Università
degli Studi
di Padova

Department of Information Engineering
Master Thesis in Computer Engineering

Time-evaluation model for live musical interaction with multiple performers

Supervisor
Sergio Canazza

Master Candidate
Paolo Laghetto

9 September 2019
Academic Year 2018/2019

To my family that has always believed in me.

Abstract

In order to assess the quality of a musical performance many aspects have to be taken into account: timing, pitch, loudness, articulation and other parameters, creating a complex environment of different variables to be evaluated. With this study we want to propose a new time-based model, that is interested in the notes duration, excluding the other parameters from the analysis. Here, the hocket rhythmic technique is particularly suitable, since it creates a shared melody between two or more performers, that requires mutual actions. In this study we applied a new methodology for this field of research, based on the predictive coding theory and implemented through a Bayesian approach. In this way we created a dynamic data-analysis model, taking into account only the IOI (Inter Onset Interval) of a performance. With this new kind of analysis it is possible to compute different types of errors, in order to assess the quality of a performance by means of objective measures and to extract useful features about the interaction between performers. The framework we built is even able to work in real-time. It can be used to control a personalized bio-feedback system, in order to create a new learning process for musicians. Moreover, at the end of a musical interaction it provides objective assessments about the quality and through useful tools, it can help musicians to analyse their errors and timing precision.

Summary

In the introduction is presented the field of research of this study, in order to better understand all the different areas that are involved in the following analysis.

In the second chapter are presented all the most important research and results of the scientific literature, that involved the analysis of musical performances. Particular attention is given to the mathematical techniques used, because later will be introduced a methodology that has never used before for this kind of musical analysis.

The third chapter will present the environment that was created to be able to analyse the timing of a musical performance, made by two musicians using a medieval hocket style. The participants, the score, the setup and procedures used to collect the data, are all described here.

Then the fourth chapter represents the core of this study, where all the analysis of the data that we have done are reported. Here it is described the time-evaluation model we created, it explains the mathematical theory behind that and all the steps that characterize the new algorithm. In this chapter are described all the 3 main objective parameters that we computed to evaluate the quality of a performance. Furthermore, we tested also some initial hypothesis, and other techniques that are usually implemented in this kind of research.

In the fifth chapter is described a real application that exploits the Bayesian method and the new evaluation parameters described in chapter four. It shows that it is already possible to implement an online application that can elaborate the data coming from a live performance. In particular it aims to use a bio-feedback system to improve the timing precision of multiple musicians while they are playing.

Finally, in chapter six we summarized all the results obtained in this study. Focusing on what has been discovered and what could be improved. Then, are reported some future applications that could take advantages of the innovative analysis approach described in this research and of the results that we have already obtained.

Contents

Abstract v

List of figures xi

List of tables xiii

1 Introduction 1

2 Quality of musical interaction 5
- 2.1 Rhythm in joint action . . . 6
- 2.2 Onset detection and IOI-analysis with cross-correlation . . . 10
- 2.3 Mean sign asynchrony and precision . . . 15
- 2.4 Cross-wavelet transform . . . 19
- 2.5 Bayesian brain hypothesis . . . 27

3 Just hock it 29
- 3.1 Participants . . . 29
- 3.2 Stimuli . . . 30
- 3.3 Setup . . . 33
- 3.4 Procedure . . . 33
  - 3.4.1 Self-assessment . . . 35
- 3.5 Hypothesis . . . 39

4 Time-evaluation model 41
- 4.1 Semi-automatic onset detection . . . 42
- 4.2 Inter-onset classes . . . 45
- 4.3 Bayesian inference approach . . . 46
  - 4.3.1 Defining prior distributions . . . 50
  - 4.3.2 Sequential Bayesian updating . . . 52
- 4.4 Fluctuation error . . . 56
- 4.5 Collapse error . . . 65
- 4.6 Narration error . . . 68
- 4.7 Moving vs non-moving conditions . . . 73
- 4.8 Learning effect . . . 75
- 4.9 Leader-Follower interaction . . . 77

5 Real-time analysis and bio-feedback system **85**
5.1 Real-time onset detection 86
5.2 Bio-feedback system 88
5.3 Real-time framework 90
5.3.1 Setup 91
5.3.2 Procedure of the real-time test 94
5.3.3 Results 95

6 Conclusion **97**

References **101**

Acknowledgments **107**

Listing of figures

2.1 Factors that affect interpersonal coordination during rhythmic joint action . . . 7
2.2 Sequences of intertap intervals (ITIs) . . . 11
2.3 Different synchronization strategies . . . 12
2.4 Cross-correlation at different conditions . . . 14
2.5 Mean signed asynchrony . . . 17
2.6 Exemples of wavelets . . . 20

3.1 Score for the first subject of each couple . . . 32
3.2 Score for the seconds subject of each couple . . . 32
3.3 Duo performance . . . 34
3.4 Annotation of the overall quality . . . 36
3.5 Annotations made by the two subjects . . . 37
3.7 Joint action sense of agency . . . 39

4.1 Inter-onset interval . . . 43
4.2 False negative onset detection . . . 44
4.3 File containing the extracted onset . . . 45
4.4 Inter-onset classes . . . 46
4.5 Weight of prior mean and evidence data mean . . . 49
4.6 IOI histogram for the first 15 seconds of the performance . . . 51
4.7 K-means clustering . . . 52
4.8 Classification . . . 53
4.9 Bayesian updating . . . 54
4.10 MAP behaviour along the performance . . . 55
4.11 Prediction-error . . . 56
4.12 Time difference between predicted and performed IOI . . . 57
4.13 Fluctuation errors . . . 58
4.14 Total fluctuation errors ranking list . . . 59
4.15 Annotation based ranking list . . . 60
4.16 Agency based ranking list . . . 60
4.17 Correlation test between annotation and fluctuation error . . . 62
4.18 Correlation test between agency and fluctuation error . . . 63
4.19 Collapse error . . . 65
4.20 Correlation test between annotation and collapse error . . . 66

4.21 Correlation test between agency and collapse error 67
4.22 Sequences of notes in A part and B part 69
4.23 Correlation test between narration error and annotation 70
4.24 Correlation test between narration error and agency 71
4.25 Moving and non-moving boxplots for fluctuation error 73
4.26 Moving and non-moving boxplots for narration error 74
4.27 Learning effect over trials 76
4.28 Learning effect of fluctuation error 76
4.29 Sequence of errors of B part for two distinct subjects 78
4.30 Sequence of errors of B part compressed 78
4.31 Sequence of errors of A part compressed 79
4.32 Cross-correlation between errors of A part 79
4.33 Cross-correlation over the 8 trials of a couple 81
4.34 Imitating and correcting behaviour of every partecipants 81
4.35 Barplot of cross-correlation values for each couple 83

5.1 Real-time onset algorithm schema 87
5.2 Errors colour management 89
5.3 Multiple musician setup 91

Listing of tables

2.1 Predictors of annotated bouts of interaction. 25

1
Introduction

For human beings music is considered as a rewarding activity which is able to empower people and connect them together [1]. People have a deep intuitive understanding of music. Musicality can be defined as a natural, spontaneously developing feature based on and constrained by biology and cognition [2]. Music has the capability to enhance social activity and it's used to regulate people's emotional arousal. It appears to be inside us, outside us, and among us, like a shared cultural phenomenon [3], offering a bio-social empowering expression that may date back to the very beginning of human evolution [2].

Since ancient times, human interaction with music has raised scientist's curiosity. The conception of music as a bio-technology reveals a double nature: a biological trait better described as musicality, on the first hand, and, on the other hand, a culturally evolved product that allows groups to express and, therefore, to distinguish themselves among each other [4].

The analysis and research-based study on music is well known as musicology. It's a field of study that require knowledge belonging to different areas. The parent disciplines of musicology include: psychology, neuroscience, acoustics, psychoacoustics, information science, computer science and mathematics.

The project developed in this thesis and the resulting applications of it, belong to the field of computational musicology. It is an interdisciplinary research field between musicology and computer science. It includes many related disciplines like music informatics, mathematical music theory, computer music, systematic musicology and computational musicology. Research in computational musicology can be seen as an intersection between humanities subjects and sciences. This field is generally considered like an extension of a long history of inquiry in music that overlaps with subject such as statistics, mathematics and information engineering. The questions raised in this area, tend to be answered either by analysing empirical data based on observation or by developing theory. The approach used in this project is a combination of both.

The use of computers that allow to study and analyse music, generally dates back in 1970 [5], but, in our days, computational musicology encompasses a broad range of research topics that deal with the multiple ways music can be represented. Audio data can be globally conceptualized as continuum set of features ranging from a lower to a higher level audio features. Low-level audio features mean taking into account loudness or spectral flux. Mid-level audio features, the target level where is concentrated the focus in this study, refers to onsets, beats and rhythm. Eventually, high-level audio features refer to style, artist, mood, and key, but, they are not taken into account in the following analysis.

In this particular study we explore the viewpoint that music is a bio-social tool that enables a group to establish an empowering joint action [6]. More specifically, at this point, the question is how, through music, a joint action such as singing or playing simple instruments, can shows particular qualities that affect group-formation. The main goal here is to discover which are the parameters that can be extracted and computed from a musical interaction, that have the property to be strictly correlated, within a certain approximation, to a subjective assessment of the quality of a musical performance.

A big challenge for scientists is the articulated and multi-faceted way in which music manifests itself. There are many different aspects and features that are characteristics of every performed song, for example, timing, dynamics, harmony, loudness, articulation and other parameters. All these different variables create a complex environment to be evaluated. A global comprehensive understanding would focus only on the fundamental principles that are at the heart of these multiple facets. For this reason, it's convenient to start from the simplest level of social interaction that music may offer: a singing dyad [7].

In this thesis, it is proposed a completely new time-based model that exploits the

Bayesian probability theory, with the aim to evaluate the quality of a musical performance respect to the regularity of the notes timing. This new model, is interested only in analysing the note duration, excluding the other parameters from the evaluation.

The timing of the musical notes in a joint action, is a feature that appear in every musical interaction and it is one of the main characteristics that should reveals information about how is perceived the performance from a human being point of view. The simplest musical interaction, in which we can test a new mathematical model to evaluate the quality of the performance, is between two people that are performing a given score. Moreover, taking into account duets performances, is also possible to analyse the leader-follower interaction, in order to establish who is showing an imitating or correcting behaviour respect to the other performer. All the research, test, data collection and analysis are made possible thanks by Institute for Psychoacoustics and Electronic Music (IPEM). It is the most advanced laboratory to conduct music-based experiment. It belongs to the University of Gent and it is located in Belgium. It is one of the most advanced centre in Europe, mainly dedicated to the study of Embodied Music Interaction (EMI). EMI manifests itself through sound-based activities, like listening, playing and dancing, with other people as in a joint action, as well as with music instruments and within the body that is considered a mediator for music playing [3]. The interactions are constrained, by acoustical structures, by cognitive activities in the sense of limitations of memory, attention, learning, by body resonances, biomechanical and energetic restrictions [3]. Many scientists believe that these musical constraints can be better understood just considering the timing of embodied interactions, like the rhythmic coordination of the human body with external musical rhythms.

During the last decade, research in EMI field, have been strongly motivated by a demand for new tools in view of the interactive possibilities offered by digital media technology. Thanks to the advanced technologies available at IPEM laboratories and the different background of the work teams that are involved in the research, complex forms of interacting, such as full-body interaction, dyadic interactions, or even the interaction of multiple musicians, have now become feasible for study.

Since the EMI is the core of the research at IPEM, in this study is taken into account how the movement of people, during a musical performance, is able to influence the ability of keeping a regular timing of notes. Moreover, this research aims also to discover how much significative could be the movement condition respect a non-moving condition for the final quality of a musical exhibition.

The entire framework of analysis that is built in this project has even the capability to work in a real-time environment. In this way the results obtained with this project open the possibility to develop and to build new applications that can work in real-time during a musical performance, to give additional information to the players. A new proposal of this kind of applications is already realized in this project. Using the time-evaluation model built in this study, it is possible to use the information extracted from a real-time music exhibition and control different kind of bio-feedback system, like for example, coloured lights or even sound, to give information to the singers/players concerning the quality of their performance. Performers can benefit from this bio-feedback system for adjusting and improving the quality of the song played and also to auto-assess themselves. At the end, the aim of this kind of applications is to improve the precision in the regularity of the timing in a performance and also to give an assessment of the overall quality of the song played by each performer.

Before going into details of this study, in the next chapter are reported the most relevant approaches and results related to this field of research and from which will be noticeable some similarities with the methodologies adopted in this work.

2

Quality of musical interaction

In a musical joint action, *quality* is an emergent feature of the interaction. It is determined by the rules set by the score, in the case of written music, and partly determined also by the musicians' culture, experience and personality. If a certain level of quality is reached during the performance, we may assume that the interaction empowers each participant and, as a result, the ensemble.

Certainly, it is not obvious to examine quality with the same tools and methods used to investigate the processes of musical interaction, hence there is the necessity to include in such study also subjective parameters, that is, the quality of the performance from both a third and a first-person point of view.

However, it is worth to summarize some of the existing methods to investigate musical interactions between two or more performers and evaluate their appropriateness in the specific study of quality.

Since the scenario under examination in this research is composed by musicians in small ensembles that have to coordinate their actions to produce sounds that form cohesive auditory melodies, it's important to understand which is the state of the art regards the cognitive-motor processes involved in rhythmic interpersonal coordination.

2.1 Rhythm in joint action

Cognitive science has taken advantage of musical tasks in order to investigate how two or more subjects build representations of a joint action, employing both behavioural and neuroscientific approaches. As it's well reported in [8], the capability to intentionally coordinate a rhythmic joint activity with others, can be seen as a specific case of joint action, that is the human behaviour that involves multiple subjects coordinating their thoughts and actions in space and time.
Ensemble musicians have to coordinate their body movements or more in general some form of action, to produce synchronous sounds and interlocking patterns, in which separate instrumental/singing parts articulate complementary rhythms. The joint action is considered rhythmic if the goal requires to produce specific patterns of relative timing between the actions made by the involved individuals. In general the prescribed score or temporal relationships require precision in the order, at maximum, of tens of milliseconds.
An important statement in [8], reports that a regular time movements facilitates the degree of precision during the musical interaction and the movement timing has to be flexible to allow rate modulations from the partners, in the order of hundreds of milliseconds, to let mutual cooperativity during the interaction.
This means that rhythmic interpersonal coordination requires simultaneous precision and flexibility on timing and it challenges the cognitive-motor system of the partners.
Keller E. P. [9] [10], proposed a theoretical description of the main factors that influence rhythmic interpersonal coordination, Figure 2.1. The framework, is formulated in the context of ensemble performance, that is exactly the showcased in this experiment. In the proposed theory, Keller established that the temporally precise rhythmic interpersonal coordination requires three core cognitive-motor skills: adaptation, attention and anticipation.

- **Mutual temporal adaptation:** rhythmic joint action, however, requires the coordination of musical sequences that could include irregular patterns of timing. Musical ensemble performances, are characterized by intentional and unintentional variations in micro-timing event and tempo, as well as systematic deviations in synchronism between parts played by different subjects. Such inconsistency in interpersonal timing must be kept

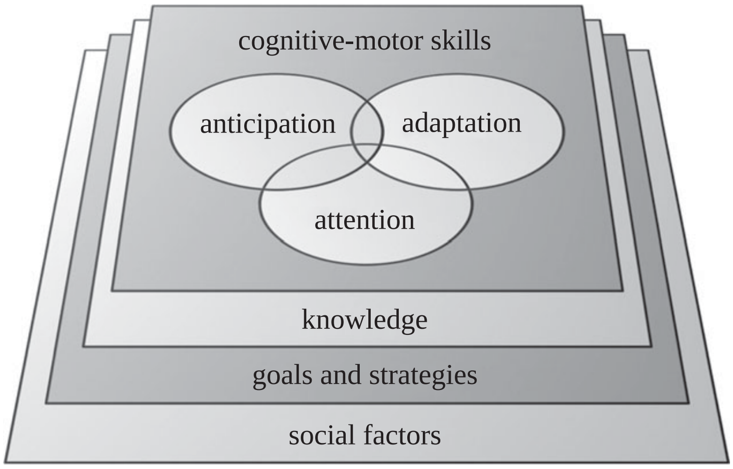


Figure 2.1: Factors that affect interpersonal coordination during rhythmic joint action [8].

under control through continuous mutual temporal adaptation [8]. Mutual adaptive timing is assisted by temporal error-correction mechanisms that enable internal timekeepers to remain entrained despite small variations in timing [11]. The ability to use temporal error correction varies across individuals [12]. The mutual temporal adaptation is considered as the glue that connect together individuals engaged in rhythmic joint action.
In musical context, mutual temporal adaptation can also contribute to ensemble cohesion by improving the similarity of partners' playing styles. Some research with experimental tasks requiring piano duets performances [13] and dyadic finger tapping [14] have demonstrated that compensatory adjustments in combination with error correction lead to co-dependencies. Eventually, mutual temporal adaptation may be, at non-conscious level, a form of behavioural imitation that facilitates ensemble cohesion by making multiple performers sound collectively as one.

- **Attention:** rhythmic joint action can be considered a form of multi-tasking. A form of split attention, which has been termed "prioritized integrative attending" in [15], includes a mixture of selective attentions.
To produce an organic sound, ensemble musicians have to pay attention to their own actions, with high priority, those of others, with lower priority, and in the meanwhile they also have to concurrently monitoring the over-

all ensemble output. In this way, during a music performance, ensemble cohesion is facilitated, since partners are allowed to adjust their actions based on the real-time comparison of the ideal ensemble sound and incoming perceptual information about the actual sound.
Prioritized integrative attending requires a cognitive effort to the extent that it involves the simultaneous separation but also integration of information from different sources [16] [17].

- **Anticipatory mechanisms:** assists precise rhythmic interpersonal coordination by allowing performers to manage the timing of their own actions with reference to predictions about the future timing of others' actions. In other words, ensemble performers use this mechanisms of anticipatory cognitive-motor to plan their own actions and to generate real-time predictions about the upcoming sounds from the partner [9].
  It has been proposed that predictions can evolve in two different ways [18]. On one side, automatic expectancies about events at short timescales, for example, the next note of the sequence played. On the other side, involves anticipating co-performers' actions by activating memory representations of shared goals [10]. Temporal prediction abilities are thus driven by the fidelity of actions simulation and mental images, and they can be acquired and strengthened through active experience and observational learning.

It has been discovered [8] that these cognitive-motor skills are influenced by: the performers' goals concerning the interaction, their knowledge about the music and familiarity with the partner, the use of regulatory strategies to facilitate coordination and psychological factors like personality. In general, articulated forms of rhythmic joint action, such as those encountered in musical ensemble performance, require pre-planning. Musicians usually train for performance through collaborative group rehearsal in order to establish shared performance goals, i.e. unified conceptions of the ideal ensemble sound [19]. Co-actors thus form memories of each other's parts and the relationship between the parts [10]. Research on ensemble performance suggest that developing shared goals means acquiring knowledge about the musical structure and the expressive intentions and playing styles of the other members [20].
Musical structure refers to the hierarchical pattern of pitch and rhythmic ele-

ments: single tones are linked into melodic motives and phrases, while rhythmic durations can be defined relative of an underlying metric framework. The way musical structure is played in performances is flavoured by micro-timing deviations and aesthetically motivated tempo variations that reflect the musician's individual expressive intentions and playing style [21].

A recent study that involved piano duos [22], has shown the importance of knowledge about both musical structure and playing style. In the cited case study, skilled pianists were required to play several duets with unknown partners. In one condition the duet performers had previously practiced both parts, while in the other condition the performers had practiced only their own part. Pianists' keystroke timing was recorded from keyboards and body movements were tracked with a motion-capture system. The results underline that the variability in interpersonal keystroke asynchronies decreased across repetitions and was generally lower in the unfamiliar condition than the familiar condition. In other words, coordination started out more precise and remained so, when pianists had not rehearsed their co-performer's part.

These results suggest that knowledge of a co-performer's part, in the absence of knowledge about their playing style, generates predictions about expressive micro-timing variations that are based instead upon one's own personal playing style, leading to a mismatch between predictions and actual events at short timescales. On the other hand, predictions at longer timescales, that is, those related to musical phrases, and reflected in body movement, are improved and facilitated by awareness of the structure of a co-performer's part.

Social and psychological factors affect rhythmic interpersonal coordination at many levels. Experimental works addressing interpersonal coordination have identified links between personality characteristics and the cognitive-motor skills involved in rhythmic interpersonal coordination. For example, as reported in [23], children with higher social skills, as assessed by their teachers, synchronized better with others in a dyadic drumming task. This could stem from a high level of awareness of others in a social context.

Interpersonal coordination can have reciprocal effects upon social outcomes concerning interpersonal affiliation, trust and prosocial behaviour. For these reasons, timing of interpersonal coordination is affected by social skills [8].

2.2 Onset detection and IOI-analysis with cross-correlation

The interaction between two or more individuals is not reducible to the respective mind reading processes, but is built by a "participatory sense-making" that spreads out in the dynamics of the interaction itself [16]. Therefore, interaction cannot be studied by analysing only one single subject at a time, but rather by analysing the emerged interaction itself, in the sense of behaviour relative to one another. Although this concept is quite spread in the theoretical field of dynamical systems [24], a deep understanding of interaction is still struggling to emerge in the filed of neuroscience, despite the growing consensus around the many facets of the "embodiment" concept [25].

Musical performance requires the skill to synchronize actions on-the-fly between performers, and it is a specific case of intentional interpersonal coordination.

Music is formed by temporally related sounds, where movements must be precisely coordinated both within and between performers. Thus, interpersonal coordination is a result of intended joint action rather than a spontaneous occurrence.

In the literature, there are some works, explained in the following paragraphs, that are not directly addressing the evaluation of the quality of performance, but they analyse some similar task, for example, joint finger tapping, to measure the synchronization between musicians and to retrieve some features like leader-follower behaviour.

Some of the approaches implemented in this research are build with the knowledge of the methods already implemented in the following works.

In a recent work in 2018 [26], has been studied a joint finger tapping task exploiting the intertap intervals (ITIs). They used a minimalistic musical paradigm to investigate the interactions between a dyad of two musicians having the same or different top-down predictions. Once extracted note on and off timestamp, all the analysis has been processed with MATLAB.

Once extracted the onset times, authors computed the intertap interval (ITI), the time between two successive taps, Figure 2.2.

In particular, they calculated the cross-correlation at lag -1, 0 and 1 between each dyad member's tapping time series. The relation between these coefficients gave an indication about leader-follower interaction.

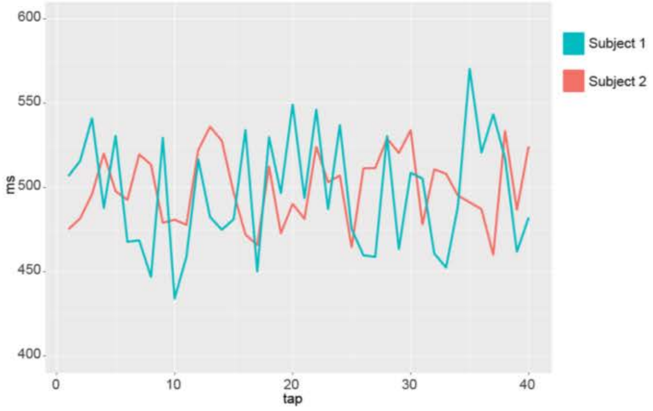


Figure 2.2: Sequences of intertap intervals (ITIs) [26].

Mutual adaptation occurs when the two members of a dyad, equally weigh the incoming auditory stimuli from the other member and their own model of the task, Figure 2.3.

Leading-following usually depends on one of the dyad members that attenuate the information coming from the other member. The leader referred as the partner who puts more confidence in his own internal model.

In the case of leading-leading, both participants exhibit leading behaviour. This means that they both discard information input from the other member and prefer to trust only on their own model.

A perfectly synchronized couple with no variation in tempo result in high correlation at all lags, whereas high synchrony with some variation in tempo produces the highest correlation at lag 0. If a leader-follower dynamic is present, a positive correlation at either lag -1 or lag $+1$ and negative or low correlation at lag 0 occurs, depending on which participant is the follower, that is, the more adaptive one.

Most dyads showed lag coefficient patterns indicative of mutual adaptation, except for a subset considering predominantly of drummers which showed a leading-leading-pattern suggesting little or even no dyadic interaction. A possible explanation for this result is that the drummer's role in an ensemble may different from other instruments because drummers serve the role of timekeeper and requires a higher degree of internal synchronization of movements.

The result is of high interest, as it complements the strategies of leading-following

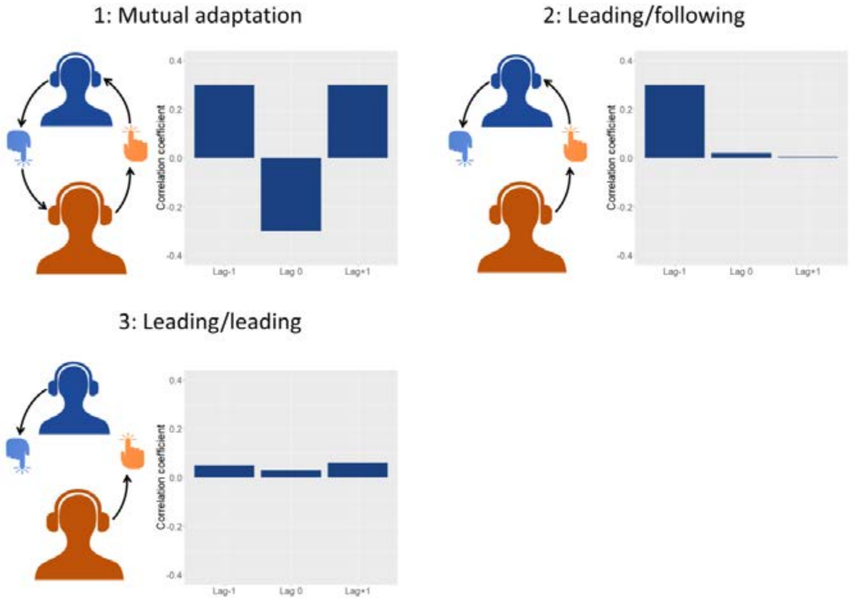


Figure 2.3: Illustration of the different synchronization strategies and their corresponding lag patterns performing cross-correlation on the ITI time series for lag -1, 0 and +1 [26].

in synchronization behaviour, underlying that the synchronization occurs also without any apparent dyadic interaction. Furthermore, it emphasizes that differences between individual musicians, such as which instrument they play, also affect interpersonal synchronization strategies [26].

Another previous work in 2010 [14], carried out a joint finger tapping experiment to study the mechanisms of coordination that are fundamentals in successful interactions. In the cited study, 16 couples of participants were asked to maintain a beat given in the first 8 seconds of each recording session, while hearing an auditory signal coming from either the other performer or generated by a computer metronome. The couples have been assessed on both synchronization and drift from the metronome. The data obtained from MIDI keyboards were imported and analysed, again, with MATLAB, taking into account only the onset tapping times.
Three types of measures were computed: windowed cross-correlations, synchronization indices, and the means of the absolute difference between the intertap intervals.

Usually, cross-correlation analysis and other similar methods assume the stationarity of the time series. Instead of assuming stationarity over the entire time series, using a windowed cross-correlation with a window size of 6 taps and a maximum of 3 lags, enabled to treat the time series as only having local stationarity. Then the values obtained for each trial were averaged to extract one only coefficient per condition.

The analysis give an indication of the directionality of the interaction, that is, whether the participants are not interacting with each other (i.e. uncorrelated), or whether there is a clear leader-follower relation where one participant lead the other towards his own tempo.

As can be seen in Figure 2.4, when members could only hear themselves, there is no correlation. In the unidirectional condition, in which they are hearing only one performer, there is small but significant negative correlation at lag zero and high positive correlation at lag $+/-1$. Lag -1 in the case both performers are hearing the first member and lag $+1$ when they are both hearing the second member. The positive correlations at lag 1 reflect the tendency of a member to adapt towards the previous ITI of the other member, by producing a shorter ITI when the other's last had been shorter and longer if the other's had been longer. In the bidirectional condition, in which both partners were hearing each other, both members showed the "follower" strategy, proven by high correlation coefficient both at lag $+1$ and -1. They formed a mutually and continuously adaptive unit with no strong evidence of a leader-follower beahaviour.

In order to address task performance, this research looked how well the individuals were able to synchronize, employing for the analysis the synchronization indices [27], based on variance of relative phase in relation to the computed metronome or the signal of the other member. The index is a number from 0 to 1 and when bigger than 0.73 it has been considered as indicating the synchronization regime [28]. The synchronization regime was found in all the conditions. The tests revealed lower synchronization indices for the unidirectional scenario respect higher values in bidirectional condition when both were hearing each other. No significant differences were found in synchronization when tapping along with the computer versus the bidirectional coupling condition. This means that couples were good to synchronize with irregular but responsive partner as with predictable but unresponsive computer. This suggests that a

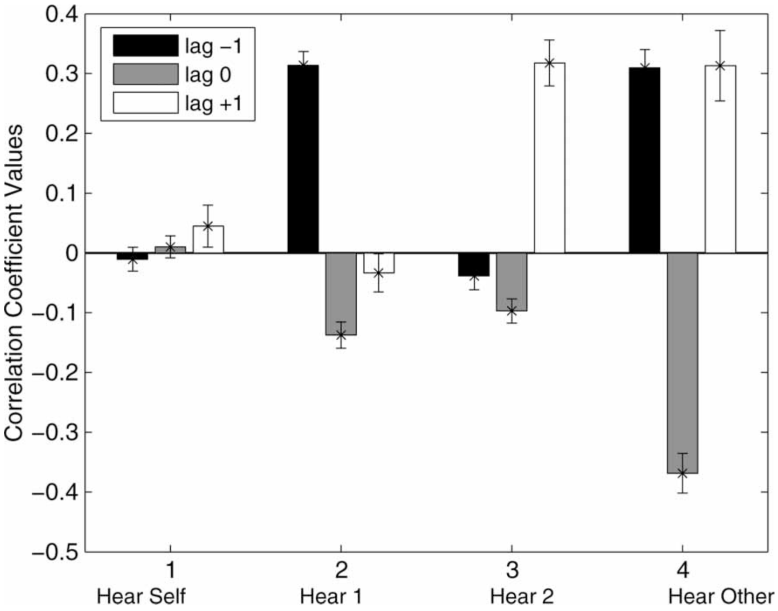


Figure 2.4: Windowed cross-correlation at lag −1, 0, and +1. Conditions 1: uncoupled condition, hearing themselves; conditions 2 and 3: unidirectional coupling, both hearing member 1 and member 2, respectively; condition4: bidirectional coupling, hearing each other [14].

successful coordination is supported by the prediction accuracy of the partner's future actions, known as anticipatory mechanism [29], and also by the mutual adaptability to the current action.

Eventually, the second and last parameter computed to evaluate the task performance, was the capability of the performers to keep the given beat.

Results showed that participants were significantly worse at keeping the given tempo when listening to the other member and even worse in the bidirectional coupling condition.

The research is well known because the authors identified a stable pattern in interpersonal coordination, which has never been reported before.

They mainly discovered that when dyads are mutually coupled to one another, they correct the duration of their ITIs in opposite directions on a tap-to-tap basis in a mutual attempt to synchronize with one another. These findings show that, in a jointly coordinated tapping task, there is evidence of a continuous mutual adaptation on a short millisecond timescale.

Finally the research conveys that interpersonal coordination is facilitated by two mutual abilities: to predict the partners' subsequent action and to adapt accordingly on a millisecond timescale.

2.3 Mean sign asynchrony and precision

The study of tapping tasks is a spread framework used in the works related to synchronization end coordination in musical interaction. However, in the latest years, many research started evaluating real and natural music instrumental ensemble.

A recent research in interpersonal entrainment in North Indian musical composition has studied the synchronization and movement coordination relate to tempo, dynamics, metrical and cadential structure [30].

The authors employed a traditional music performance method to evaluate sensorimotor synchronization and coordination between two or more musicians. Sensorimotor synchronization (SMS), can be defined as a process by which the sound-producing actions of musicians can become tightly synchronized and seems to operate over relatively short timescales, typically hundreds of milliseconds, occurring spontaneously. Indeed, the term 'coordination', refers to the management of performances over longer timescales, from a few seconds up to several minutes and is more accessible to conscious control, including visual cues such as head nods or upper body sway. Compared to SMS, coordination refers more to culturally-specific knowledge.

We have already mentioned some of the main works, concerning synchronization, but there is still a large literature of experimental research.

In [31], it's reported that analysing performances by piano duos, synchronization accuracy is not significantly affected by visual contact, although synchronization is more accurate when the movements of the pianist playing the melody part preceded those of the accompanist.

A related study [32], found that players of an accompaniment part were able to synchronize using only visual information about movement but they were more accurate when audio information was also available.

The Indian musical interaction study [30] aims to add more knowledge on this works by examining aspects of SMS and movement coordination in natural performances, to present the most extensive quantitative study of entrainment based on natural performance data.

The music genre used in [30] can be regarded as generally performed by a duo: plucked lute (string instrument) player and tabla (Indian drum) accompanist, sometimes there could also be a third musician playing an unchanging reference part "drone" on the tanpura (long-necked plucked string instrument).

This simple scenario allows to consider a number of different factors in entrainment, while at the same time limiting to a single genre and bounding the analysis to the interactions within a dyad.

Analysis started with onset detection, and was particularly challenging due to the highly complex auditory signals produced by the stringed instruments, and the very wide dynamic range of all instruments. Manual labelling and correction (to add, remove or reassign the automatically extracted onsets) were required. The proportion of onsets requiring intervention was typically 5-15% of the total automatically assigned by an onset detector algorithm.

Sensorimotor synchronization was analysed using the pre-computed event onset times and the measures of asynchronies were calculated as onset time differences where onset belonging to different instrument were assigned to the same metrical position. Using this pairwise asynchronies, were quantified two aspects: mean relative position and precision.

The mean relative position of the two instruments was measured as the mean signed asynchrony. It basically indicates whether and how much an instrument tends to play ahead (negative asynchrony) or behind (positive asynchrony) in time with respect to the other one.

Precision is given by the standard deviation of the sign asynchrony (lower mean relative position leads higher precision). In the case the standard deviation is not an appropriate statistical measure, for example, whether the raw asynchrony data is more appropriate as input for analysis than a summary measure such as standard deviation, they used the mean absolute (unsigned) asynchrony as a measure of 'precision', that can be used to judge how "successful" synchronization is within duos. The overall distribution of signed asynchronies between the melody instrument and tabla resulted symmetrical, as reported in Figure 2.5, with a mean close to zero: mean = -2.34 ms, SD = 27.87 ms.

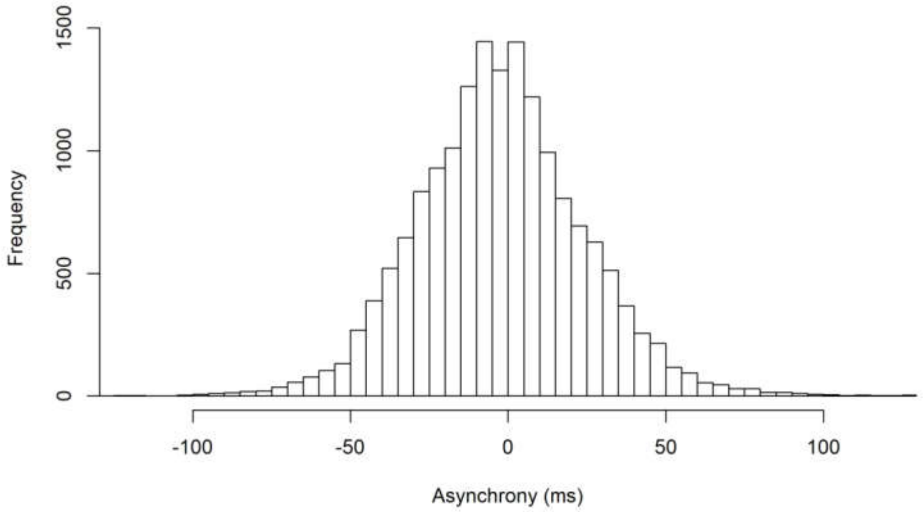


Figure 2.5: Pairwise asynchronies between plucked strings and tabla. Negative values indicate melody lead, positive value means melody follow the tabla [30].

In order to explore more deeply the variability of the data, the distribution of asynchronies was computed splitting each performance by 2-minutes segments. This reveals considerable variation within and between performances. Furthermore, the authors investigated which was the leader in instances in which the tabla takes solos and the string player provides accompaniment, i.e. the normal roles reversed. As result, across the corpus the lead instrument played ahead of the one performing an accompanying role. For that reason the "melody lead" phenomenon is in fact the lead instrument, no matter whether that instrument is melodic or tabla. The asynchronous behaviour indicates a tendency to reduce (i.e. precision to increase) over the course of repeated performances. This is to be expected also because performances increase in tempo after one another, and previous studies have found that variability of asynchrony increases proportionally with the duration of the inter-onset interval (IOI) [33].

The experiment showed that in the genre taken into account, synchrony becomes tighter and the melody instrument tends to play further ahead, with increasing tempo.

An other method that was employed to check the relationship between speed of playing and synchrony used event density. For each onset, a local event density

for that instrument was calculated over a window of 2 seconds, as the number of detected onset and then was averaged over each segment of the played song to produce mean event density parameter. Correlations were computed between these event densities (for each instrument separately, and also for the sum of the two) and the asynchrony data. Thus, synchronization resulted significantly more precise at higher tempi and even at higher event densities.
Many other factors which affect synchronization could be considered. In this research, the authors focused specifically on loudness, cadential figures and also reported observations on the occurrence of metrical errors.
It might be thought that loudness is positively correlated with speed, instead from this work emerged that peak levels are not significantly correlated with mean event density levels neither with asynchronization. Peak levels were computed using peak RMS amplitude during the onset detection phase.
At the end of the tests to establish whether asynchronies are correlated at cadences, resulted that mean absolute asynchronies is not significantly different at cadences than non-cadences.
Regards the errors, in this corpus the authors identified three different metrical performance errors. The first type of error is associated with any noticeable uncertainty, which can be due to a momentary loss of attention following a cadential downbeat. The second type of error occurs during a bridge between melodic improvisation, it was the most difficult to find exactly, and the third type of error occurs in the accompaniment of a solo.
As result, emerged no evident disruption of synchronization, so it suggests that it is possible to play through a mistake without compromising synchronization.
At the end, from this research, we can summarize that synchronization between musicians is therefore affected by factors including tempo, dynamics, and leadership, with only small effects of metrical position.

The limits of the works analysed before is that they assume the processes under examination to be stationary for a certain amount of time, with a homogeneous variance across conditions, characteristics that are hardily met by a musical performance, and by human behaviour in general.
On the other hand, there are also studies interested in the presence or absence of joint periodicities during interaction, that have removed such assumption.

Actually, it turns out to be a good method, for example, when there is a clear periodicity in the actions of the performers. Although, for complex interactions involving different periodicities makes it harder to derive a feature that captures coherence.

Different periodicities at different locations of a piece are reflected in the wavelet analysis. Therefore, since it worth drawing the most from this new kind of analysis, in view of an effective method to understand interaction quality, let's have a look in more details to some of the most recent studies exploiting cross-wavelet transform (CWT).

2.4 Cross-wavelet transform

In the latest years, wavelet analysis has attracted much attention in signal processing. A few studies have recently investigated the interaction between two musicians employing wavelet theory [34] [35]. It is an interesting tool for different fields, including engineering, mathematics and physics. It has been successfully applied in many areas such as transient signal analysis, image processing, communication systems, and other signal processing applications [36]. There are opportunities for further developments of the mathematical understanding of wavelets in a wide range of applications in science and engineering. Any application using the Fourier transform can be formulated using wavelets to provide more accurately localized temporal and frequency information. This analysis technique can resolve some of the difficulties inherent in Fourier analysis, like finding the relation between the Fourier coefficients to the global or local behaviour of a function. Unlike Fourier analysis, wavelet analysis expands functions not in terms of trigonometric polynomials but in terms of wavelets, which are generated in the form of translations and dilations of a fixed prototype function, called the analysing wavelet or mother wavelet. Different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are [37], Figure 2.6.

The wavelets belonging to these families have special scaling properties and they are localized in time and frequency. Every application using the fast Fourier transform (FFT) can be formulated using wavelets to provide more localized

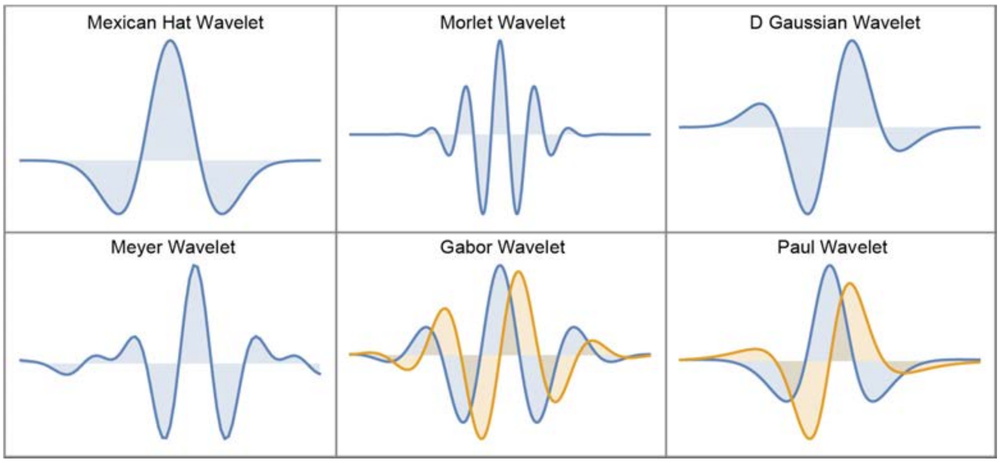


Figure 2.6: Example of different families of wavelet [38].

temporal (spatial), and frequency information, useful especially in case of non-stationary signals.

For these reasons, wavelet transform (WT), is a technique employed in a lot of different studies concerning astronomy, acoustics, nuclear engineering, sub-band coding, signal and image processing, neurophysiology, music, magnetic resonance imaging, speech discrimination, optics and earthquake-prediction.

Wavelet algorithms process data at different scales or resolutions. If we look at a signal with a large “window,” we would notice gross features. Similarly, if we look at a signal with a small “window,” we would notice small features. This makes wavelets interesting, because they are able to mark both the big characteristics of a signal and also small little details. An example of a simple non-stationary signal is shown in Figure 2.7 and its corresponding continuous wavelet transform is reported in Figure 2.8.

During a musical interaction, of particular interest, is the emergence or absence of joint periodicities at different time scales. Accordingly, a melody may have a nested time structure that descends from the whole composition to its different parts, phrases, single notes and these periods may be performed by different musicians, each having their own specific part in the timing.

However, when timing interleaves, there should be a correlation in periodicities. Here it is where cross-wavelet transform become handy. Similarly, body movements may be segmented at different timescales, from big movement like walking

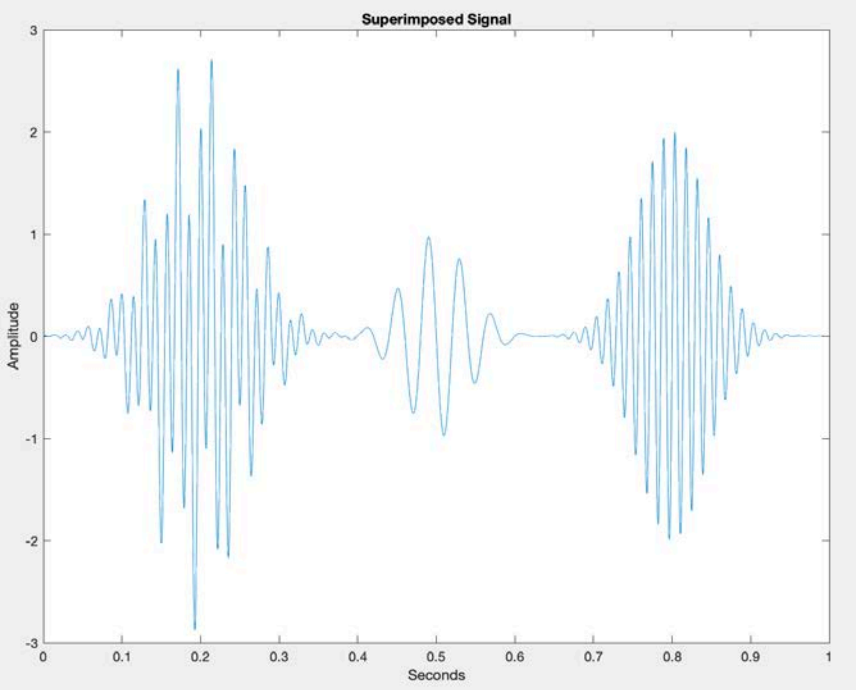


Figure 2.7: Signal consisting of exponentially weighted sine waves. There are two 25 Hz components, one centered at 0.2 seconds and one centered at 0.5 seconds. There are two 70 Hz components: one centered at 0.2 seconds and one centered at 0.8 seconds. The first 25 Hz and 70 Hz components co-occur in time.

to the oscillation of limbs or fingers. Each of these body parts is typically characterized by a distinct range of periodicities that interleave and correlate with each other.

The benefit from the use of cross-wavelet transform are shown in [35], were three couples of professional piano players improvised over three different backing tracks, half performances with visual information about their co-performer, half without, while their body movements and musical composition were recorded. The forearm and head movements of two pianists were recorded by a motion capture system. Couples of pianists were instructed to develop 2-minutes improvised duets on an *ostinato*, a *swing* and a *drone* backing track. The *drone* track was a uniform alternation of just two chords, the *ostinato* track was a complex four chords ascending progression and the *swing* was a bass line of a jazz.

XWT (Cross-wavelet transform) was used to assess coordination between two time series through spectral decomposition. It exposes regions with high com-

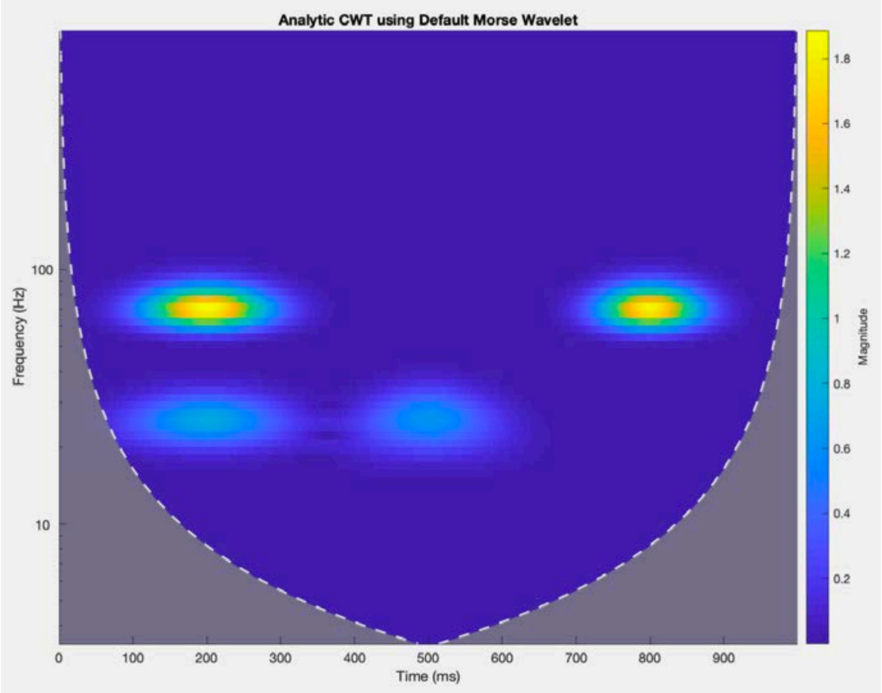


Figure 2.8: Continuous wavelet transform of the signal shown in Figure 2.7.

mon power and further reveals information about the phase relationship of the coordination that occurs between participants across multiple time scales. The results from a cross-wavelet analysis of the movement coordination that appear between the lateral movements of the right forearms of two piano players playing with the *ostinato* backing track are shown in Figure 2.9. The level of coherence between the movements of the performers' over time is denoted by the color: red for high coherence, dark blue for low coherence, and it is displayed as a function of period, in units of seconds, on the y-axis. Right arrows mean equal in-phase coordination, that is, the two systems are visiting the same states in perfect synchrony, instead, left arrows equal anti-phase coordination, perfect opposition. For example, Figure 2.9 (A) shows the results of the cross wavelet analysis when a couple was instructed to perform in synchrony with *ostinato* backing track.
It is noticeable that there is much less coherence and in-phase stability behaviour in Figure 2.9 (B) which is a cross-wavelet plot of the musicians improvising with one another, over the same track.

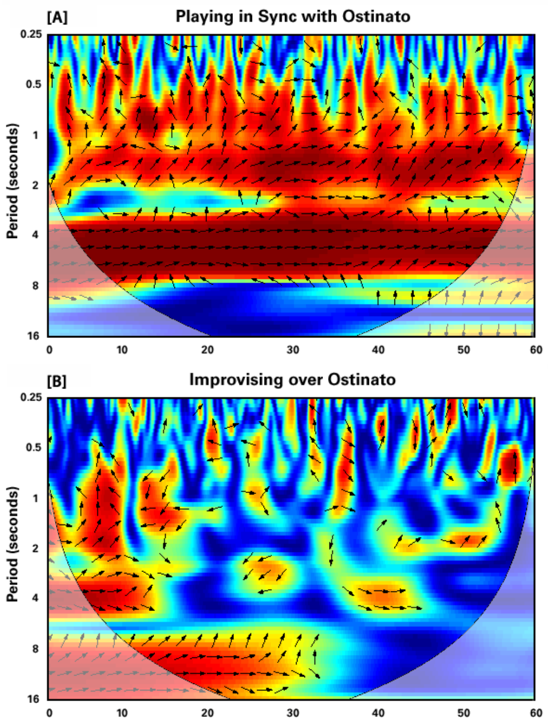


Figure 2.9: Cross-wavelet plots of the lateral movements of the musicians' right forearms, displaying the coordination while the two players played the exact same part, in synchrony with the *ostinato* backing track (A) and when the musicians just improvise over the same *ostinato* backing track (B).

Thanks to the deployment of cross-wavelet analysis, the time series of these movements showed different periodicities based on the features of the musical track. This analysis was used to capture how the musicians' movement coordination relates to shorter, longer-term temporal structure and phrasing of the musical context, also exploring how this coordination varies across different musician's bodies parts, and finally the effects of the visual information manipulations.

The *ostinato* track, for example, has a melodic phrase that is repeated every 4 seconds. Accordingly, the cross-wavelet plot reveals a high degree of coherence caused by musician's movement coordination, denoted by red colour, and in-phase coordination, right pointing arrows, at the four-seconds interval.

On the contrary, the *drone* track didn't show any specific periodicity, probably due to its simpler structure, that offered the musicians more variety of

motion. Instead, for the *swing* condition, wavelet coherence showed musicians' head movements coordination resulted at the faster time scale, between 0.25 and 0.5 seconds, respect the longer time scales of 4 seconds due to the right arm movements coordination.
The authors suggested the conclusion that expressive interactions are guided not only by brain processes, but also by bodily dynamics emerging on the fly, in accordance with the tenets of music cognition [39].

Closer to the research questions of this thesis, [34] investigated whether visual bouts of interaction in sequences of recorded jazz-duet-performances, evaluated by four experts watching them (without audio feedback), could be predicted by the musicians' movements, analysing the body movement coordination at different time scales by means of cross-wavelet transform.
Such bouts were defined as periods of interaction arising from the behaviour of the performers, where the characteristic movement patterns of the two musicians indicated a degree of correspondence in the eyes of the annotator and with the annotation of also the body part of the performers that were implicated in each bout. The work made use of two datasets of video recordings: one in which performances comprised a regular beat and metrical structure and the other in which the performance style is characterized by the avoidance of a regular, predictable beat and metre, in other words, free improvisation. This allowed to investigate how the degree of predictability afforded by the metrical structure and the pulsed or not nature of the music might differentially influence the types and timescales of movement coordination between the partners. The authors defined in total 12 possible predictors of visually bouts of interaction (Table 2.1) to be extracted from all the performances: 9 potential movement predictors and 3 more predictors computed from the audio data of the duo performances.

To assess the classification accuracy of the set of predictors were used two complementary classification techniques: logistic regression and random forest classification. The results suggests that at least four of the predictors failed to predict interaction bouts in any of the datasets, specifically: Movement CWT Phase, Audio Pulse Clarity, Audio RMS, and Audio WT Energy (Broad) were discarded.
Another remarkable result from this analysis is the difference in prediction accuracies between the non-pulsed and pulsed duos. In the pulsed duos, the annotated

Table 2.1: Predictors of annotated bouts of interaction [34].

	Predictor name	**Description**
1	Movement CWT Energy (Broad)	Energy of the cross-wavelet transform over frequency band (0.3–2.0 Hz).
2	Movement CWT Phase	Phase of the cross-wavelet transform, indicating the momentary lead/lag relationship between the performers.
3-7	Movement CWT Energy (CF)	Energy of the cross-wavelet transform, computed using frequency bands where the centre frequencies were 0.3, 0.4, 0.6, 0.9 and 2.0 Hz.
8	Movement WT Energy (Any)	Energy of the wavelet transform, computed for each individual performer.
9	Movement Quantity	Summed quantity of motion from both performers, computed using frame differencing.
10	Audio RMS	Amplitude envelope of the audio signal in terms of the root mean square energy.
11	Audio WT Energy (Broad)	Energy of the wavelet transform, computed from the audio envelope over a broad frequency band (0.25–10 Hz).
12	Audio Pulse Clarity	Clarity of the pulse sensation.

interaction bouts were usually more difficult to classify with the individual predictors compared to the non-pulsed duos. Here, as well as in [35], the answer depended on the kind of music the duets performed.
The best prediction rate were obtained using logistic regression models for the non-pulsed duos with only a single predictor, but these models surprisingly failed to deliver statistically significant improvements when additional predictors were added. This suggests that there were either interactions between the predictors not correctly specify in the models, or that there were non-linear relationships between the predictors.
At the end, the annotated bouts of interaction were successfully predicted for the non-pulsed condition, primarily by the CWT energy of the movements, across a broad frequency range, and followed by moderately fast co-occurring movements, as indexed by Movement CWT Energy in frequency bands centred around 0.9 Hz and 0.6 Hz. This means that non-pulsed duo performances were characterized by shared periodic movements across a broad frequency range, although there was also some tendency for mid-range-frequency.
Instead, different optimal predictor emerged in the pulsed duo performances, in which interactions were characterized by slower shared periodic movements (Movement CWT Energy centred in 0.4 Hz) and periodic movements at unrelated frequencies (Movement WT Energy (Any)). Furthermore, simple additive logistic regression models failed to improve the model classification rates.
The authors explained this difference pointing to the structure of the tunes: while jazz standards exhibit clear formal boundaries that allows for clear and easy coordination, in particular in moments of transitions from one musician to the other one, on the other hand free improvisation requires to musicians more relevant visual communication established by their bodily movements. The outcomes of the presented work [34], are important for developing new computational methods with the aim to approximate human judgements of meaningful coordination between co-performers.

As we have seen, despite its wide use, cross-wavelet analysis hardly captures coherence when different periodicities are involved, as in a musical interaction. Actually, most of the studies evaluated the body interaction of musicians, indeed, it is considered as a way to enhance their musical joint action, but, as shown in the free improvisation in [34] and non-pulsed duos in [35], some formal structure profited of it more than others. In general, the studies on embodied cognition suggest that empowering effects of such an interaction are dependent on quality, that is, on properties of the performance. Although no one of the previous reported works was explicitly addressing the expressive quality of a performed musical interaction, they can be interpret as steps toward a definition of such a quality.

In this thesis we address exactly the question whether there are measurable markers of a qualitative performance/expressive joint action building a model based on the Bayesian statistics.

2.5 Bayesian brain hypothesis

A "predictive brain" hypothesis theory ranks among the most promising and the most challenging ideas ever to emerge from computational and cognitive neuroscience [40]. The Bayesian brain hypothesis [41] employ the Bayesian probability theory, to draw up perception as a constructive process based on generative models. It is a nowadays largely debated theory that sees the brain as a generator of predictions, rather than a passive receptor of stimuli from the external world, that are processed and used as premises for action. The underlying idea is that the brain represents sensory information probabilistically, in the form of probability distributions and for which has a model of the world that it tries to optimize using sensory inputs [42].

Bayesian approach to brain functions explore the capacity of the nervous system to manage situations of uncertainty in a way that is close to the optimal prescribed by Bayesian statistics.

This recent theory assumes that the brain maintains internal probabilistic models that are updated by neural processing of sensory information using approximation methods characteristic of Bayesian probability [43].

Therefore, brain can be seen as an inference machine that continuously predicts and explains its sensations [44]. The core of this hypothesis is a probabilistic model that can generate predictions, against which sensory samples are evaluated to update beliefs about their causes and so update the model.

This generative system is mainly composed by a prior probability of those causes (given by a prior knowledge) and a likelihood, that is, the probability of sensory data, given their causes. Every time new data are available, the likelihood is responsible to access the posterior probability distribution by an update of the prior probability distribution. In a social context, the feedback is provided by the other interacting subjects' behaviour.

As reported in [45], music is a perfect case against which the predictive model may be tested, caused by its very syntactic structure implies rhythmic, melodic and harmonic expectancies, that is, prediction. Bayesian regression seems to be a good method to study our problem and to extract measures of performance relative to a Bayes' optimal observer.

There is a lack of studies of musical interaction deploying Bayesian methods to investigate expressive quality of musical interaction to date.

In this thesis, believing that accurate prediction of the partner's action is crucial in musical ensembles [46], we explore a new model based on Bayesian coding hypothesis to find significative parameters that can denote the quality of a musical interaction.

3
Just hock it

In order to investigate the quality of musical interaction in duets performances, it was necessary to create the environment to develop the tests.

In this section is described the experiment that was created in the laboratories of IPEM, and in which the analysis of this study will focus on.

As we have said, since the timing of the musical notes is one of the main characteristics in a joint (musical) action, the target of the study is to understand how and which parameters that we can extract and compute from a live performance, can be significative to evaluate the quality of the musical interaction.

3.1 Participants

The ethics committee of the Faculty of Arts and Philosophy of Ghent University approved the study and the consent procedure.

Starting in 2018 and ending in 2019, a total of 15 couples of musicians were recruited, 16 men and 14 women with an average age of 30.7 years old. Each duo could be formed by 2 men, 2 women or both gender together, the only requirement was that participants of each pair knew each other already.

As musicians we meant people currently playing an instrument or singing, with at least 5 years of regular (formal or informal) musical training, that had to be capable of singing a simple melody written on a music sheet.

The gender and the age resulted well balanced in the sampled couples.
By means of a questionnaire given before the recording session started, were collected also information about subjects' musical background, familiarity with music ensemble and with the partner.
Participants that took part in the experiment, were informed in advance about the task, the procedure and the technology used for the measurements. They had the opportunity to ask questions and were informed that they could stop the experiment at any time. There was the necessity to exclude 3 other couples from the analysis, because they couldn't perform the assignment and asked to stop the experiment or because of technical problems with the equipment.

3.2 Stimuli

Since the most relevant parameter of the score that we have to choose for this case study, is the length of the notes that will characterize the singing dyad, we explore this topic taking the advantages of the medieval style called *hocket*.
In music, hocket (whose etymology has to do with "hiccup") is a rhythmic technique using the alternation of notes between the voices of the ensemble, such that one voice sounds while the other rests. In other words, it requires that two or more musicians build a melody together by singing parts that never overlap. In origin, the term referred to an inuit throat-singing but is just an instance of the hocket.
The choice of the hocket style is due to its intrinsic intersubjective nature according to which, even the fundamental musical elements, timing, rhythm and melody, are made possible only by the interplay of two (or more) voices.
For these reasons hocket singing is a notable example of joint musical action.
During a singing dyad that employ this medieval style, several things can be observed, like the regularity of the length of the notes played, synchronization, tuning, the movement and pressure made by the body of the performers and physiological measures (stress and heart rate). These are likely to be components of an optimal interaction, or better, like reported in [4], of homeostasis, which is a state of being where cognitive and motivational brain mechanisms reinforce each other. Thereby, adopting the hocket rhythm style, the quality of

the outcome strongly depends on the interaction between performers, and thus on the contribution of each part to the whole.

In hocket polyphonic style, indeed, a single melody is split into two or more parts in which the notes alternate almost regularly. Here timing is a key issue and regularity is needed to build the required melody.

Musicians were asked to sing an interleaved melody "on stage" and due to a short and limited rehearsal time, the task was expected to be challenging, in order to lead to different outcomes in performances quality.

The musical stimulus (the score) that people involved in the experiment had to sing, was created by Alessandro Dell'Anna from a universally known pop tune that the majority of participants have recognized after the introductory bass riff. The song we are talking about is a modified version of "Billie Jean" by Michael Jackson. Nevertheless, as a self-encouragement and word pun, the melody used for this experiment was named "Just (like) beat it" (reminiscent of another Jackson's greatest hit) and split it into two interlacing parts: "Just" and "Beat", assigned to each of the two participants to be sung in a hocket style, hence "Just hock it". Figure 3.1 shows the first part of the melody "Just" assigned to the first individual of each ensemble and Figure 3.2 report the "Beat" part, the score assigned to the other partner.

The two parts are complementary, this means that the notes never overlap, forming a single melody. Each part is divided in 2 sections called segment A, constitute by the first 8 bars (first two rows) and segment B, constitute by the last 8 bars (latest 2 rows). Each of the 2 scores ("Just" and "Beat") contain 40 notes, distributed in 16 bars, with a 4/4 meter and a 120 bpm tempo indication. The only note used is the eighth note, or a quaver. It is a musical note played for one eighth the duration of the whole note (semibreve) that is the same amount of twice the value of the sixteenth note (semiquaver) or half the duration of a quarter note (crotchet) or still, one quarter the duration of a half note (minim). It can be easily found in Figure 3.1 and Figure 3.2.

Furthermore, in order to create a sort of song, the "Just" and "Beat" parts had to be repeated four times in total, without interruptions. This allows to create a performance that lasts approximately 120 seconds. For these reasons, it was required that the performers have to be musicians with at least 5 years of experience, with the capability to read a score and to perform in an ensemble.

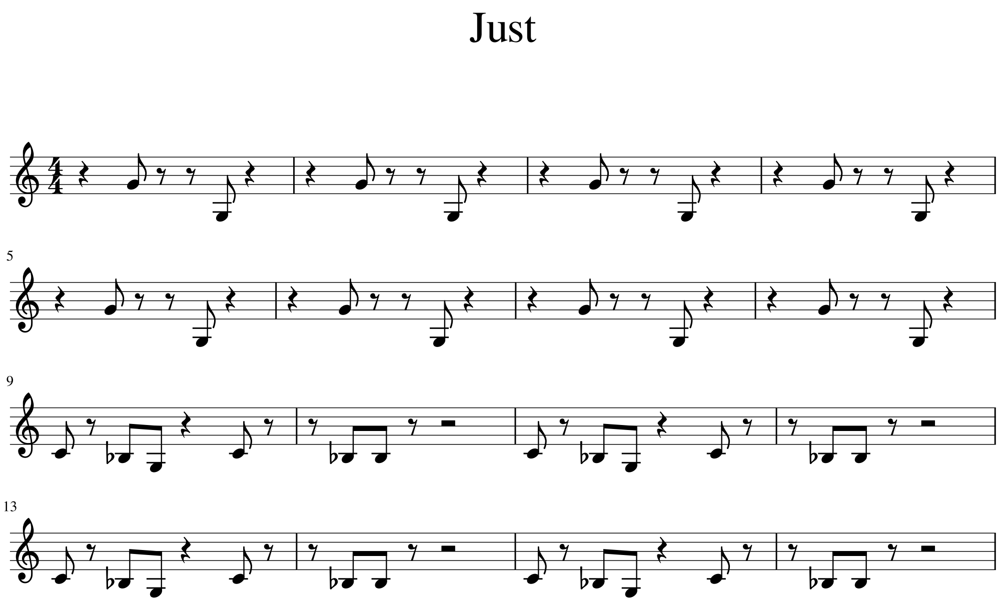


Figure 3.1: Score assigned to the first subject of each couple.

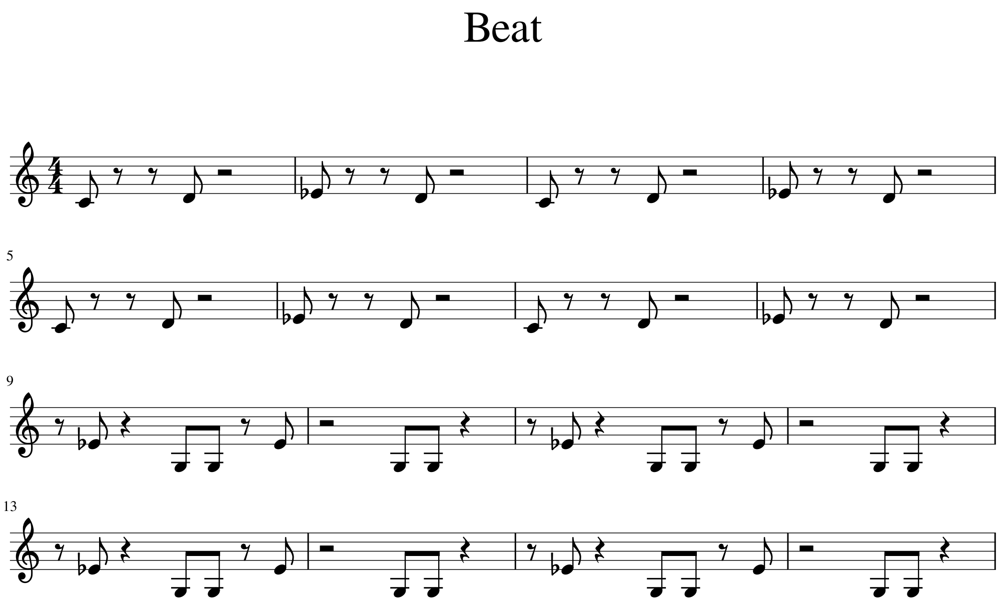


Figure 3.2: Score assigned to the second subject of each couple.

3.3 Setup

The entire experiment was performed at the ASIL, the most advanced laboratory in Europe, to conduct music related studies. It belongs to IPEM research centre of the University of Ghent and was built by the desire of Marc Leman in 2017. It provides a big stage with 3D audio speakers and motion capture system.

However, the setup for this experiment was quite simple. It was composed by two microphones, attached directly to the head of the participants, to be really close to the mouth. The ASIL is soundproofed and the acoustic is dry. In this way having the microphone close to source it ensure neither reverb nor eco in the recordings.

The performers have to stand on top of two balance boards facing each other at 1.5 m. The balanced board is able to capture the pressure and the movement of the performer's body, thanks by 4 sensors positioned on the four corner of the boards. The data relative to the pressure that each sensor was detecting, were stored as MIDI data.

The entire session was recorded with a Logitech webcam positioned in the middle of the subjects in order to record the entire scene with the whole body of the performers in frame.

Microphones, balance boards and webcam were all connected to the main computer running Ableton to record all the signals generated from each performance. An example of the stage and the setup used, is shown in Figure 3.3.

3.4 Procedure

Upon arrival of each couple at the IPEM laboratories, they were firstly asked to fill the informed consent. An oral explanation of the task followed, in which participants were informed that they had to build a melody together, combining the notes from the two voices, stressing that they will never overlap, just alternating one note after one another, passing from one subject to the other.

Moreover, individuals were not allowed to read the partner's score.

The whole experiment was divided in 5 subsequent time sessions.

A the beginning the subjects were free to rehearse their part in two different

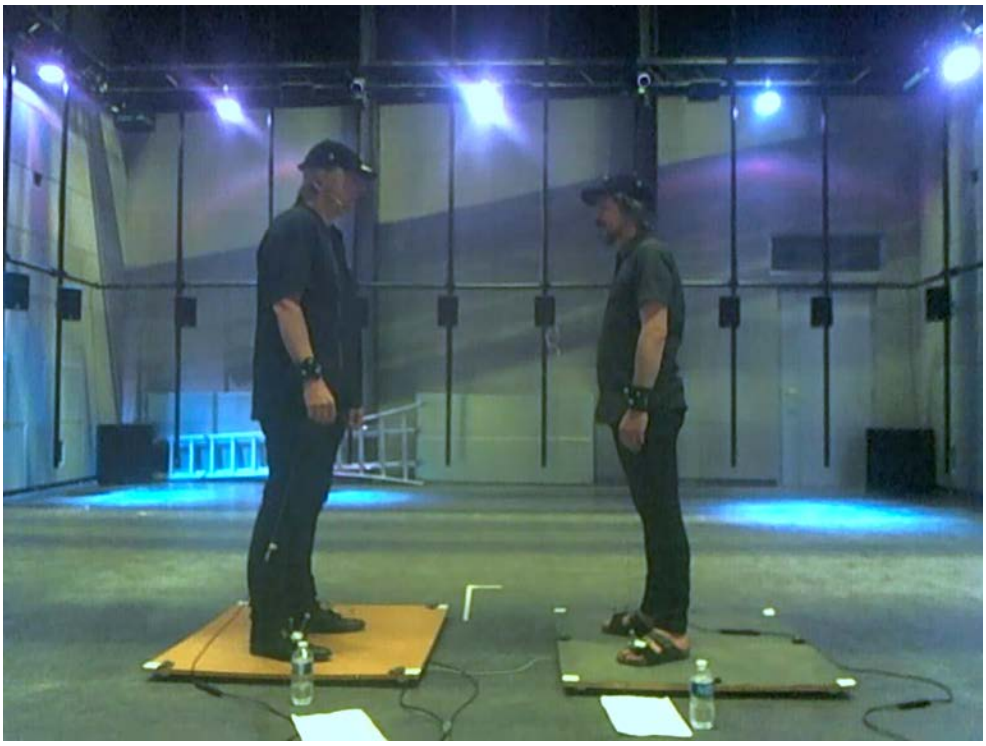


Figure 3.3: Recording session of couple 11 in the ASIL. Performers standing on balance boards in order to measure the movements and the voices were recorded with microphones directly mounted near the mouth.

rooms for 5 minutes and they were allowed to listen one or two times their own melody to capture the right pitch.

Then, they were gathered in the Art-Science-Interaction Lab (ASIL) and invited to get on the stage.

When all the devices and tools were correctly connected, 15 minutes were reserved to allow performers to rehearse, for the first time, together, before the beginning of the real recording session. However, also the training phase was recorded, so that information about the process of learning and coordination could be available as well.

Even if one or both participants were not entirely smooth in the execution of their performance, the training phase stopped after 15 minutes and the two performing condition were explained. They were required to perform 8 distinct trials, each one spaced by a little break of 1 or 2 minutes.

Before the start of each trial they were told which one of the 2 conditions they

had to perform: the moving condition, in which subjects were encouraged to move according to the music in whatever way the liked, always staying on top of the balance board, and the still condition (or non-moving), in which no kind of movement were permitted (except for those strictly necessary to sing).
In each single trial, after 4 beats of a metronome, subjects had to start singing, repeating four times the score assigned, with a suggested rate of 120 bpm.
In this way one single trial lasted about 120 seconds. The total recording session consisted in 4 moving trials and 4-non moving trials in a randomized order, and lasted about 20 minutes.
A final requirement was to keep going even if one or both performers got something wrong, like some missing notes, and to continue until the experimenter have said "stop". Furthermore, a recording of each participant alone in the non-moving condition was taken as a baseline before and after the 8 experimental trials. In total, this phase lasted about 45 minutes.
A second phase followed, in which participants were separately asked two questions for further investigation. Musicians were asked whether they could sing the whole melody alone, this means the entire song made by the union of the 2 music sheets, and whether, after a listening session, they were able to identify the correct melody between the right one and another similar one.

3.4.1 Self-assessment

After about 20 minutes of synchronization of the audio and video signals by means of a script, a recording of all the trials of a couple was made available for a manual annotation. Participants were then placed in front of two different monitors, without having the possibility to look at the monitor of the partner, for the self-assessment phase.
From the subjective point of view we explored two factors. Firstly, they were given two headphones and they were asked to assess the quality of their performance in its development over each single trial. They had to listen and watch each entire trail and in the meanwhile they had to rate the musical coordination of the pair. The suggestion was to assess the quality of the interaction, rather than the quality of one single performer. Through a proper software they could use the mouse to move a slider and rate the quality of the interaction with a

score from 0 to 120, while the video recordings was going. As you can see in Figure 3.4, all along the video appears a line that is the corresponding evaluation of the interaction in each moment.

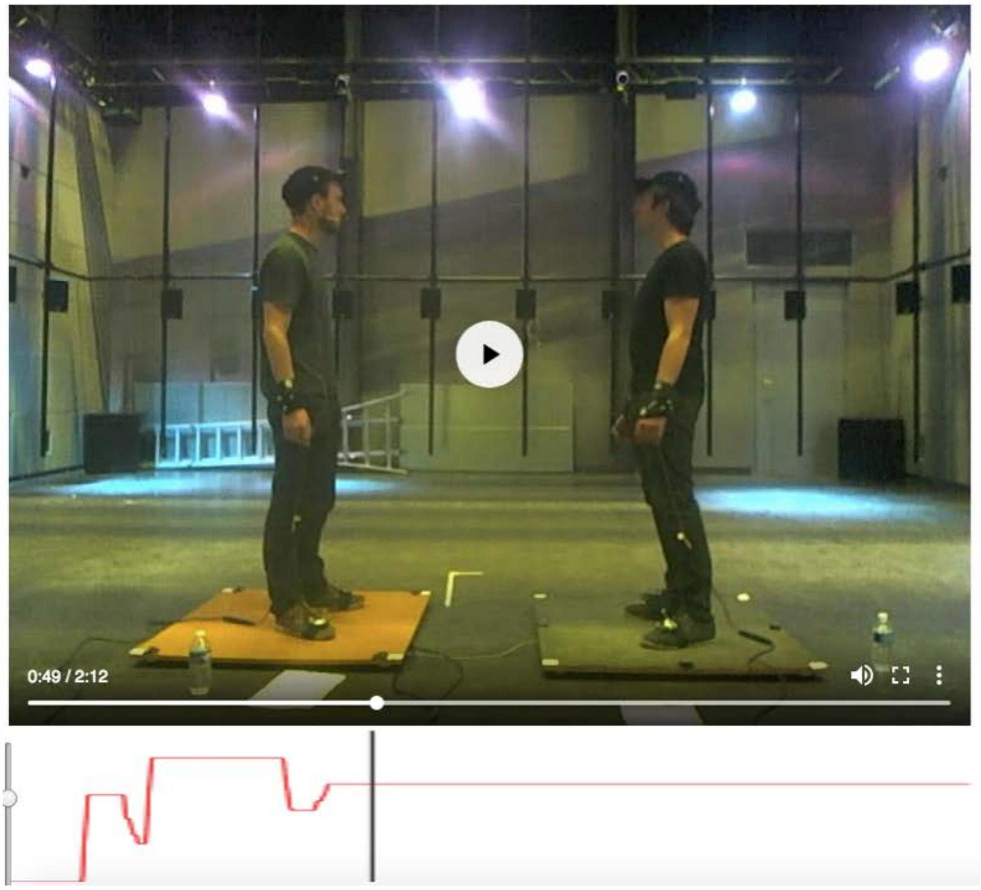


Figure 3.4: Manual annotation of the quality of the interaction. While watching and listening to the recorded trial, the slider at the bottom on the left, can change the score given to a specific moment of the performance.

In the case musicians were unsure about their assessment, they could stop the video, go back and reassess it. At the end of the process, for each trial, they just had to save their assessment.

The subjects had to assess the trials independently, without being affected by the partner's judgment, looking at her/his monitor.

This type of performance evaluation is particularly useful because it allows to see the development of the assessment throughout an entire exhibition. As you can see from Figure 3.5, it's easy to see where errors occur, with drops in the score evaluation, or if the interaction is stable, marked by a continuous line.

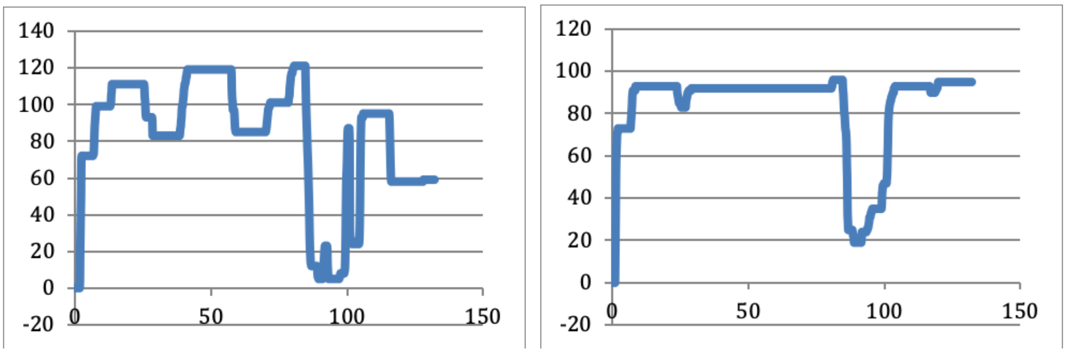


Figure 3.5: Two example of annotations. On the left there is the annotation of one subject and on the right the annotation of the partner for the same trial. X-axis represent time in seconds, y-axis is the score of the quality from 0 to 120.

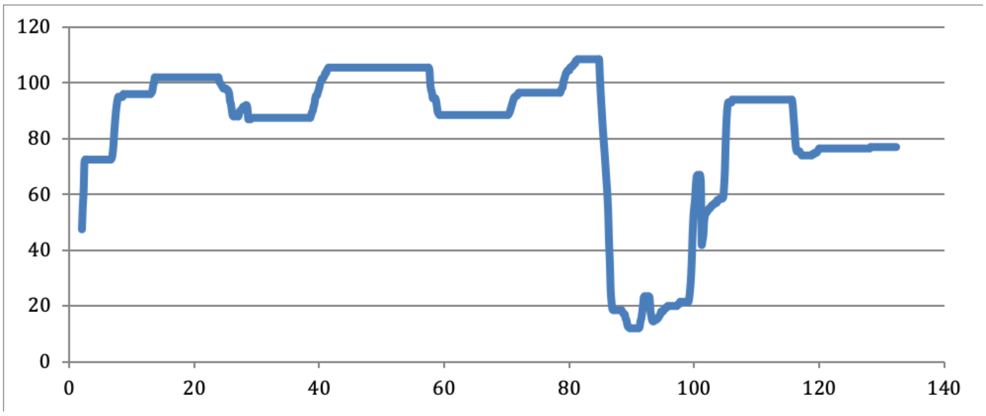


Figure 3.6: Average of the two subjects annotations. Moments of optimal interaction between 95 and 120, if lasting at least 10 seconds.

Nonetheless, from this over time evaluation can be extracted the mean score as a single value to sum up the entire annotation of each trial.
A drawback could be the lack of more specific details about how to establish the score to the quality in a scale from 0 to 120. This may cause some different interpretations among assessors, as it can be seen in Figure 3.5. Even though the evaluation method may seem a bit vague, we will see, in the results, that these assessments show some degree of coherence and can be compared to objective measures.

After this first self-assessment phase, musicians had to assess the second subjective factor.

They were required to listen again all the trials one by one and measure the joint sense of agency, that can be summarize as how much the performers are perceived as one single unity.

The complex issues of agency have been examined by researchers in the related fields of psychology and neuropsychology.

At the individual level, the sense of agency is defined as the subjective feeling of being in control of a given action, compared to being subjected to an action or event [47]. In performing music, this is the awareness that a certain action will produce a sound with this pitch or that intensity of the tone. This process implies a sensorimotor prediction because a given motor action is connected to a given sensory outcome [48].

The phenomenon of sense of agency can also be generalized from the individual to the group, distinguishing a "shared" from a "we" sense of joint agency [49].

Pacherie, in [49], reported that if an action is a joint action, the relative sense of agency should be a joint sense of agency, that means, a feeling of being in control of at least a part of the joint action outcome.

Furthermore, in the case of joint agency, Pacherie [49] stresses the importance of predictability: whereas a shared sense of joint agency should results from a low predictability of the partner's action, a we-agency should presuppose high similarity among partners' actions and, as a result, high predictability.

In this phase, musicians were asked to look at the highest moments of their assessment for each trial and evaluate the joint sense of agency, explained as the feeling of control over the process on a scale between 0 (independent), 3 (shared) and 6 (complete unity with the partner), as shown in Figure 3.7.

This question was explained saying that the interaction could be either the product of two actions not really well coordinated between them (independent) or the product of two coordinated but distinct actions (shared) or the product of two actions that are not felt as different, but rather as the accomplishment of a single subject.

Whatever the limits of such a distinction, the interest lies in the possibility of identifying a level in which the contributions of the (two, in this experiment) partners are clearly distinct and another level in which they blend into one, lead-

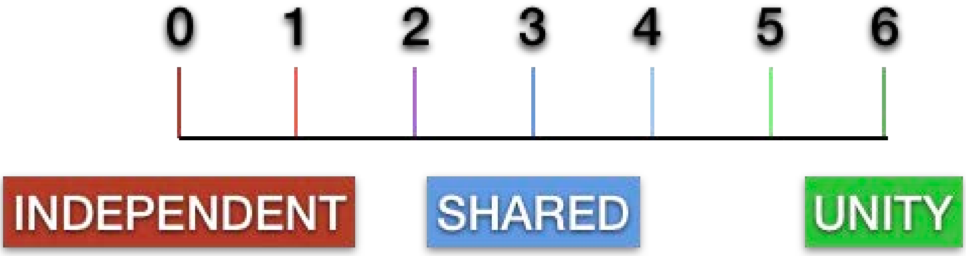


Figure 3.7: Scale to evaluate the joint action sense of agency.

ing to a kind of boundary loss between agents.
After listening every trial, musicians had to evaluate the joint sense of agency writing the relative number in a form.
Therefore, the self-assessment phase was composed by two different evaluations: annotation and agency, for each trials, so that subjective annotations could be compared with an objective assessment of the timings.
The self-assessment phase took about one hour of time.
Finally, at the end of this process subjects completed the questionnaire leaving some general comments and opinions about the experiment.
In total, each couple required between 1.5 and 2 hours of work to complete the whole experiment.

3.5 Hypothesis

Interaction quality is observable in bio-social markers, for example, in acoustical expressions, such as timing, and in bodily expressions, like moving.
We assumed that the interplayed melody required a constancy in timing. Due to the fact that the timing of one singer affect the timing of the other singer, both timings had to be predictable. This means that partners should be able to predict each other's timing in order to perform correctly according to the musical score, although expressive timings, such as change in tempi, were possible.
The literature and the background developed so far lead us to the following hypotheses:

- Musical rules, described through a musical score, define a framework for interactive performances but not all performances are empowering, or have an empowering effect. To reach that empowering effect, it is necessary to establish an interaction quality. Optimal interaction moments are characterized by synchronization between the performers, that means regularity of the notes duration.

- The empowering effect of a performance can be experienced by the performers themselves. Performers can estimate and evaluate their own performance quality, from a third-person perspective, through annotation values and they can access their own experience, from a first-person viewpoint, in terms of the agency experienced during the interaction.
- We hypothesize that timing and movement are correlated with annotations and with estimations of agency. In particular, we assume that more accurate timing is correlated with higher annotation values, and with higher values for joint-agency experiences. Given the high similarity of the singers' parts, our hypothesis is that the highest quality moments of interaction would be characterized by higher sense of joint-agency values, that is, by we-agency.
- We hypothesize that movement would allows subjects to embody their timing, improving the overall quality of the performance respect the non-moving condition and in general allows them to make their timing more accurate. We hypothesized that allowing performers to move, in accordance with the major tenet of the embodied cognition framework [50], should bring better performances.
- Of particular interest is also the question whether learning plays a role in reaching interaction quality. We expect improvements in synchronization and timing over the repetition of trials.

After tested 15 couples of musicians and stored audio files, MIDI movement data, self-assessments and video recordings of a total of 120 trials, we are ready to discover the new time-evaluation model based on Bayesian statistics, used to create objective measures about the quality of the performance.

4

Time-evaluation model

In this chapter is described the analysis of the data previously collected during the experiment with the couples of musicians.

The core of the analysis, comes from an innovative idea of Prof. Marc Leman head of IPEM, and employ a new methodology for this filed of research, based on Bayesian inference. As we will see, it allows to define 3 objective measures, about the quality of a performance, that we can compare with the subjective measures, the self-assessments, already collected.

The analysis are mainly done using MATLAB R2018b.

As we have said, in order to evaluate the quality of a musical performance we limit ourselves to the analysis of timing, excluding other parameters like pitch, loudness and articulation. Unlike previous data-analysis methodology which assumed a stationary tempo during the performance, or strict hocket execution based on reciprocal actions, here it is adopted a more dynamic data-analysis approach.

We gave to the performers the indication of a 4/4 meter and 120 bpm, but we assume that the timing is based on the tempo that emerges from the interaction. There is not one or the other performer that is the reference for the timing-analysis.

We assume non-stationarity because we do not require a stable tempo during the performance and musicians are considered as components of an interaction dynamics. Each subject is part of a joint action and its relative timing characteristics.

To regulate that dynamics we assume that each musicians makes a prediction of the global timing of the joint action (the interaction). According to the predictive coding theory we assume that performers constantly adjust their predictions about the joint timing. The data-analysis approach is based on the idea that performers try to reduce performance errors based on predictions. The fundamental prediction in the domain of timing is about tempo and note durations.

Given the particular nature of the hocket music style adopted in this study, we assume that the interaction quality largely depends on the regularity of the notes duration sang during the interaction.

Furthermore, our task could be considered as a semi-hocket style because the performance is based on mutual actions rather than reciprocal actions. Indeed, according to the score used in this experiment, performers could sing two consecutive notes, rather than only one note as in a reciprocal action.

Since we are interested in the joint timing of each performance, with the audio file collected from the experiment, we can now extract onset and inter-onset intervals.

4.1 Semi-automatic onset detection

An inter-onset interval, or IOI, is the time elapsed between two consecutive onset. More specifically, it is the time between the beginning of one note and that of the next note. It can be determined by one performer, that is singing two notes after one another, or by two performers, each singing a note in sequence.

The IOI is generally considered to be the strongest contributor to accent and pulse perception.

IOI are independent to the duration of the sounding notes, which in this experiment are always shorter than the IOI, because the notes of the employed scores never overlap, resulting in a silence before the next onset.

The corresponding term used in acoustics and audio engineering to describe the initiation of a sound is onset. It refers to the beginning of a musical note, or in general of a sound, in which the amplitude rises from zero to an initial peak.

Onset is related to the concept of transient: all musical notes have an onset,

but do not necessarily have an initial transient, that is characterized by a high amplitude, short-duration sound at the beginning of a waveform.
Audio onset detection concerns itself with finding the time-locations of all sonic events in a piece of audio. An example is reported in Figure 4.1.

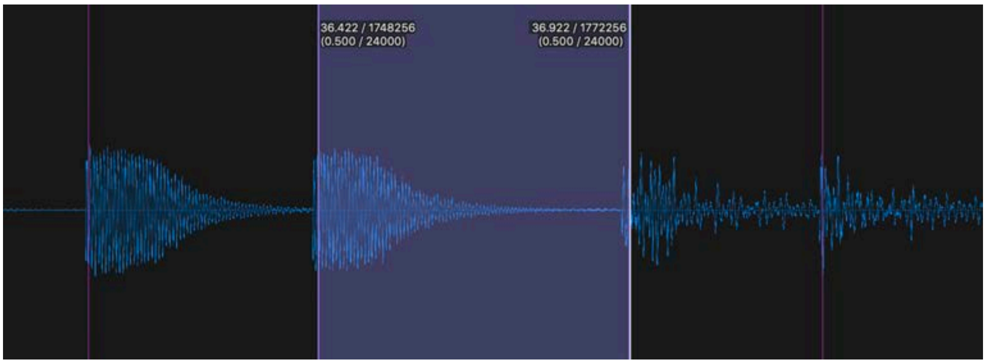


Figure 4.1: Example of onset detection applied to an audio file. The vertical purple lines determine the onset found and the highlighted area is the IOI, showing an interval of 500 ms.

In signal processing, onset detection is an active research area. The different approaches for the detection can operate in time domain, frequency domain or complex domain. For example, they can look at the increasing amplitude of a signal in time-domain, the increasing of spectral energy, changes in spectral energy distribution or they can even look at spectral patterns recognizable by machine learning techniques such as neural networks. Generally every technique should allows to adjust some parameters in order to adapt the onset detection algorithm to a specific audio file, in order to minimize the amount of false positives and false negatives. Moreover, the effectiveness of an onset detection algorithm increase when the notes on the audio file have a clear and strong attack.
In music, the term attack refers to the manner in which a note is performed by the musician, whether decisive and quick, or smooth and slow.
For that reason, we asked performers to sing every note emitting a sound similar to the pronunciation of “ta” or “pa”, in order to ease the onset detection phase.
For the onset analysis we used Sonic Visualizer. It is an application for viewing and analysing the contents of music audio files, developed at the Centre for Digital Music, Queen Mary, at University of London. It is particularly handy to use

and it has an onset detection function built in. Once an audio file is imported in Sonic, it is possible to apply the onset detection and adjusting some thresholds based on the amplitude of the signal to analyse.

Although many parameters, of the onset detector algorithm, are adjustable, there are some conditions in which the onset detector is not able to correctly find all the notes.

In this experiment, for the majority of the trials, the onset detector struggle to identify two really closed onset, as shown in Figure 4.2. This happened when there wasn't a clear gap between the amplitudes of signals relative to two different notes.

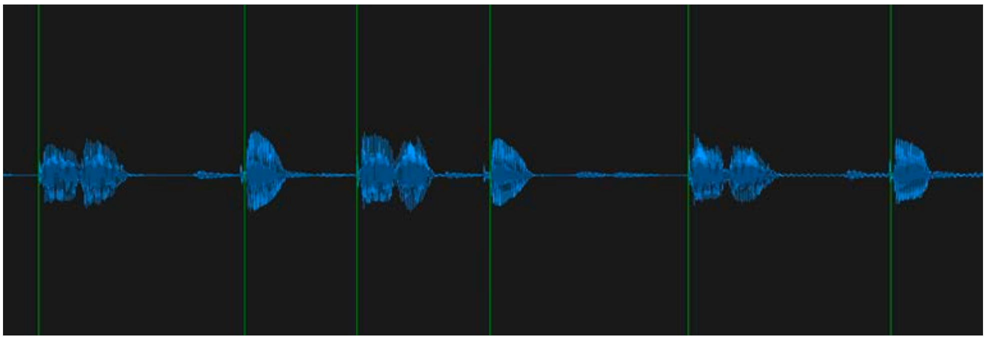
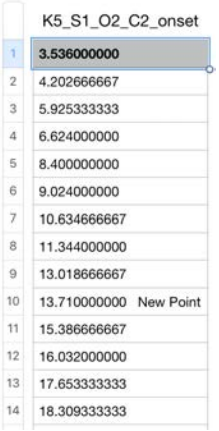


Figure 4.2: Example of false negative. Two consecutive notes, in which the audio signal is not clearly separated, are captured as one single onset by the transformation function of Sonic Visualizer.

The false positive case happened rarely and was solvable just tuning the onset detector's thresholds.

As you may have noticed, this process always required a manual check of the onset detected and eventually a manual correction, this means, adding onset that the algorithm wasn't able to find or removing those in excess. This is the reason why we can refer to this process as semi-automatic onset detection.

After manually checked the onset, with Sonic Visualizer was possible to export the timestamp of the onset in a CSV file, as shown in Figure 4.3, that was directly manageable by MATLAB. These operations were repeated for each performer of every couple for all the trials. This phase required a lot of time, but it was necessary to pay attention and to analyse carefully every audio file because the further analysis will employ mainly the onset data extracted at this step.



	K5_S1_O2_C2_onset	
1	3.536000000	
2	4.202666667	
3	5.925333333	
4	6.624000000	
5	8.400000000	
6	9.024000000	
7	10.634666667	
8	11.344000000	
9	13.018666667	
10	13.710000000	New Point
11	15.386666667	
12	16.032000000	
13	17.653333333	
14	18.309333333	

Figure 4.3: Example of the first part of a CSV file with the timestamps of the onset extracted for the second musicians of couple 5 during the second trial. “New Point” indicates a manually added onset.

4.2 Inter-onset classes

It is important to notice that the CSV files extracted, contains only the onset of each individual. To compute the IOI of the musical interaction of every trial, it is necessary to merge the onset of both performers and then calculate the difference between the consecutive onset. In this way, using MATLAB, it is possible to compute the IOI of the joint action.

Looking at the scores used in this experiment, it can be noticed that merging together the two music sheets (“Just” and “Beat”), as it results from the joint action in the performances, the time intervals between two consecutive notes can be of 3 different durations.

There are only 3 different intervals length that, according to the score, should appear in the performances and those are: half note, quarter note and eighth note, as reported in Figure 4.4. Each one is called a duration class or better: an inter-onset class. With the specifications given with the scores, the fastest note, the eighth note, should be caught in each performance within the range from 200 ms to 400 ms. Then the second interval, the quarter note, should be performed as the double of the eighth note, at around 600 ms and finally the half note at around 900 ms.

We didn’t use any metronome to help the couples to align to the perfect tempi and then to maintain it, because musicians had to be free to express themselves

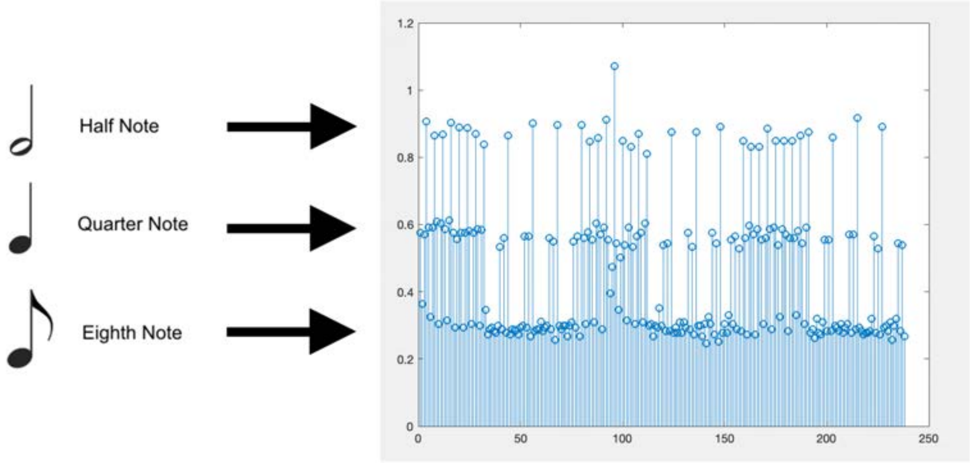


Figure 4.4: Here it's easy to notice the 3 different IOI that appear during a performance that respects the notes length of the scores. On the x-axis there is the sequence of IOI all over the performance and on the y-axis their relative value, i.e. the duration of each interval.

and to adjust the timing based on their interaction capabilities in a musical ensemble. Fo this reason we have to create a classification mechanism that is able to categorize each IOI emerged in a performance, assigning it to one of the 3 inter-onset classes, memorizing the development of each classes to finally predict the future behaviour of each one of them such as reflecting the behaviour of the human brain. For this purpose we exploited the Bayesian brain hypothesis.

4.3 Bayesian inference approach

Bayesian inference is an important technique in mathematical statistics. In general, (statistical) inference is the process of deducing properties about a population or probability distribution from data. Bayesian inference is therefore the process of deducing properties about a population or probability distribution from data using Bayes' theorem.

The key feature is that this method allows to update a probability for a hypothesis as more evidence data becomes available.

Bayesian updating is widely used in a broad range of activities such as engineering, medicine, philosophy and it is computationally convenient, but it is not the only updating rule that might be considered rational.

Moreover, Bayesian probability matches particularly well the problem we are facing in this study, because it interprets the probability as an expectation representing a state of knowledge.

Bayesian inference derives the posterior probability as a consequence of a prior probability and a likelihood function derived from new observed data. Therefore, it requires to specify a prior probability, that is updated to a posterior probability in light of new data (evidence).

Our goal is to update our knowledge of μ, the expected value of an inter-onset class, with a new IOI emerged from the performance; i.e., we wish to find $p(\mu|X)$ where X is the new IOI data. Using the general Bayes' updating rule [51]:

$$p(\mu|X) \propto p(X|\mu) \times p(\mu) \tag{4.1}$$

where $p(X|\mu)$ is the likelihood function for the current evidence data and $p(\mu)$ is the prior probability for the inter-onset class mean.

Assuming that the each class of IOI is Normally distributed with a mean of μ and a variance of σ^2, then the likelihood function for X is:

$$p(X|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \tag{4.2}$$

Assuming also the prior distribution for each class of IOI as Normally distributed, with mean M and standard deviation τ, the prior distribution results:

$$p(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{(\mu-M)^2}{2\tau^2}\right\} \tag{4.3}$$

Plugging the likelihood and prior into Bayes' rule gives us:

$$p(\mu|X) \propto \frac{1}{\sqrt{\tau^2\sigma^2}} \exp\left\{\frac{-(\mu-M)^2}{2\tau^2} + \frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\} \tag{4.4}$$

As we will prove, the posterior can be re-expressed as a Normal distribution. Since the terms outside the exponential are normalizing constants with respect to μ, we can drop them. We therefore focus on the exponential.

Let's re-write the terms inside the exponential:

$$-\frac{1}{2}\left[\frac{\mu^2-2\mu M+M^2}{\tau^2}+\frac{\sum_{i=1}^n x_i^2-2n\bar{x}\mu+n\mu^2}{\sigma^2}\right] \tag{4.5}$$

Any term that does not include μ can be viewed as a proportionality constant, can be factored out of the exponent and can be dropped (recall that $e^{a+b}=e^a e^b$). Therefore, we obtain:

$$-\frac{1}{2}\left[\frac{\sigma^2\mu^2-2\sigma^2\mu M-2\tau^2 n\bar{x}\mu+\tau^2 n\mu^2}{\sigma^2\tau^2}\right] \tag{4.6}$$

$$-\frac{1}{2}\left[\frac{\mu^2(n\tau^2+\sigma^2)-2\mu(\sigma^2 M+\tau^2 n\bar{x})}{\sigma^2\tau^2}\right] \tag{4.7}$$

Then:

$$-\frac{1}{2}\left[\frac{\mu^2-2\mu\frac{(\sigma^2 M+n\tau^2\bar{x})}{(n\tau^2+\sigma^2)}}{\frac{\sigma^2\tau^2}{(n\tau^2+\sigma^2)}}\right] \tag{4.8}$$

$$-\frac{1}{2}\left[\frac{\left(\mu-\frac{\sigma^2 M+n\tau^2\bar{x}}{(n\tau^2+\sigma^2)}\right)^2}{\frac{\sigma^2\tau^2}{(n\tau^2+\sigma^2)}}\right] \tag{4.9}$$

In other words, $\mu|X$ is Normally distributed with mean:

$$\frac{\sigma^2 M+n\tau^2\bar{x}}{n\tau^2+\sigma^2} \tag{4.10}$$

and variance:

$$\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2} \tag{4.11}$$

We can also re-written the mean in the following form:

$$\frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}M+\frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\bar{x} \tag{4.12}$$

and the variance can be re-written as:

$$\frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}} \tag{4.13}$$

This is an important result. We can notice two things. First, the variance of $\mu|X$ is smaller than the variance of the prior mean (τ^2) and smaller than the variance of the data mean (σ^2/n). That is, combining the information from the prior and the new data, gives us a more precise estimate than if we used either information source by itself.

Second, the posterior mean is a weighted average of the prior mean M and the new data mean $\bar{x}$. The weight on the prior mean is inversely proportional to the variance of the prior mean ($1/\tau^2$), and the weight on the data mean is inversely proportional to the variance of the data mean (n/σ^2).

Therefore, the amount of weight on prior and likelihood depends on the relative uncertainty between the two distributions. This concept is represented graphically in the Figure 4.5.

Three Normal distribution are shown: blue represents the prior distribution, gold the likelihood and pink the posterior. In the left graph in the figure, you can see that the prior (blue) is much less spread out than the likelihood (gold). Therefore the posterior resembles the prior much more that the likelihood. The opposite is true in the graph on the right.

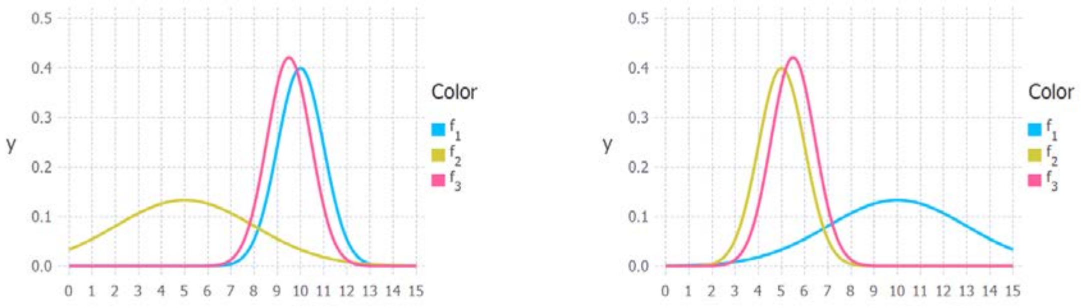


Figure 4.5: Bayesian inference for Normal distribution. In blue is represented the prior, gold is the likelihood and pink is the posterior. The variance of prior and likelihood assume the role of weight to generate the posterior [52].

If the prior mean is very precise relative to the data mean, then we should weight it highly. Alternatively, if the data mean is more precise, then it should be assigned a bigger weight.

We have shown the Bayesian update of a prior normal distribution with new sample information [51].

Involving Gaussian distributions makes the maths a lot easier and when applied to Bayesian inference it requires computing the product of the two Normal distributions.

Even without doing the maths, we knew that the posterior would be a Gaussian distribution. This is because Gaussian distribution has a particular property that makes it easy to work with. Gaussian family is conjugate to itself with respect to a Gaussian likelihood function. This means that whether we multiply a Gaussian prior distribution with a Gaussian likelihood function, we will get a Gaussian posterior function. The fact that the posterior and prior are both from the same distribution family (they are both Gaussians) means that they are called conjugate distributions and the prior distribution is called conjugate prior for the likelihood function [52].

There are many other inference situations in which priors and likelihoods are chosen such that the distributions are conjugate because it makes maths easier. For example, in data science in Latent Dirichlet Allocation (LDA), that is an unsupervised learning algorithm for finding topics in several text documents.

The advantage for using Bayesian framework is that it allows to update your beliefs iteratively in real-time as new data comes in. In this case it works as follows: we have a prior belief about the mean value of an IOI class and then we receive some new data from a performance. We can update the belief by calculating the posterior distribution like we did above. Afterwards, we get even more data comes in. So the posterior becomes the new prior. We can update the new prior with the likelihood derived from the new data and again we get a new posterior. It's a cycle that can continue indefinitely, updating the prior every time new data comes in.

4.3.1 Defining prior distributions

As we have said, there are 3 parameters that we want to estimate during a performance. These are the 3 types of IOI that we can find in a performance according with the score used.

Since in this experiment, musicians weren't using a metronome, but they just received suggestions about tempi, we have firstly to detect which are the 3 in-

tervals length in which they are singing.
For these reasons, instead of just using the fixed intervals that emerged from the score as prior beliefs, we based the prior exploiting the first 15 seconds of each performance. In this way we take into account all the IOI that are part of the first part of a song to define the prior distributions for each of the 3 inter-onset classes.

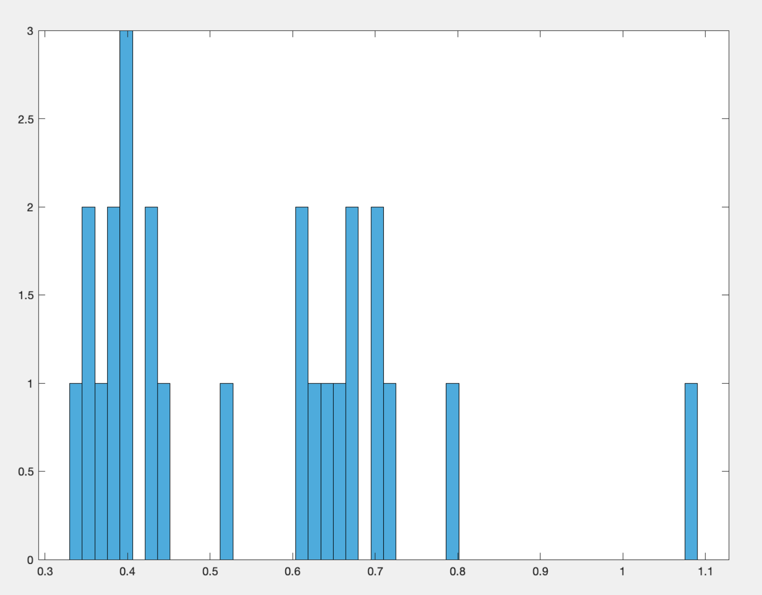


Figure 4.6: Histogram of the IOI in the first 15 seconds in trial 4 of couple 4. X-axis is the length of IOI and y-axis is the number of IOI having a certain value. The 3 inter-onset classes are not well separated. Index of bad performers.

As we can see in Figure 4.6, determining which are the 3 main intervals played in the first 15 seconds, could be difficult to understand it from the IOI's histogram, especially when performers don't clearly respect the intervals between the notes. However, since we knew in advance that we wanted to divide the data in 3 categories, we employed the k-means clustering algorithm.
The first values for the centroids are taken near the median of the IOI, knowing that based on the score, the IOI in the first 15 seconds should have the same number of IOI belonging to class 1 (the shortest interval) and class 3 (the longest interval), and half of them should be of class 2.
This algorithm showed to perform well also in case of bad performers, in which

there was not a clear separation between different IOI classes.
Since k-means clustering could not converge we assure that after a maximum fixed number of iterations it stops, giving an approximate information about the 3 classes.
In Figure 4.7 is shown the plot of the 3 clusters founded by the k-means algorithm, given the initial 15 seconds of a trial.

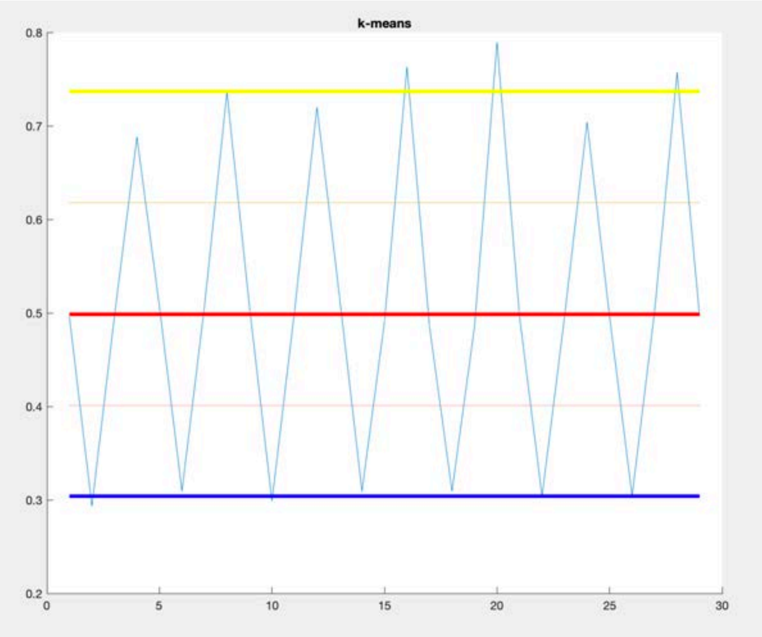


Figure 4.7: Clusters obtained from the data in the first 15 seconds in trial 1 of couple 11. Blue line, red line and yellow line highlight the clusters centroids. X-axis is just the sequence of IOI encountered and the y-axis is the length of the IOI.

Now that we have found the center of each cluster (or IOI class), we can use the centroid of each class to estimate the duration of the next IOI of its specific class. Then, around the 3 expected values we can know build the prior distribution. In other words, we have built the priors beliefs.

4.3.2 Sequential Bayesian updating

Now that we have a prior distribution for each class we can adjust the expectation of each one of them along the performance. Since performers are human being,

there will be little changes of the intervals if they are good to maintain regularity of the timing, indeed, we will discover big gap between the expectation values of each class and the next IOI interval, in the case musicians are not able to maintain timing regularity.
The first thing that we do when we analyse a new IOI (after the first 15 seconds), is to classify it into one of the three classes. To do this we just assign it to the nearest expectation value of the 3 classes.

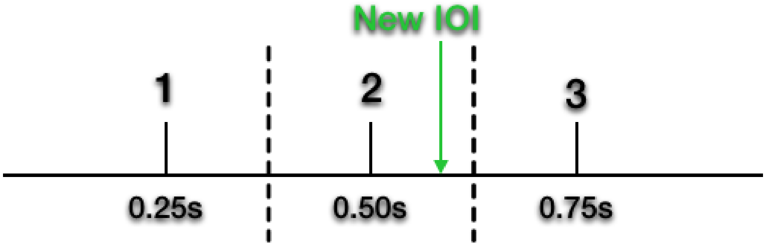


Figure 4.8: Classification of a new IOI into one of the three possible classes. In this case, the new IOI is assigned to class 2 because it has the nearest local IOI mean.

At this point, the likelihood function for the chosen class can be computed. It reflects the new evidence data contained in a short (shifting) 5 seconds window. Based on the performed IOI data in the buffer, we calculate an evidence-based IOI (using expectation-maximalization).
Then the posterior distribution, as we have shown before, can be computed multiplying the prior and the likelihood obtaining another Gaussian distribution, as shown in Figure 4.9.
This cycle is repeated for every new performed IOI. It's important to remember that the 3 classes are always independent, this means that each one of them maintains its own prior distribution that is updated whenever a new performed IOI is classified within that class and so it is included in its buffer (5 seconds length) that contains the data to compute the relative likelihood function.
We decided to maintain a buffer of 5 seconds, instead of, for example, including a fixed number of performed IOI, to have a smooth adjustment of the IOI expectation over the performance, that reflects the behaviour of the most recent performed IOI of each class.
At every cycle, one prior distribution of a selected class is updated with the like-

lihood to get the posterior, that will become early the prior of the next cycle. This mechanism continue until the last IOI performed.

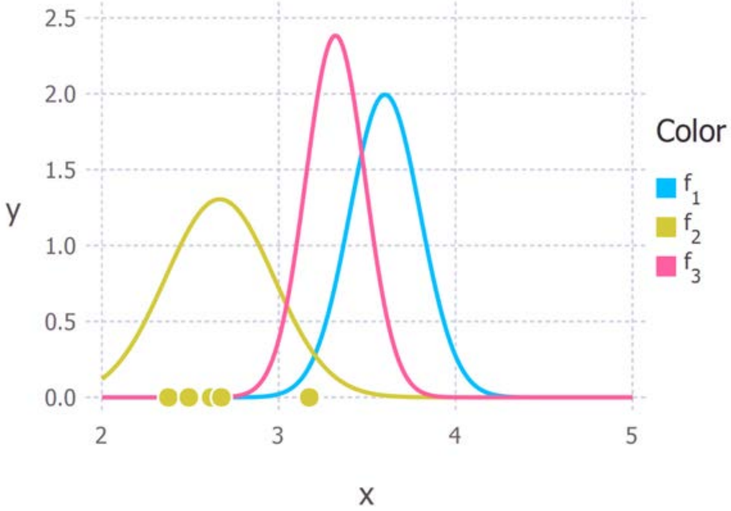


Figure 4.9: In blue is represented the prior distribution, in gold is represented the likelihood function computed with the new IOI data encountered in the last 5 seconds (represented by the gold balls on the x-axis) belonging to a specific class. Multiplying these 2 distributions we obtain the posterior distribution (pink line) [52].

Therefore, at each interaction there are 3 maximum a posteriori probability (MAP) estimates, that are represented by the mean (or mode) of the 3 posterior distributions. We assume that these local means IOI works as a predictors for generating (through joint action) a new IOI of each class.

This innovative mechanism based on Bayesian updating can be considered as the mathematical expression of the predictive coding theory which assumes that performers, in order to regulate the dynamics, constantly adjust their predictions about the joint timing.

Therefore, running the bayesian updating method through an entire performance, allows us to obtain the progression of the MAP estimate for each one of the 3 IOI classes, as can be seen in Figure 4.10. In this way it is noticeable the behaviour of means values of the posterior distributions along the performance, as we have shown before, the shifting in time are due to the new IOI that adjust the predictions over time.

As you can notice from Figure 4.10, with this representation is possible also to recognize some repetitive behaviour. For example, for the couple five in trial one which is the performance take into account in the Figure 4.10, we can see that the class one (shortest interval) shows a progression that repeats over the same segments. In this case, we can easily see the trend to increase the duration of the first IOI class during A segment, indeed, performers usually decrease the duration of it during the B segment.

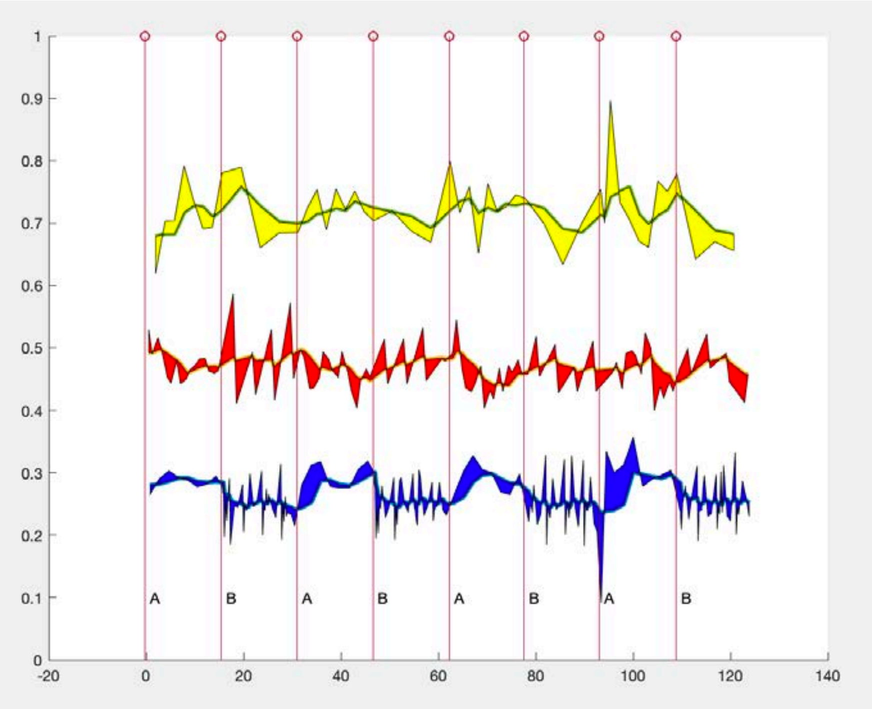


Figure 4.10: X-axis represents the seconds, vertical lines split A and B segments of the score. Y-axis represents the length of the intervals. The 3 levels at different intervals length represent in blue the behaviour of the first class, in red the second class and in yellow the third class of intervals. The black bold lines in the middle of each of the 3 classes are the MAP estimates of each category.

Moreover, the areas colored in yellow, red and blue are the areas that separate the prediction values from the evidence IOI. When a new IOI is analysed, it is not only used to update the inter-onset class predictor of its own class but we can also compute the distance (in milliseconds) with the previous estimated value. Thus, the time difference of the new IOI compared with the local IOI mean is then defined as a prediction-error.

A graphical representation is reported in Figure 4.11.

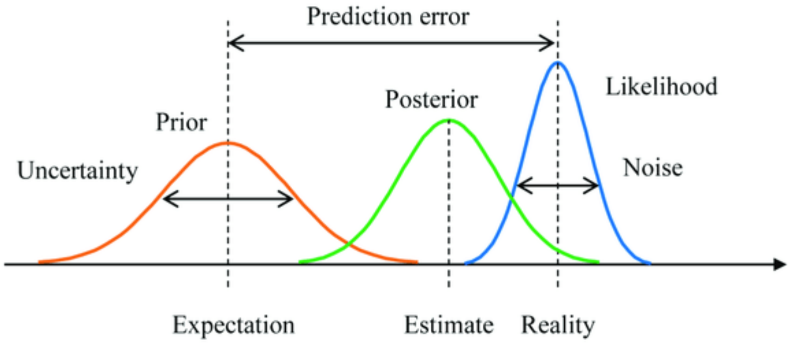


Figure 4.11: The time difference between the MAP estimate (mean of the prior distribution) and a new IOI is called prediction-error [53]. Using this new evidence we can compute the likelihood to update the prior distribution and get the posterior which mean value is the new estimate duration for future IOI.

Given our innovative dynamic framework, based on the predictive coding theory, we want to compute some objective measures that can reflect the quality of a performance. For this reason, exploiting the new prediction method, we considered 3 different types of prediction-errors, related to:

- **Fluctuation error**: micro-timing prediction-errors.
- **Narration error**: meso-timing prediction-errors.
- **Collapse error**: macro-timing prediction-errors.

4.4 Fluctuation error

Fluctuation errors are the micro-timing predictions-errors, this means that the time difference between the predicted IOI and the successive actual value is very closed. In particular, an error is classified as fluctuation error when the IOI deviate from its expected class value of at maximum two times the standard deviation of the prior distribution (2σ), generally in the order of tens of milliseconds.
These fluctuations can be caused by different sources such as micro-corrections due to small mistakes, or even small onset measurement errors due to the analysis. However, fluctuations can be considered necessary in order to maintain a

performance state, even of high quality.

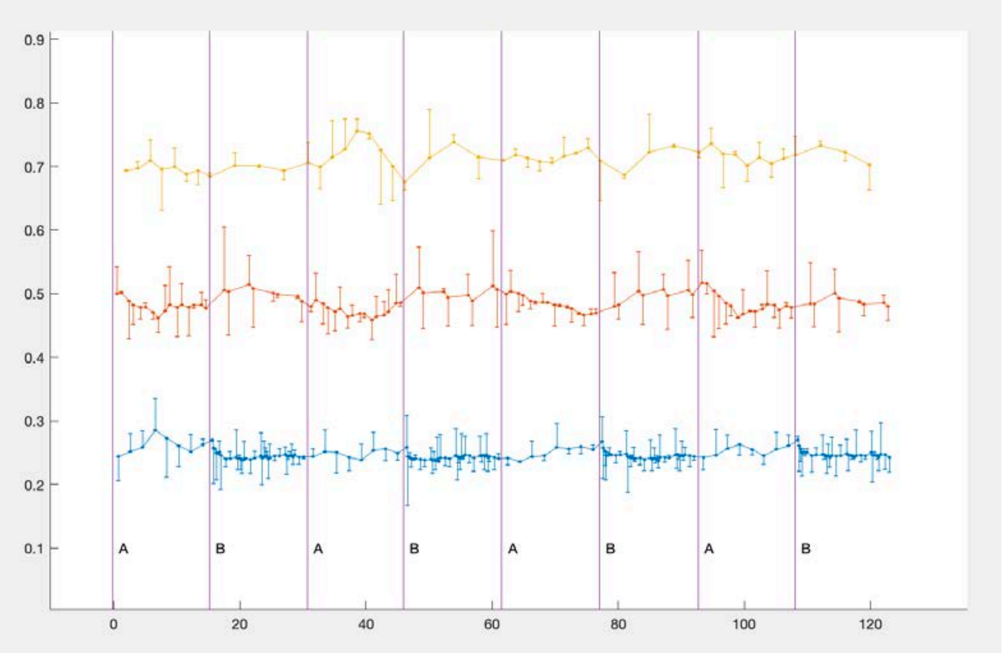


Figure 4.12: The 3 horizontal lines are the progression of the MAP estimates for each class and at correspondence of the new IOI, the vertical lines represent the fluctuation errors. X-axis represents the duration of the performance in seconds and y-axis represents the length of the intervals in seconds. The long black vertical lines, split A and B segments of the performance.

In Figure 4.12 you can see all the fluctuation errors found over a performance and the relative distance from the expected values. You can observe also how the prediction value of each IOI class adapt at every new IOI evidence. As expected, in case a new IOI is longer than its predicted value, than the MAP estimate will increase its value, the opposite if a new IOI is shorter than the value predicted. Storing all the fluctuation errors that we encounter in a performance, allows us to compute the final Root Mean Square Error (RMSE) for each IOI class, where low fluctuation values are considerate sign of good interaction, because they mean regularity over the performance, on the other hand, high fluctuation values indicates that, at least analysing micro-timing, musicians didn't maintain constant notes intervals.
In Figure 4.13 is plotted the sequence of fluctuation errors that are computed

over 120 seconds of a performance, with different colours to differentiate the classes and at the bottom are also reported the final RMSE values for each IOI class.

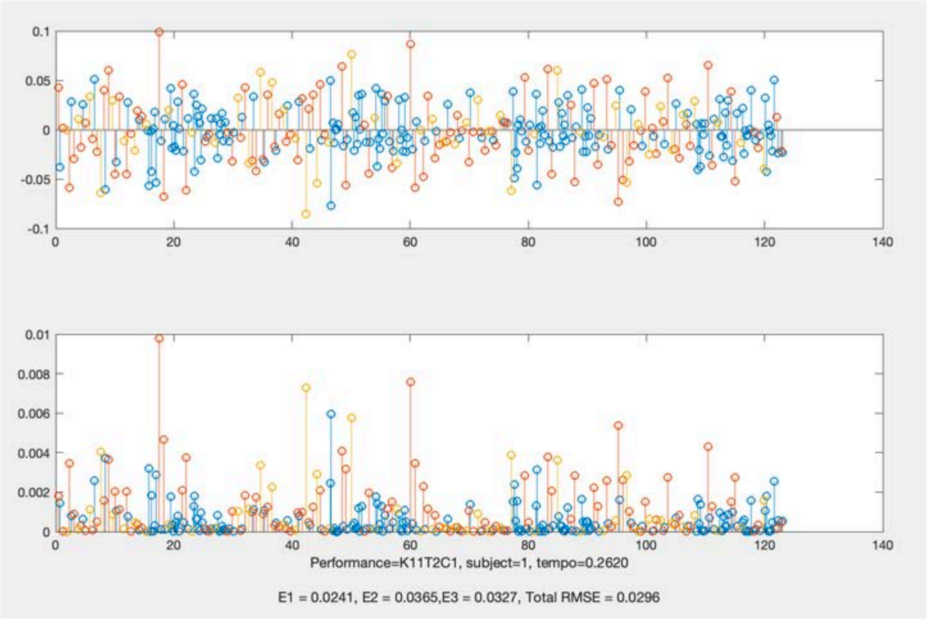


Figure 4.13: X-axis represents the seconds of the performance, instead y-axis is the fluctuation error in seconds. This is the entire sequence of micro-errors found in the trial 2 of couple 11. Fluctuation errors of class 1 are drawn in blue, class 2 in red and class 3 in yellow. The bottom plot reports the absolute value of the errors. Finally you can see the actual RMSE values computed for each IOI class and also the total error calculated combining all the errors of the 3 classes.

Since we have the singing part of each individual and then the parts of the subjects in the same couple are merged together to analyse the joint action, we are able to compare the fluctuation errors of each single musician of the ensemble.

Furthermore, thanks to manual annotations about the starting and ending points of A and B segments of each trial, it is also possible to compute the fluctuation errors for just a specific part of a performance.

Probably the most useful parameter that we can extract from the overall fluctuation errors in each single performance is the total RMSE fluctuation error. It summarize in a single value, the regularity of the musicians during the interaction. It is computed as the RMSE of all the fluctuation errors belonging to all the 3 classes. With it, we can now draw up the raking list of the couples, starting from the couple with the lowest combine fluctuation error until the couple with the highest value. The ranking list is reported in Figure 4.14 with respect to the mean RMSE fluctuation error obtained from the 8 trials of every couple.

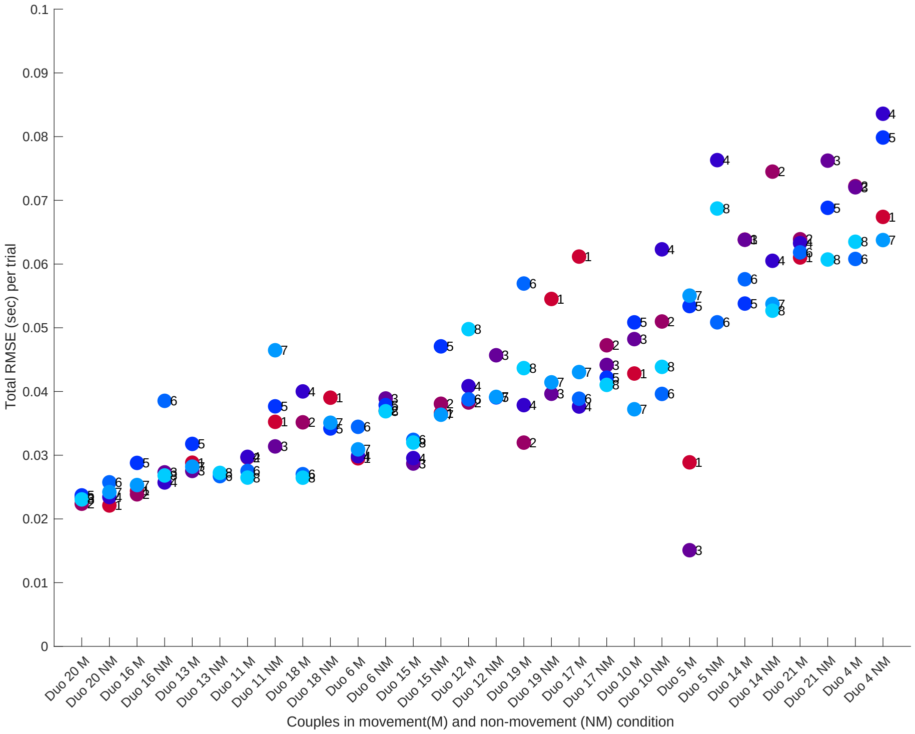


Figure 4.14: Couples' ranking list based on the mean RMSE fluctuation error computed over the 8 trials. Here it is reported the total fluctuation error of each single trial and each couple is also subdivided in two columns: trials performed with movement condition and without.

We can obtain a couples' ranking list also based on the previous subjective measures that we have taken. Respectively the annotation and the agency self-assessments. In Figure 4.15 is shown the ranking list based on annotation values of the couples, starting with the pair with the highest mean evaluation, over the 8 trials, until the couples with lowest mean annotation values.

Then in Figure 4.16 there is the ranking of the couples based on the agency values, again computing the mean over the 8 trials of every pair of musicians.

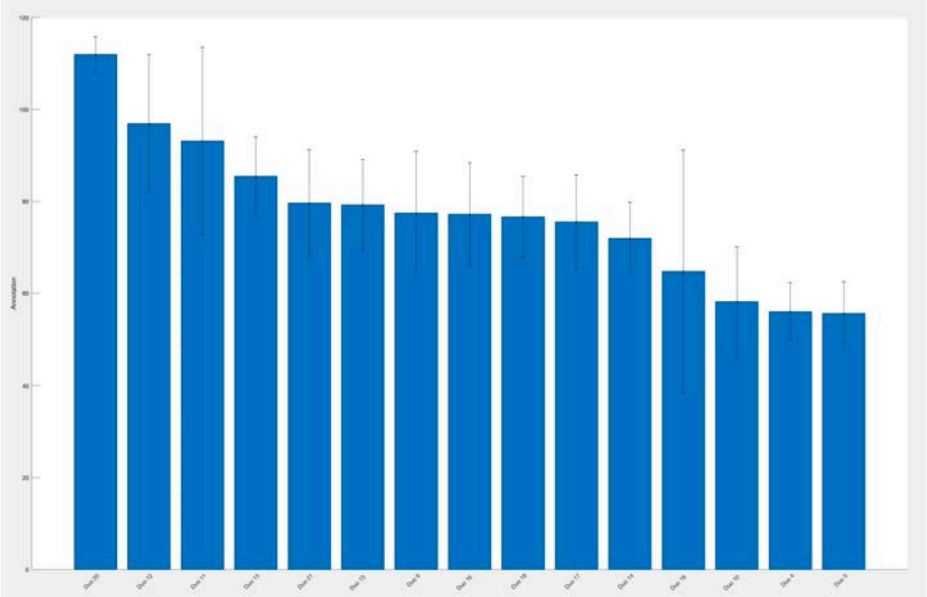


Figure 4.15: Ranking of the couples based on the mean Annotation values computed over the 8 trials. Y-axis is the scale used for the annotation measure that goes from 0 to 120. Starting from the left we find the best couples, for the annotation measure, going towards the right side of the plot where there are the worse couples.

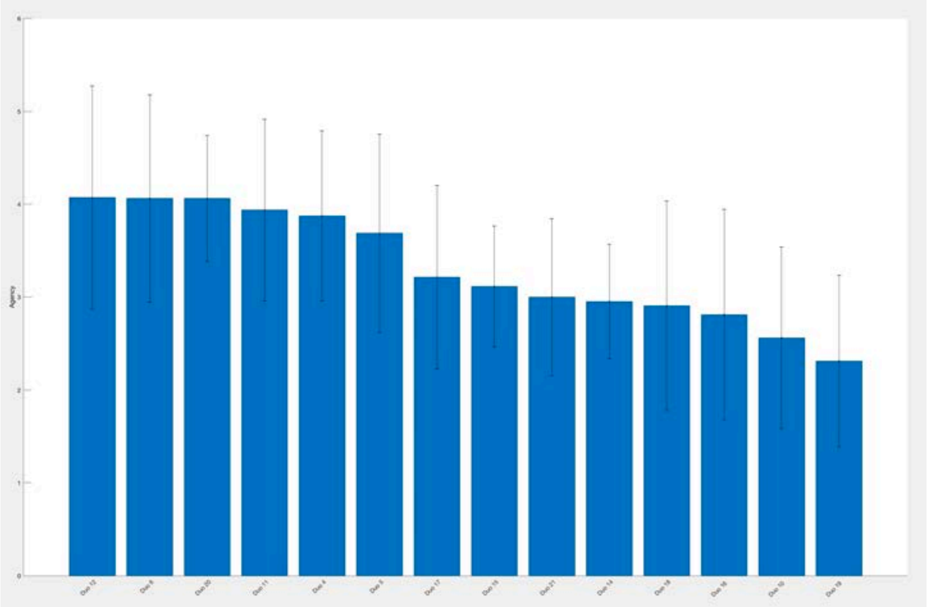


Figure 4.16: Ranking list based on the mean agency values computed over the 8 trials. Y-axis represents the different level of agency, from 0 to 6. Starting from the left towards the right side of the plot we find the couples with a decreasing agency value, this means from the best until the worse couples.

Now we have 2 subjective measures, annotation and agency, that we can compare with the first objective measure, the fluctuation error.

Since annotation and agency are considered two different measures of the quality of a musical performance, we are interested to understand, whether and how much correlated these measures are with the fluctuation error, that is, the objective measure computed exploiting the innovative approach based on Bayesian probability.

Our aim is to find whether there is significant relationship between low value of fluctuation error and high values of annotation or agency, that characterize good performances and also whether high fluctuation error corresponds to low annotation and agency values.

For these reasons, we computed Kendal's rank correlation coefficient. It is a statistical measure for calculating the degree of similarity between two rankings, and can be used to assess the significance of the relation between them.

Computing the Kendal's correlation in MATLAB, we got 2 outputs: RHO and P-value. RHO is the correlation coefficient that goes from -1 to 1 and P-value test the hypothesis of no correlation against the alternative non-zero correlation. If P-value is smaller than 0.05, then the correlation value RHO is considered significantly different from zero, this mean that it is found a certain degree of correlation between the measures analysed.

As you can see in Figure 4.17, the correlation test between fluctuation error and annotation found significative correlation, P-value under 0.05 that means accepting the hypothesis of correlation and with a RHO value of 0.37, both for the movement and non-movement condition. This confirm the relationship between our first objective measure and the subjective measure.

To compute the correlation are taken into account all the trials of every recorded couple.

Further analysis shows us also more details about the correlation within these measures.

Splitting the couples in two groups, on one side the best couples and on the other side the worse couples, based on their value of the fluctuation error measure, we found the following results. Computing the correlation analysis, within annotation and fluctuation error belonging to good performers we found significant correlation (P-value equal to zero) and RHO equal to -0.35, but within bad per-

formers' fluctuation error and annotation we got a P-value of 0.7, much higher than 0.05, led us to accept the hypothesis of no correlation.
The last result suggests that the fluctuation error is significantly correlated to the annotation measure in case the performers are good to maintain timing regularity but in case of bad musicians it does not reflect well the qualitative evaluation of a performance.

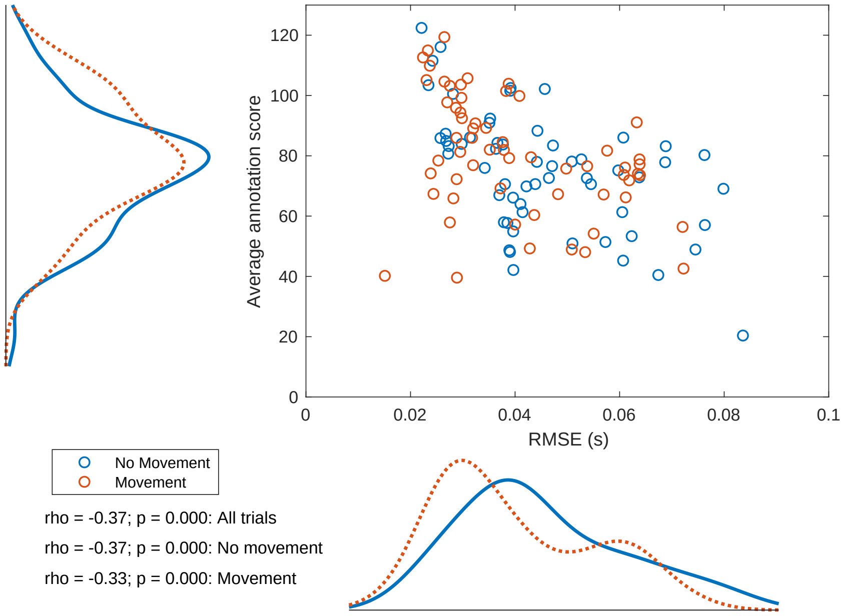


Figure 4.17: Scatter plot of the trials where in x-axis there is the value for the fluctuation error and on y-axis there is the average annotation score. Blue circles mark performances with non-movement condition, while red circles mark the movement condition performances.

The difference that emerged between good performers and bad performers can be caused by the level of details that the fluctuation error is capable to capture. Since it is a really fine measure, in the order of tens of milliseconds, this seems to characterize well the good couples but bed performers could be characterized better by other parameters.
The second correlation analysis that we have done is between agency and fluctuation error. The scatterplot is reported in Figure 4.18.
Overall it shows a significant correlation within the measures, but with lower

RHO value compare to the previous correlation analysis. Indeed, no correlation is found whether movement and non-movement conditions are analysed separately, with a P-value greater than 0.05. Agency correlations analysis reported always lower correlation values respect the correlation analysis for the annotation measure.

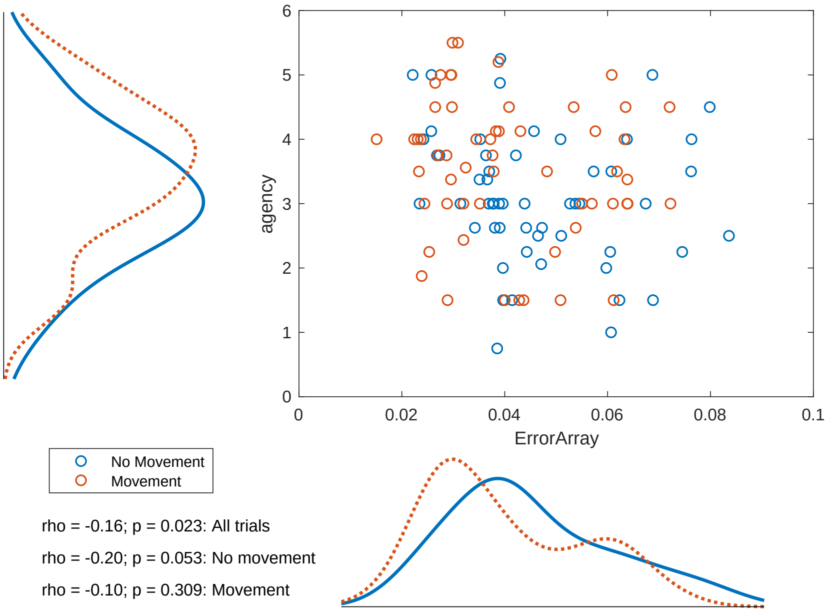


Figure 4.18: Scatter plot of the trials, where x-axis represents the fluctuation error and y-axis is the agency value. Red circles mean movement performances and blue circles non-movement performances.

Moreover, computing the correlation with the couples divided in two groups like before: good performers and bad performers, the result reflect what we have found also for annotation values. This means that also agency measure seems to be well correlated with the fluctuation error values belonging to good performers (P-value equal to 0.009 and RHO of 0.24) but it shows no significant correlation with bad performers (P-value equal to 0.423 and RHO of 0.09).

Summarizing, the fluctuation error is the first objective measure that we can extract from the predictive dynamic framework we have built. It can captures the micro-timing prediction-errors. It takes into account only the IOI that are really closed to the means inter-onset classes, less than 2σ of distance from the means. We know that fluctuation errors can't be the single parameter able to perfectly reflect the quality of a performance, but it can give us an evaluation of the performers precision, in the orders of tens of milliseconds, to keep the IOI constant over time.

As a first test, we have taken the annotation and agency values made by the performers, as 2 different measures of quality of each performance. These were the data we had already collected so they were just ready to use for our analysis. Further tests, could use quality evaluation data taken by third subjects that might assess multiple performances based on some particular quality aspects, like annotation and agency used in our experiment or even based on new ideas.

Overall, we have found that the objective measure that we have built, the fluctuation error, shows significant high correlation values with annotation measure and slightly lower but still significant correlation values also with the other subjective measure, agency.

This is the proof that fluctuation error can be used as a reliable measure to evaluate the quality of a performance and in this case we have proven that it is connected with the values of annotation and agency.

Moreover, in both cases, this error representation seems to correlate better with good performers than bad performers. The results shows that fluctuation errors it's a reliable objective measure to evaluate the timing regularity and it is particularly significant in case of overall good performers.

It's important also to notice that the values given by the performers about agency, should be taken evaluating only the best moments of the each performance, so it is not an evaluation of the overall performance.

Finally, we suggest that bad performers could be characterize better taking into account other measures of errors. Since fluctuation errors are collected only when small errors happen, we thought that we could also take into account, in another separate parameter, the big errors of performers with respect to the predicted inter-onset means. For this reason now we will explore the collapse error, the second objective measure.

4.5 Collapse error

Collapse errors are the macro-timing prediction-errors, this means that time difference between the predicted IOI and the actual value is not so close.
They are catastrophic in the sense that they can break down the interaction, although it was expected to continue. This also means that the dynamic systems' state is not longer the same. An error is classified as collapse error whenever the time difference between a new IOI and its expected value is longer than two times the standard deviation of the prior distribution of its class, generally in the order of hundreds of milliseconds.
An explicative image is reported in Figure 4.19.

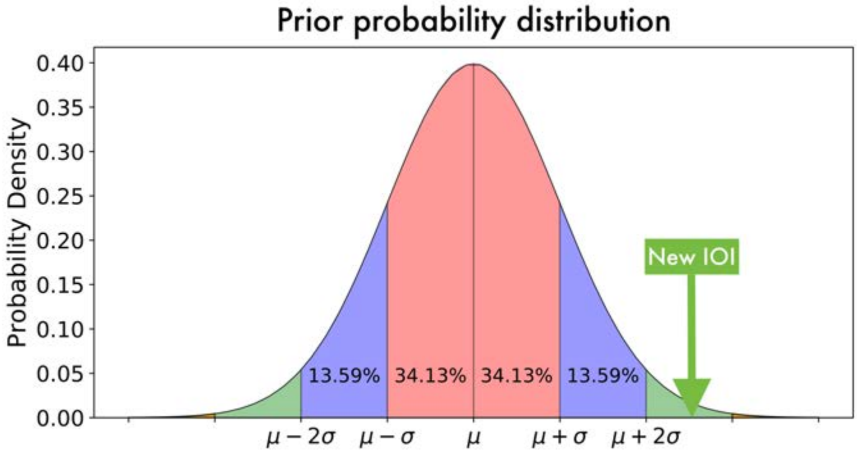


Figure 4.19: Collapse error is defined as the time difference between the predicted mean of the prior distribution and the IOI, in the case that distance is greater than 2σ, where σ is the standard deviation of the probability distribution.

These are bigger errors respect the fluctuation errors, and can be more easily detected also by a human being listening a performance.
They can be caused by mistakes or could be also interprets like a misclassification, this means that an IOI is so far from it's expected mean class that is recognized belonging to an adjacent class.
Collapse errors do not influence the update of the prior probabilities, that are influenced only by micro-errors.

Since there could be also pauses during the performances or very long IOI caused by mistakes, just like the manual annotations done to mark A and B segments along the sequence of IOI, it has been marked also the moments where the interaction broken, in order to exclude these moments from the analysis.
Unlike fluctuation errors, there could be trial that don't have any collapse error, if the performers kept constant IOI.
As we have done before, the total collapse errors founded in a trial are merged calculating the RMSE and obtaining in this way one single value for every trial. Then, we are interested to find if this second objective measure that we have computed, is correlated with the subjective measures annotation and agency. So, like we did before, we computed Keller's correlation test for both the measures.

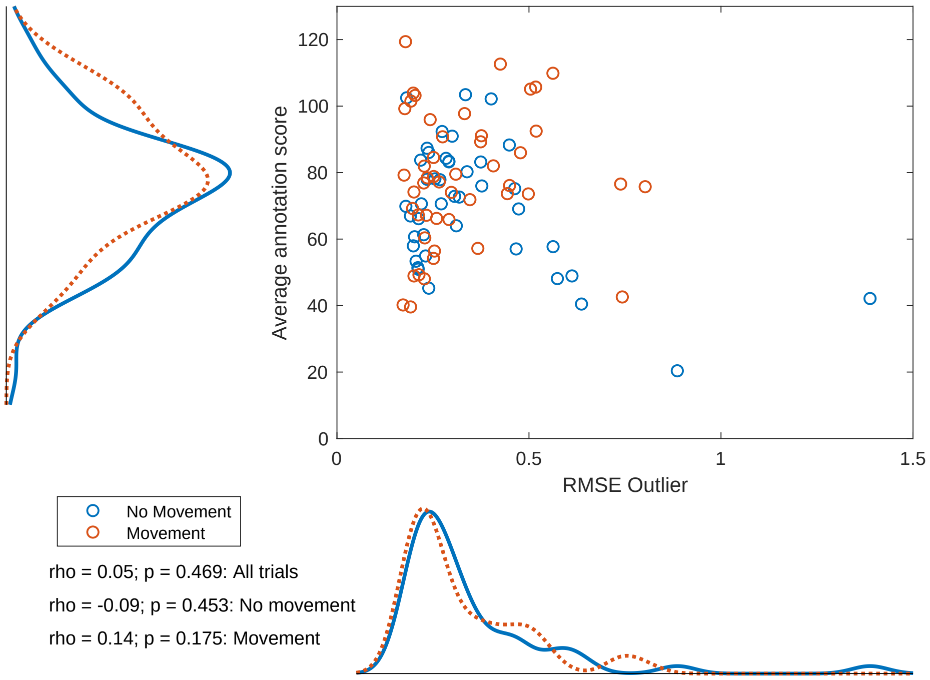


Figure 4.20: Scatter plot of the trials, where x-axis represents the collapse error and y-axis is the annotation value. Red circles mean movement performances and blue circles non-movement performances.

The correlation test with annotation is shown in Figure 4.20 and, unlike the fluctuation error, results no significant correlation.

Even considering movement condition and non-movement condition separately, the test suggests still no significant correlation. All the values of the test are reported in Figure 4.20.

As we did before, we computed also the correlation splitting the couples in 2 group separately, good performers and bad performers, but still P-values are respectively 0.130 and 0.867, a lot over the threshold for considering the hypothesis of correlation.

This results show that collapse error are not reflecting the annotation measure. The second step is to test the correlation with the other subjective measure, agency. Results are reported in Figure 4.21.

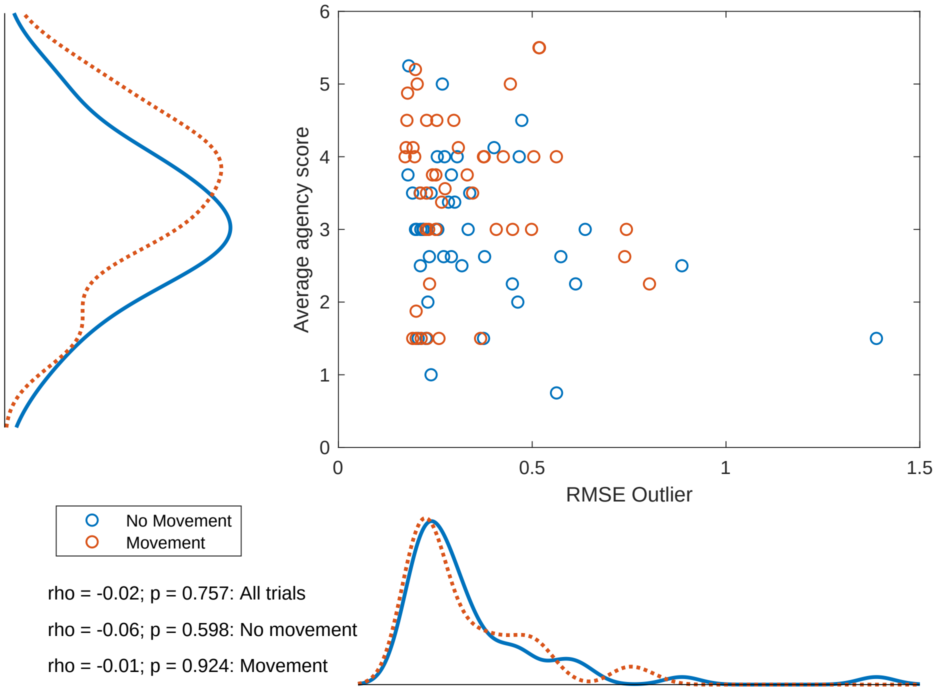


Figure 4.21: Scatter plot of the trials, where x-axis represents the collapse error and y-axis is the agency value. Red circles mean movement performances and blue circles non-movement performances.

Again, collapse errors doesn't show significant correlation even compared with agency measures, P-values are always way beyond the threshold.

Whether we try to repeat the test only with good performers or bad performers, the results don't change, reporting P-values of respectively 0.956 and 0.547, still no significant correlation.

This second objective measure that we have computed, didn't show any significant correlation in all the tests performed.

The method with which it is computed is the same used for fluctuation errors, the only difference is that collapse errors are not taken into account to update the prior because they should appear rarely and they could cause quick changes of the predicted intervals.

The results suggest that measuring only the bigger errors of performers is not enough to establish a measure of quality of an entire performance. This is the results we got using as comparison the ranks of annotation and agency.

We suggest that collapse errors, could be considered such as a further information to supplement other measures, to have a better overall understanding of the total errors of a performance, but it can't be used as the only parameter to rely and from which extracting an overall evaluation of a musical performance.

4.6 Narration error

Narration errors are the meso-timing prediction-errors. Up to now, we only have been looking at the length of the IOI and how much the performers maintained the three inter-onset classes as much as constant over the performance, completely ignoring the melody.

For these reason, we want also to take into account the narrative, that is defined by sequences of notes duration.

An error is detected when a predefined sequence is interrupted by the omission of a note or an IOI is so far from the expected duration that is recognized as IOI of another class, breaking the correct execution of a sequence.

Since for this case study, the score is characterized by two main short sequences of notes, one over the A part and the other over the B part, we want to analyse whether the performers are repeatedly singing two sequences of notes as requested by the score.

In particular, according to the score used in this experiment, the class intervals sequences that should emerge from the interaction are the following:

- Repeated sequence over A part: 2-1-2-3

- Repeated sequence over B part: 1-1-1-1-1-1-1-2-1-2-1-3

Where 1-2-3 are the possible inter-onset classes that we have defined. During the execution of the A part, its sequence is repeated 8 times in row, while during B part the second sequence is repeated 4 consecutive times.
The correct sequence, according to the score can be seen from the fluctuation errors in Figure 4.22, where we can see that the IOI are correctly classified in the expected classes according with the given score.
With this third objective parameter, based on narration, we want to measure how much the musicians are able to repeat the same sequences of IOI over a performance.
In order to easily adapt the script to any other score and make the analysis more flexible, we decided not to use the fixed IOI sequences imposed by the score adopted in this experiment, but recognizing the sequences that musicians are singing in a performance and looking at how many correct repetitions of the same sequences there are in the performance.

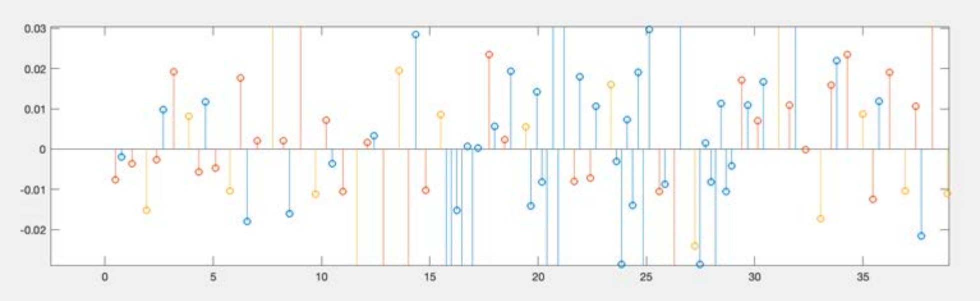


Figure 4.22: Here you can see the classification of the IOI, marked with different colours for each class. Blue line in case of class 1 (shortest interval), red line in case of second class (middle interval) and yellow line for the third class (longest interval). In this plot, the sequences of the IOI that emerged from the interaction correspond with the ones expected from the score.

For this purpose, is needed to find the 2 smallest sequences that are repeated during each trial.
First of all, the sequence of IOI that composes a performance, is converted in a sequence of numbers, where each number corresponds to the class of an IOI. Then, for searching the most repetitive pattern we set some handy boundaries,

such as, minimum length of a pattern is 4 IOI and maximum length is half of the total IOI of a performance.

The searching algorithm accepts a pattern when the following conditions are founded: the exact sequence of a pattern has to be repeated at least 2 times in row and it has to cover at least the 30% of the total amount of IOI in a performance.

With these conditions, we were able to find the two most repetitive pattern of each trial.

After the pattern detection phase, we just had to check how much of an entire performance was covered by the two most common pattern.

At the end, foreach trial, we obtained a single value indicating the percentage of the performance in which the musicians correctly repeated the pattern found in the previous phase, that, in order to be more flexible, could be completely different from the ones established by the score used in this experiment.

Moreover, we can now test the correlation with the other subjective measures: annotation and agency. The results of the test with annotation measures are shown in Figure 4.23.

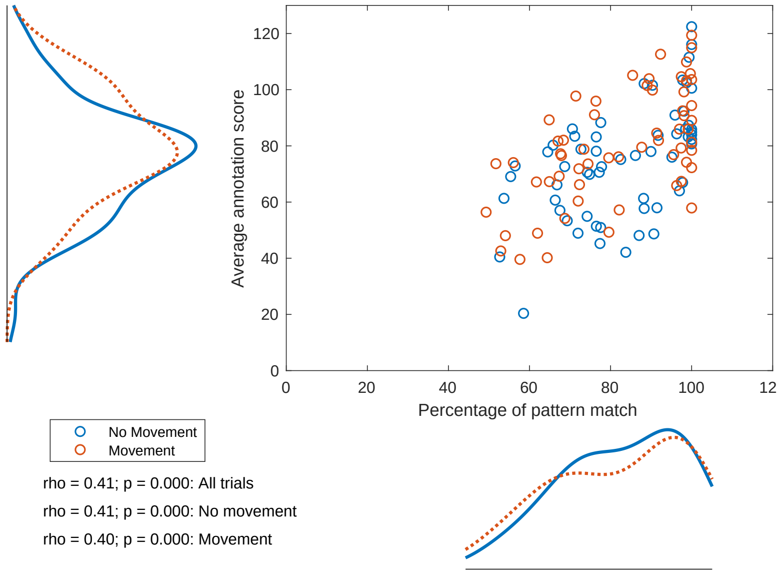


Figure 4.23: Scatter plot of the trials, where x-axis is the percentage of the performance correctly covered by the two most repetitive pattern, it can goes from 0% to 100%, and on y-axis are reported the annotation values from 0 up to 120.

Keller's correlation test accepted the correlation hypothesis (P-value equal to 0) with a correlation value RHO equal to 0.41, even in case the data were split in movement and non-movement conditions the test showed the same correlation results. We ran the test also splitting the couples in good performers and bad performers, obtaining in both cases significative correlation with narration error with values of RHO equal to 0.25 in case good performers and RHO equal to 0.28 for the bad performers.
Then, we ran the correlation test with agency, results are shown in Figure 4.24.

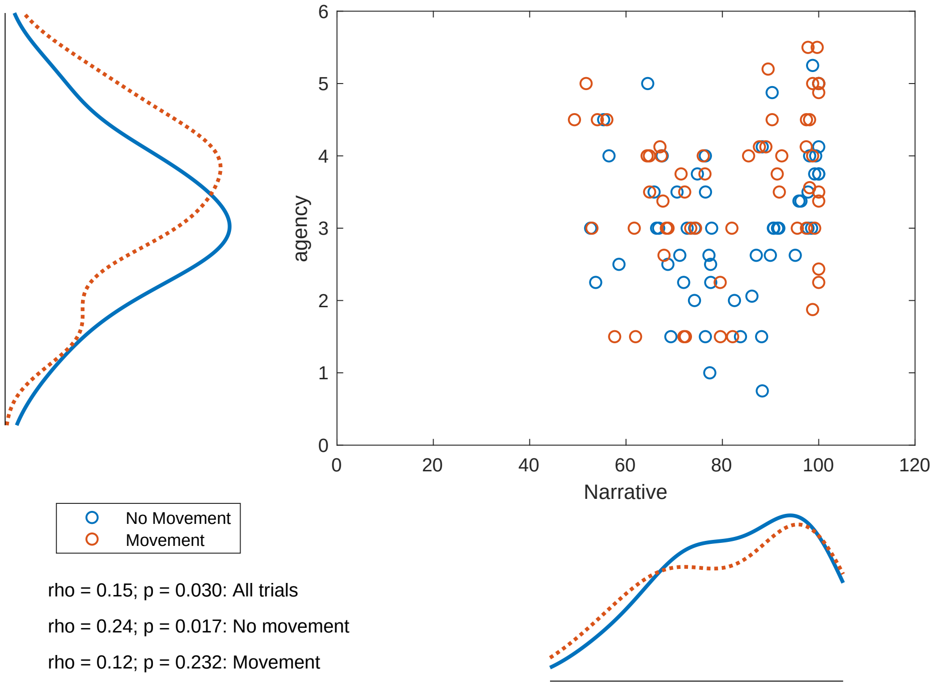


Figure 4.24: Scatter plot of the trials, where x-axis is the percentage of the performance correctly covered by the two most repetitive pattern, it can goes from 0% to 100%, and y-axis represents the agency values from 0 to 6.

Also the correlation test with agency shows significant correlation with a RHO value of 0.15. Narration error is still significantly correlated to agency considering no movement condition performances, but this time the test reports no significant correlation for the movement condition.
Whether we run the test for the good performers and bad performers separately, it shows significant correlation in both cases, with RHO of 0.28 for the good

performers and 0.24 in case of bad performers.

In this way we have built the third objective measure and, as we have done for the fluctuation error and collapse error, we can build the couples' ranking list based on the narration errors made by the couples in the 8 performances.

Narration error comes from the classification of each IOI in a specific class, based on the Bayesian regression, and then, after the pattern detection phase, it is calculated as the percentage of the total performance correctly covered by its most repeated pattern.

Respect to annotation measure, the narration error reached the maximum correlation degree compared to fluctuation error and collapse error, so it is the parameter that we can compute objectively that better reflects the quality of a performance like a manual annotation could do.

This parameter is not taking into account how much the IOI are far from the predicted values, but it is just checking whether the sequences of IOI classified over the performance (thanks to the innovative Bayesian method) are creating always the same pattern over the song.

The only test that showed no significant correlation is between narration error and agency in case of movement condition. This is in contrast with the results obtained with the annotation in case of movement, from which emerged high correlation. This suggests that correct execution of the narrative of a song could be present also in case of not high level of agency between the performers, that based on the result obtained for the fluctuation error, instead, can be considered necessary to perform high precision in terms of micro-timing errors.

In this experiment narration error has shown the highest values of correlation with both annotation and agency measures, therefore it is the best objective measure that we built that reflects the values of such measures. Moreover, from the results we can extract that this third objective parameter can be related to a human being point of view about the quality of a performance, because it has high correlation values in almost all the test computed.

4.7 Moving vs non-moving conditions

Since half of the trials are recorded in moving condition and the other half in non-moving condition, we can observe whether movement significantly influences changes of the 3 objective parameters that we have computed.
For this reason we computed the boxplot for each condition and then we compared the two boxplot with the *signrank* function (Wilcoxon signed rank test) in MATLAB that returns the p-value of a paired, two-sided test for the null hypothesis that difference between two paired samples comes from a distribution with zero median.

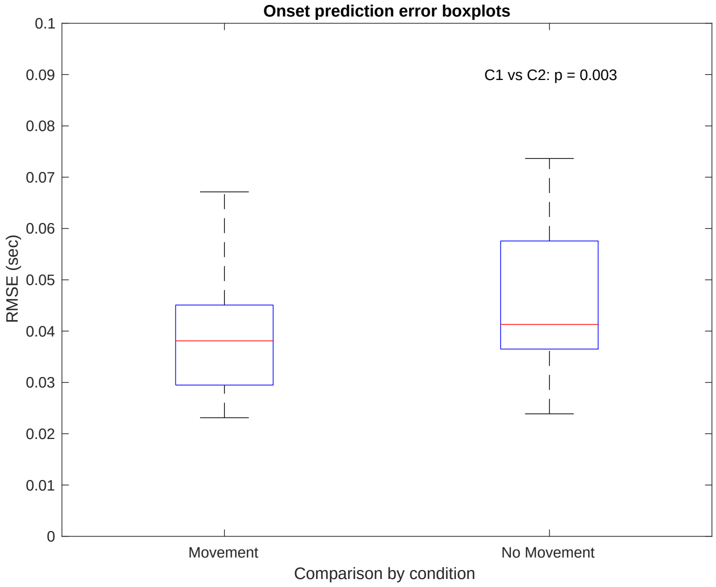


Figure 4.25: Boxplot comparison between movement and non-movement conditions for the fluctuation error computed for all the trials. X-axis is divided on the two conditions and y-axis represents the RMSE values for the fluctuation error.

Figure 4.25 shows the result of the Wilcoxon signed rank test that reports, at the default 5% significance level, a p-value of 0.003 that indicates that the test rejects the null hypothesis of zero median between pair samples.
This result suggests that the movement of the performers influences positively the fluctuation error, showing lower fluctuation errors in movement condition against higher errors when performers are not moving.

We are now interested to do the exact same test also for the narration and collapse errors.

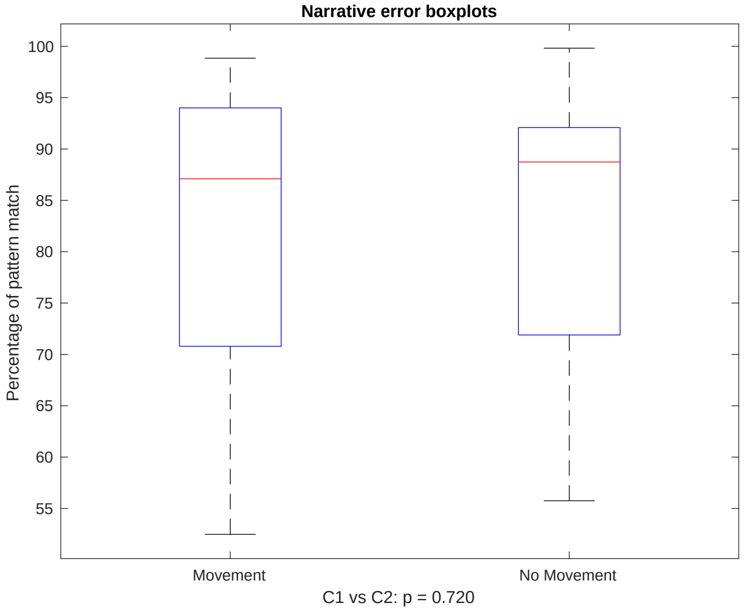


Figure 4.26: Boxplot comparison between movement and non-movement conditions for the narration error computed for all the trials. X-axis is divided on the two conditions and y-axis represents the narration error.

Figure 4.26 reports the boxplot for movement and non-movement conditions relative to the narration error and the result of the Wilcoxon signed rank test shows a p-value of 0.72, over the threshold of 0.05. The result reports that there is no significant difference between the two conditions.

The same result is obtained computing the test considering collapse error. In this last case the p-value is equal to 1, that means no significant improvement in movement condition also for the collapse error.

Therefore, we can summarize that the movement is not helping performers to reduce narration errors and also collapse errors, because we didn't find significative difference between narration errors in movement condition compared to non-movement condition. For the collapse error we got the exact same result, again no significant difference, or improvement, when performers were singing in the movement condition respect non-movement condition.

Indeed, we found significative difference between movement condition and non-movement condition in the fluctuation errors. The Wilcoxon signed rank test reported that the boxplots of the two different conditions are significantly different. Therefore, we can say that moving while singing can help performers, probably in subconscious way, to maintain a more precise rhythm, i.e. constant duration of the IOI performed.

Since all the performances are assigned to one condition (moving or non-moving) just relying on the instruction given to the performers, further analysis could check if performers were actually moving and not moving in the assigned conditions. For example, there could be some performances of non-moving conditions in which the body movements of musicians are significant. With this other check we could exclude performances that were not executed as requested.

4.8 Learning effect

We are interested to discover whether the repetitions of 8 times the same trial, have significant effects that we can notice in the objective parameters studied.

To test the learning effect over the 8 trials we used an approach similar to the one employed to compare the movement versus non-movement condition.

If we look at the development of the fluctuation errors over the 8 trials, we can see that there is not a clear behaviour shared by the different couples. In the plot reported in Figure 4.27 you can see the development over the 8 trials of the fluctuation errors, for all the couples. Sometimes there is an improvement over the trials of the fluctuation error, but is not an evident trend.

To test whether there is a significant improvement of the performances over the 8 trials we decided to compare the values (for the 3 parameters created: fluctuation error, collapse error and narration error) of the first trial versus the results that musicians obtained in the last trial.

As can be seen in Figure 4.28 we computed the boxplots of the trials for all the couples, taking into account the total fluctuation error of each trial and grouping by the number of the trial for the entire recording session.

From the plot it is easily noticeable that the boxplots are really close to each other.

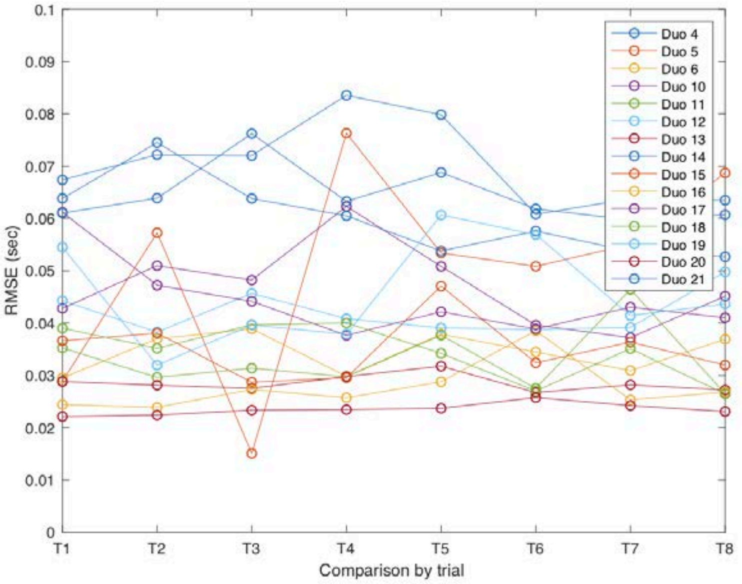


Figure 4.27: Developments of the fluctuation error over the 8 trials, for all the couples.

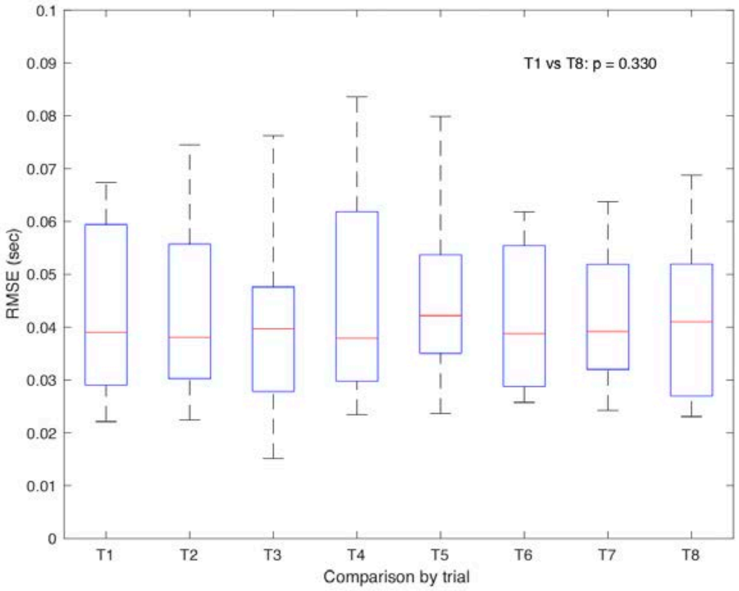


Figure 4.28: Boxplot comparison between the sequence of trials. X-axis represents the number of the trial and y-axis represents the fluctuation error.

To verify whether there is a significant improvement over the trials, we calculated the Wilcoxon signed rank test between the first bloxplot (of trial 1) and the last one (of trial 8).
Figure 4.28 report a p-val (from the signrank function computed) equal to 0.33 that means no significant difference between the performances analysed. This means that there is not a significant fluctuation error improvement between the first trial and the last one.
We have repeated the exact same test also for the collapse error and the narration error. The first reported a p-val of 0.742 and the latter a p-val of 0.808. Still we didn't find any significant improvement between the first and the last trial.
Therefore, the hypothesis of having a learning effect over the 8 trials of the experiment, is rejected. In this experiment there is not an evident learning effect. The reasons could be that the trials may not be enough to notice such effect and people struggle to improve with this genre of music that could sound unfamiliar for many of them. It's important to notice that the trials were recorded all in the same hour without individual rehearse between them, it was a demanding challenge for the musicians so probably in order to notice evident learning effect is needed more time for individual and ensemble rehearse.

4.9 Leader-Follower interaction

Another challenge that we tried to face, was to recognize the leader-follower behaviour between two performers based on the errors values. In particular we wanted to analyse the sequences of fluctuation error made by the two distinct subjects, to find correlation information that can suggest a leader or follower trend.
This analysis is based on the micro-timing error. As we have seen, fluctuation error has a good degree of correlation compared to subjective qualitative measures but more important, for every IOI, it shows the error made by the performer. It can be positive or negative with respect to the mean of the IOI class and comparing it with the following errors of the other partner we want to find if the behaviour of a musicians (her/his errors) influenced the errors of the partner.
As the literature reported in chapter two has shown us, this topic is usually explored computing the cross-correlation between the time-series of the two dif-

ferent subjects.
First of all, we can notice that the examined performance is divided in 2 different segments: A part and B part. The first part is characterized by a reciprocal interaction, this mean that the actions of the 2 subjects are always alternating. Instead, B part, that is slightly longer and complex compared to A part, shows a mutual interaction, in which the same subject sometimes has to sing two notes in row. This can cause misalignment during the cross-correlation of two fluctuation errors arrays that belong to each single performer. For this reason we decided to merge errors, of an individuals, that appear consecutively and don't have an error of the other subject in between.
To better understand the results of this process, in Figure 4.29 you can see the two sequences of errors of two participants before the compression and in Figure 4.30 the sequences of errors after the compression.

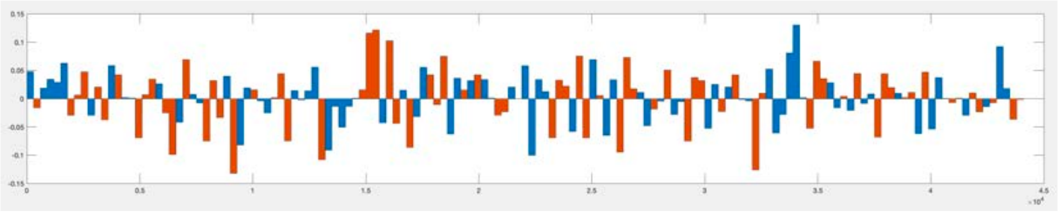


Figure 4.29: Here are represented the sequences of errors of two performers, one subject in red and the other in blue. Every error has the same width along the x-axis, instead the y-axis represents the fluctuation error in seconds. We can notice that there are consecutive errors both for the red subject and for the other one.

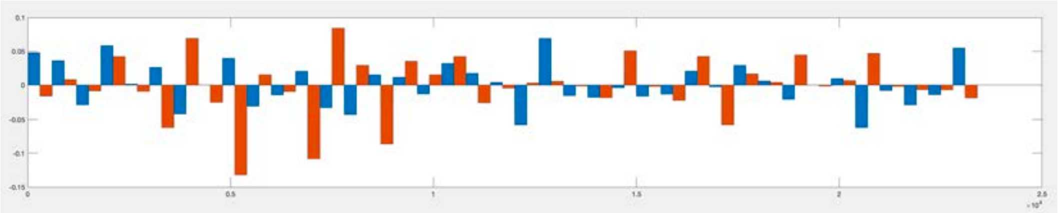


Figure 4.30: Here we have the same sequences of errors of Figure 4.29 but consecutive errors belonging to the same subject are merged together, in order that the width is the same for all the errors and the hight is the mean of the merged errors.

The compression of consecutive errors allows to execute a cross-correlation function with an exact match between the errors of two performers.

The changes that we made at the sequences of errors reflect our idea to test how much an error of a subject influenced the error of the other subject that comes later.

At this point we performed the cross-correlation test between the errors sequences of each trials.

Since A part and B part are really different and we could find different leader-follower behaviour in each one of them, we calculated the cross-correlation separately. This means that the errors in all the A part of a trial were linked together in a single sequence, the same happened for the B part.

Moreover, we analysed the output of the cross-correlation function for lag +1 and -1 only. In other words, the two sequences of errors are shifted of maximum one position. In this way the error of a subject is compare with the error of the other subject that comes before and after.

An example of the sequences used for the cross-correlation are reported in Figure 4.31 and a visual explanation of the output obtained is shown in Figure 4.32.

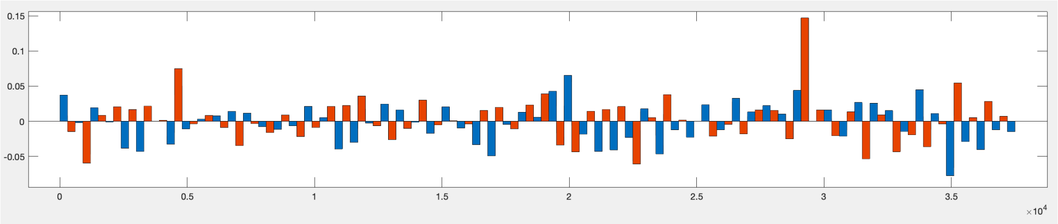


Figure 4.31: Here are represented the sequences of errors of two performers for A part, they don't require any compression, they are ready to compute the cross-correlation function.

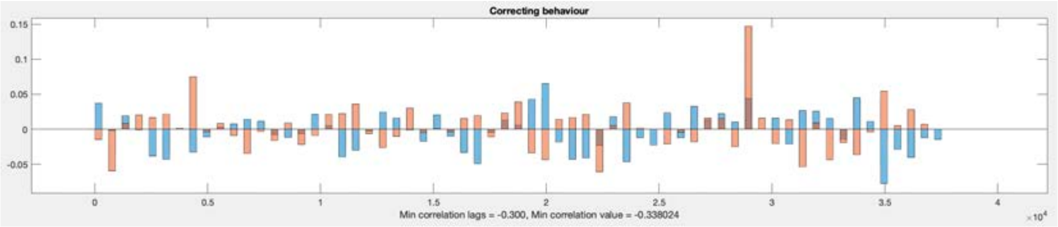


Figure 4.32: Calculating the cross-correlation between the sequences of errors shown in Figure 4.31, we got a minimum value of -0.338 at lag -1.

The plot of Figure 4.32 shows that running the *xcorr* function in MATLAB, with the two arrays containing the sequences of fluctuation errors reported in Figure 4.31, we obtained a negative value for the lag -1.

It is shown only the output at lag -1 because is the strongest cross-correlation value between the two possible shifting. Indeed, we considered only the maximum of the absolute values obtained at lag -1 and +1. At lag 0 we have always a cross-correlation value of zero because, as you can see in Figure 4.31, the errors of the two subjects are naturally shifted of one position, as they appear during the execution of a performance.

Moreover, we also compute the correlation test to make sure, with the value RHO and P-val, that the output that we get from the cross-correlation is significative. We stored only the cross-correlation values that resulted significant.

There are two parameters that we used to set the correcting or imitating behaviour to one of the performers. First, we looked at the cross-correlation value: in case of negative output, such as the example in Figure 4.32, there is negative correlation between the errors and we can interprets such as correcting behaviour, i.e. when one subject is producing longer notes, the other one shows a correcting behaviour, this means, the partners is emitting shorter notes.

In case of positive cross-correlation, we interpret it as imitating behaviour. That is, if one subject is making longer notes, also the other one is singing longer notes. In other words, they are copying the errors of each other.

Then we ran the test for all the trials of every couple. In Figure 4.33 is plotted the output of the cross-correlation for A and B part of the 8 trials of couple 5. For each trials is shown the behaviour of the subjects in A and B parts. Is not specified which subject of the couple has a certain behaviour, but we can understand from the colour of the geometric shape, whether there is an imitating or correcting behaviour. Moreover, we can notice that A and B part have always different cross-correlation values and also sometimes show different behaviour on the same trial. If we want a more general view, about the behaviour of both musicians over the 8 trials, to understand which one of the two is showing an imitating or correcting behaviour, we have summarized all this information in the plot in Figure 4.34. For every couple it counts all the cross-correlation output for A part and B part. Then, for each part it shows which was the musician that had a certain behaviour in the majority of the trials.

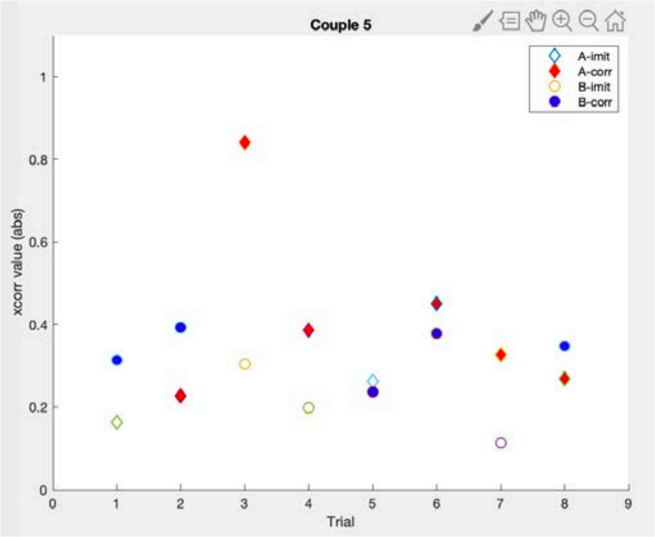


Figure 4.33: A and B parts are distinguished by circles and rhombus. Whether the geometric shape is empty, means imitating behaviour, if it is filled, means correcting behaviour. Y-axis reports the absolute value of the cross-correlation function.

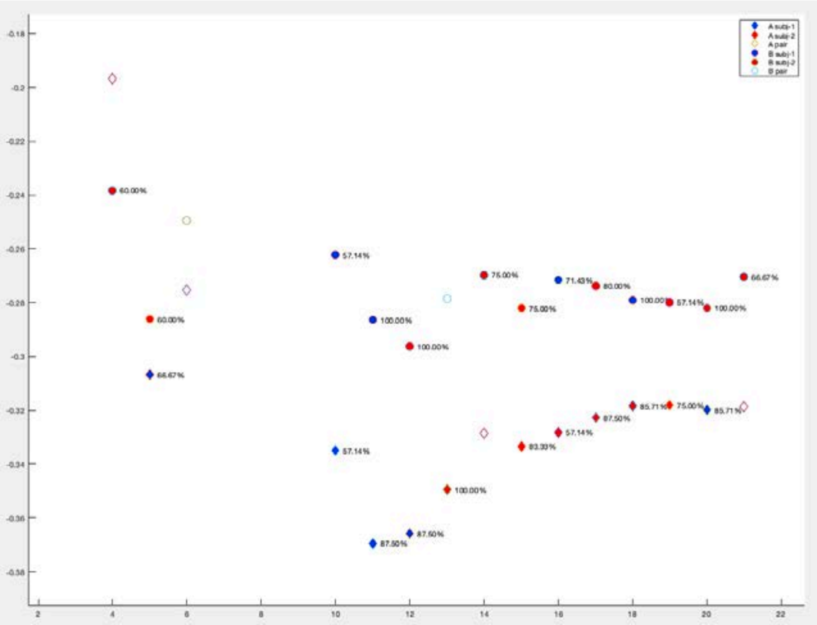


Figure 4.34: Cross-correlation results of each couple divided in A and B part. The colour specify which subject has a more evident behaviour in the 8 trials. Instead the geometric shape is empty whether there isn't a specific behaviour of one subject that is repeated more than 50% of the trials. X-axis represents all the recorded couples and y-axis represents the average cross-correlation value calculated on the 8 trials.

Averaging all the trials of each couple, emerged always a higher correcting behaviour compared to imitating behaviour. Indeed, the average cross-correlation values (y-axis in Figure 4.34) calculated over the 8 trials, shows always negative values, that we interpret as correcting behaviour, from one of the participants in every couple.

The main result that we obtained can be summarized in Figure 4.35, where we can see that the A part achieved always higher cross-correlation values compare with B part, but in couple 4.

We were also interested to see whether a higher correcting behaviour could be the cause of an improvement over some of the objective errors previously calculated, but as you can see from Figure 4.35 there is not evident difference between the cross-correlation values of the different couples, indeed the correcting behaviour has not any significant correlation with any other measures previously computed.

Here we have described the approach that we used to compare the errors of two performers. Based on the common used method of cross-correlation, we applied it to the errors that musicians made. For this reason we have classified two possible conditions: imitating and correcting behaviour.

In case the cross-correlation highlighted a positive value, we treated it as an imitating action, because it means that the errors of one of the two performers are following the behaviour of the errors made by the other musician, so in this case we could have marked one subject as leader and the other as follower. Instead, the majority of performances highlighted a negative value for the cross-correlation function. We referred to this action as correcting behaviour, because when the error of a performer is positive, the next error of the other musician appears negative, so they are correcting each other.

The results show that averaging the output of every performance we got, for all the couples, the predominance of negative cross-correlations values, this means correcting behaviour. Moreover, results suggest that in the music parts where the rhythm is simple to maintain, for musician with years of experience, we have a more negative cross-correlation values respect more difficult parts. We interpret more negative values for the cross-correlation functions as stronger correcting behaviour. This means that it is more easy to have a correcting behaviour between performers over simpler parts of the performed song.

Overall, since we don't have any imitating behaviour, in average over the performances, we can say that in this case study there is not an evident leader-follower interaction.

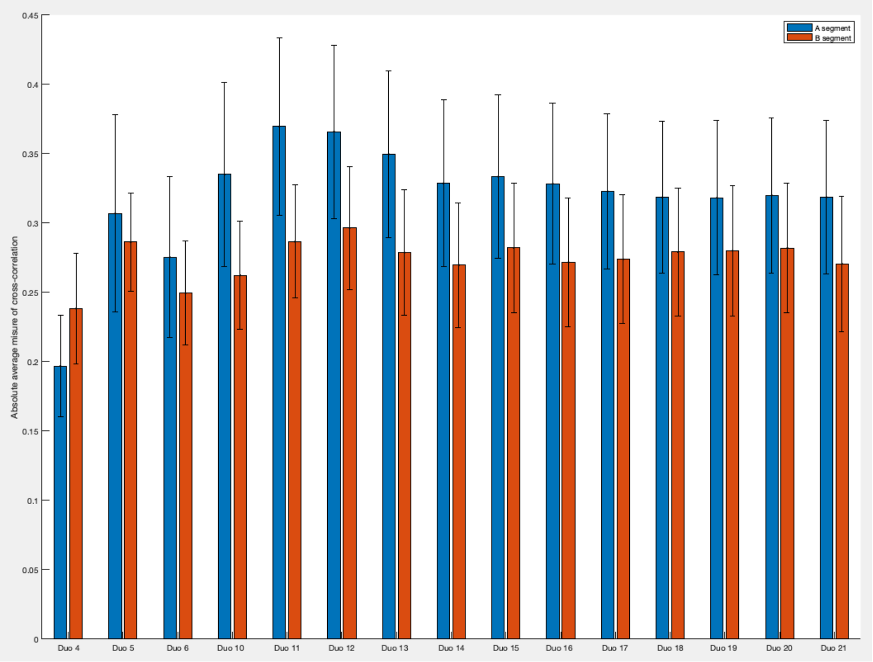


Figure 4.35: For each couple here it is shown the absolute value of the average cross-correlation output and the variance, for A part and B part (across all the 8 trials).

5 Real-time analysis and bio-feedback system

This chapter describes a real application that we built, which exploits the information that the analysis we have previously done can give us.

A possible implementation aims to help musicians, during the execution of a performance, to maintain a regular rhythm. In the performance requested in this experiment, constant rhythm means constant IOI with low fluctuation errors.

The idea from which this real-time analysis and bio-feedback system comes from, is to use the 3 main objective measures we computed before: fluctuation error, collapse error and narration error, to create an application that can be used by performers in order to learn and to improve the execution of the correct notes duration during a performance.

We think that the measures of quality that we have created could be useful not only at the end of a performance to assess it but, whether they could be computed in real-time, but they can be used to create an application that helps musicians to improve the timing precision.

For these reasons there is the need of some sort of bio-feedback system, such as different colored lights or even a sound system, that can provide a reward when the subjects are performing well (according to the objective parameters) and a penalty in case they are changing the timing and are not constant over the execution of the performance.

5.1 Real-time onset detection

First of all, since our experiment is based on the calculation of the notes duration, we need to detect the onset in real-time, while musicians are performing.

In signal processing, onset detection is an active research area and there are annual competitions, for example the MIREX (Music Information Retrieval Evaluation eXchange) that holds annual meeting and features also an Audio Onset Detection contest.

Since this field is well explored we decided to employ the *Madmom* audio signal processing library created in 2015. It is a library written in Python with a strong focus on music information retrieval (MIR) tasks. It implements a lot of functions that can be applied to audio signals, among which there are also an offline and online onset detector. The onset detector program detects all the onset in an audio file according to the algorithm described in [54] which reports a method that incorporates a recurrent neural network that operates in real-time with minimum delay.

Many onset detection methods have been proposed over the years. Traditional methods usually incorporate only spectral and phase information of the signal but top-performance algorithms employ machine learning techniques and use probabilistic information. However, approaches like [55] use neural network and in [56] a Hidden Markov model but they all have in common, that they work only in offline mode because the peak-picking methods used, rely on future information to determine the location of an onset.

Only few algorithms have been designed specifically for online scenarios, for example [57]. In real-time the aim is to minimize the delay between the occurrence of the onset in the audio signal and its detection.

The method explained in [54] and implemented by the Madmom library is based on a revised version of the onset detector algorithm that won 2 years the MIREX onset detection contest, with some modifications in order to enable the system to work in real-time scenarios. The algorithm is structured in three main processing step: signal pre-processing, neural network onset prediction and peak post-processing.

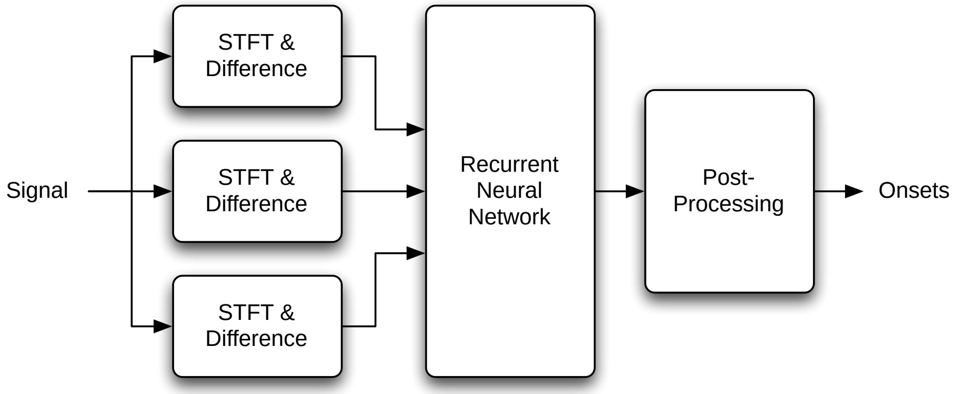


Figure 5.1: Real-time onset detection system overview [54].

As input, the system takes a discretely sampled audio signal with a rate of 100 frames per second and transformed it to the frequency domain with three parallel Short-Time Fourier Transforms (STFT) with different windows lengths to capture both recent and also “older” information.

The data obtained is then fed into the recurrent neural network to detect the next occurring onset in the audio stream. The network was trained as a classifier with supervised learning and early stopping on a 75% portion of the total dataset that consisted of 327 audio excerpts taken from different previous works. The data set included annotations manually checked using spectrograms obtained with different STFT lengths used together to capture the precise timing of an onset without missing any one due to insufficient frequency resolution. The dataset included a total of 28067 onset.

The output of the neural network is an onset activation function with values in the range from 0 to 1. It represents the probability of an onset at a given position.

Finally, simple post-processing is used to report the onsets instantaneously while minimizing the number of false detections.

The main advantage of using a neural network, compared to simple signal-based onset detection methods, is that its onset activation function has a very low noise floor with high peaks only at the onset positions. In this way, can be applied a very low threshold to detect the onsets as early as possible without risking many false detections. Moreover, there is the possibility to merge closed onset, such as 10 ms far from each other, as only one onset detected.

The online algorithm called OnsetDetectorLL (recall to Lucky Luke, the cowboy known to “shoot faster than his shadow”, because it is able to detect an onset before a human can hear it) compared to its offline variant and also to all the detection methods implemented in *aubio* (dedicated library for audio tasks), showed excellent results [54]. It falls short off the performance the offline version of the same algorithm but clearly outperforms other onset detection methods based on spectral flux, complex domain, high frequency content and combinations of them.

After having installed the Madmom package there are some executable programs, such as OnsetDetectorLL, that can be ran in offline mode or online mode. To use the offline mode it needs to specify the audio file to analyse and then it create a CSV file with the timestamps of all the onset detected. Whereas, in the online mode, it takes as input the audio signal that it is captured by the first channel of the computer audio board and then, it dynamically update in real-time a CSV file with all the onset detected.

In both cases it is possible to adjust the threshold for the peak-picking in order to modulate the sensitivity of the onset detection.

In this way, while the onset detector script is capturing and writing the onset timing on a dynamic file, it is possible to read all the new values from this file through MATLAB and using the same methods explained in chapter four, we can compute the three objective parameters (fluctuation, collapse and narration errors) and update their values while performers are playing.

5.2 Bio-feedback system

Our aim is to exploit the objective measures about quality and time precision, while performers are playing, and give them useful information about timing. Therefore, the system analyses the onset detected by the real-time onset detector program, it computes the IOI when a new onset is available and based on the Bayesian approach explained in chapter four, it computes the fluctuation and collapse errors while musicians are playing/singing.

In particular, since we want to inform musicians about how regular they are maintaining the timing, we decided to use different colored lights to advise performers about their timing.

Since each subject is a component that creates the joint action, the error computed for an onset is a useful information for the musicians that played that specific notes. Therefore, we have to reserve as many bio-feedback sources as the number of participants in the ensemble. For example, in this last experiment in order to test the real-time bio-feedback system we used some RGB stage lights of the ASIL.

The set of lights sources reserved for each performers give her/him information about her/his timing. So, we can compute their errors and they can receive different information based on the colours of the lights.

The schema of colours used to advise about the type of error that they are committing is reported in Figure 5.2.

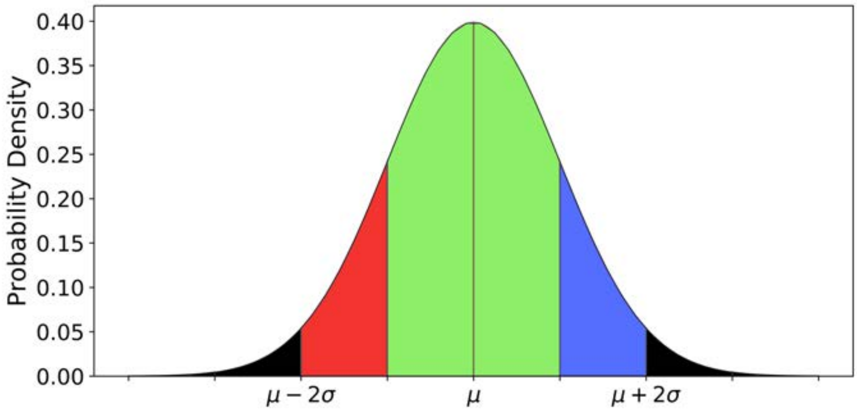


Figure 5.2: Probability distribution of a certain inter-onset class. Based on how far will be the next IOI from the expected value (mean μ and standard deviation σ), the lights will switch to the colour of the assigned area.

As can be easily seen in Figure 5.2 we have defined four different colours that report four different information. Green light is associated to a regular performed IOI. This means that the IOI is closed to the expected duration of its class. Red light is associated to an IOI that is faster (shorter duration) respect to the expected value. The threshold that defines how far an IOI has to appear to be

considered as anticipatory, can be decided based on the kind of application that we want to build. In the experiment conducted in this study, we can consider the thresholds as $\mu \pm \sigma$. On the other side, blue light is associated to an IOI that is slower (longer duration) respect to the predicted value.
Finally, in case the error is bigger than 2σ respect the expected value, what we have previously called collapse error, a black out of the lights reveals the bad execution of the performance.
This is the schema we have used in order to inform in real-time the performers about their timing. The colour of the lights reflect the value of the fluctuation error and in case of black out they report the presence of a collapse error.
Moreover, the system provided also a reward for the participants in case both of them were maintaining a correct execution of the interval durations for a certain amount of time. Concretely was created a game of coloured lights that started when performers reached a pre-fixed number of consecutive IOI with low fluctuation error, i.e. IOI that fall inside the green region of Figure 5.2.

5.3 Real-time framework

The entire framework represents one possible real application that can benefit from this study. The final test aims to employ the new objective qualitative measures, described in chapter four, and thanks to the real-time onset detector and the bio-feedback system, it aims to improve the learning process, to help performers keeping the correct duration of the notes during a musical performance. Finally, the computed measures such as fluctuation error, collapse error, and narration error could also be used as reference to asses the quality of the performance. As we have seen from the results at the end of chapter four, the most reliable parameters that best reflect the quality of a performance are the narration error and the fluctuation error.
We tried to implement this framework in which we have multiple performers singing/playing together monitored in real-time, with the lights that change based on the errors of each performers and with the final computation of the overall quality of the performance.

5.3.1 Setup

This new application was tested and executed in the ASIL. We tested the framework with an ensemble formed by even more performers than before, involving a total of three musicians. They were positioned at the center of the laboratory, in order that they could see each other but more important they had a set of three or four RGB LED lights dedicated to each one of them, as a help for them to adjust the timing in case of mistakes.

We noticed that using the onset detector over a voice audio signal, as we have done before, could be difficult to recognize with high precision all the notes sang by the musicians. In order to increase the reliability and the precision of the real-time onset detector, we decided to use components of the drum to play, instead of singing. Therefore, performers had to use a bass drum, a snare drum and a floor tom, as you can see in Figure 5.3.

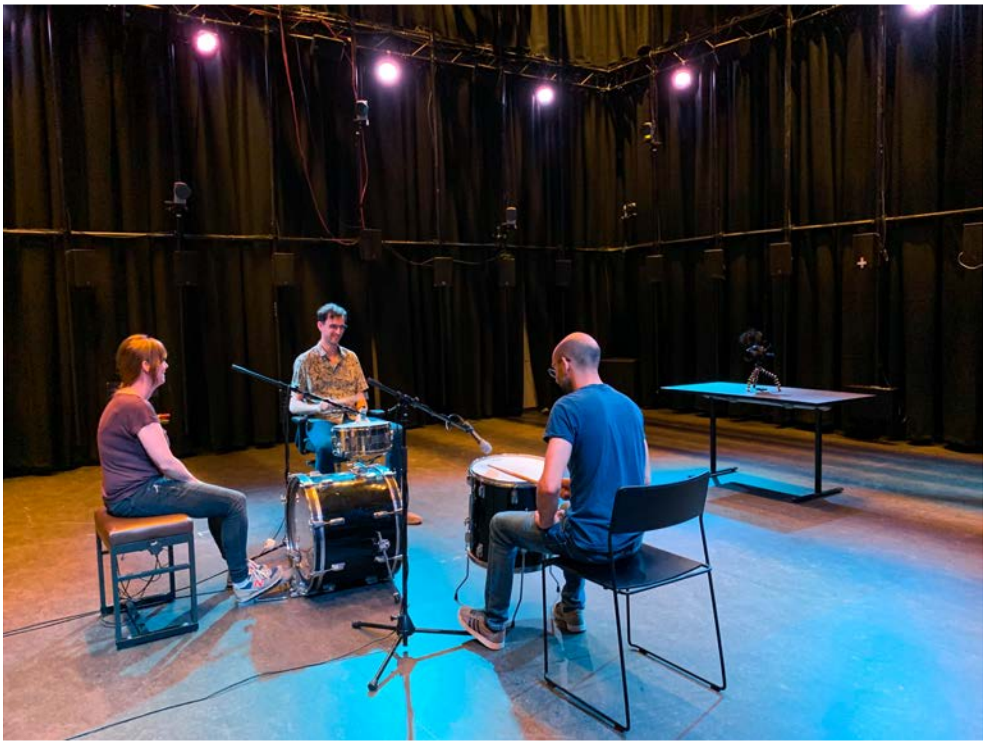


Figure 5.3: Setup with three musicians in the ASIL. Every one with an individual component of the drum and a set of lights that during the performance change the colours based on the information to give to each one of the performers.

Musicians were free to chose the drum part with which they felt more comfortable. In Figure 5.3 you can see the setup with three musicians in the ASIL. As you may have noticed, each drum's component has a directional microphone to capture the sound and they were positioned as closed as possible to their sound sources. Every microphone was connected to the rack PC of the ASIL, that can be seen as computer with advanced audio boards in order to manage multiple audio inputs (with the availability of phantom power for the microphones sources) and multiple audio outputs. The three microphones were connected to the mic. inputs of the rack PC.

Using Ableton live 10 software, we created four output channels to redirect the microphones sources. In the first output channel we mixed together the signals coming from all the microphones. This will be the channel used by the onset detector program. Then, in the output channels 2, 3 and 4 we redirected the sound coming from each microphone independently. Using Ableton we could adjust the volume of each source in order to normalize them to the same level. In this way at the output of the audio board of the rack PC we got 4 channels that we sent to the M-Track Eight by M-Audio, an 8 channel audio board that allowed us to connect the four channels to the computer were all the computations and analysis were done.

Notice that the role of Ableton software in the rack PC could be replace by an analog mixer. At the end we just needed to mix all the microphones signals in one channel for the onset detector program and then control the output volume and redirect the single sources of sound to independents output channels.

The need of a channel with all the mixed sound sources was necessary because the Madmom library analyse the signals that it receives in the first channel of its computer's audio board. For this reason mixing together all the sources it's possible to analyse and to detect the onset of multiple sources that are playing together.

Since there was the need to recognized which of the three musicians was the actor of the onset that was detected, we created a python script that using *pyaudio* library could evaluate the volume of the input microphones and understood which of the three performers was actually playing at every detected onset. In this way we could associate every onset to the musicians that generated it.

Finally, to manage the lights that have to advise the musicians about their errors, we used the Enttec DMX USB PRO interface for sending and receiving a physical DMX512 (Digital MultipleX) signal to control LED lights from PC and Mac programs. Through the DMX python library we were able the control the colour of the lights based on the errors.

Since there are many libraries and different scripts to use in this kind of experiment the final setup was managed as follow: the onset detector python library works independently, it analyse the signal that contain the combined sources and when an onset is detected it writes the timestamp in a CSV file. The same CSV file is read by the MATLAB script that compute the IOI and using the Bayesian approach update the distribution of the 3 inter-onset classes. For each IOI it computes also its error from the predicted value and write the error value in a shared file.

The last program we created is written in python and has 2 threads: the first detects which microphones is working, analysing the amplitude of the signals that comes from the microphones, in this way the other thread, that controls the lights through the DMX protocol, can take the value of the error given by the MATLAB script and the number of the microphone that has generated the last note, to adjust the colour of the lights for the musicians who generated a specific onset.

However we observed quite soon that whether the lights change the colour at each single error, could be too fast and difficult for the musicians to adapt her/his tempo based on the advice. For this reason we created a simple filter that had the aim to create a smoother change between the colours of the lights. In particular, it is important to remember which are the previous errors of a subject and their amplitudes. The algorithm for the smoothing works as follow: it simply sum up all the errors of one subject, no matter if they are positive or negative errors, respect the expected inter-onset mean class. When the sum is higher or lower than a pre-fixed threshold, the lights turn red or blue based on whether the timing is getting faster or slower. Then the filter reduce the sum of errors proportionally to the time passing.

With this method there is the need of a series of consecutive errors to reach a status that expect to alert the performer. Moreover, we can have multiple consecutive small errors that summing up overtake the threshold and the user is alerted that there is an evident change of the notes duration.

Overall this smoothing algorithm seems to work well. Adjusting the parameters such as the thresholds for the alert of the performers and how much decreasing the sum of the errors over time, it can be adapted to different applications.

5.3.2 Procedure of the real-time test

The aim of this first real application, was to test the real-time onset detector with three performers playing different parts of the drum and to try also a first implementation of a new sort of learning method, that exploits the errors extract from the Bayesian approach to change the colour of the lights that should be used as timing reference by musicians.

As a first test we analysed a trio of musicians. They were musicians with more than 5 years of experiences in music, but none of them was a drummer.

They were instructed to perform the exact same melody created from the “Just” and “Beat” scores of Figure 3.1 and Figure 3.2 but this time split in three parts. Each one of the musicians were allowed to chose the part of the drum they liked the most and they were positioned at the center of the ASIL.

The three microphones were already positioned closed to the point where the musicians had to hit the drum.

Since not all of them felt comfortable to play the drum, we allowed 30 minutes of free rehearsing.

The first thing we needed to do was the calibration of the volume of every microphone. Based on the pressure the musicians used to play the drum, we had to equalize all the volumes to be as similar as possible to allow the onset detector to caught all the onset.

Second, in order to correctly detect which microphone was the source of each note, we set a threshold for every microphone. The threshold allowed to distinguish when the sound captured by a microphone came from its user or whether it came from a different musician. Since the sounds emitted by the drum’s parts are quite loud and they can be easily captured by all the microphones in the stage, with this method a python script could recognize which was the correct source of the sound just checking which of the three input signals (one for each microphone) had an amplitude higher than its threshold.

When all the microphones were correctly calibrated the session could start.

Musicians had to alternate the same A and B parts of the melody that we have described in chapter 3, for a total of 120 seconds. As in the analysis reported in the chapter 4, the first 15 seconds of the performance were used to create the prior knowledge about the duration of the 3 inter-onset classes. During this initial period all the lights were white to illuminate the stage, and then when the priors were formed the colours of the lights of each performers started to change based on their errors. In case all of them received green lights (that means good tempo regularity) they received the reward in the form of a game of colored lights.

At the end of the performance the lights stopped and the program revealed the final fluctuation error, collapse error and narration error.

From the analysis in chapter 4 we have discovered that narration error and the fluctuation error are the best parameters that can summarize the quality of the performance.

With the fluctuation error we can observe the precision to follow a regular timing. Usually, best performances of the previous experiment, that we can now call the offline version, got a total fluctuation errors around 30 milliseconds. Instead, for the fluctuation error, best couples of offline version reached a perfectly execution over the 90% of the total song. Moreover, the same plots we did for the offline version are computed also in the online version. In this way, at the end of a performance musicians can have a look at the plots such as the development of the 3 intervals over time, pictured in Figure 4.10, and understand whether there are some repetitive errors that occur over the performance looking at the plot with the sequence of types of errors shown in Figure 4.13.

The plots and the results of the errors, can be used in this way as a reference for the musicians to evaluate themselves and to improve their timing precision during the performance.

5.3.3 Results

The online version of the test with the real-time onset detector and with an ensemble formed by three musicians has proved to work really well.

First of all we compared the results, that are, the total errors obtained with the real-time analysis, with the correct errors obtained with an offline analysis.

For this reason, we took the recorded performances and we calculated all the onset with Sonic Visualizer, as for the previous experiment with the couples of musicians. After a manual check, we used the extracted onset to re-compute the (offline) analysis and obtained again the fluctuation, collapse and narration errors.

Overall the offline results differ from the online version of only few milliseconds. We have proved that the onset detection can work in real-time with a general error of only 1% of the total onset missed. This is caused of course by the use of the drum, that is a particularly suitable instrument for the goals of this experiment, but the correct analysis of the onset detector is mainly determined by the precision with which the microphones and the signals are adjusted.

Regards the learning process, the filter we applied to provide a smooth effect in the change of lights colours works fine, but some performers could prefer a more quick response to the note played.

Since the lights are just a bio-feedback with the aim to help performers, their behaviour and responsiveness can be adjusted based on the preference of each musician.

Some performers reported that having to look at the lights could distract them from the execution of the performance and from maintaining an eye contact with their partners. For these reason another solution could be studied implementing a bio-feedback that use for example a sound system instead of lights feedback.

We even tried the same experiment with only one person playing the drum and it works as with multiple performers.

This test was just the first implementation that exploited the analysis reported in this study. We created an application that works in real-time during a performance and which can elaborates all the data without delay, even from multiple performers. Here we have shown that the results of the analysis computed in real-time can be used, for example, to create a new learning method for musicians.

6 Conclusion

In this study we faced the manner of evaluating the quality of a musical performance, through objective measures. The research focused on the analysis of the timing that emerges from a musical interaction.

We created the environment that set the basis for this new kind of analysis and that allowed us to collect data about the timing of a joint action. Therefore, exploiting the medieval hocket music style, we defined the rules and created a test in which a performance was made by the joint action of two performers.

From this simple and basic musical interaction we collected the data of 15 different couples of musicians, for a total of 120 performances recorded. For each performance we collected also the self assessments of the participants, namely the annotation and agency values.

We could use these two measures, that overall evaluate the quality of a performance, as our first subjective parameters of reference and from which starting elaborating and comparing other objective measures, computed from the collected data.

Relying on the predictive coding theory, we created a system that using the Bayesian probability theory, allows to emulate the behaviour of the human brain, that, according with this theory, it uses the sensory input to constantly generating and updating a mental model. The Bayesian approach that we built follows exactly this theory. It can generate an initial prior knowledge and then it can update the model when new evidences are available.

The key aspect that we considered, to later defining some measures of quality, is the prediction error. It can be measured as the distance (over time) from the value predicted by the model and the real evidence.

Based on these assumptions, we defined two types of errors: fluctuation error and collapse error. The first one includes the micro-timing errors that are necessary, even in good performances, to maintain a stable rhythm. Instead, the second represents the macro-timing errors that should significantly affect the quality of the performance.

Checking whether these two measures computed from the data, reflect the values about the quality collected with annotation and agency evaluations, the results suggest that fluctuation error is highly correlated to both the subjective measures, instead collapse error is not correlated with any one of them.

From this results we can consider the fluctuation error as a good approximation of a personal qualitative measure of a musical performance but we can't say the same thing for the collapse error. Using this last parameter we can not evaluate the quality of a performance as a human being would do it.

Maybe a different computation of this type of error could bring to different results. A different ways to compute this error could be to sum all the collapse errors instead of computing the RMSE of all the collapse errors detected on a performance. In this way the number of times that this mistakes appear would have a higher impact in the final error value that we take into consideration at the end of each performance.

The third objective parameter that we computed takes into account the classes that the Bayesian approach assigned to each onset and it checks how well the performers have repeated the same patterns of intervals over the song. This could appear as a simple measure but the results showed that it reached the highest correlation with the annotation and agency measures. Therefore we can rely on this parameter to evaluate the quality of a musical performance even more than the fluctuation error.

Then, we tested also the hypothesis that we made at the beginning of the study, about the increasing of the performance due to the moving condition respect the non-moving condition and the learning effect that should be present while repeating the same performance 8 times.

Comparing the behaviour of the three parameters that we have built between the moving condition and non-moving condition, it emerged that the movement helps only to maintain a lower fluctuation error. Collapse error and narration error didn't show significant difference between the two conditions.

To test whether the learning effect leads to a reduction of the errors along the 8 trials of each couple, we tested whether the results of the first trial are significantly different from the results we got on the last trial. The Wilcoxon signed rank test showed that the two trials didn't bring significantly different results. The lack of free time to rehearse together, discuss and try the critics parts of the song could be the cause of these results.

The last parameters we studied were the imitating and correcting behaviour, through the use of the cross-correlation between the sequences of errors of each couple. We considered an imitating behaviour the case in which one of the musicians was copying the actions (in our case the errors) of the partners. Instead, we associated a correcting behaviour, whether one of the partners compensated the errors of the other. The results reported a correcting behaviour in the majority of the performances. We calculated also that the degree of cross-correlation is not significantly correlated with any of the computed subjective measures. This means that having a higher or lower imitating behaviour doesn't correspond to a better or worse performance in terms of errors.

The deployment of the Bayesian probability is something that has never been used before in this kind of musical research. However, applying this technique we obtained two parameters, namely fluctuation error and narration error, that have shown to be correlated to measures of quality given by human beings. These parameters confirmed that we reached the initial goal of this study, but we went further more.

We built a real application that implements the objective measures we have created. Its aim is to help performers to maintain a correct timing during the performance. Moreover, at the end of a performance it provides other tools to help musicians to understand which errors they committed, to give them some insight about the rhythm that they kept and finally it offers an overall evaluation of the quality of the performance, thanks mainly to the fluctuation error and the narration error.

This was just a possible application that can benefit from the results achieved

in this study. The last experiment proved also that this kind of analysis can be done in real-time and can involve also multiple performers.
There are many improvements that could be done starting from this project. The Bayesian approach could be refined to adapt also to data the are not normally distributed. Could be modified the calculation of the collapse error in some different ways as we have seen before, because, as the fluctuation error it could be related to a measure of quality.
Could be created a different subjective measure about quality, for example, to evaluate the performance by a team of experts, in order to have a more homogeneous criterion of evaluation over different couples.
However, the application of the methodologies developed in this project can open new paths especially in the employment of the Bayesian statistics. Moreover, as we have proved, this study could have also immediate and practical applications, for example, to develop a better learning process for musicians and for a fast and easy objective assessment of a musical performance.

References

[1] M. L. Chanda and J. Levitin, “The neurochemistry of music,” *Trends in Cognitive Sciences*, vol. 17, no. 4, pp. 179–193, April 2013.

[2] H. Honing, C. ten Cate, and I. Peretz, “Without it no music: cognition, biology and evolution of musicality,” *Phil. Trans. R.*, 2015.

[3] M. Leman, P. J. Maes, and M. Lesaffre, *What is embodied music interaction*, January 2017.

[4] M. Leman, *The Expressive Moment.* MIT Press, 2016.

[5] P. Manning, *Electronic and computer music.* Oxford University Press, 2013.

[6] N. Sebanz, H. Bekkering, C, and G. Knoblich, “Joint action: bodies and minds moving together,” *Trends in Cognitive Sciences*, vol. 10, no. 2, February 2006.

[7] A. D’Ausilio, G. Novembre, L. Fadiga, and P. E. Keller, “What can music tell us about social interaction?” *Trends in Cognitive Sciences*, vol. 19, no. 3, March 2015.

[8] E. P. Keller, G. Novembre, and M. J. Hove, “Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination,” *Phil. Trans. R.*, 2015.

[9] P. Keller, “Joint action in music performance,” *Enacting intersubjectivity: a cognitive and social perspective to the study of interactions*, pp. 205–221, 2008.

[10] ——, “Ensemble performance: Interpersonal alignment of musical expression,” *Expressiveness in music performance: empirical approaches across styles and cultures*, pp. 260–282, 2014.

[11] E. Large and S. Grondin, “Resonating to musical rhythm: Theory and experiment,” *Psychol Time*, pp. 189–232, 01 2008.

[12] B. H. Repp, P. E. Keller, and N. Jacoby, “Quantifying phase correction in sensorimotor synchronization: empirical comparison of three paradigms.” *Acta psychologica*, vol. 139 2, pp. 281–90, 2012.

[13] W. Goebl and C. Palmer, “Synchronization of timing and motion among performing musicians,” *Music Perception - MUSIC PERCEPT*, vol. 26, pp. 427–438, 06 2009.

[14] I. Konvalinka, P. Vuust, A. Roepstorff, and C. Frith, “Follow you, follow me: Continuous mutual prediction and adaptation in joint tapping,” *Quarterly journal of experimental psychology*, vol. 63, pp. 2220–30, 11 2010.

[15] P. Keller, “Attending in complex musical interactions: The adaptive dual role of meter,” *Australian Journal of Psychology*, vol. 51, no. 3, pp. 166–175, 1999.

[16] H. De Jaegher and E. Di Paolo, “Participatory sense-making: an enactive approach to social cognition,” *Phenomenology and the Cognitive Sciences*, vol. 6, pp. 485–507, 12 2007.

[17] E. Bigand, S. McAdams, and S. Forêt, “Divided attention in music,” *International Journal of Psychology*, vol. 35, no. 6, pp. 270–278, 2000.

[18] J. Phillips-Silver, “Searching for roots of entrainment and joint action in early musical interactions,” 02 2015.

[19] P. E. Keller, “Joint action in music performance.” 2008.

[20] J. Ginsborg, R. Chaffin, and G. Nicholson, “Shared performance cues in singing and conducting: A content analysis of talk during practice,” *Psychology of Music*, vol. 34, no. 2, pp. 167–194, 2006.

[21] C. Palmer, “Music performance,” *Annual review of psychology*, vol. 48, pp. 115–38, 02 1997.

[22] M. Ragert, T. Schroeder, and P. E. Keller, "Knowing too little or too much: the effects of familiarity with a co-performer's part on interpersonal coordination in musical ensembles," *Frontiers in psychology*, vol. 4, p. 368, 2013.

[23] A. Kleinspehn, "Goal-directed interpersonal action synchronization across the lifespan: A dyadic drumming study," Ph.D. dissertation, 2008.

[24] A. P. Demos, R. Chaffin, and V. Kant, "Toward a dynamical theory of body movement in musical performance," *Frontiers in Psychology*, vol. 5, p. 477, 2014.

[25] M. Lesaffre, P.-J. Maes, and M. Leman, Eds., *The Routledge companion to embodied music interaction.* Routledge, 2017.

[26] O. A. Heggli, I. Konvalinka, M. L. Kringelbach, and P. Vuust, "Musical interaction is influenced by underlying predictive models and musical expertise," 2019.

[27] K. V. Mardia and P. E Jupp, *Directional Statistics*, 01 2000.

[28] E. Tognoli, J. Lagarde, G. C. DeGuzman, and J. A. S. Kelso, "The phi complex as a neuromarker of human social coordination," vol. 104, no. 19, pp. 8190–8195, 2007.

[29] G. Knoblich and J. S. Jordan, "Action coordination in groups and individuals: Learning anticipatory control." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 29, no. 5, p. 1006, 2003.

[30] M. Clayton, K. Jakubowski, and T. Eerola, "Interpersonal entrainment in indian instrumental music performance: Synchronization and movement coordination relate to tempo, dynamics, metrical and cadential structure." *Musicae Scientiae*, vol. 23, pp. 304–331, 07 2019.

[31] P. E. Keller and M. Appel, "Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 1, pp. 27–46, 2010.

[32] L. Bishop and W. Goebl, "When they listen and when they watch: Pianists' use of nonverbal audio and visual cues during duet performance," *Musicae Scientiae*, vol. 19, no. 1, pp. 84–110, 2015.

[33] B. Zendel, B. Ross, and T. Fujioka, "The effects of stimulus rate and tapping rate on tapping performance," *Music Perception*, vol. 29, pp. 65–78, 09 2011.

[34] T. Eerola, K. Jakubowski, N. Moran, P. E. Keller, and M. Clayton, "Shared periodic performer movements coordinate interactions in duo improvisations," *Royal Society Open Science*, 2 2018.

[35] A. Walton, M. Richardson, P. Langland-Hassan, A. Chemero, and A. Washburn, "Musical improvisation: Multi-scaled spatiotemporal patterns of coordination." in *CogSci*, 2015.

[36] D. T. Lee and A. Yamamoto, "Wavelet analysis: Theory and applications," *Hewlett Packard Journal*, 1994.

[37] A. Graps, "An introduction to wavelets," *IEEE computational science and engineering*, vol. 2, no. 2, pp. 50–61, 1995.

[38] B. Howard, "Dynamics of controlled systems," 2013.

[39] M. Leman, *Embodied Music Cognition and Mediation Technology.* The MIT Press, 2007.

[40] A. Clark, *Embodied prediction.* Open MIND. Frankfurt am Main: MIND Group, 2015.

[41] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *TRENDS in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[42] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, p. 127, 2010.

[43] S. Deutsch, "Bayesian brain: Probabilistic approaches to neural coding," *IEEE Pulse*, vol. 1, no. 3, pp. 64–65, Nov 2010.

[44] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, “The helmholtz machine,” *Neural computation*, vol. 7, no. 5, pp. 889–904, 1995.

[45] S. Koelsch, P. Vuust, and K. Friston, “Predictive processes and the peculiar case of music,” *Trends in Cognitive Sciences*, 2018.

[46] M. Rohrmeier and S. Koelsch, “Predictive information processing in music cognition. a critical review,” *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, vol. 83, pp. 164–75, 02 2012.

[47] P. Haggard and B. Eitam, *The sense of agency.* Social Cognition and Social Ne, 2015.

[48] D. M. Wolpert, K. Doya, and M. Kawato, “A unifying computational framework for motor control and social interaction,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1431, pp. 593–602, 2003.

[49] E. Pacherie, “The phenomenology of joint action : Self-agency versus joint agency,” 2010.

[50] L. De Bruyn, M. Leman, and D. Moelants, “Quantifying children’s embodiment of musical rhythm in individual and group settings,” in *The 10th International Conference on Music Perception and Cognition*, 2008.

[51] R. Jacobs, “Bayesian statistics: Normal-normal model,” *Department of Brain & Cognitive Sciences University of Rochester*, December 2008.

[52] J. Brooks-Bartlett. (2018) Probability concepts explained: Bayesian inference for parameter estimation. Medium. [Online]. Available: https://urly.it/32jmj

[53] H. Yanagisawa, O. Kawamata, and K. Ueda, “Modeling emotions associated with novelty at variable uncertainty levels: A bayesian approach,” *Frontiers in Computational Neuroscience*, 01 2019.

[54] S. Böck, A. Arzt, F. Krebs, and M. Schedl, “Online real-time onset detection with recurrent neural networks,” 09 2012.

[55] A. Lacoste and D. Eck, “A supervised classification algorithm for note onset detection,” *EURASIP J. Adv. Signal Process*, vol. 2007, no. 1, pp. 153–153, Jan. 2007.

[56] N. Degara, M. E. Davies, A. Pena, and M. D. Plumbley, “Onset event decoding exploiting the rhythmic structure of polyphonic music,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1228–1239, 2011.

[57] D. Stowell and M. Plumbley, “Adaptive whitening for improved real-time audio onset detection,” in *Proceedings of the 2007 International Computer Music Conference, ICMC 2007*, 2007, pp. 312–319.

Acknowledgments

I would first like to thanks Prof. Dr. Marc Leman, head of IPEM of the University of Ghent, for offering me the opportunity to work in the most advanced centre for music embodied studies and also for driving me and inspiring me in this fascinating project.

I would like to acknowledge Prof. Sergio Canazza of the University of Padova, for the encouragement and appreciation that I received since the first moment I have contacted him. I am gratefully indebted for his valuable comments on this thesis.

Special thanks to Jeska Bauhmann and Alessandro Dell'Anna who have collaborated with me in this research and who were always able to help me.

I would also like to thanks the people of the amazing team of IPEM, for creating a comfortable and collaborative working environment.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement.