigned. By experimentation we found

[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.

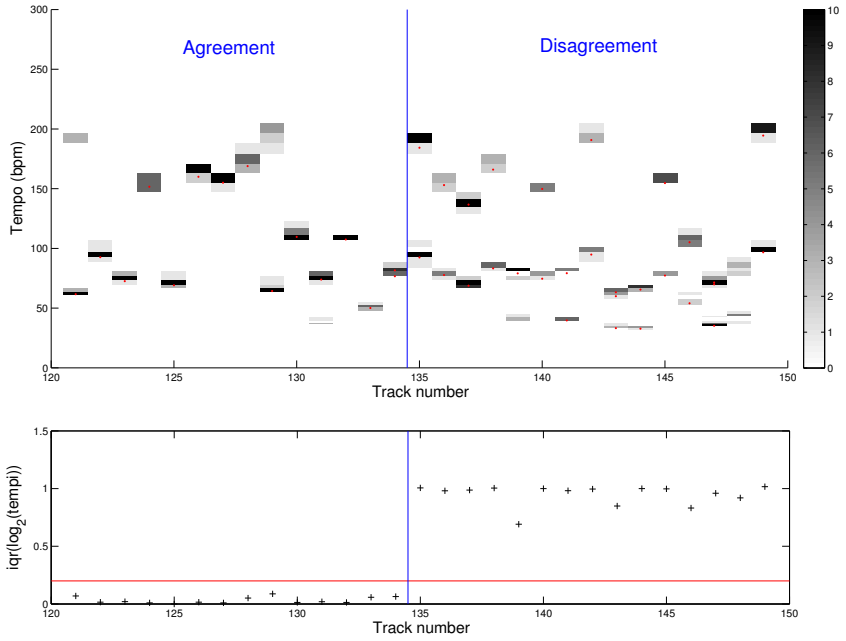


Fig. 5. [Top part] For each track a we represent the various annotated tempi $t_{a,u}$ in the form of a histogram. [Bottom part] For each track a, we represent the computed IQR_a. We superimposed to it the threshold τ that allows deciding on the assignment of the track to the agreement (left tracks) or disagreement (right part).

$\tau = 0.2$ to be a reliable value. This process leads to a balanced distribution of the test-set over classes: $\#(\mathcal{A})$=134, $\#(\mathcal{D})$=115.

Illustration: In Figure 5 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 is 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 ($\mathcal{A}$ and $\mathcal{D}$) 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

The 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), Tempo-GMM and Tempo-SVM lead to results above a random classifier.

The best results are obtained with the Tempo-GMM and Tempo-SVM models (predicting the agreement/disagreement using four individual tempo predictions). Their performances largely exceed the other models.

In terms of Mean Recall, the Tempo-SVM outperforms the Tempo-GMM classifier (74.9% instead of 70.1%). However this is done at the expense of the distribution between the agreement and disagreement Recalls: while the Tempo-GMM has close Recalls for the two classes (73.7% and 66.5%), the Tempo-SVM model clearly recognizes more easily the class $\mathcal{A}$ (87.3%) than the class $\mathcal{D}$ (44.3%, i.e. less than a random classifier). This unbalancing of Recall makes us prefer the Tempo-GMM model over the Tempo-SVM model.

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($\mathcal{A}$)	Recall($\mathcal{D}$)	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%**
Tempo-SVM	**87.35%**	**44.35%**	**74.85%**

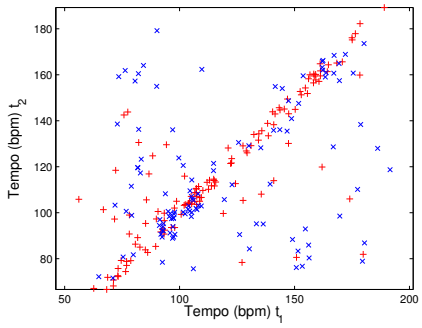
3.4 Discussions on the model Tempo-GMM

The Tempo-GMM model relies on the agreement between the four individual tempo estimations $T_{ener}, T_{sim}, T_{specbal}, T_{harmo}$. In Figure 6 we represent the relationship between these four estimated tempi for data belonging to the classes

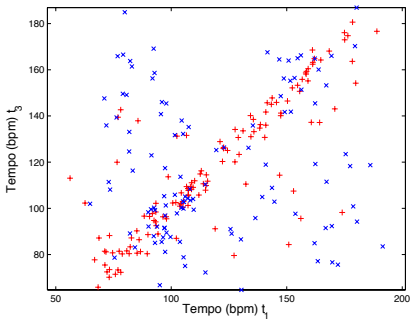
[6] $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$

[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.

(a) $t_1 = T_{ener}/t_2 = T_{sim}$



(b) $t_1 = T_{ener}/t_3 = T_{specbal}$



(c) $t_2 = T_{sim}/t_3 = T_{specbal}$

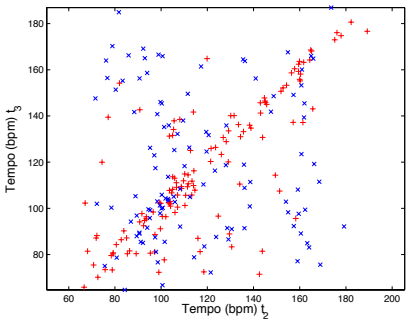


Fig. 6. Each panel represents the relationship between the estimated tempo for (a) $t_1 = T_{ener}/t_2 = T_{sim}$, (b) $t_1 = T_{ener}/t_3 = T_{specbal}$, (c) $t_2 = T_{sim}/t_3 = T_{specbal}$. Red plus signs represent data belonging to the agreement class, blue crosses to the disagreement class.

agreement (red plus sign) and disagreement (blue crosses)[8]. 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.

We now investigate the usefulness of each of the four tempi estimation $T_{ener}, T_{sim}, T_{specbal}, T_{harmo}$ for our agreement/ disagreement estimation. As a reminder, T_i is the tempo estimation obtained with $d_i(\lambda)$ using GMM-Regression: $T_i = f(d_i(\lambda))$. The question is twofold: are the values we expect to have for T_i

[8] It should be noted that we didn't plot the relationship between T_{harmo} and the other estimated tempi because the effect we wanted to show was less clear. We will investigate why in the next paragraph.

the correct ones ? Is T_i useful ? In order to test the first, we only consider the subset of tracks for which people agree on the tempo (the 134 items belonging to the class $\mathcal{A}$). In this case, $T_i = f(d_i(\lambda))$ should be equal to the shared perceptual tempo t. Table 2 indicates the tempo accuracy at 4% obtained with each $d_i(\lambda)$. The best results are obtained with the Energy variation f(78.3%), followed by the Short-term event repetition (55.0%) and the Spectral balance variation (47.0%). The Harmonic variation is strongly inaccurate (only 20.5%). A similar observation has been made by [17]. Because its estimation is strongly inaccurate, it is likely that T_{harmo} is actually not useful for the prediction of tempo agreement/ disagreement. Actually, using only $T_{ener}, T_{sim}, T_{specbal}$ as input to our Tempo-GMM model allows increasing the classification into agreement ($\mathcal{A}$) and disagreement ($\mathcal{D}$) by 1% (71.2% without using T_{harmo} compared to 70.1% when using it).

Audio Feature	Correct tempo estimation
$T_{ener} = f(d_{ener}(\lambda))$	78.3%
$T_{sim} = f(d_{sim}(\lambda))$	55.0%
$T_{specbal} = f(d_{specbal}(\lambda))$	47.0%
$T_{harmo} = f(d_{harmo}(\lambda))$	20.5%

Table 2. Correct tempo estimation (in %) of the 134 tracks of the class agreement by a GMM-Regression algorithm, using $d_i(\lambda)$ as input ($i \in [ener, sim, specbal, harmo]$).

3.5 Discussion on Moelants and McKinney preferential tempo assumption.

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 if their preferential tempo assumption holds for our test-set. For this, we compute the histogram of all annotated tempi for the tracks of our test-set. This histogram is represented in Figure 7 (blue vertical bars). We compare it to the one obtained in experiments 1 and 2 of Moelants and McKinney [14] (represented by the green dotted curve). Their distribution is uni-modal with a peak centered on 120 bpm while our distribution is bi-modal with two predominant peaks around 87 and 175 bpm. Since these distributions largely differ, Moelants and McKinney preferential tempo assumption does not hold for our test-set.

We then tried to adapt their assumption to our test-set. We did this by adapting their resonance model. In [20], they propose to model the tempo annotations distribution by a resonance curve: $R(f) = \frac{1}{\sqrt{(f_0^2 - f^2)^2 + \beta f^2}} - \frac{1}{\sqrt{{f_0}^4 - f^4}}$, where f is the frequency, f_0 the resonant frequency and β a damping constant. The resonant model that best fits our distribution has a frequency of 80 bpm (instead of 120 bpm in [14]). It is represented in Figure 7 by the red curve.

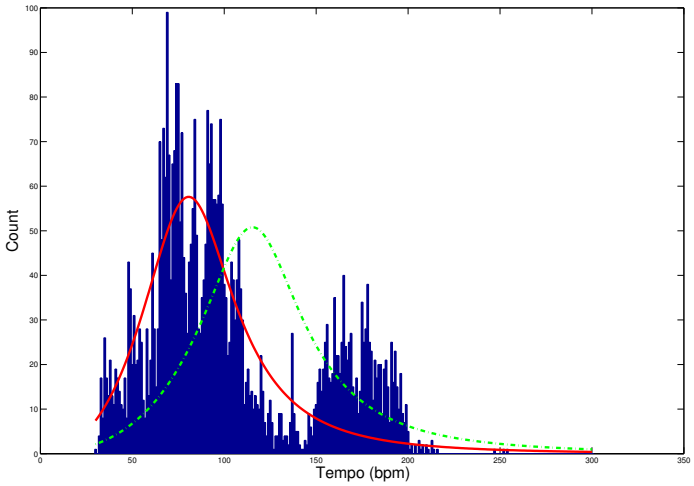


Fig. 7. Histogram of tempi annotation for the tracks of the Last-FM test-set. We superimposed to it the resonant model as proposed by Moelants and McKinney [14] with a frequency of 80 bpm (red line) and with a frequency of 120 bpm (green dotted line). The 80 bpm model has been fitted from our test-set. The 120 bpm model corresponds to the McKinney and Moelants experiment.

We then re-did our experiment changing the preferential tempo interval in our prediction model to $[60 - 100]$ bpm (instead of $[110 - 170]$ bpm in [14]). Unfortunatelly it didn't change our results in a positive way: mean-Recall(MM-Ener)=50.39%, mean-Recall(MM-Sim)=42.49%.

Note that, the difference of resonant frequency may be due to the different test-sets, experimental protocols and users[9]. Note also that the bad results we obtained with Moelants and McKinney model may also be due to our audio features that are not suitable for this kind of modeling. These acoustical cues are more adapted to a tempo-estimation task since they have a lot of peaks (at the fundamental tempo and at its integer multiples). It makes the tempo estimation more robust but hampers the selection of the two pre-dominant peaks..

[9] Firstly the test-set for our experiment and the one of [14] largely differ in their genre distribution. In [14], the tracks are equally distributed between classical, country, dance, hip-hop, jazz, latin, reggae, rock/pop and soul. In our test-set, most of the tracks are pop/rock tracks (50%), soul and country (about 10% each). The other genres represent less than 5% each. The experimental protocols also largely differ. Our test-set comes from a web experiment, done without any strict control on the users, whereas McKinney and Moelants had a rigorous protocol (lab experiment, chosen people). Users have then very different profiles. In McKinney and Moelants experiment, the 33 subjects had an average of 7 years of musical education. In our case, we reckon that almost nobody had a musical training.

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 features and the two classes.

The first model is based on Moelants and McKinney [14] assumption that agreement is partly due to the presence of a main periodicity peak 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 [14] 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 tempo 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 (then SVM) are then used to model those four tempi. The corresponding model leads to a 70% mean-Recall (and 74% for the SVM). Although SVM classifier has better overall results, the class-result are far from being equally-distributed (87% for the agreement class against 44% for the disagreement one). This made us prefer the GMM classifier (which has well-distributed results by class). 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.

In a post-analysis, we found out that our harmonic variation feature, because of its inaccuracy, was not beneficial for predicting tempo agreement and disagreement. Further works will therefore concentrate on improving this feature. Future works will also concentrate on studying the whole model, i.e. introducing the user variable u in the tempo estimation $f(a, u) = T_{a,u}$. 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 and by the French government Programme Investissements d'Avenir (PIA) through the Bee Music Project.

References

1. Mark A Bartsch and Gregory H Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 15–18. IEEE, 2001.
2. Ching-Wei Chen, Markus Cremer, Kyogu Lee, Peter DiMaria, and Ho-Hsiang Wu. Improving perceived tempo estimation by statistical modeling of higher-level musical descriptors. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.
3. Bee Yong Chua and Guojun Lu. Determination of perceptual tempo of music. In *Computer Music Modeling and Retrieval*, pages 61–70. Springer, 2005.
4. Taoufik En-Najjary, Olivier Rosec, and Thierry Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *INTERSPEECH*, 2003.
5. Patrick Flandrin. *Time-frequency/time-scale analysis*, volume 10. Academic Press, 1998.
6. Jonathan Foote. Visualizing music and audio using self-similarity. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 77–80. ACM, 1999.
7. Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *ICME*, volume 1, pages 452–455. IEEE, 2000.
8. Aggelos Gkiokas, Vassilios Katsouros, and George Carayannis. Reducing tempo octave errors by periodicity vector coding and svm learning. In *ISMIR*, pages 301–306, 2012.
9. Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1832–1844, 2006.
10. Jason Hockman and Ichiro Fujinaga. Fast vs slow: Learning tempo octaves from user data. In *ISMIR*, pages 231–236, 2010.
11. Jean Laroche. Efficient tempo and beat tracking in audio recordings. *Journal of the Audio Engineering Society*, 51(4):226–233, 2003.
12. Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. In *ISMIR*, pages 317–322, 2011.
13. Martin F McKinney and Dirk Moelants. Extracting the perceptual tempo from music. In *ISMIR*, 2004.
14. Dirk Moelants and M McKinney. Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous. In *Proceedings of the 8th International Conference on Music Perception and Cognition*, pages 558–562, 2004.
15. G. Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.
16. Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *ISMIR*, pages 35–40, 2007.
17. Geoffroy Peeters and Joachim Flocon-Cholet. Perceptual tempo estimation using gmm-regression. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 45–50. ACM, 2012.
18. Geoffroy Peeters and Helene 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, 2011.

19. Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer. From rhythm patterns to perceived tempo. In *ISMIR*, pages 519–524, 2007.
20. Leon van Noorden and Dirk Moelants. Resonance in the perception of musical pulse. *Journal of New Music Research*, 28(1):43–66, 1999.
21. Linxing Xiao, Aibo Tian, Wen Li, and Jie Zhou. Using statistic model to capture the association between timbre and perceived tempo. In *ISMIR*, pages 659–662, 2008.
22. José R Zapata, André Holzapfel, Matthew EP Davies, João Lobato Oliveira, and Fabien Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *ISMIR*, pages 157–162, 2012.