

Analysis of Ensemble Expressive Performance in String Quartets: a Statistical and Machine Learning Approach

Marco Marchini

TESI DOCTORAL UPF / 2014

Thesis advisors

Dr. Mestre Esteban

Dr. Ramirez Rafael

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona, Spain





Copyright © Marco Marchini, 2014

Dissertation submitted to the Department of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfilment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA.

Music Technology Group (<http://mtg.upf.edu>), Dept. of Information and Communication Technologies (<http://www.upf.edu/dtic>), Universitat Pompeu Fabra (<http://www.upf.edu>), Barcelona, Spain.



*In loving memory of Riccardo Venier (1982-2005),
violin performer and brilliant unconventional mind.*



Acknowledgments

First and foremost I would like to express my deep gratitude to Xavier Serra: initially for giving me the opportunity to access the Sound and Music Computing Master, and later for supporting me and accepting me as a Ph.D. candidate.

I am also greatly thankful to my two Ph.D. advisors. Esteban Maestre taught me how to measure live musicians performances, demonstrated me how to use complex electronic sensors and enriched my knowledge on Signal Processing among other things. He transmitted me the enthusiasm and the urge to improve every little detail. It has been an honor to be his first Ph.D. student. Without his contributions in time, ideas and funding this work would have never existed. Rafael Ramirez has been a great guide in Machine Learning, especially in the last two years of my Ph.D. program. He provided me with fresh new ideas right in the moment I needed them and with the encouragements that moved me to complete this dissertation, even in the tough times of the Ph.D. pursuit.

I also want to express my gratitude to my Ph.D. colleague Panos Papiotis. It has been a great pleasure working together with him on this project. He has been the source of many good advices, and an enjoyable person to share the office with. Not only has he been a good colleague but also a good friend in several life situations. I want to send special thanks to Hendrik Purwins, who was Panos' and my advisor during the Sound and Music Computing Master. He introduced me to the Art of R&D and taught me how to constantly upgrade my abilities, and how to translate them into results and publications. But most of all, he taught me how I can have fun doing all this. Without him, I would have never even applied for a Ph.D.

position. Further special thanks go to Alfonso Perez. He has been of great support both for the definition of the experiments and their organization. During the first year of the Ph.D. he organized a workshop on musical expression in ensembles that planted the seeds of our forthcoming research experiments. I also take the occasion to acknowledge Guillermo Navarro for participating to the workshop and providing us with scores to use in the recording sessions as well as many valuable suggestions.

I would like to thank Prof. Marcelo Wanderley and Erika Donald for the support in organizing the experiments, as well as CIRMMT and BRAMS labs at Montreal (Quebec, Canada) for hosting them. I am also grateful to the other McGill researchers who provided material support for the correct development of the experiments: Carolina Brum Medeiros, Marcello Giordano and Vincent Freour.

I want to express my heartfelt gratitude to some people that directly or indirectly contributed to this work. First of all I want to thank Oscar Mayor for his contributions to the Repovizz project and for always being there to help me with any technical problems. Special thanks go to Quim Llimona, an incredibly talented student, who made Repovizz even more visual. Then I want to thank Genis Caminal for the incredible help in designing and constructing our custom synchronization board. I additionally thank Merljin Blaauw for leaving us with a functioning VST-plugging for acquiring violin performance and for helping me in extending its functionalities from time to time. I thank Rafael Caro for devoting hours of his unconditional help to the musicological analysis of the music scores. Grateful thanks go to Jordi Bonada for discussing with me the details of the score-performance aligning algorithm and for giving me valuable advices. I want to thank Graham Coleman, who promptly replied to my question on the MTG newsletter with a valuable advice about the integration of Weka into Matlab making it possible for me to automatize all the machine-learning tests. Lastly, I would express my deep gratitude to my friend Andrea Bensi, who offered me some extra cloud computing power at the beginning of my machine-learning experiments.

During these years at the UPF I have got technical help from the CAU department (Centre D'atenció a l'Usuari) although I never imagined how much I would have been grateful to them. I want to acknowledge Ivan Jimènez Rodà who set up a new open grid scheduling service from UPF, which was crucial in speeding up (by a factor of at least 10) my late experiments. He always promptly replied to my lengthy emails with questions

about the usage of the service. Also a special thank goes to my guardian angel Enric Cañada Muñoz who, when I thought I was doomed, succeeded in recovering data worth months of my work from my broken hard drive (and for free!!). Another special thanks goes to the people from the administration that always knew which button I had to push: Alba Rosado, Cristina Garrido, Sonia Espí, Vanessa Jimenez and Lydia Garcia.

During these years I felt part of the Music Technology Group, and had the opportunity to meet many valuable researchers and friends. This experience was not only formative for me from a professional standpoint but also from a personal standpoint. I would like to thank many of the past and the present MTG members that I met during my stay in Barcelona. I thank Emilia Gomez, Jordi Janer, Perfecto Herrera and Sergi Jordà (together with X. Serra) who have been wonderful teachers. I thank all the people who have accompanied me socially and academically. I attempted to make a list of some, with a high chance of forgetting someone: Marti Umbert, Pratysh, Frederic Font, Andrés Bucci, Carlos De Los Santos, Sebastian Mealla, Alvaro Sarasua, Mathieu Bosi, Mohamed Sordo, Nicolas Wack, Gerard Roma, Dmitry Bogdanov, Jose Ricardo Zapata, Justin Salamon, Carles F. Julià, Daniel Gallardo, Laia Cagigal de Frias, Zuriñe Resa, Joan Pablo Carrascal, Imanol Gomez, Jorge García, Juanjo Bosch, Leny, Vinceslas, Sergio Giraldo, Zacharias Vamvakousis, Panagiotis Melidis, Andreas Neocleous, Alastair Porter, Ricard Marxer, Saso Musevic, Sertan Şentürk, Stanislaw Gorlow, Martin Haro, Cyril Laurier, Agustín Martorell, Piotr Holonowicz, Waldo Nogueira, Sankalp Gulati, Koduri Gopala Krishna.

Last but not least, sincere thanks go to my family and friends for the loving support. Warmth.



Abstract

Computational approaches for modeling expressive music performance have produced systems that emulate human expression, but few steps have been taken in the domain of ensemble performance. Polyphonic expression and inter-dependence among voices are intrinsic features of ensemble performance and need to be incorporated at the very core of the models. For this reason, we proposed a novel methodology for building computational models of ensemble expressive performance by introducing inter-voice contextual attributes (extracted from ensemble scores) and building separate models of each individual performer in the ensemble. We focused our study on string quartets and recorded a corpus of performances both in ensemble and solo conditions employing multi-track recording and bowing motion acquisition techniques. From the acquired data we extracted bowed-instrument-specific expression parameters performed by each musician. As a preliminary step, we investigated over the difference between solo and ensemble from a statistical point of view and show that the introduced inter-voice contextual attributes and extracted expression are statistically sound. In a further step, we build models of expression by training machine-learning algorithms on the collected data. As a result, the introduced inter-voice contextual attributes improved the prediction of the expression parameters. Furthermore, results on attribute selection show that the models trained on ensemble recordings took more advantage of inter-voice contextual attributes than those trained on solo recordings. The obtained results show that the introduced methodology can have applications in the analysis of collaboration among musicians.



Resum

L'estudi de l'expressivitat musical ha produït models computacionals capaços d'emular l'expressivitat humana, però aquests models encara no es poden aplicar al cas dels conjunts musicals. Per estudiar l'expressivitat dels conjunts musicals, s'han de tenir en compte dues característiques principals: l'expressió polifònica i la interdependència entre veus. Per aquesta raó, proposem una nova metodologia que es basa en la introducció d'una sèrie d'atributs intervocals, que hem extret de la partitura, que es poden utilitzar per construir models d'expressivitat individuals per a cada un dels músics. Hem colleccionat un conjunt de peces musicals a partir de l'enregistrament multipista i de la captura de moviments d'un quartet de cordes, en un corpus que recull peces concretes tocades tant en grup com individualment. D'aquestes dades, hem extret diversos paràmetres que descriuen l'expressivitat per a cada un dels músics d'un conjunt de corda. El primer pas ha estat estudiar, des d'un punt de vista estadístic, les diferències entre l'actuació d'una mateixa peça tant en solitari com en grup. Després, hem estudiat les relacions estadístiques entre els atributs intervocals i els paràmetres d'expressivitat. A continuació, hem construït models d'expressivitat a partir de la utilització d'algoritmes d'aprenentatge automàtic amb les dades colleccionades. Com a resultat, els atributs intervocals que hem proposat han millorat la predicció del paràmetres d'expressivitat. Hem pogut demostrar com aquests models que han après d'actuacions en grup utilitzen més atributs intervocals que aquells que ho han fet d'actuacions en solitari. Aquests resultats mostren que la metodologia i models introduïts es poden aplicar en l'anàlisi de la col·laboració entre membres d'un conjunt musical.



Resumen

El estudio de la expresividad musical ha producido modelos computacionales capaces de emular la expresividad humana, pero estos modelos todavía no se pueden aplicar al caso de los conjuntos musicales. Para estudiar la expresividad de los conjuntos musicales, se deben tener en cuenta dos características principales: la expresión polifónica y la interdependencia entre voces. Por esta razón, proponemos una nueva metodología que se basa en la introducción de una serie de atributos intervocales, que hemos extraído de la partitura, que se pueden utilizar para construir modelos de expresividad individuales para cada uno de los músicos. Hemos coleccionado un conjunto de piezas musicales a partir de la grabación multipista y de la captura de movimientos de un cuarteto de cuerdas, en un corpus que recoge piezas concretas tocadas tanto en grupo como individualmente. De estos datos, hemos extraído varios parámetros que describen la expresividad para cada uno de los músicos de un conjunto de cuerdas. El primer paso ha sido estudiar, desde un punto de vista estadístico, las diferencias entre la actuación de una misma pieza tanto en solitario como en grupo. Después, hemos estudiado las relaciones estadísticas entre los atributos intervocales y los parámetros de expresividad. A continuación, hemos construido modelos de expresividad a partir de la utilización de algoritmos de aprendizaje automático con los datos coleccionados. Como resultado, los atributos intervocales que hemos propuesto han mejorado la predicción de los parámetros de expresividad. Hemos podido demostrar cómo los modelos que han aprendido de actuaciones en grupo utilizan más atributos intervocales que aquellos que lo han hecho de actuaciones en solitario. Estos resultados muestran que la metodología y modelos introducidos se pueden aplicar en el análisis de la colaboración entre miembros de un conjunto musical.



Contents

Abstract	xi
Resum	xiii
Resumen	xv
Contents	xvii
List of Figures	xx
List of Tables	xxiv
List of Algorithms	xxvii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	4
1.3 Scientific Context	5
1.4 Summary of contributions	9
1.5 Thesis Outline	10
2 Scientific Background	13
2.1 Expressive music performance	13
2.2 Ensemble performance	16
2.3 Computational models of expressive performance	18
2.3.1 Empirical models of expressive performance	19
2.3.2 Models based on Machine Learning	20

2.3.3	Models addressing polyphony	22
2.4	Expressive capabilities of bowed instrument	24
2.5	Gesture acquisition in bow instruments	25
2.6	Conclusions	26
3	Data Acquisition	29
3.1	Music material	30
3.2	Recording setup	36
3.3	Audio features	39
3.4	Bowing motion features	40
3.5	Bow Force Estimation	42
3.5.1	Measurement of applied force	43
3.5.2	A simplified physical model of hair ribbon deflection	43
3.5.3	Optimization Procedure	48
3.5.4	Results	50
3.5.5	Bow force estimation in string quartet recordings	51
3.6	Conclusions	55
4	Score & Performance Analysis	57
4.1	Score Analysis	58
4.1.1	Melodic Descriptors	60
4.1.2	Rhythmic Descriptors	62
4.1.3	Harmonic Descriptors	64
4.2	Performance Analysis	66
4.2.1	Note profiles	67
4.2.2	Score-performance alignment via dynamic programming	80
4.2.3	Performed expression parameters	89
4.3	Conclusions	100
5	Analysis of Ensemble Expressive Performance	101
5.1	Introduction	101
5.1.1	Horizontal and Vertical score context	103
5.1.2	Auto-regressive behavior in expressive performance	103
5.1.3	Verifying hypotheses and analyzing EEP models	104
5.2	Statistical analysis	106
5.2.1	Experiment I: timing in <i>solo</i> vs <i>ensemble</i>	106
5.2.2	Experiment II: timing and dynamics with increasing expressiveness	112
5.2.3	Limitations of statistical methods in the analysis of EEP	118

5.3	Machine-learning approach	118
5.3.1	Learning Task	119
5.3.2	Algorithms	120
5.3.3	Feature Sets	121
5.3.4	Feature Selection	122
5.3.5	Building the Datasets	123
5.3.6	Results	124
5.3.7	Discussion	132
5.4	Conclusion	133
6	Conclusion	135
6.1	Introduction	135
6.2	Summary of contributions and key results	136
6.2.1	Multi-modal data acquisition	136
6.2.2	Automatic music transcription	137
6.2.3	Music Representation	137
6.2.4	Ensemble expressive performance modeling	137
6.3	Critique	139
6.4	Outlook	140
	<hr/>	
	Bibliography	143
	A Publications by the author	151

List of Figures

1.1	Creative cycle involving composers and performers. Expressive performance modeling focuses the right part of the circle.	3
1.2	Research areas related to ensemble expressive performance modeling: machine-learning, gesture recognition, signal processing, source separation, automatic music transcription and music representation.	6
2.1	Model of current systems for expressive music performance. . .	19
2.2	Schematic visualization of some bowing parameters (namely tilt, bow displacement and bridge distance) that depend on the relative position of the bow on the violin.	24
3.1	Musical scale recorded with different articulations. The same ascending scale is repeated three times, each time with a shorter notes.	32
3.2	Phrasing exercise, where the musicians are divided in two groups and have to follow each other to form a scale in thirds.	33
3.3	Starting bars of the Beethoven's String Quartet No. 4 (Opus 18) 4th movement.	35
3.4	A picture from one of the recording sessions. Musician's faces have been blurred out for privacy.	37
3.5	Schematics of the system used to synchronize the audio card with the Polhemus EMF MoCap system.	38
3.6	Screen-shot of the VST plug-in user interface for recording audio and EMF motion data.	39

3.7	Schematic representation of relevant positions and orientations tracked in the extraction of bowing motion parameters with EMF sensing technology.	42
3.8	Measured string and hair ribbon segments, computed from their extracted end points, versus their actual configuration. Deformations have been exaggerated in order to illustrate the importance of segment S_P	44
3.9	Single elastic thread attached at the extremities A and B and pushed at a point C.	44
3.10	Decomposition of the internal forces exerted by a single elastic thread.	47
3.11	I. Non touching ribbon. II. Partially touching. III. Fully touching.	47
3.12	The function Map applied to a rectangle. For this example exaggerated parameters used are $r = 0.5$, $b = 1.5$, $\theta = 0.1\text{rad}$	49
3.13	An excerpt of the recorded force signal (<i>continuous gray line</i>) along with its prediction (<i>blue dashed line</i>) after the optimization has been performed.	51
3.14	Interface used to manually calibrate the individual parameters of each string. At the top, we show the trajectory of bowing parameters for a short segment of the recording in the space of bow displacement and pseudo force. Lower in the middle the current force estimation is shown, a different color is assigned to each string. Finally, at the bottom the corresponding sound waveform and bow displacement are shown. On the right one can see a set of parameters that can be adjusted interactively and, on the bottom right, the trajectory on the velocity-estimated force space.	54
4.1	An example demonstrating the Horizontal and Vertical contexts as they are derived from the score.	59
4.2	Prototypical Narmour structures.	60
4.3	The circle of fifths. The pitch classes are organized by intervals of fifths. Notes further away from the tonic degree sound more dissonant, the distance can be counted as number of steps in the circle. As an example, in the tonality of C major, the note D has melodic charge 2 since there are two steps to reach D from C in the circle: $C \rightarrow G \rightarrow D$	62
4.5	Legato mean note trajectories, violin 1.	70
4.6	Martelé mean note trajectories, violin 1.	71
4.7	Staccato mean note trajectories, violin 1.	72

4.8	Detection of onsets using the gestural trajectory in the velocity-force space. The figure shows the three types of onsets defined and their relative thresholds for automatic onset detection. . . .	74
4.9	An example from the database showing bow velocity, force, and gesture onset likelihood along with the sound waveform. The gestures are here sub-sampled at 20Hz. The range for the likelihood is between zero and one since we do not apply here the logarithmic scale.	76
4.10	Sound waveform of a note with marked sustain period considered and the time instants t_k, t_k^+ and t_{k+1}	77
4.11	Some of the derived audio features and the score template for the chroma.	78
4.12	Template chroma vector $\text{ChrTempl}(p)$ for the C major chord, $p = \{C, E, G\}$	79
4.13	Visualization of the matrix of accumulated likelihood M of Algorithm 2. The dots connected by lines show the most likely alignment obtained in the backtracking process of Algorithm 3. . . .	84
4.14	Root mean squared deviation for the performance sampling database a number of iterations has been executed. We compare the results obtained using only audio with the results obtained by also employing gestural data.	88
4.15	Root mean squared deviation obtained by running the aligner on the test dataset. We used the weight parameters obtained from the two optimizations on the performance sampling dataset: both audio only and audio plus gestures. Keeping the same weight parameters we perform several alignments using different sampling rates.	89
4.16	An example of the Sound Level, Pitch and Bow Velocity time series along with the descriptors extracted at the note level. . . .	90
4.17	Collective tempo curve computed on the performance of recording I.2 with three different smoothing factors: at the level of the section, phrase, and bar. The vertical bars mark boundaries of the sections in the rondó structure.	94
4.18	Representation of the onset times of the notes in performance time versus score time. The vertical lines mark phrase boundaries and the red line marks the collective time which was obtained with a locally cubic regression with imposed continuity at phrase boundaries. In the plot, two fermatas are evident at time 20, and at time 28 seconds.	95

4.19	Visualization of asynchrony of each group notes with same score onset position. In red the collective time is displayed as obtained through the regression procedure.	96
4.20	Visualization of participatory asynchrony note in the score. The participatory asynchrony is positive if the arrow points downward and negative if the arrow point upward. Its absolute value is the difference in milliseconds from the collective time which is displayed by the red line.	97
4.21	Estimated vibrato extent for a short sequence of note from violin 1. The red rectangles mark detected vibratos and their height is proportional to the (detected) vibrato extent of each note.	99
5.1	Individual tempo curve of the four instruments for the solo (left) and the ensemble case (right). Vertical grid lines mark the boundaries of the repetitions (full line) and the beat start time (dashed line).	108
5.2	Tempo curve of the joint ensemble performance. Vertical grid lines mark the boundaries of the repetitions (full line) and the beat start time (dashed line).	108
5.3	The plot shows p-values of the observed 0.56 correlation coefficient (in the joint ensemble) for increasing values of σ . The standard deviation σ by affecting the variance of the Gaussian noise introduced to the score onset times for the empirical significance test.	110
5.4	Duration of pauses and semiquavers blocks for violin one: solo vs ensemble.	111
5.5	SQDR of solo vs ensemble.	112
5.6	Narmour class histogram of eighth notes in the datasets I.1-I.3.	115
5.7	Tempo curves for the beginning of the piece in the three expressive intentions.	116
5.9	We represent the note sound level of the four musicians along with the predicted value (using model trees with the FS1 and FS5). The example is taken from Beethoven's String Quartet No. 4 (Opus 18) 4th movement. We show the exaggerated expressive execution.	125
5.10	Resulting correlation coefficient for each expressive intentions, targets and musicians.	128
5.11	Resulting correlation coefficient for different learning tasks and musicians in the "solo vs. ensemble" scenario.	129

List of Tables

3.1	List of recordings analyzed in this dissertation.	31
3.2	Correlation and relative error (both in percentage) between the force prediction and the measured force for each training set and test set coupling.	52
4.1	Horizontal context descriptors.	61
4.2	Vertical context descriptors.	63
4.3	Score context of other voices.	65
4.4	Heuristic formulation of the terms in Equation (4.5)	86
5.1	Categorization of different types of contexts for the prediction of expression parameters. Some examples are included in each category.	105
5.2	Correlation between pitch and joint ensemble tempo curve. . . .	109
5.3	P-values of the 6-way ANOVA analysis of the eighth notes in dataset I.1-I-3. We highlight in bold the p-values lower than $p=0.05$ below which the effect of each factor is considered statistically significant.	117
5.4	Correlation coefficient on each expressive intention. Each cell contains the best correlation coefficient across feature sets for the specific musician and task. The highest value of each column is shown in bold.	127

5.5	Results of feature selection. We report the mean percentage of horizontal features (PHF) across musicians for the <i>solo</i> and the <i>ensemble excerpts</i> dataset after the feature selection. Notice how PHF is higher in the <i>solo</i> than in the <i>ensemble</i> case (except for bow velocity). We report the one-tailed p-values for each percentage.	131
-----	--	-----



List of Algorithms

- 1 Algorithm to extract the log-likelihood of onsets from bow velocity and force. 75
- 2 Forward step algorithm to accumulate the likelihood sequentially. 83
- 3 Backtracking step to find the most likely alignment. 84



Introduction

“My freedom will be so much the greater and more meaningful the more narrowly I limit my field of action and the more I surround myself with obstacles. Whatever diminishes constraint diminishes strength. The more constraints one imposes, the more one frees one’s self of the chains that shackle the spirit”

Igor Stravinsky, *Poetics of Music*

1.1 Motivation

Music performance plays an important role in our culture. Crowds of people fill up concert halls to listen to performers play scores by acclaimed composers. In classical music, performers are required to respect the written score; however, even in this case, each performance sound somehow different, i.e. the interpretations of the same piece performed by two different musicians may differ considerably. The score only codifies a part of the instructions for the performer; most details of the performance are instead implicitly deduced from the context (historic period, composer and performance praxis) or provided by performer’s artistic interpretation of the piece (Gabrielsson, 2003; Palmer, 1997). Many of these implicit instructions are known as expressive deviations.

In the western tradition, music notation arose from the need to specify music on paper so that it could be preserved, stored, transported and performed.

The ancients used to write down the choir's melodic line by specifying pitch sequence without necessarily specifying note duration (Cresti, 1970). The purpose was not to specify every detail of the music piece but just to write down on paper the main musical idea. Improvisation was, indeed, a common practice and on the basis of few notes an entire performance would be built. Later on, in the middle ages, it was together with the rise of polyphony that music notation became an essential tool to handle the complexity of multi-voice compositions.

During the course of history, music notation evolved so to include more and more information about the music. The manuscripts gradually started to look similar to modern scores including meter, key, and the exact pitch and duration of each note. This was achieved also through a series of treatises on music theory rationalizing and revolutionizing the art of performing music (e.g. the Music treatises by Jean-Philippe Rameau's in the 18th century). In this process, also performance practice evolved together with the way music was thought and encoded. A creative cycle, represented in Figure 1.1, gradually emerged in which the composer and the performer assumed different roles but still interacted with each other. In the latest centuries, it is possible to observe an increase in the number of annotations included in the score. Composers seemed to claim control over the performance. It was not enough including in the score the exact pitch and duration of each note, or even how the instrument should be tuned; composers started to specify how the performance should "feel", also providing indications for how such effect could be realized with modulations of tempo and dynamics.

The study of *expressive performance* focuses on the right part of the artistic cycle of Figure 1.1: it studies the transformation process from written score to music performance. In some sense, this process arises from the need to compensate for missing information in symbolic scores. At the same time, performers' renderings of a score are themselves an artistic form of expression as well as an attempt to *interpret* the original intent of the composer.

In the last 30 years, with the advent of computers and MIDI instruments, we have learned a lot about expressive performance though an increasing amount of studies. Technology has allowed us to record very accurately the intensities of individual notes played on different instruments. Nowadays, Music Information Retrieval (MIR) techniques enable ever more sophisticated analysis and transcription of audio signal. Moreover, we have been able to extract distinctive performing patterns employed by musicians, and build systems capable of emulating human expression.

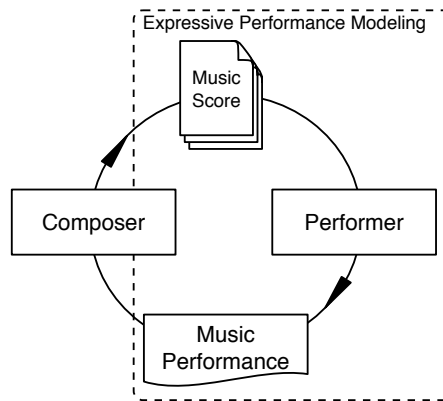


Figure 1.1: Creative cycle involving composers and performers. Expressive performance modeling focuses the right part of the circle.

However, most of the studies on expressive performance have been limited to solo expressive performance and have not addressed the problem of playing expressively in an ensemble. In approaching the new problem, research could be inspired by existing systems that address the problem of playing polyphonic solo performances, i.e., with the presence of harmony and multiple-voices. Some studies of piano expressive performance have already addressed the problem of polyphony (see Chapter 2). Nonetheless, *Ensemble Expressive Performance* (EEP) raises many more complex interactions between different musicians (who play several harmonies and melodic lines concurrently). It is thus unclear whether such discoveries can be extended to ensemble performances.

In the case of an orchestra the conductor takes charge of controlling tempo and dynamics (among other things). All of the music events somehow seem to be triggered by his/her movements. In the case of smaller conductor-less ensembles, such as string quartets, each musician, in principle, is supposed to follow all the others in an equal manner. It is during rehearsals when a shared interpretation of the piece emerges: musicians discuss verbally how each part should be played, how to coordinate attacks at specific points, which notes to stress, which articulations to use, and so on. Whilst this verbal agreements play an important role for structuring the performance, repeated ensemble rehearsals builds up in musicians' minds a solid implicit knowledge of what to expect from their fellows and how much they can

rely on each other at each moment. When playing together, each performer plays his own part in synchrony with his fellow's parts so as to blend in a coherent interpretation. Performing expressively together requires keeping a balance among the individual expressive performance actions to achieve an artistic rendering of the piece.

A series of interesting issues arises when considering the problem of playing expressively in an ensemble (e.g., representation, synchronization, and leadership). Probably for this reason, most of the literature has focused on very specific synchronization music skills, in an attempt to reveal hidden mechanisms underlying ensemble performance. However, to the best of our knowledge, the relations between score and expressive choices in ensemble performance have never been investigated using a data-driven approach.

1.2 Objectives

In this dissertation we aim at studying EEP by building models capable of predicting how and when musicians employ expression parameters in the context of ensemble performance. Our goal is not only to obtain better than chance predictions, but also, more importantly, to obtain new insights about how musicians make expressive choices when playing in an ensemble. We ask question such as: can we detect pattern arising from ensemble dynamics inside the quartet? Can we find statistical evidence of differences between playing solo and playing in ensemble? How do different playing conditions affect the models?

We focus on string quartets since in this type of ensemble there is no conductor; there are enough instruments to carry on a meaningful analysis but not too many as to render the analysis unfeasible. In order to acquire the necessary data we employ audio and motion capture acquisition systems. We investigate how to extract expression parameters specific to bowed instruments by employing multi-modal recording technologies.

We introduce a methodology for computational modeling of EEP from audio and gesture recordings of ensemble music performance. We propose a method to deal with two main issues arising in ensemble music:

- *polyphonic expression*: each musician plays their melody with possibly a different expression respect to the one of the other concurrent voices;
- *inter-dependence among musicians*: each musician takes into account information about concurrent voices to shape their expression.

Firstly, to deal with polyphonic expression, we assign separate models of expression to each musician. This means that, as it happens in human performances, simultaneous voices might have different expression (e.g. the case where a voice containing the leading melody is played loud while the accompaniment voices are played soft). Secondly, to deal with inter-dependence among musicians, we extend the music analysis procedure in order to include information about accompanying voices as input of each model. The models can thus take advantage of information about accompanying voices (including score melodic/rhythmic information and also expression parameters used by other) to predict outcome of expression parameters. Our hypothesis is that we can achieve a better prediction of expression when musicians are considered integrated in the ensemble; and thus by considering also the extended context derived from fellows' voices.

The purpose of training expressive performance models on each musician separately and introducing inter-voice dependencies is two-fold. First, this work can be viewed as a proof-of-concept application of expressive performance modeling in music ensembles. And second, by comparing models obtained on different expression parameters, different musicians and different playing conditions we investigate on several aspects of ensemble performance. Rather than presenting final answers to broad ensemble performance aspects, which might require a considerably larger amount of data, we discuss results on specific aspects, which highlight the potentials of using this method to study ensemble expressive performance. In particular, the difference between playing solo and in ensemble was considered both as a sanity-check test — to see whether the models make sense of the data — and as a mean to understand in which part of the trained models should we look for signs of inter-dependence. Additionally, we look for differences between the considered expression parameters to understand which of them was shaped as a result of inter-dependence among ensemble voices, and which only a product of the individual voices.

1.3 Scientific Context

The core investigation on EEP lies at the intersection of several areas of research (Figure 1.2). Data acquisition is challenging, and not only because of the technical requirement to build an acquisition system. The correct processing and storage of data might indeed require techniques derived from the fields of *gesture recognition*, *signal processing*, *source separation* and *automatic music transcription*. Additionally, we need to consider prob-

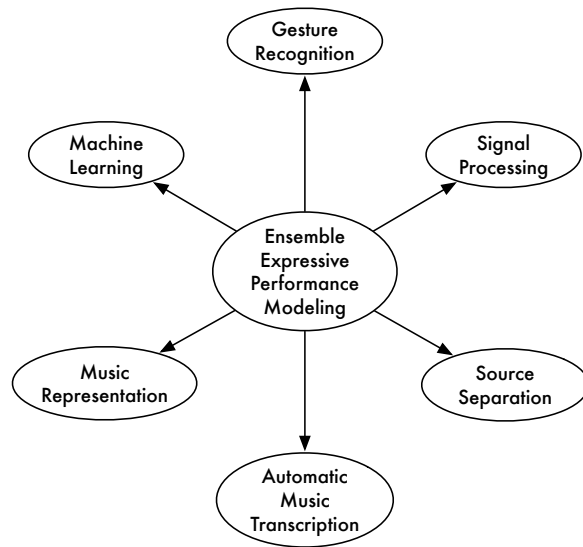


Figure 1.2: Research areas related to ensemble expressive performance modeling: machine-learning, gesture recognition, signal processing, source separation, automatic music transcription and music representation.

lems of *music representation*, in order to store the music in a way that it algorithms can process it and understand it. Lastly, we use techniques of machine-learning to “give sense” to the data, finding recurrent patterns which describe the phenomena underlying EEP.

One of the main problems for the analysis of ensemble performances is the need to acquire the expressive actions of each musician independently. For acoustic instruments (without a MIDI output) this is problematic since musicians have to play together in the same room. This means that any ambient microphone placed in the room to record a single musician will receive leaking sounds from the other musicians. One solution to the problem is to employ techniques of source separation.

Source separation is an active research field whose aim is solving the problem of stream segregation. We, humans, perform stream segregation constantly, for example when we walk in the streets and we are able to distinguish the sound of a car from the sound of a crying baby. Humans are good at this task; however, it is still a very challenging task for computers. A good source separation system would enable research in EEP to be carried out

on commercial audio recordings in which only one (possibly stereo) audio track of the whole performance is available.

Now assume that the problem of source separation is solved and we have the sound of each musician in a separate audio track. There is still the problem of transcribing each audio track to a sequence of note events in order to understand the underlying music content. This task, called *Automatic Music Transcription*, is also one of the most challenging tasks of Music Information Retrieval (MIR). What makes this task difficult is not only the problem of recognizing note events from sound, but also the quantization required to make a representation meaningful. The analysis of sound with computational tools often reveals many details about the transitions between notes, which do not necessarily produce note event splitting. This happens because sound is a continuous phenomenon, but through the process of listening, we quantize it in time and in frequency (Bregman, 1990). Some trajectories of pitch within notes are perceived as embellishments and do not produce the splitting into smaller note units. When we listen to music we spontaneously perceive note events and thus compress complex sounds into compact note sequences. It is at this level of representation that we generate expectations on upcoming note events and we compare melodies one with each other.

In this dissertation we overcome the problems of source separation by limiting the analysis only to custom recorded performances, aided by the use of specifically designed multi-modal acquisition technologies. Additionally we do not address the problem of automatic music transcription since we have access to the score of each of the recorded material. The problem of music transcription is thereby restricted to the simpler problem of score-performance alignment. In other words, we just need to find the time of onset and offset of each note on each analyzed recording.

Music Representation is an important area of research studying how music is represented symbolically. Music Representation plays an important role in EEP since it provides a very convenient shorthand representation to describe the musical content with few relevant descriptors. The music content becomes obvious in its symbolic representation, by discarding all of the unnecessary details of the sound. Note events are ordered in time and have pitch and duration attributes. The events might also be organized in a hierarchy consisting of multiple time spans, from motives to phrases, from phrases to sections. In multi voice scores, notes are organized into parts (i.e. first violinist, second violinist, viola and cello) with an implicit relation of precedence and simultaneity among note events. In this disser-

tation we propose a new way to represent the local musical context in order to characterize each note in a multi-part score. We use such representation to derive a set of independent variables, which are then processed by machine-learning algorithms.

Machine Learning is a very important research area within the field of Artificial Intelligence. It is also integrated in Statistics since its main goal is to study independent variables and dependent variables, relating one with the others by means statistical models. In this dissertation we use machine-learning algorithms as a tool for learning models of ensemble expressive performance, relating score contextual variables with performance variables. The *performance variables* are the dependent variables of the models of EEP i.e., what such models are supposed to predict. Those can include any aspect of the performance that is not already included explicitly in the score e.g., intensity of each note, timing deviations from score, intonation deviations, and articulation.

We use machine-learning algorithms to build models of EEP. The algorithms make use of training data from real performances to output the models. This means that we need to provide the machine-learning algorithms with example values of each dependent variable from real performances. In order to measure the expression parameters, the earliest studies used analog machines that would produce continuous measurement on paper or photographic film. Seashore (1938) employed a piano camera system to record gestural data from hammer and foot-pedal movements. For voice or violin sounds he used the Henrici's Harmonic Analyzer, a *Signal Processing* analog machine that would track the pitch of a recorded waveform. In the last decades signal processing has been extended to the digital domain providing methods to execute ever more sophisticated analyses of audio in the digital domain. This research area has grown significantly in the digital era accumulating a consistent set of algorithms to extract features from audio and gesture signals. Needless to say, signal processing nowadays enables a much faster, cheaper, widespread and automatized processing of audio signals.

In some cases the analysis of music performance aims at detecting certain patterns in time and space. This task consists of recognizing a continuous signal by assigning it to one of several possible categories. *Gesture Recognition* aims at analyzing human gestures in order to extract meaningful information and patterns. This research area takes advantage of several pattern recognition algorithms. The study of gestural communication among musicians is beyond the scope of this dissertation. Nevertheless, we adopt some

basic gesture analysis tools to detect sound producing gestures.

1.4 Summary of contributions

This thesis extends both research in expressive performance and research in ensemble performance proposing a methodology where both fields can benefit from each other. Due to the multidisciplinary character of the topics addressed we can divide the contributions into three categories, each dependent on the others: multi-modal data acquisition, automatic music transcription and expressive performance modeling. The main contributions of the thesis can be therefore summarized as follows:

Multi-modal data acquisition

- Proposal of a recording setup for acquiring multi-modal performance (audio and gesture) in string quartets (Chapter 3).
- A public database of string quartets multi-modal performance recordings.
- A mathematical model of bow deflection based on physics of hair ribbon.
- Parameter optimization of the bow deflection model based on recordings of sensor data.
- A novel methodology for estimating bow force from motion capture data avoiding thus the need to introduce additional intrusive force sensors.

Automatic music transcription

- An heuristic-based extension of previous methods for score-performance alignment based on acquired gestural data and estimated force.
- Parameter optimization for score-performance alignment based on real performances.
- Quantitative evaluation of gesture-based score-performance alignment method with respect to a baseline approach (i.e., audio only).

Music analysis

- Methodology for the extraction of contextual descriptors in multi-voice music scores.

Expressive performance modeling

- Statistical analysis of timing, dynamics, vibrato and inter-musicians synchronization.
- Comparison of machine-learning EEP models in solo and ensemble context, based on feature predictive power and feature selection.
- Comparison of machine-learning EEP models in mechanical, normal and exaggerated expressive intentions based on predictive power.

1.5 Thesis Outline

The organization of this Ph.D. dissertation is the following. In Chapter 2 we introduce the topics of expressive performance and ensemble performance reviewing the related existing research. We thereby focus on computational models of expressive performance with a special attention to the points of contact between polyphonic expressive performance and ensemble performance. Additionally, we give account for the literature on the expressive capabilities of bowed instruments and the acquisition of bowing gestures.

In Chapter 3 we present the corpora that was acquired and analyzed within the scope of this Ph.D.. We describe the recording set-up procedure employed in order to capture multi-modal data of each performance. We therefore present the data processing procedure employed to extract meaningful time-series from the raw acquired data. These data include bowing gestures and low-level audio features. We follow the work of Maestre (2009) to extract bowing performance descriptors; we extend his work by presenting a novel methodology to estimate the bow pressing force (based on Marchini et al. (2011)).

In Chapter 4 we extensively describe the attributes extracted from the score and introduce inter-voice contextual descriptors. Contextual descriptors are thereby categorized depending on their music significance (melodic, harmonic, rhythmic. . .) and/or depending on the voices that they span (horizontal and vertical contextual descriptors). Additionally, we describe and

evaluate the multi-modal (audio and gestures) score-performance alignment procedure that we employed for extracting note on-set/off-set times in the performance data. We then describe the procedure to extract note-by-note expression parameters on which we focused our analysis.

In Chapter 5 we analyzed the following aspects in the acquired corpora: the relations between score contextual descriptors and expression parameters; the differences between solo and ensemble performances; and the difference between various degrees of expressiveness (mechanical, normal and exaggerated performances). We start with a statistical analysis of timing on a short exercise performance, which highlights several fundamental differences between solo and ensemble. We then analyze statistically a set of pieces to find the difference between the degrees of expressiveness. We demonstrate the statistical validity of vertical score contextual features. We then perform a further analysis based on machine-learning modeling in which we incorporate a multitude of contextual descriptors and discuss on the predictive power of the derived models. In such analysis, rather than focusing on one contextual descriptor at a time, we compare the predictive capabilities of a multitude of models.

Finally, Chapter 6 summarizes the contributions of the thesis and draws directions for future research and applications of this work.



Scientific Background

In this chapter we introduce the relevant scientific background for this dissertation. Expressive music performance is introduced first, which is a central topic in this thesis (Section 2.1). We then give a general summary of the studies on ensemble performance (Section 2.2). We then specifically address the literature on computational models of expressive performance (Section 2.3). Since in this dissertation we address expression in string quartets, we summarize literature on expressive capabilities of bowed instruments (Section 2.4) and acquisition of bowing gestures (Section 2.5).

2.1 Expressive music performance

The path that leads a musician from an initial reading of the score to a well-rehearsed stage performance involves many different interacting phenomena (social, psychological, perceptual, physiological, musicological and didactic to cite a few). There are several different performance-related tasks like sight-reading, performing well-learned music from memory or from notation, improvising, playing by ear. Depending on the task, the musician's attention might be concentrated on one aspect rather than another. As described in Palmer's (1997) review on expressive performance, the performance process can be summarized by the following sentence:

During a performance, musical structures and units are retrieved from memory according to the performer's conceptual interpretation, and are then prepared for production and transformed into appropriate movements. (Palmer, 1997)

Whereas there is a wide interest from a psychological standpoint to understand all those underlying cognitive and motor mechanisms, research on expression in music performances (Gabrielsson, 1999, 2003) investigates the manipulation of sound properties such as pitch, timing and amplitude in an attempt to understand and recreate expression in performances. Although we tend to highlight the differences among performers, there are indeed many commonalities across performances arising from similar way to process the music content: note grouping, unit identification, thematic abstraction, elaboration, and hierarchical nesting.

Many studies have focused on *expressive performance actions* such as timing, dynamics, intonation and vibrato. Swedish psychologist Carl Emile Seashore carried out the first studies during the 1930s in Iowa. Some of those early measurements can be found in his book “Psychology of Music” (Seashore, 1938), which also includes measurements of violin performances. Seashore divided the psychology of music in three large fields dealing with the musician, the music, and the listener, respectively. By including the music as a field per se, he highlighted the idea that psychology of music needs to include a good understanding of the musical medium in order to understand how this affects the performer and the listener. He stated as fundamental proposition the following sentence, which can be seen as the germ of future research in expressive music performance.

The artistic expression of feeling in music consists in esthetic deviation from the regular—from pure tone, true pitch, even dynamics, metronomic time, rigid rhythms, etc. (Seashore, 1938)

Following the Seashore’s early directions, studies of expressive performance have therefore analyzed such deviations. The strategies and changes which musicians apply to the music, but that are not marked on the score, are referred to as *expressive performance actions*: expressive dynamics, expressive timing, expressive articulation, expressive vibrato and more. Music is intrinsically a multi-dimensional phenomenon and thus several expressive performance actions might co-exist and interact during the performance. In this dissertation, the term *expression parameter* refers to a continuous value specifying of how much a performed note deviates from the score nominal value. For each note, we might consider many different kinds of expression parameters: note lengthening, intonation deviation, vibrato rate, vibrato extent, peak intensity, attack slope, sustain time, bow velocity etc. . . .

Early researches have shown that musicians instructed to play in a strictly mechanical way would still exhibit timing variations of the same type as in expressive performance but on a reduced scale (Bengtsson and Gabrielsson, 1983; Palmer, 1989). This sparked a debate in the research community regarding to which extent musicians are able to really control independently each of the expression parameters. There are probably many different phenomena causing the observed deviations (Gabrielsson, 2003). The discussion in the literature led to the emergence of three main hypotheses. The hypotheses are (a) the use of performance actions to highlight the musical structure to the listeners, (b) a compensation for perceptual bias for which the perception of one expression parameter might be affected by another independent (e.g. pitch and timing), (c) the existence of certain preferred, biomechanical motor patterns and constraints which are also dependent on the instrument. No research has been able to exclude any of the three hypotheses suggesting, on the contrary, that all of the three factors might contribute with different degrees of importance to the emergence of the deviations. The performing task and the amount of expertise of the musician might in fact affect the relative contribution of each factor. This debate was especially directed towards explaining timing variations, but the previous hypotheses could be applied to other expressive performance actions alike.

The relation between music structure and expression (the hypothesis (a) above mentioned) is in the literature one of the most supported factors for introducing expressive deviations (Palmer, 1997; Gabrielsson, 2003). Analysis of music structure is one of classical objectives of systematic musicology. Such an analysis is classically based on harmony, motives, melodic transformations and repetitions. Jackendoff and Lerdahl (1983) have introduced one of the most influential theories in their book *A generative theory of tonal music* (GTTM) where they described the procedure of understanding the music through a set of hierarchical rules. There have been computational approaches attempting to automatize this task although with big problems of combinatorial explosion, which have favored approaches based on cognition and heuristics. The approach by Narmour (1992) employs gestalt expectation to derive a classification of pitch sequences.

In the late 1960s, expressive performance research revived with studies on systematic rhythmic deviations (Bengtsson and Gabrielsson, 1977). Lastly, it was during the 1980s, with the advent of MIDI and personal computers, that research in expressive music performance gradually started drawing the attention of larger number of researchers. The release of the first Disklavier MIDI piano by Yamaha in 1987 made the piano the test bed instrument of

most of the research in expressive music performance.

Repp (1992) carried out extensive measurements of timing and dynamics on piano performances of Schumann's *Träumerei*. He compared students' and expert pianists' timing patterns showing that students were much more homogeneous among themselves than the experts. The use of dynamics in *Träumerei* was also showed to be similar and consistent over repeated performances (Repp, 1996). Palmer (1996a,b) also studied pedal timing and arpeggio performance by a concert pianist, showing consistent pedal patterns before note onsets to avoid dissonances with successive note-events.

In this dissertation we are especially interested in computational models of expressive music performance. For this reason, we will devote Section 2.3 to computational approaches in expressive music performance. But first, since we will pose a special focus on the models addressing ensemble performance issues, we preliminary introduce the literature on ensemble performance in the following section.

2.2 Ensemble performance

Research in ensemble performance has been carried out from different viewpoints. Probably one of the most general approaches is provided by the literature of joint action. Clark (1994) described the problems arising in day-life conversations (vocalization, attention, misunderstandings etc.) and the main dynamics individuals employ for preventing and correcting them. One of the main hypotheses of this research field is that when the problems are shared among a group of individual, their solution requires joint action. In other words, each member of the group needs to take decisions based on their fellow's actions and, at the same time, pro-actively provide them with instructions to solve encountered problems. This idea is applicable in many contexts such as sports, self-management teams, dance performers, theaters actors, and music ensembles.

One of the main aspects considered in the studies of joint action is the social hierarchy underlying certain coordination dynamics (Pacherie, 2012). Whilst this idea can be fruitfully applied to certain music ensemble, it is less clear how this can be applied on small conductor-less ensembles such as string quartets. In the case of orchestras we can easily imagine the conductor on top of the hierarchy followed by the main section of violinists, and then the rest of the instruments. In this paradigm the individuals at a

higher node of the social hierarchy need to control their lower fellows and provide them with directions to solve encountered problems.

In the case of musical ensembles, the musicians are required to direct their attention continuously to the performance to keep in-sync with their fellows. Theoretical research (Keller, 2008) has pointed out three key cognitive processes present in ensemble musicians: auditory-imagery, where the musicians have their own anticipation of their own sound as well as the overall sound of the ensemble, prioritized integrative attention, where musicians divide their attention between their own actions and the actions of others, and adaptive timing, where each musician adjusts the performance to maintain temporal synchrony. This perspective makes it possible to identify the following important points: first, that each musician incorporates the ensemble score as well as the performance of the rest of the ensemble into an anticipation of the produced result. Second, the musician defines the saliency of each performed note with respect to the ensemble as a whole, shaping their performance so that it integrates both with the ensemble's actions as well as personal expressive choices. Lastly, the above choices must be made while maintaining ensemble synchrony at the same time.

Most of the literature has focused on studying timing asynchrony among performers to understand the synchronization mechanism. Such studies have either reduced the task to tapping or have neglected the role of score context. Repp (2005) has extensively studied the synchronization on the task of tapping together developing a theory of synchronization based on simple phase and frequency correction mechanisms.

More recently, Wing et al. (2014) applied first-order linear phase correction model for studying synchronization in string quartets. The study shows how each musician adjust tempo with respect to the fellows. This revealed two different contrasting strategies employed by two string quartets going from a first-violin-led autocracy versus democracy. Moore and Chen (2010) studied two string quartet members synchronizing bow strokes at high speeds. He showed how an alternating renewal process could model such specific musical skills. Goebel and Palmer (2009) studied the synchronization among musicians also taking into account musician's role and the effect of auditory and visual feedback.

The literature did not address systematically the problem of playing expressively in ensemble. To the best of our knowledge the only (pioneering) work considering this issue was the one by Sundberg et al. (1989). In this dissertation we have taken inspiration from that paper, especially for the

derivation of ensemble descriptors (see Section 4.1). However, Sundberg's approach was carried out from an analysis-by-synthesis perspective (we explain it in the next section) whereas here we use a data-driven approach.

2.3 Computational models of expressive performance

Expressive music performance was also studied from a computational perspective. Building computational models of expressive music performance means to develop algorithmic procedures capable of predicting or emulating expressive music performance. Most current computational models of expressive music performance can be described using the diagram of Figure 2.1 (see Kirke and Miranda (2013) for a detailed overview). The core of any performance system is the *Performance Knowledge* module, which encodes the ability to generate implicitly or explicitly an expressive performance. This module can be in any mathematical form taking as input the *Music Analysis* module and giving as output a representation of the generated performance, including expressive performance actions.

There are several reasons for building computational models of expressive music performance. The most obvious reason is for synthesizing expressive performance from music data files. There are many files available on the Internet (MIDI and MusicXML being the most popular) containing non expressive scores, which are used by musicians for storing and sharing pieces as well as for having a preview listening of the rendering. The rendition with common software is robotic and rather unattractive and could be easily improved using computational models of expressive music performance.

Similar tools could aid composers by improving realistic playback on music typesetting and composition tools. Additionally, computer generated music systems could exploit expressive performance models for gaining naturalness of synthesized rendition. Another motivation for computational models of expressive music performance is automatizing the task of music accompaniment. A system for automatic music accompaniment would have a range of applications from music didactics to experimental concert scenarios.

RenCon¹ (contest for performance rendering systems) is a formal competition happening each year since 2002 where a jury of experts evaluates computational models of expressive performance. The evaluation is based

¹<http://www.renconmusic.org>

on rendering of a compulsory score chosen by the jury. There is also an open section where researchers can propose their own pieces. The jury gives scores for “humanness”, “expressiveness” and “preference” to each generation and the consequent final ranking dictates the winner of the competition.

We left out the most straightforward motivation for building computational models of expressive performance: investigating human expressive performance. In fact, studying the *Performance Knowledge* module of Figure 2.1 can provide great insights on the nature of expressive music performance. This is also the motivation underlying this thesis dissertation. We indeed develop models of ensemble expressive performance in an attempt to understand simultaneously the problem of playing expressively and the problem of playing in ensemble.

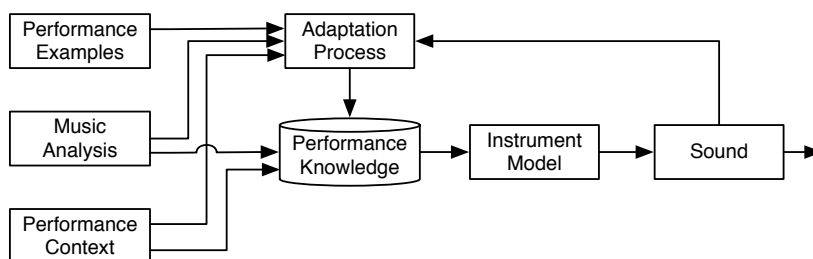


Figure 2.1: Model of current systems for expressive music performance.

We present here some of the models of expressive performance in the literature. We distinguish between empirical models, where the adaptation process is manually obtained through expert observations on re-synthesis, and models based on machine-learning, where the adaptation process is handled by an algorithm fitting live music performance examples.

2.3.1 Empirical models of expressive performance

A long-term research on expressive performance has been carried out at Royal Institute of Technology (KTH) in Stockholm. Director Musices (DM) has been an ongoing project since 1982 devoted to the development and study of rules of expressive performance. The approach of DM distinguished itself from the previous studies of expressive performance based on measuring expressive performance. DM has been based on an analysis-by-synthesis approach since the beginning (Sundberg et al., 1983; Friberg, 1995). This

approach is based on intuitively defining rules of expressive performance and then verifying their correctness by listening to synthesized performances of several scores. The whole system has accumulated a set of around 30 rules defined by simple mathematical equations. Such rules can be fine-tuned by tweaking their parameters to produce performances with different “characters”, which can also be more or less plausible. Johnson (1992) proposed another similar expert system where the rules are devised from interviews with expert musicians.

The approach based on analysis-by-synthesis has an additional application, not present in the machine-learning based approaches. Analysis-by-synthesis, by manually adjusting the rules parameters, can lead to producing completely new pleasing expressive performances that would even be physically impossible to reproduce by humans due to motor constraints. This allows for a series of applications to performance creativity, allowing the user to create novel performances with low effort. This is less likely to happen in the machine-learning approaches since the algorithms generally learn motor constraints together with expression contained in the provided recording examples.

One of DM’s rules is the *phrase arc rule* which results in a deceleration of tempo near phrase boundaries. Applying these rules might require some amount of manual work for annotating the start and the end of each phrase in the score, since DM provides no music analysis algorithm to perform this task. Todd (1992) proposed structure-level models of timing and dynamics based on hierarchy of phrases, thus extending the phrasing idea even further. Todd Model (as it is collectively called) contains a greater expressive representation since the transformation that is finally applied to each note is the results of multiple layers of transformations acting on each level of the hierarchy. The system was based on GTTM rules to derive the multi-level hierarchy and proposed a simple rule the-faster-the-louder linking together timing and dynamics performance actions. The model, though interesting, lacked of a thorough evaluation on real performances and lacked of flexibility since the relation between the hierarchy and the expressive transformation was complex and this made it difficult to “hack” it for applications in different contexts.

2.3.2 Models based on Machine Learning

Machine learning (ML) is a branch of artificial intelligence concerning the construction of systems that can learn from data. Systems of Machine

Learning differ significantly from each other but they all attempt to solve the following problem (at least the ones for which we are concerned in this thesis). Suppose to have a set of input variables x_1, \dots, x_n and having to predict the outcome of an output variable y . The problem is to find a function f such that $f(x_1, \dots, x_n)$ is equal or, at least, as close as possible to y . In other words, the function f encodes the dependency relationships between the input variables x_1, \dots, x_n and the output variable y . An explicative toy example is weather forecast, which can be obtained using the input variables x_1 =temperature, x_2 =pressure, x_3 =wind speed direction. Now assume that we want to predict y =true/false, the answer to the question “it going to rain tomorrow?”. To find such f one could apply a set of complex models of weather forecasting, which involves a lot of knowledge about how weather works. Instead, the approach of machine-learning system is to automatically learn the function f from a series of m training examples $X = (x_i^{(k)})_{m \times n}$ with known output $Y = (y^{(k)})$; $i = 1, \dots, n$ and $k = 1, \dots, m$. Depending on the machine-learning system the function f might be encoded in different ways (a linear function, a set of rules or more complex function). In some cases the knowledge encoded by f is explicit and thus can be understood from a human; on other cases, the function f can be very good at predicting the variable y although the knowledge is implicit and thus difficult to understand.

Machine-learning models of expressive music performance are the ones that make use of machine-learning systems to implement the adaptation process of Figure 2.1. Several machine-learning techniques have been used as an approach to investigate expressive performance. Widmer and Goebel (2004) and others have trained expressive models from real piano-performance data. Widmer’s work is supported by the largest corpora of recordings ever studied in research of expressive performance (Widmer, 2002). The corpora consists of hundreds of thousands of notes played on a Bösendorfer grand piano allowing for high precision recording of every key and pedal movement. The derived system was used both for deriving explicit knowledge about simple and general “performance rules” employed by musicians and for implementing systems that can generate expressive renditions of input score files.

Some works have considered other instruments and genres (e.g., De Man-
taras and Arcos (2002); Ramirez and Hazan (2006)), therefore addressing
additional expressive manipulations that are absent in piano performances
(e.g., vibrato and glissando). Bresin (1998) developed a hybrid system
combining artificial neural networks with KTH rules obtained through an

analysis-by-synthesis procedure. Other works have studied expressive manipulations as a mean to carry emotions. Kendall and Carterette (1990); Gabrielsson and Juslin (1996) have approached the concept of expressive intention and emotion in expressive performance. These works led to the proposition of the GERM model (Juslin et al., 2002), which enables the generation of expressive performance with a continuous set of emotions in the valence-arousal space.

Computer systems of expressive music performance generally employ music analysis procedures in which score is translated into a sequence of feature vectors, each describing one note (Kirke and Miranda, 2013). Each feature vector describes the nominal properties of a note as well as the score context in which the note appears (e.g. melodic/rhythmic contours and harmony). Computational models of expressive performance take feature vectors as input to produce predictions of expression parameters (e.g. intensity and timing transformations). In data-driven systems, the computational models are built automatically by feeding machine-learning algorithms with training examples. Training examples are collected by extracting note expression parameters from recordings of human music performances along with feature vectors from the score. Such computational models can therefore be used to reproduce expressive performance of new untrained music scores by applying the predicted expression parameters to (otherwise robotic) computer score renderings. Computational models are also used as a mean to investigate how humans play music expressively. This is possible either when the knowledge of the trained models is explicit and can thus be easily translated to human readable rules revealing recurrent performance patterns (Widmer, 2002), or when models trained on different musicians and/or different performing conditions are compared (Poli, 2004).

All the models of expressive music performance based on Machine Learning have been implementing some music analysis procedure to derive the input variables and attempt to predict the expression parameters. However, most systems apply a music analysis based on one single part and it is not clear how this can be extended to the case of ensemble score where more parts are playing concurrently. At the same time, there are some interesting questions that arise when analyzing ensemble performance and could be addressed by building EEP models.

2.3.3 Models addressing polyphony

Vernon (1937) showed for the first time that piano players introduce sys-

tematic delays of finger pressure between main melodic line and accompaniment. In compositions with polyphonic voices such as Bach's canons, one of the required skills for the musician is exactly to attempt to recreate the impression of ensemble orchestration by giving a different musical character to each of the simultaneous voices. Palmer (1996a,b) studied the so-called melody lead in piano performance: melody notes in chords come somewhat earlier (20-50 ms) than other notes in the chord; this effect was more pronounced for expert performers than for student performers.

Most of the systems of music expressive performance do not address explicitly the problem of polyphony, or are limited to monophonic/homophonic scores. To the best of our knowledge, little work has been devoted to address the problem of polyphony in an exhaustive way. This problem is present in instruments capable of producing harmonies (such as piano or guitar) where different voices can be reproduced simultaneously, although they are all triggered by the same musician. The challenges are similar to the ones we have to address in the case of EEP.

Hashida et al. (2007) considered the problem of integrating individual rule of expression in a multi-voice score. The system renders expression on a polyphonic score by applying timing and dynamics transformations to each of the individual voices. As a consequence of applying note time stretching to each individual voice the voices accumulate time lags among them. This problem is solved in a post-processing step identifying "synchronization points" and reinforcing them. Hashida et al.'s (2007) system involves an interesting approach for addressing the issue of rendering ensemble scores. However the basic assumption of the system is that each musician is focused on *his own* voice, trying to make it expressive, and then on top of it a between-voices synchronization procedure is added in a post-processing step. In this way the system does a good job in rendering different musical characters on each job, but does not consider the possibility that the expression of each of each voice may also be dependent on other voices.

Polyhymnia is an automatic piano system developed by Kim et al. (2011) that won the first place in the autonomous section of RenCon 2010. This system takes into account *polyphonic expression* that can occur in a polyphonic instrument such as the piano. This means that two concurrent melodies played on the left and right hand can be realized with different expression parameters (e.g. one melody played staccato and the other legato). The system employs a probabilistic approach in modeling dependencies among voices, although it is not clear to us how to scale this ap-

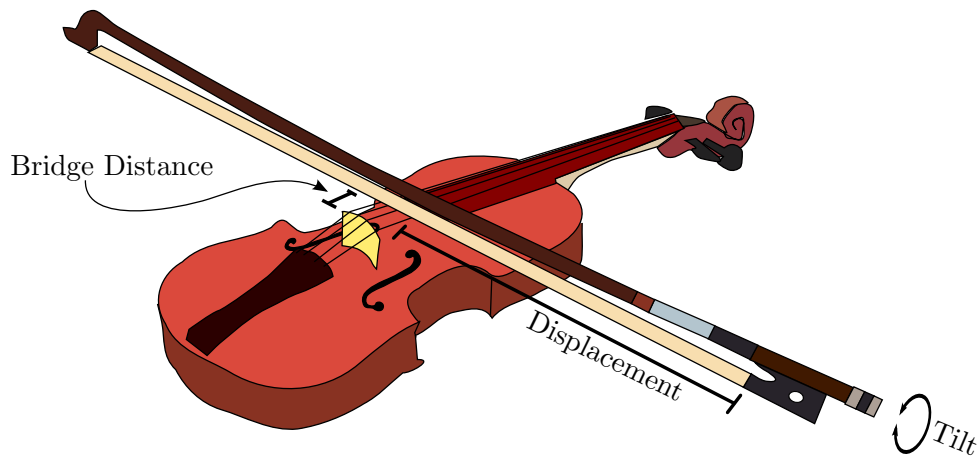


Figure 2.2: Schematic visualization of some bowing parameters (namely tilt, bow displacement and bridge distance) that depend on the relative position of the bow on the violin.

proach to more than two voices. The good performance of the system at the RenCon context confirms that considering dependencies among voices is a missing part of current systems of expressive performance which should be explored further not only for ensemble but also for polyphonic instruments.

2.4 Expressive capabilities of bowed instrument

Violin is regarded as among the most complex musical instruments, making different control parameters available for the performer to freely shape rich timbre characteristics of produced sound. Seashore (1938) included several pioneering measurements from the audio recordings of violin performance showing a degree of expressiveness comparable with human voice. Not only the performer is able to give a particular character to each note, but also he/she can continuously move in the control parameter space realizing smooth transitions from one note to the next (Schoonderwaldt, 2009a; Woodhouse and Galluzzo, 2004).

Both performer's hands are important in the production of the sound. The left hand controls the length of the string played thus modulating the pitch while the right hand excites the strings through *bowing gestures* (see Figure 2.2). The bowing gestures can be described as an evolution in time of a set of bowing control parameters which have been described in the litera-

ture: bow velocity, bow-bridge distance, bow force, bow position, bow tilt, and bow inclination (Maestre, 2009; Pérez, 2009; Schoonderwaldt, 2009a). With the left hand, string players can modulate continuously pitch curve since such instruments are fretless. This not only allows for adjusting of the tuning on-the-fly, but also to produce voice-like vibratos (Prame, 1997; Seashore, 1938).

2.5 Gesture acquisition in bow instruments

The expressive capabilities of string instruments derive from the physical properties of the instrument and how this is played. A significant research effort has been devoted to methods for accurate acquisition of control parameter signals (Schoonderwaldt et al., 2006; Schoonderwaldt, 2009b; Askenfelt, 1986; Young, 2003; Maestre et al., 2007; Demoucron et al., 2009; Gnaus et al., 2009; Paradiso and Gershenfeld, 1997). In those studies, various acquisition systems based on placing sensors on the bow and/or the violin have been developed demonstrating how it is possible to acquire right-hand control parameters with motion capture techniques. The left hand yet remains difficult to be tracked without great intrusiveness. (Kinoshita and Obata, 2009) showed that tracking force applied by the left hand on the neck is highly informative about the score, the speed and dynamics of the performance.

Within the different bowing control parameters, only the bow transversal velocity could be considered as of comparable importance as the bow pressing force exerted by the player on the string Cremer (1984). The measurement of bow pressing force not only has received special attention because of its key role in timbre control, but also because of a number of measurement-specific issues that appear as harder to overcome, as it is accuracy, robustness, or intrusiveness. An early attempt to pursue the measurement of bow pressing force from real violin practice dates back to 1986. Askenfelt (1986) used wired strain gages at the frog and the tip in order to infer the bow pressing force applied on the string. Although useful for the instrument-modeling purposes of the authors, significant intrusiveness would make difficult the use of such a system in a real performance scenario.

Intrusiveness improved significantly by a first wireless acquisition system proposed by Paradiso and Gershenfeld (1997), who attached a resistive strip to the bow that was driven by an antenna mounted behind the bridge of the cello. A measurement relative to the bow pressing force was carried out by

using a force-sensitive resistor below the forefinger. Despite the reduction on intrusiveness, the obtained measure resulted rather unrelated to the actual force exerted on the string, as it happened to Young's approach (Young, 2003), who measured downward and lateral bow pressure with foil strain gages permanently mounted around the midpoint of the bow stick.

The first effort to relate the strain of the bow hair as a measure of force was carried out by Rasamimanana (2003), although the technique reached its first state of maturity (in terms of accuracy) with the technique introduced by Demoucron and Caussé (2007) and more recently reused and improved by Guaus et al. (2009): the deflection of the hair ribbon is measured at the frog (and also at the tip in one of the earlier versions) by using a strain gage attached to a plate laying against the hair ribbon which bends when the string is pressed. This technique, while providing surprisingly good estimations of bow force, suffers from remarkable intrusiveness and reduced robustness, making difficult its prolonged use in stage or performance contexts.

In Chapter 3 we provide a novel mathematical model of bow deflection that we used to estimate force from motion capture data without the need for introducing further intrusive sensors. In the scope of this dissertation, due to the complex recording setup, we favored easy setup and minor intrusiveness to high accuracy in bowing force estimation. The derived estimated bowing force proved useful for the alignment of the performance to the score.

2.6 Conclusions

Expressive music performance is a highly multidisciplinary research field that has been approached in the past from cognitive psychology, sensory-motor psychology, musicology, and artificial intelligence standpoint of views. The multidisciplinary nature of this field can be traced back to Seashore's (1938) ternary sub-division into the study of the performer, the music and the listener. Although there is still a debate over the nature of expressiveness in music, this field has reached the sufficient maturity to produce systems that emulate human expression.

The study of ensemble music performance has been approached in more recent years, thanks to ever-more-ubiquitous sensor technologies. The studies mostly focused on the performers and their skills, without considering the musical dimension and the listeners. We believe that ensemble performance

studies could be highly enriched by integrating methods coming from the expressive music performance literature.

We therefore focused on computational models of expressive performance and highlighted the points of contact between polyphonic performance and expressive performance from which our research could draw some inspiration.

Lastly, we introduced the literature on the expressive capabilities of bowed instruments and the acquisition of bowing gestures that is the core technological tool through which we carried out the work in this Ph.D. dissertation.



Data Acquisition

In this chapter we introduce the music material and the methodology for data acquisition of *Ensemble Expressive Performance*. The data presented in this dissertation is part of a larger collection of recordings which we recorded during two research stays at McGill University in a collaboration between the Music Technology Group¹ (Universitat Pompeu Fabra, Barcelona), the Centre for Interdisciplinary Research in Music Media and Technology², the Input Devices and Music Interaction Laboratory³ (McGill University, Montreal), and the International Laboratory for Brain, Music, and Sound Research⁴ (Université de Montréal and McGill University). All the data analyzed in this dissertation were acquired from performances by professional string quartets, and consist of both audio and gestural data. The members of the quartet had been performing together for more than a year, had more than 10 years of music experience, and had an average age of 30 years.

We start presenting the music material (Section 3.1) and then the procedure employed for data acquisition. In Section 3.2 we describe the recording setup, which included acquisition of audio and motion data. We then overview the procedures for extracting bowing (Section 3.4) and audio descriptors (Section 3.3) that are relevant in this dissertation. Finally, in Section 3.5 we focus on the specific problem of *bow force estimation*, which represents a contribution of this work to the field of bowing gesture ac-

¹<http://www.mtg.upf.edu>

²<http://www.cirmmt.org>

³<http://www.idmil.org>

⁴<http://www.brams.org>

quisition that we will use for improving note-by-note segmentation of the recordings (see Sections 4.2.2 and 4.2.2). We describe a procedure that we devised for estimating the force applied by the bow on the string without the need of additional intrusive sensors. We evaluate such procedure on a dedicated set of force measurement recordings that we present later in this chapter (see Section 3.5.4).

3.1 Music material

The music material used in this dissertation consists of solo scales played with different articulations, short exercises for string quartets, excerpts from the repertoire of the string quartet, and an entire movement from a Beethoven's concerto. We had access to the digital version in MusicXML file format of all the scores we recorded.

Both the scales and the exercises were selected in a preliminary workshop with a professional string quartet player (not taking part in the recordings). Those excerpts are easy enough to be played after a short rehearsal. In the case of the pieces, we asked the performers to give us their current repertoire as well as propose which pieces they felt most comfortable with when performing in a concert.

We recorded the musicians in an *ensemble* condition, where they would interpret together all the parts an ensemble score. For some recordings, we also recorded the musicians in a *solo* condition. In the solo condition each musician would play his own part alone; in the case of the exercises, without even knowing the parts of the others (see . In the ensemble condition they would play together; in this case we also asked the musicians to play with non-normal expressive intentions such as *mechanical* and *exaggerated*.

We mixed the different experimental conditions, although timing constraints in the recording schedule limited the possibility of realizing all the condition combinations for all music material. For example, recording the solo condition would lengthen the recording time since the score had to be recorded four times (one of each musician playing solo).

We provided labels to refer to each of the recordings. The scale exercises were all recorded in solo conditions and were realized by the musicians with legato (S.1), martelé (S.2), détaché (S.3), staccato (S.4) and staccato on string (S.5) articulations. The phrasing exercise (EX.1) was recorded both in solo and ensemble conditions. Beethoven's movement is played in ensemble three times with increasing degrees of expressiveness: mechanical

(I.1), normal interpreted (I.2), and exaggerated (I.3). Finally the pieces excerpts were recorded both in solo and ensemble condition (P.1-P.4). In Table 3.1 we summarize all of the music recordings used in this dissertation.

Table 3.1: List of recordings analyzed in this dissertation.

Label	Name	Condition	Number of pitched note events					Aprox. recording time
			V11	V12	V1a	Cello	Total	
S.1	Scale Legato	solo	288	288	288	288	1152	15 mins
S.2	Scale Martelé	solo	288	288	288	288	1152	15 mins
S.3	Scale Detaché	solo	288	288	288	288	1152	15 mins
S.4	Scale Staccato	solo	288	288	288	288	1152	15 mins
S.5	Scale Staccato on String	solo	288	288	288	288	1152	15 mins
EX.1	Phrasing Exercise	solo/ens.	64	64	64	64	256	5 mins
I.1	Beethoven's <i>Allegro-Prestissimo</i> mechanical	ens.	840	630	561	472	2503	5 mins
I.2	Beethoven's <i>Allegro-Prestissimo</i> normal	ens.	840	630	561	472	2503	5 mins
I.3	Beethoven's <i>Allegro-Prestissimo</i> exaggerated	ens.	840	630	561	472	2503	5 mins
P.1	Solo vs. Ensemble (bars 54-78 from <i>Allegro ma non tanto</i>)	solo/ens.	135	104	106	100	445	5 mins
P.2	Solo vs. Ensemble (bars 138-151 from <i>Allegro ma non tanto</i>)	solo/ens.	94	144	164	192	594	8 mins
P.3	Solo vs. Ensemble (bars 8-50 from <i>Menuetto</i>)	solo/ens.	156	204	206	196	762	5 mins
P.4	Solo vs. Ensemble (bars 28-45 from <i>Allegro-Prestissimo</i>)	solo/ens.	55	83	73	56	267	5 mins

Articulations on musical scales

As a first approach to the analysis of gestural and audio performance we started recordings musicians playing simple scales. The main idea behind

Figure 3.1: Musical scale recorded with different articulations. The same ascending scale is repeated three times, each time with a shorter notes.

these recordings was to sample a standard way of executing simple notes with various articulations, durations and dynamics.

Each musician played alone the score in Figure 3.1 in several takes, each with a different articulation: *legato*, *martelé*, *detaché* and two types of *staccato*. Such articulations were decided together with the musicians. For each articulation the musician repeated the score of Figure 3.1 three times in order to realize three increasing levels of dynamics: *piano*, *mezzo forte*, *fortissimo*.

Ensemble phrasing exercise

In order to obtain reliable results using computational means, the existence of valid hypotheses is of very high value; for that reason, we devised an experimental framework that provided a set of recordings where the studied relationships among the musicians are well defined and unambiguous. The corpus is based on an exercise handbook for string quartets⁵, intended for improving the “ensemble skills” of the quartet members. The material is divided into six categories, with each category containing a number of short exercises dealing with a different aspect of ensemble performance: *Intonation*, *Dynamics*, *Unity of Execution*, *Rhythm*, *Phrasing*, and *Tone production/Timbre*. An exercise consists of a simple, low difficulty score, together with annotations on what is the specific goal that must be achieved by the quartet.

⁵Mogens Heimann - Exercises for the String Quartet.

Figure 3.2: Phrasing exercise, where the musicians are divided in two groups and have to follow each other to form a scale in thirds.

In Section 5.2 we will present the analysis the phrasing exercises EX.1, where the challenge for musicians is to play together “as one instrument”. The exercise consists of the ascending and descending D major scale in thirds shown in Figure 3.2. In this exercise, the quartet is divided in two sub-groups (violins in the first sub-group, viola and cello in the second sub-group) performing short sixteenth note sequences in alternation. Musicians were instructed to play the score as if it was played by one instrument. We did not impose on them further constraints such as to follow an external metronome. The goal of this exercise is self-evident in the score. The notes within each group have to blend together while allowing the blocks of semi quaver notes (sixteenths notes) formed by each group to slot together in a temporal order. In addition to that, the requirement of achieving a good “unity of execution” (as suggested by the title chosen by Mogens Heimann for this exercise) means that the parts played by each group have to be connected to the part of the subsequent section without disruptions in terms of tempo and dynamics. It is also worth to notice that the slurs contained in the scores, by requiring the musicians to keep a certain bow direction might also pose some constraints to the synchronization process.

For the EX.1 exercise, and for the other exercises that are not analyzed in this dissertation⁶, we recorded the musicians in three experimental condi-

⁶We employed an alternative approach to measuring inter-dependence among musicians for studying this set of exercises. The approach, based on time series analysis, investigates on the (linear and non-linear) correlations among musicians’ performance. We compared correlations among musicians playing their exercise part alone and in en-

tions: solo (first sight), rehearsal, and ensemble. In the first condition (solo), each musician had to perform their part alone without having access to the full ensemble score nor the instructions that accompany the exercise. In this way we wish to eliminate any type of external influence on the performance, be it restrictions imposed by other voices of the ensemble or instructions by the composer that are not in relation to the individual score of the performer. In the second condition (rehearsal), following the solo recordings of each quartet member, the group of musicians was provided with the full ensemble score plus the composer instructions; they were then left to rehearse the exercise alone until they were able to fulfill the requirements of the exercise. In the third condition (ensemble), following the rehearsal, the quartet was finally recorded performing the exercise as a group.

For each case we recorded 4 consecutive repetitions of the score in Figure 3.2. The analyzed material thus consists of 512 notes as each of the four musicians played 64 notes both in solo and in ensemble.

Expressive intentions

In addition to the exercises we wanted to acquire the performance of already rehearsed scores, to acquire data from the most natural performing condition. We were not able to give them the time to learn a new piece because, on one hand, it would have made the performance less natural given the timing constraints of the recording schedule. We contacted the musicians few weeks before the recording and asked them to propose a list of pieces from their rehearsed repertoire.

We selected the longest (and most recently performed) piece in the quartet's current repertoire: Beethoven's String Quartet No. 4 (Opus 18) *Allegro-Prestissimo* movement. The musicians were instructed to perform the entire piece three times, once for each of the following expressive intentions:

- **Mechanical**, where the musicians stayed as "faithful" to the score as possible without introducing expressive deviations.

semble. We defined a different time series for each exercise which related with the goal of the exercise, e.g. we used pitch deviation for the intonation exercise, we used sound level for the dynamics exercise. Lastly, we found correlation measures for which the overall quartet interdependence increased in the ensemble performance with respect to the solo performance. We refer the reader to Papiotis et al.'s (2014) article for the details of this approach, and to Panos Papiotis' Ph.D. dissertation (Papiotis, in preparation).

Quartet No. 4 in C Minor - Opus 18 No. 4
Allegro - Prestissimo

Ludwig van Beethoven

Allegro ma non tanto

Figure 3.3: Starting bars of the Beethoven’s String Quartet No. 4 (Opus 18) 4th movement.

- **Normal**, where the musicians performed the piece as they would in a concert scenario.
- **Exaggerated**, where the musicians introduced “extreme” expressive deviations from the score, in comparison to the “normal” condition.

The instructions regarding the performers’ expressive intent have been shown to be clearly understood by musicians and non-musicians alike (Kendall and Carterette, 1990). The purpose of the above scenario was to capture different degrees of “deviating from the score”, i.e., introducing personal choices which are not explicitly stated.

In total, the three repetitions consisted of 7508 notes (2520 notes for violin 1, 1889 for violin 2, 1683 for viola and 1416 for cello).

Solo vs. Ensemble conditions

As we did for the short phrasing exercise we wanted to compare solo and ensemble conditions in selected excerpts from the string quartet’s repertoire. We selected short excerpts from the quartet’s repertoire: four excerpts from the same Beethoven’s String Quartet No. 4 (Opus 18). The selected excerpts correspond to bars 54-78 (P.1), 138-151 (P.2) from *Allegro ma non*

tanto, bars 8-50 (P.3) from *Menuetto*, and bars 28-45 (P.4) from *Allegro-Prestissimo* respectively (see Table 3.1). These short excerpts were then recorded under the following two conditions:

- **Solo**, where the musicians each performed their own part alone (without listening to pre-recorded material or a metronome signal).
- **Ensemble**, where the entire quartet performed together following a brief rehearsal.

The purpose of the above scenario was to observe whether the musicians would perform their parts differently from one condition to the other, given that in the *solo* condition there were no external perturbations to the performance. The *solo* and *ensemble* recordings were carried out on different days.

In total, the four excerpts comprised of 2068 notes (440 notes for violin 1, 535 for violin 2, 549 for viola and 544 for cello). Considering then that each excerpt was played twice by each musician, the total amount of measured notes amounts to 4136.

3.2 Recording setup

The recording setup required interconnecting several devices to acquire the performance of the four musicians simultaneously. Besides the sensing technology used in this dissertation, As we are collecting a database of ensemble performances to be shared with the research community, we acquired not only the sound produced by each musician in a separate tracks and the instrumental motion that were used in this dissertation, but also body motion and ambient audio.

The complete setup involved acquiring (1) **sound** through two ambient microphones and four piezoelectric contact microphones, (2) **motion** through eight Electro-Magnetic-Field (EMF) motion capture wired markers and 84 wireless infra-red markers, and (3) **video** through an HD video-camera. In Figure 3.4 the reader can see a picture taken during one of the recording sessions, we blurred out musician's faces for privacy. In this dissertation we focus on analyzing data coming from contact microphones and EMF motion capture system only, although we briefly describe in the following paragraphs the whole recording setup.



Figure 3.4: A picture from one of the recording sessions. Musician’s faces have been blurred out for privacy.

The audio data include four individual audio tracks (one for each musician, all sampled at 44100 Hz) from piezoelectric bridge pickups (models: Fishman V-100 and Fishman C-200) installed on the instruments. Pickup gains were manually set with the aid of level meters in the recording equipment to avoid clipping of individual audios. To balance the sound level of the quartet as a whole, we employed an iterative procedure, based on listening to the mix of the pickups on the headphones while adjusting the four gains. Additionally, we acquired the ambient audio with a cardioid microphone and a binaural microphone.

The video was acquired with a Canon VIXIA HF R200 camera to which we fed the SMPTE signal. This allowed synchronizing start and stop recording position in a post-processing step.

Bowing and body motion was acquired through two motion capture systems. The first is a Qualisys system⁷ that we used to acquire the position of 84 markers, 7 placed on each instrument, and 14 placed on each musician. The second system is a Polhemus Liberty⁸ wired motion capture system (based on EMF sensing). This last system acquired bowing motion data

⁷<http://www.qualisys.com>

⁸<http://polhemus.com/motion-tracking/all-trackers/liberty>

at a sample rate of 240 Hz and was used as detailed in Maestre's (2009) and Pérez's (2009) Ph.D. Theses. We extended software running as a VST plug-in for the acquisition of audio and EMF motion data simultaneously (see Figure 3.6 for a screen-shot).

The acquisition of infrared MoCap was handled via proprietary software (the Qualysis Tracker Manager) developed for the Qualisys system that was hardware synchronized to the audio card via SMPTE time-code signal. For this reason, we the audio and the infrared MoCap did not need any additional synchronization procedure. The synchronization of audio and EMF motion capture data was instead handled by a custom synchronization system. We integrated into our VST plug-in software a module that generate periodic pulses synchronized with the audio card clock and outputted to a dedicated analog audio output line. The analog signal is converted to a standard 5 Volts on/off signal pulse by a custom sync circuit realized at Universitat Pompeu Fabra. The on/off pulse is fed to the Polhemus board that, as a result, returns time-stamped data. Our software compares output pulses with incoming time stamps to continuously adjust potentials drifts between Polhemus and audio frames by dropping or repeating Polhemus frames when necessary. The schematics of the synchronization system is illustrated in Figure 3.5.

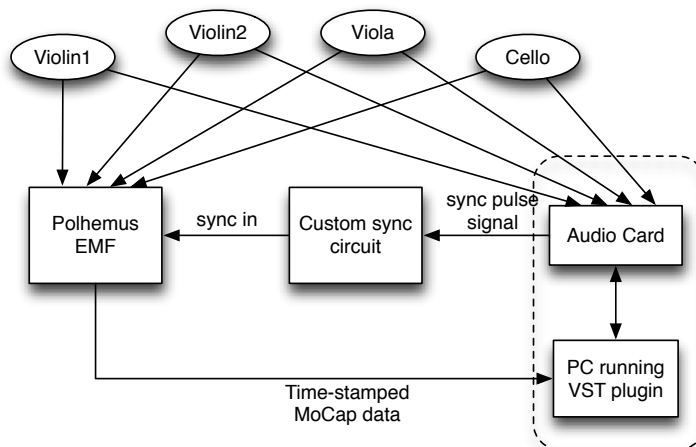


Figure 3.5: Schematics of the system used to synchronize the audio card with the Polhemus EMF MoCap system.

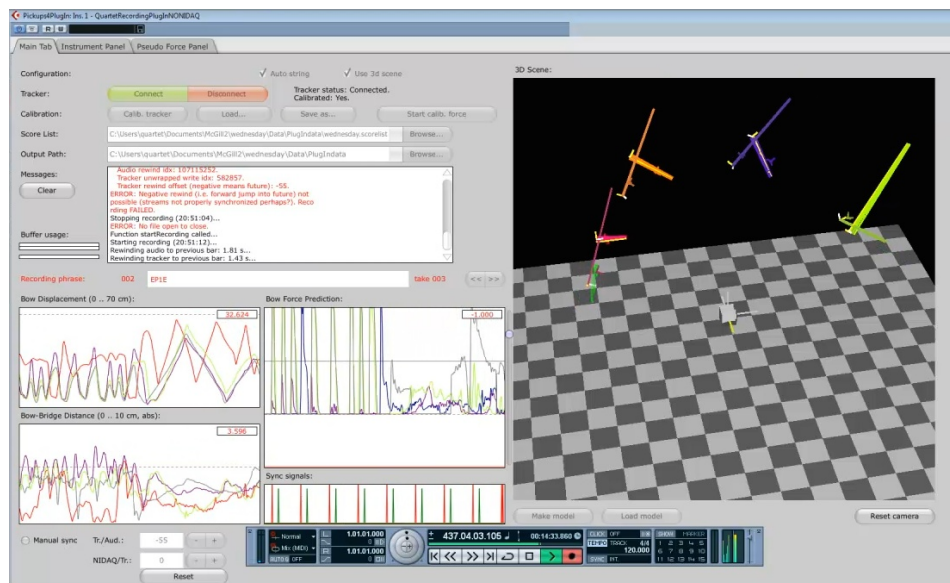


Figure 3.6: Screen-shot of the VST plug-in user interface for recording audio and EMF motion data.

3.3 Audio features

Audio data post-processing included the extraction of audio features on each individual instrument track. The purpose of those features was twofold. First, we used them to automatically segment the performance using the score-alignment algorithm that will be explained in Section 4.2.2. Second, we use the audio features to describe each note with a set of expression parameters, such as note sound level and vibrato, that will be explained in Section 4.2.3. For simplicity we extracted all the audio features at the same sampling rate as the EMF motion data. Thus we have both audio features and motion at 240 Hz.

The extracted audio features are the following:

- Sound pressure level (SL)
- Pitch fundamental frequency (F0)
- Pitch aperiodicity (AP0)
- Harmonic Pitch Class Profiles (HPCP)

Sound level was estimated for each note as follows. First, the audio signal corresponding to each voice was divided into frames of 23ms with an 82% overlap so as to obtain 240 Hz. Let $(x_1^{(i)}, \dots, x_N^{(i)})$ be the samples of the i -th frame. The energy of the frame was computed as the root mean square of its samples:

$$\text{RMS}(i) = \frac{1}{N} \sqrt{(x_1^{(i)})^2 + \dots + (x_N^{(i)})^2}$$

We then obtained the sound level $\text{SL}(i)$ by expressing the RMS energy in dB:

$$\text{SL}(i) = 20 \log_{10}(\text{RMS}(i))$$

Fundamental frequency and aperiodicity were obtained via an autocorrelation-based method as described by De Cheveigné and Kawahara (2002). The F_0 was used as the pitch performed by each musician at each frame instant i . The aperiodicity measures by how much the signal is non-periodic. We used the aperiodicity as a measure of how pitched the sound is at each instant, which was useful to discard non-pitched transient for vibrato extent estimation (see Section 4.2.3).

Chroma is a common audio feature useful for extracting information about active pitch classes at each instant in time. More specifically here we compute the Harmonic Pitch Class Profiles (HPCP) as described by Gómez (2006) in her Ph.D. thesis. In order to extract such feature we used the HPCP vamp plug-in available at the MTG website⁹. We used the default parameters of the vamp plug-in with the exception of the hop size that was set so to obtain 240 frames per second (the same number as the motion data). At each frame i , corresponding to an instant in time, the algorithm returns a 120-dimensional vector $\text{Chr}(i)$ of positive weights summing to one; these weight represent the amount of activation of a certain pitch class at that instant.

3.4 Bowing motion features

Bowing motion features were important in our work as they enabled to improve automatic performance alignment (see Section 4.2.2) and to use bow velocity as and additional expression parameter (see Section 4.2.3). A complete set of continuous bowing descriptors can be acquired using EMF motion sensing technology. Position and orientation of the two 6DOF

⁹www.mtg.upf.edu/technologies/hpcp

sensors are tracked at 240Hz. The first is attached to the violin back plate, in the upper bout, at the neck edge. The second one is affixed to the bow wood, close to its center of gravity. From the data provided by these sensors, a set of motion descriptors is extracted by means of the data processing steps outlined in this section. For a more detailed description of the procedure for obtaining relevant bowing motion parameter streams, refer to Maestre's (2009) Ph.D. Thesis.

Initially, a calibration of the string and hair ribbon ends is performed. The exact position of the eight string ends (four at the nut and four at the bridge) can be tracked by previously annotating (during calibration) their position relative to the coordinate system defined by the position and orientation of the 6DOF sensor placed in the violin. In a similar manner, the positions of the two hair ribbon ends (at the frog and at the tip) are estimated from the position and orientation of the 6DOF sensor placed on the bow. Both the violin plane and the bow plane (respectively defined by their normal vectors v_n and b_n) are estimated. The former is estimated from the eight string ends, and the latter is estimated from both the sensor position and from the two hair ribbon ends (see Figure 3.7). With this information for each instrument, we are additionally able to extract a 3D image of the performance to be visualized in real-time as displayed in Figure 3.6.

We estimate the bow tilt by measuring the angle between the violin plane normal vector v_n , and a vector b_o being simultaneously parallel to the bow plane and perpendicular to the bow plane normal vector b_n (the vector b_o is obtained as the vectorial product of the hair ribbon vector h and b_n). In a similar manner, the string being played is estimated by measuring the angle between v_n and h (see Figure 3.7). By defining a line between the ends of the string being played (depicted as s_b and s_n), and another line between the ends of the hair ribbon (depicted as h_f and h_t), a segment P is defined by a line perpendicular to both p_h and p_s . The bow transversal position is defined as the distance between p_h and h_f , and the bow-bridge distance is defined as the distance between p_s and s_b . Bow velocity is obtained as the time derivative of the bow transversal position.

The force that the bow exerts on the string is a further important feature of bowed instruments playing but could not be captured directly. However, we can extract from the MoCap data a correlated variable that comes from the deflection of the hair ribbon: the *pseudo-force*. The derivation of the pseudo-force also follows simple geometric projections that are displayed in Figure 3.8. The pseudo-force is defined as the shortest distance between

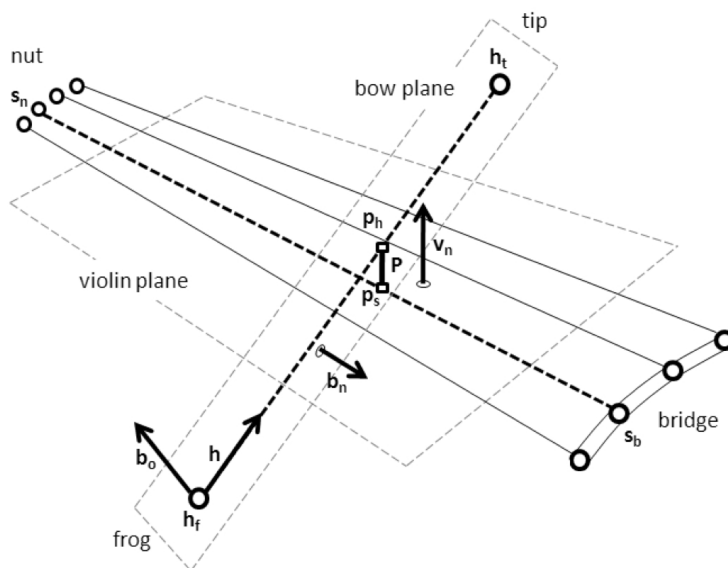


Figure 3.7: Schematic representation of relevant positions and orientations tracked in the extraction of bowing motion parameters with EMF sensing technology.

the virtual position of the hair ribbon (assuming it as a rigid object that do not interact with the string) and the string that is being played. For more information about the procedure employed for the extraction of bowing features we refer the reader to Maestre's (2009) Ph.D. Thesis.

3.5 Bow Force Estimation

Whereas most of the extraction of bowing motion features was based on reproducing existing state-of-the-art systems, the extraction of estimated bow force represents a contribution of our work. In this section, we present the methodology for the estimation of *bow pressing force* by using a simplified physical model of the hair ribbon deflection which makes use of only position and orientation (6DOF) measurements on the bow and violin. The motivation is to minimize the intrusiveness by avoiding the use of additional sensors, and therefore construct a more reliable system that can be used spontaneously by musicians and for longer periods of time. The principal source of information comes from measuring the pseudo-force, as it was

already proposed by Maestre et al. (2007). The distance between the ideal (no deflection) segment defined by the ends of the hair ribbon, and the segment defined by the ends of the string being played. A simplified physical model of the hair ribbon deflection is constructed and calibrated from real data measurements using a load cell, and used later for estimating the bow force in real performances. For more information about the estimation of bowing force we refer the reader to our NIME paper (Marchini et al., 2011).

3.5.1 Measurement of applied force

In order to both design and evaluate our system for the estimation of bowing force, we used a linear load cell transducer to measure the actual force being applied by the bow, as suggested by Schoonderwaldt (2009a) and implemented in Gaus et al. (2009). The cell is fixed to a wooden support, while a thin methacrylate cylinder is placed over the cell to simulate a virtual string. By using another EMF sensor attached to the wooden support, we are able to track the ends of the cylinder and thus acquire a number of simultaneous bowing parameters (including *bow displacement* and *pseudo-force*) along with the output of the load cell (the actual force in Newtons).

The output of the linear load cell itself is calibrated using a set of precision weights; the force produced by these weights on the load cell is derived from Newton's second law of motion, $F = Mg$, with the gravity constant $g = 9.8m/s^2$. The voltage output of the load cell is post-processed to match the corresponding unit of Newtons by applying a simple linear transformation of the form $y = \gamma x + \phi$, where γ equals the voltage gain and ϕ equals the voltage offset.

3.5.2 A simplified physical model of hair ribbon deflection

In this section we present a simplified physical model of a flexible thread, and then we extend it to the case of multiple hairs and generalize it to describe the complete hair ribbon. We use such physical model in order to approximate, given solely information extracted from 6DOF sensors, the force exerted on the string regardless of the displacement or tilting of the bow. An important simplification was to assume the bow stick as rigid.

The thread

The simplest approximation of the bow hair-ribbon is a single elastic thread stretched between two points A and B (see Figure 3.9). At its rest position

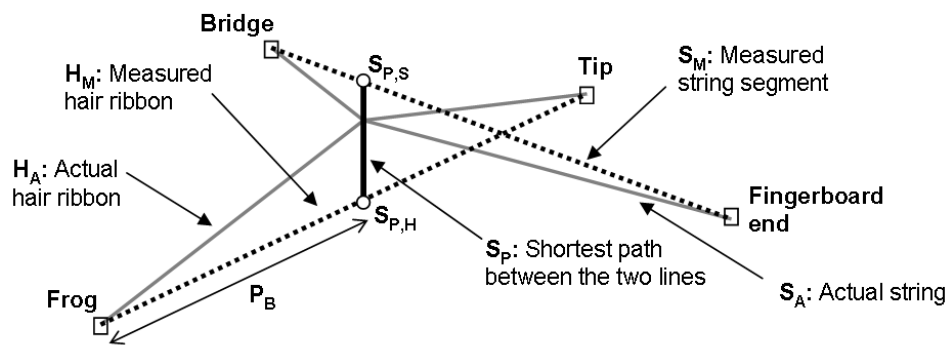


Figure 3.8: Measured string and hair ribbon segments, computed from their extracted end points, versus their actual configuration. Deformations have been exaggerated in order to illustrate the importance of segment S_P .

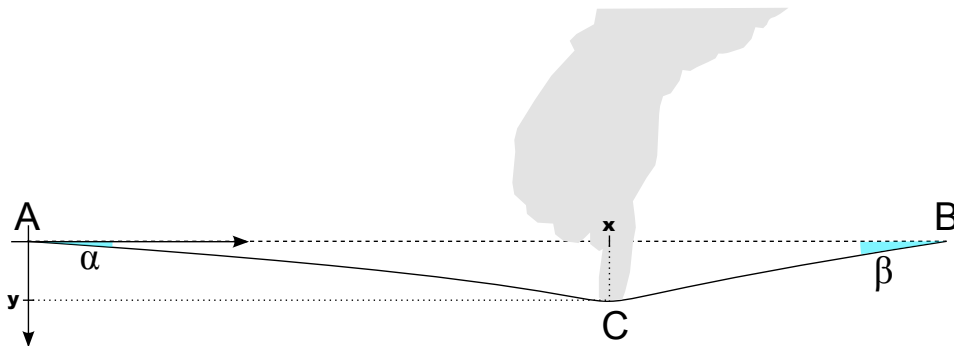


Figure 3.9: Single elastic thread attached at the extremities A and B and pushed at a point C.

the thread has length ℓ , which coincides with the distance between the points A and B. When a force is applied on a point C, the thread stretches and is elongated until an internal equilibrium of the system is reached.

In its rest position, we consider such thread as the limit of an array of masses connected by springs, presenting a mass-to-mass distance approaching zero. We parameterize the thread by a function $u : [0, 1] \rightarrow \mathbb{R}^2$, and express the

potential energy of the thread as

$$\frac{1}{2} \frac{T}{\ell} \int_0^1 u'(t)^2 dt,$$

where T is the tension of the thread. If u is the parametrization of the thread of Figure 3.9 with $u(0) = A$, $u(\nu) = C$, and $u(1) = B$ where $\nu \in]0, 1[$, the potential energy is given by

$$\frac{1}{2} \frac{T}{\ell} \left(\frac{x^2 + y^2}{\nu} + \frac{(1-x)^2 + y^2}{1-\nu} \right).$$

The internal equilibrium, i.e. the minimum potential energy, is reached for $\nu = \nu_{\text{eq}}$, where

$$\nu_{\text{eq}} = \frac{x^2 + y^2}{x^2 + y^2 + \sqrt{((1-x)^2 + y^2)(x^2 + y^2)}}. \quad (3.1)$$

In this equilibrium state, the point C is subject to two forces \vec{f}_1 and \vec{f}_2 in the direction CA and CB respectively. Their magnitude is the following:

$$\|\vec{f}_1\| = T(\ell_1 - \nu_{\text{eq}}\ell) \quad (3.2)$$

$$\|\vec{f}_2\| = T(\ell_2 - (1 - \nu_{\text{eq}})\ell) \quad (3.3)$$

Let's denote ℓ_1 and ℓ_2 the length of the AC and CB parts respectively, α and β the angles CAB and ABC respectively, and $\Delta\ell = \ell_1 + \ell_2 - \ell$ the total en-lengthening of the thread. The total force that the thread exerts on C is $\vec{F} = \vec{f}_1 + \vec{f}_2$. If we set a coordinate system at point A oriented as shown in Figure 3.9, the point C will be described by its coordinates (x, y) . Now, writing $\vec{F} = (F_{\text{horz}}, F_{\text{vert}})$ where F_{horz} is the horizontal component of \vec{F} and F_{vert} the vertical component, the vertical component of the force can be considered as the force applied to the string, and written (observing that $\sin(\alpha) = \frac{y}{\ell_1}$ and $\sin(\beta) = \frac{y}{\ell_2}$) as

$$F_{\text{vert}}(x, y) = \|\vec{f}_1\| \frac{y}{\ell_1} + \|\vec{f}_2\| \frac{y}{\ell_2}. \quad (3.4)$$

The Hair Ribbon

A more precise approximation of the hair ribbon is to consider it as a strip of parallel threads, assuming that the force exerted by the ribbon is the sum of the contributions of each thread. Considering an homogeneous distribution

of threads with density ρ and defining w to be the width of the strip, we can define the force applied to the string as

$$\text{Force} = \rho \int_0^w f(z) dz, \quad (3.5)$$

where $f(z)$ is the force density of the thread situated at position z on the strip.

Let $m := y(0)$ and $M := y(w)$ be respectively the measured left-hand side and right-hand side *bow-string distance*. We then have

$$y = m + \frac{(M - m)}{w} z. \quad (3.6)$$

Figure 3.11 schematically depicts the possible relative positions (displacements) that we considered for the hair ribbon (transversal view) as relative to the string. The displacement of the string with respect to the bow is determined by the linear relation $y(z) = az + b$. Only the threads where y is positive are contributing to the force (as they are in contact with the string). This happens¹⁰ when $z > -\frac{b}{a} =: c$. Having the diagram of Figure 3.10 as a reference, we considered the variable x as constant with respect to z while, depending on the changes of y , we reduce the problem to three main cases, defined as

Case I : The y are negative (the ribbon is not touching) with respect to all $z \in [0, w]$, having

$$f(z) = 0 \quad \forall z;$$

Case II : y is positive for $0 < z < c < w$ reaches 0 for $z = c$ and is negative for $z > c$, with

$$f(z) = \begin{cases} 0 & \text{for } z \in [0, c] \\ F(x, y(z)) & \text{for } z \in (c, w] \end{cases}; \quad (3.7)$$

Case III : y is positive for all $z \in [0, w]$, so

$$f(z) = F(x, y(z)). \quad (3.8)$$

¹⁰Considering, for the moment, the case where $a > 0$ without any loss of generality.

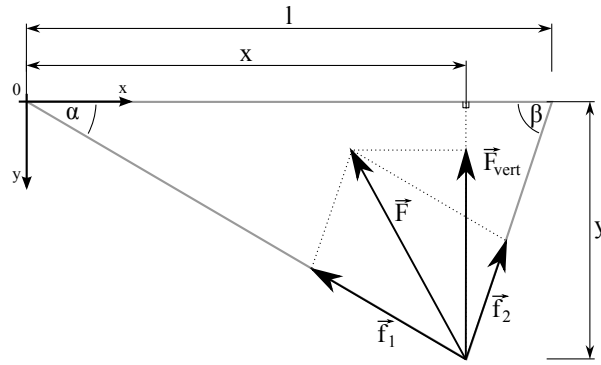


Figure 3.10: Decomposition of the internal forces exerted by a single elastic thread.

Case I is of little significance, since the force is zero. In the other two cases, applying equation (3.6) in the equations (3.7) and (3.8), we may completely rewrite equation (3.5) using the definition of f and $\Delta\ell$ canceling thus all the implicit dependencies to the variable y . Applying then the substitution $z = \frac{w}{M-m}(y - m)$ to the integral we can write the results in term of the function

$$\tilde{F}(z) := \frac{1}{T} \int_0^z F_{\text{vert}}(x, y) dy, \quad (3.9)$$

where we divide by T so that \tilde{F} do not depend on the tension. We will handle this parameter in the further formulae.

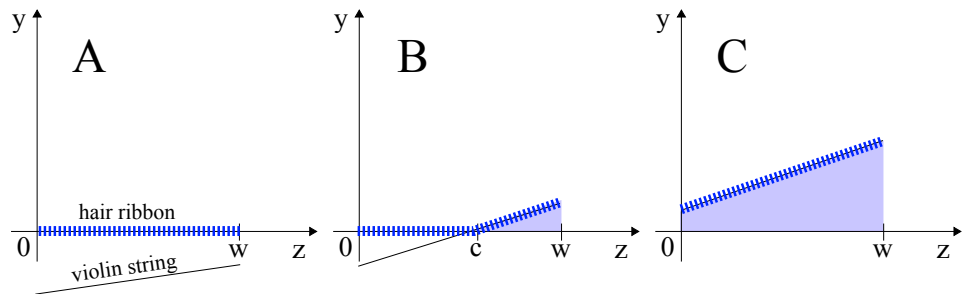


Figure 3.11: I. Non touching ribbon. II. Partially touching. III. Fully touching.

For the fundamental theorem of calculus plus taking into account the term $\frac{w}{M-m}$ of the substitution, we conclude, for the three considered cases, as

Case I : The y are negative (the ribbon is not touching) with respect to all $z \in [0, w]$, thus

$$\text{Force} = 0;$$

Case II : y is positive for $0 < z < c < w$ reaches 0 for $z = c$ and is negative for $z > c$, having

$$\text{Force} = T\rho w \frac{\tilde{F}(\max(M, m)) - \tilde{F}(0)}{|M - m|}; \quad (3.10)$$

Case III : y is positive for all $z \in [0, w]$, so

$$\text{Force} = T\rho w \frac{\tilde{F}(M) - \tilde{F}(m)}{M - m}. \quad (3.11)$$

Finally, we observe that:

1. The final force **only** depends on the variables x , M and m plus the constants T , ρ and w .
2. The cases I, II and III can be identified only looking at the value taken by M and m . Case I holds when both are negative, case II when they differ in sign and case III when both are positive.
3. Considering $a > 0$ did not cause a loss of generality. Indeed, for $a < 0$, due to a symmetry of the problem, we could just switch M with m but this, thanks to the way equation (3.10) has been expressed, does not change the result. Finally we can interpret the case $a = 0$ as a limit case of equation (3.11) when $(M - m) \rightarrow 0$. The results of the limit is, in fact, $\text{Force} = \rho w F_{\text{vert}}(x, M)$ corresponding to an equal contribution of all the threads to the final force.
4. The function \tilde{F} has an analytical formulation which can be obtained expanding the integral in Equation (3.9) using Equations (3.1) to (3.4).

3.5.3 Optimization Procedure

The described model is parameterized by a single scalar value given by the product $T\rho w$ of the tension T , the thread density ρ and the width w of the hair ribbon. The whole force will be scaled by the factor $T\rho w$. It is not necessary to estimate the three constants since we are interested

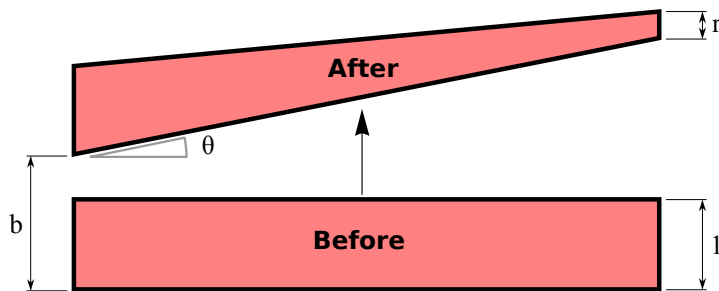


Figure 3.12: The function Map applied to a rectangle. For this example exaggerated parameters used are $r = 0.5$, $b = 1.5$, $\theta = 0.1\text{rad}$.

only on their product. Thus, we need to infer the size of this factor from an experiment; however, there are additional conditions in the real case which are not addressed by the physical model. First, the motion tracking sensor placed on the bow stick might rotate by a small angle θ after the calibration has been performed, causing a rotation of all the data. Secondly, due to the movement of the sensor, it might be necessary to adjust the bow displacement adding an offset constant a , and the pseudo-force with an offset constant b . Lastly, in order to address the problem of the bending of the stick, another constant r is added defining a transformation which will compensate, the effect of the stick bending by dividing the pseudoforce by a value depending on the bow displacement. The final transformation is given by the following formula:

$$\begin{cases} x' = a + x \\ M' = \text{Map}_{(x,b,\theta,r)}(M) \\ m' = \text{Map}_{(x,b,\theta,r)}(m) \end{cases} \quad (3.12)$$

where

$$\text{Map}_{(x,b,\theta,r)}(y) := \left(b + y \left(\frac{1+r}{2} + \frac{(-1+r)(-\frac{l}{2} + x)}{l} \right) \right) \text{Cos}[\theta] + x \text{Sin}[\theta] \quad (3.13)$$

Description

Suppose we have a training set $\{(x_i, M_i, m_i)\}_{i=1,\dots,n}$ where x_i is the bow displacement, M_i is the pseudo-force of left side and m_i is the pseudo-force of right side at time i . Given the parameters T , θ , a , b and r we consider the prediction

$$\text{Force}_i(T, \theta, a, b, r) = \text{Force}(x_i + a, \text{Map}_{(x_i, b, \theta, r)}(M_i), \text{Map}_{(x_i, b, \theta, r)}(m_i)). \quad (3.14)$$

We want to find the better values for the parameters in order to minimize the absolute error:

$$J(T, \theta, a, b, r) = \frac{1}{2} \sum_{i=1}^n (\text{Force}_i(T, \theta, a, b, r) - \text{nidaq}_i)^2.$$

where nidaq_i is the load cell measured force at time instant i . We thus aim at finding the parameters:

$$(T^*, \theta^*, a^*, b^*, r^*) = \arg \min_{(T, \theta, a, b, r)} J(T, \theta, a, b, r)$$

We use the Nelder-Mead simplex method Lagarias et al. (1999) in order to find a local minimum, starting from the identity transformation parameters: $T = r = 1$, $\theta = a = b = 0$. In order to reduce the computation time, we down-sampled the signal to 8 samples a second. Before the optimization of the parameters we also filtered the dataset, to remove noisy data. We removed the samples where the measurement of the Force cell was less than 0.2. In fact the sensitivity of the sensor for small forces was reduced and noisy.

3.5.4 Results

Using the acquired gesture parameters along with the Force cell data, we recorded three evaluation datasets. In the *dataset 1*, an almost constant force was applied with different bow transversal positions and different tilts. In the *dataset 2*, the pseudo-force was changing constantly from positive to negative while changing tilt and bow transversal position in order to simulate the way violin is normally played. In the *dataset 3*, bow transversal positions was kept fixed while the force and the tilt where changing. This was done for many different bow transversal positions. Each recording was around one minutes long. We created a *Joint Dataset*, with the samples

of the three recordings and we performed a 10-fold cross validation, with a performance going as low as 84.01 mean correlation.

Finally, in order to investigate on the best calibration procedure, we also compared the three different datasets. The table 3.2 shows the results of each possible combination of training-set with test-set. It clearly shows that the third dataset is good enough to predict the rest. We can conclude that, a short calibration of one minute, with static bow displacement positions results sufficient for general-purpose force estimation. This last type of calibration was indeed eventually used before each performance to calibrate the musician's bows.

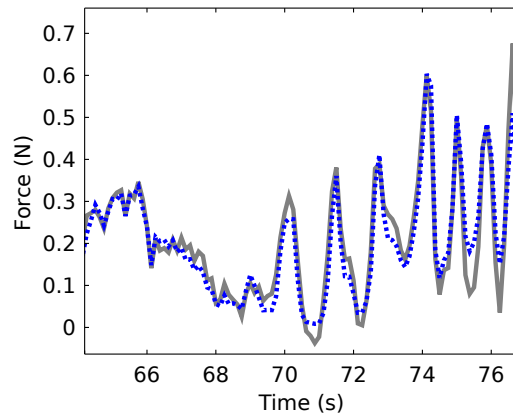


Figure 3.13: An excerpt of the recorded force signal (*continuous gray line*) along with its prediction (*blue dashed line*) after the optimization has been performed.

3.5.5 Bow force estimation in string quartet recordings

The introduced physical model hair ribbon deflection produced good results in laboratory tests. One of the main advantages of this method respect to a black-box machine-learning approach is that the model depends on very few parameters that can be adjusted in a post-processing step if something goes wrong during the recording. Given the results of our investigation on bowing force estimation and on the optimization of the parameters, we decided to adopt the following procedure in each of the carried out recording sessions:

Table 3.2: Correlation and relative error (both in percentage) between the force prediction and the measured force for each training set and test set coupling.

		Test Set							
		Dataset 1		Dataset 2		Dataset 3		Joint	
		<i>corr</i>	<i>rel err</i>	<i>corr</i>	<i>rel err</i>	<i>corr</i>	<i>rel err</i>	<i>corr</i>	<i>rel err</i>
Training Set	Dataset 1	89.21	16.75	96.18	23.68	98.84	14.67	97.29	17.03
	Dataset 2	95.03	13.78	97.76	16.75	98.37	20.13	97.4	16.88
	Dataset 3	84.95	23.61	95.87	22.64	98.88	12.18	96.13	19.82
	Joint	94.1	13.04	97.3	19.73	98.94	12.19	98.12	13.75

1. At the beginning of the session we asked the musicians to set the tension of their bows to the desired amount.
2. We explicitly asked them not to change the tension of the bows during the whole session.
3. At the end of the session we calibrate each bow on the load cell. Each calibration involves a recording of around two minutes where we record the force obtained with different degrees of tilt, pseudo-forces, and bow displacements.

In a post-processing step we employed each calibration recording to derive the optimal tension parameter T^* . The remaining parameters θ^* , a^* , b^* and r^* are specific of the calibration procedure and they could not be applied equally well to the playing on the instrument. Those parameters were therefore adjusted manually in a post-processing step in which we compared resulting force estimation with audio and other bowing parameters. We adjusted the parameters individually on each string to account for the distorting effects of the electromagnetic field sensor at various orientations. This procedure was realized on each of the recording sessions focusing on small parts of the recording to derive the parameters on each string.

There was no ground truth measurement of the bowing force applied by the musicians on their instruments, so that a complete evaluation of the method was impossible. However, the audio track provided a guideline to our manual parameter adjustments since transitions from silence to sound could be spotted out and compared with spikes of force. In order to carry out the manual estimation of the force parameters we developed an interface (showed in Figure 3.14) in which it was possible to interactively see the effect of manipulating the parameters. Whereas the transition between no-force to force could be spotted out relatively well by our model, the amount of force undergoes slight fluctuations along the recording due to the movement of the musician in space. For the scope of this dissertation we used force estimation to improve the score-performance alignment algorithm by exploiting the force and bow velocity transitions (see Section 4.2.2).

In the rest of the dissertation we will use bow velocity and estimated bow force for the analysis of expressive performance. Since we used the sampling frequency of 240Hz for both audio and gesture descriptors, we have 240 measurements of bow velocity and estimated bow force per second. Those values are parameterized by the frame number $i = 1, \dots, m$ where m depends on the duration of the recording. We will use the following shorthand notations:

$BV(i)$ = Bow velocity along the direction of the bow, measured at frame i
in cm/s

$EBF(i)$ = Estimated bow force, measured at frame i in Newtons

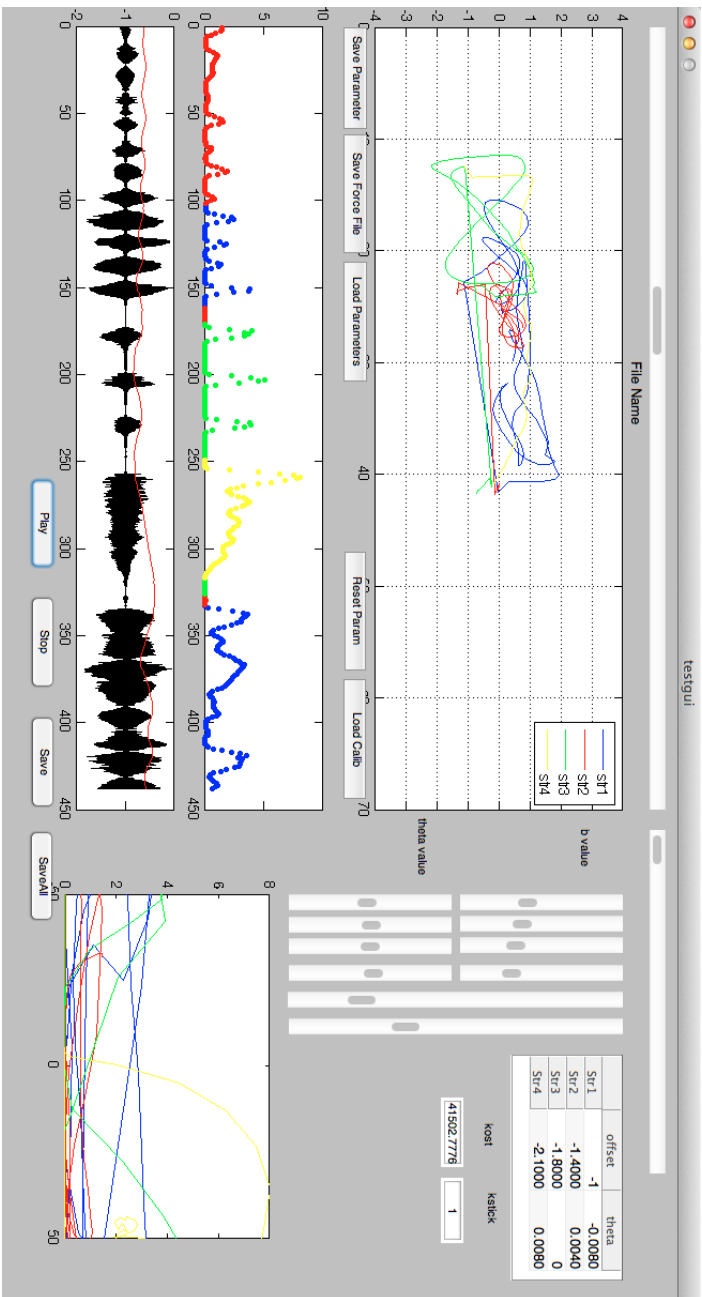


Figure 3.14: Interface used to manually calibrate the individual parameters of each string. At the top, we show the trajectory of bowing parameters for a short segment of the recording in the space of bow displacement and pseudo force. Lower in the middle the current sound waveform and bow displacement are shown. On the right one can see a set of parameters that can be adjusted interactively and, on the bottom right, the trajectory on the velocity-estimated force space.

3.6 Conclusions

Data acquisition was crucial in this work and represents a contribution per se. To the best of our knowledge, no work had before acquired string quartet performance to such level of detail. We believe that the collected data represents a valuable dataset that could be used not only to study ensemble expressive performance but also other aspects of ensemble performance such as non-verbal communication or synchronization.

In this chapter we reported the procedure followed to acquire the data and discussed on the choice of the music material and on the time scheduling constraints that forced some of our decisions. We listed the recordings analyzed in this dissertation along with the underlying motivations that pushed us to choose such scores. We presented the *solo* and *ensemble* conditions specifying which excerpts were recorded in one or both conditions.

We took great advantage from previous work on the extraction of bowing parameters and extended it to also include bowing force estimation. A great advantage of our dataset is given by the possibility to focus on the performance of each musician separately. Performance actions of musicians were, in fact, acquired simultaneously and independently one from the others. Furthermore, the use of individual piezoelectric pickup allowed an individual extraction of audio feature on each musician, even in the ensemble condition when they shared the same room.

The release of the open access dataset for research purposes enables additional studies and the possibility for each researcher to test their models. We hope that the free availability the dataset will foster research in ensemble performance.



Score & Performance Analysis

One of the main goals of research on expressive performance is relating score contextual information to expression parameters. However, both the MusicXML scores and the raw performance data need to be interpreted from a musical perspective before any analysis of expressive performance can be carried out. In this chapter we cover the extraction of score contextual descriptors and performed expression parameters, which we will then use in Chapter 5 to analyze expressive performance.

In the first part of the chapter (Section 4.1), we cover the extraction of score context. On the surface, each note is defined by its basic properties such as pitch and duration. However, as soon as a note appears within a context (e.g. a melody), it can assume a wide variety of connotations. This is analogous to the language domain; the very same word can assume different meanings when we place it in alternative sentences. In this latter case, the context helps discriminating between different acceptations of the same word. Similarly, in music, the score context helps understanding the function of a particular note in the melody.

In music, the context is not only given by previous and following notes. In fact, due to polyphony, there might be notes played simultaneously which also participate in defining the function of a particular note. For these reason we distinguish between two main types of context in music scores: the *horizontal* context and the *vertical* context. The *horizontal context* is defined by the information about previous and following notes constituting the melody where the note appears. The horizontal context can be seen as analogous to the context in written text because of its sequential nature. The *vertical context* is instead typical of music polyphony as it is defined

by the information about concurrent melodies being played.

In this first part of the chapter (Section 4.1), we show that in machine-learning algorithms both horizontal and vertical contexts can be represented and encoded as feature vectors. We provide a methodology to extract horizontal and vertical context features from string quartet scores. Features can be numeric, categorical or boolean. The procedure could be easily extended to other types of ensemble music.

In the second part of this chapter (Section 4.2), we preliminarily focus on the procedure employed for deriving score-performance alignment of the recorded material. As already mentioned in Chapter 1, such procedure is a critical part of our analysis and replaces an otherwise impractical music transcription procedure. Thereafter, we define a set of *note expression parameters* computed from note durations and from the acquired data on each note. These parameters represent a varied sub-set of the expressive capabilities of the bowed instruments. The expression parameters that we employed in this dissertation are represented by numeric values.

4.1 Score Analysis

The recorded material was accompanied by corresponding scores in MusicXML, an XML-based file format for representing Western musical notation. This format, although proprietary, can be freely used under Public License. In a MusicXML, data are organized in *parts* (which we also refer to as *voices*) comprehending parts for first violin, for second violin, for viola, and for cello. In this section we describe the process we have employed to analyze the parts and extract a set of score descriptors.

In *ensemble* scores we have more than one concurrent voice being played simultaneously. In the analysis of expressive performance we focus on one voice at a time (e.g. first violin) characterizing the context of each note in that *reference voice*. The note that is being characterized is referred to as the *reference note* (see Figure 4.1). By analyzing the reference voice we can already extract horizontal melodic/rhythmic context of the reference note (see Figure 4.1). However, the same melodic line can assume different connotations if the accompaniment is altered and, in this case, the musician might decide to render it employing different expression. For this reason, contextual information that only describes one melodic line is not sufficient for predicting the performance. In this case we deal with four voices, which is a good compromise between a duet and an orchestra; there being enough

inter-voice relations and contextual information to make a specific analysis meaningful, but not too many to dramatically increase the complexity and make the analysis impossible.

The figure shows a musical score for four instruments: Violin I, Violin II, Viola, and Violoncello. The time signature is 4/4. A specific note in the Violin I part is highlighted with a solid box and labeled 'REFERENCE NOTE'. A vertical dashed box around it is labeled 'VERTICAL CONTEXT OF REFERENCE NOTE'. Horizontal dashed boxes around the note in each voice are labeled 'HORIZONTAL CONTEXT OF REFERENCE NOTE'. A horizontal dashed box around the note in the Violoncello part is labeled 'HORIZONTAL CONTEXT OF CONCURRENT NOTE'.

Figure 4.1: An example demonstrating the Horizontal and Vertical contexts as they are derived from the score.

We introduced a set of *ensemble* context descriptors which can be extracted, in theory, from any ensemble score. Figure 4.1 graphically depicts the kind of score contextual descriptors we considered on a real score example. *Horizontal context descriptors* were computed based solely on a musician's individual voice, ignoring the voices of the others (see Table 4.1). These include both properties of the note itself, as well as properties of the neighboring notes (preceding and subsequent) in that voice. Different temporal context window sizes can be considered by adding more or less neighboring notes to the *feature set*.

Vertical context descriptors include information from the score about the notes played by other musicians concurrently with the reference note (see Table 4.2). In each of the three other voices, the corresponding *concurrent note event* was selected as the one note (or rest) that is active (or starts) at the beat position where the reference note starts. Vertical descriptors are extracted from the list of concurrent notes by either relating their prop-

erties to the reference, or combining their properties to compute resuming descriptors. The horizontal context of each concurrent note event was also considered as a vertical context of the reference note (see the selected context from the voice of the viola in Figure 4.1).

In Tables 4.1 to 4.3 we list a set of score contextual descriptors, which we will use extensively in Chapter 5. We provide the formula of each descriptor employing the following mathematical notation. Assume that the reference note N_0 is fixed and has its onset at beat b_0 in the score, has pitch p_0 , and duration d_0 . At beat b_0 the other three musicians are playing the concurrent note events $N_0^{(1)}$, $N_0^{(2)}$ and $N_0^{(3)}$ respectively. Note how we use the superscripts to identify the concurrent notes in other voices. We analogously use superscripts to refer to properties of such notes; e.g., $p_0^{(2)}$ is the pitch value of the second of the other musicians. We use the subscripts to indicate neighboring notes; e.g., d_{-1} and d_1 refer, respectively to the duration of the previous and the following note in the voice line of N_0 . We combine both subscripts and superscripts to refer to neighboring notes in the voice of other musicians; e.g., $p_{-1}^{(1)}$ is the pitch of $N_{-1}^{(1)}$, which is the previous to the concurrent note, in the voice line of the first of the other musicians. In the following sections we explain how melodic, rhythmic and harmonic descriptors are defined and extracted from the score.

4.1.1 Melodic Descriptors

We include the nominal pitch of the reference note and a set of melodic descriptors to characterize the melodic line of the target voice. The *Intervallic contour* (see Table 4.1, 6th row) is represented by a collection of melodic intervals (signed difference of pitch semitones) for each couple of consecutive notes.



Figure 4.2: Prototypical Narmour structures.

Furthermore, we implement the *Narmour implication realization model* in each group of three consecutive notes (Narmour, 1992). The Narmour group $Narmour(x, y, z)$ of three consecutive pitches x , y and z describes how the expectation that is built over the sequence of pitches (x, y) is fulfilled by

Table 4.1: Horizontal context descriptors.

<i>Horizontal Context Descriptors (HCD)</i>					
	Descriptor	Units	#	Formula	Range
Nominal	Pitch	Semitones	1	p_0	$[36, \dots, 96]$ (corresponding to C2–C7 midi range)
	Duration	Beats	1	d_0	$]0, +\infty[$
	Melodic Charge	Circle of Fifth Steps	1	c_0	$[0, 1, 2, 3, 4, 5, 6]$
	Metrical Strength	Label	1	$\text{MetricalStrAt}(b_0)$	{‘strongest’, ‘strong’, ‘weak’, ‘weaker’, ‘weakest’}
Neighboring notes (± 2 notes context)	Narmour Group	Label	3	$\text{Narmour}(p_{-1}, p_0, p_1)$	{‘P’, ‘D’, ‘ID’, ‘IP’, ‘VP’, ‘R’, ‘IR’, ‘VR’, ‘P.’, ‘ID.’, ‘IP.’, ‘VP.’, ‘R.’, ‘IR.’, ‘VR.’, ‘none’}
				$\text{Narmour}(p_{-2}, p_{-1}, p_0)$	
				$\text{Narmour}(p_0, p_1, p_2)$	
	Intervallic Contour	Semitones	4	$p_1 - p_0$	$\{-60, -59, \dots, 0, \dots, +59, +60\}$
				$p_2 - p_1$	
				$p_0 - p_{-1}$	
				$p_{-1} - p_{-2}$	
	Rhythmic Contour	Ratio (positive real)	4	d_1/d_0	$]0, +\infty[$
				d_2/d_1	
				d_0/d_{-1}	
				d_{-1}/d_{-2}	
	Rest	Boolean	4	r_1	true/false
r_2					
r_{-1}					
r_{-2}					

the pitch z . We encoded this as class labels; of which there are 15 possible labels (see Table 4.1) to which we add an additional “none” label for the case that one of the three notes is a rest. Figure 4.2 shows some prototypical Narmour structures.

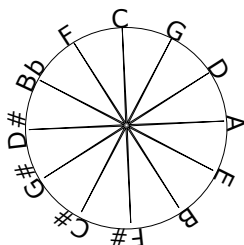


Figure 4.3: The circle of fifths. The pitch classes are organized by intervals of fifths. Notes further away from the tonic degree sound more dissonant, the distance can be counted as number of steps in the circle. As an example, in the tonality of C major, the note D has melodic charge 2 since there are two steps to reach D from C in the circle: $C \rightarrow G \rightarrow D$.

The *melodic charge* c is a descriptor of note salience, which is defined as the smallest number of steps required to reach the note in the *circle of fifths* (see Figure 4.3) from the tonic note (a number from zero to six). The melodic charge is considered a horizontal descriptor since it can be computed independently from the concurrent melodic voices by knowing the tonality alone.

4.1.2 Rhythmic Descriptors

Rhythmic information of the target voice was included by describing the durations of neighboring notes, relations to the meter and eventual pauses. We employed the metrical strength and rhythmic contour as well as a Boolean variable indicating the presence of neighboring rests (see Table 4.1).

Metrical strength depends on the position of the note relative to the bar meter and is encoded by a single class label from strongest to weakest depending on the beat position b_0 of the reference note. For the 4:4 meter we define the metrical strength, per case, as follows, where the first case that matches apply:

Table 4.2: Vertical context descriptors.

<i>Vertical Context Descriptors (VCD)</i>					
	Descriptor	Units	#	Formula	Range
Nominal	Rest	Boolean	3	$r_0^{(1)}$	true/false
				$r_0^{(2)}$	
				$r_0^{(3)}$	
Relative to Target	Inter-Beat Interval (IBI)	Beats	3	$b_0^{(1)} - b_0$] $-\infty, 0$ [
				$b_0^{(2)} - b_0$	
				$b_0^{(3)} - b_0$	
	Relative Pitch	Semitones	3	$p_0^{(1)} - p_0$	{ $-60, -59, \dots$ $\dots, 0, \dots$ $\dots, +59, +60$ }
				$p_0^{(2)} - p_0$	
				$p_0^{(3)} - p_0$	
	Relative Duration	Ratio (positive real)	3	$d_0^{(1)}/d_0$] $0, +\infty$ [
				$d_0^{(2)}/d_0$	
				$d_0^{(3)}/d_0$	
Resuming Descriptors	Harmonic Descriptors	Real, and a Boolean	2	HarmChAt(b_0)	$[0, 6$ [
				$c_0 \geq \max(c_0^{(1)}, c_0^{(2)}, c_0^{(3)})$	true/false
	Minimum IBI	Beats	1	$\max(b_0^{(1)}, b_0^{(2)}, b_0^{(3)}) - b_0$] $-\infty, 0$ [
	Minimum Relative Duration	Ratio	1	$\min(d_0^{(1)}/d_0, d_0^{(2)}/d_0, d_0^{(3)}/d_0)$] $0, +\infty$ [
	Maximum Relative Duration	Ratio	1	$\max(d_0^{(1)}/d_0, d_0^{(2)}/d_0, d_0^{(3)}/d_0)$] $0, +\infty$ [
	Minimum Relative Melodic Charge (minRMC)	Difference (integer)	1	$\min(c_0^{(1)} - c_0, c_0^{(2)} - c_0, c_0^{(3)} - c_0)$	{ $-6, -5, \dots$ $\dots, 0, \dots$ $\dots, +5, +6$ }
Maximum Relative Melodic Charge (maxRMC)	Difference (integer)	1	$\max(c_0^{(1)} - c_0, c_0^{(2)} - c_0, c_0^{(3)} - c_0)$	{ $-6, -5, \dots$ $\dots, 0, \dots$ $\dots, +5, +6$ }	

$$\text{MetricalStrAt}(b_0) = \begin{cases} \text{“strongest”}, & \text{if } b_0 \text{ is at the 1st beat of the bar} \\ \text{“strong”}, & \text{if } b_0 \text{ is at the up-beat (3rd beat of} \\ & \text{the bar)} \\ \text{“weak”}, & \text{if } b_0 \text{ is a quarter division of the bar} \\ & \text{(that is not already stronger)} \\ \text{“weaker”}, & \text{if } b_0 \text{ is an eighth division of the bar} \\ & \text{(that is not already stronger)} \\ \text{“weakest”}, & \text{in all other cases} \end{cases}$$

The pieces in our recordings have the following meters: 2:2, 4:4 and 3:4. We treated 2:2 and 4:4 in exactly the same way. In the case of the ternary meter we applied the same definition as 4:4 except that there are two up-beats in the 2nd and the 3rd positions. In the ternary meter, since quarter divisions are all marked as “strong” or “strongest” there are no “weak” notes, but there can be “weaker” or “weakest” notes.

Rhythmic contour was characterized by the ratios between consecutive nominal durations of neighboring notes. For example the sequence ♪♪♪♪ has rhythmic contour $[1, 4, \frac{1}{8}, 3]$. We applied this definition since multiplying each note value by a factor produces the same rhythmic pattern at a different speed. For example the rhythmic pattern of ♪♪♪ would produce the same the rhythmic contour as the one of ♪♪♪.

Employing a similar approach, we also introduced vertical rhythmic information derived from the other voices of the ensemble (see Table 4.2, *Relative Duration*). For each concurrent note, we computed the *relative duration* with respect to the reference note (ratio of nominal durations). In order to quantify the phase difference between notes we compute the *inter-beat interval* (IBI) as the difference of start times between the reference note and each of the concurrent notes (in beats). Additionally, as resuming descriptors, we computed maximum and minimum values across all the concurrent relative durations and the minimum IBI.

4.1.3 Harmonic Descriptors

Harmonic relations between voices are very important in *ensemble* performances, especially where string instruments are concerned given that into-

Table 4.3: Score context of other voices.

Score Context of Other Voices (SCOV)					
Descriptor	Units	#	Formula	Range	
Neighboring notes (± 1 notes context)	Intervallic Contour	Semitones	6	$p_1^{(1)} - p_0^{(1)}$	$\{-60, -59, \dots, 0, \dots, +59, +60\}$
				$p_0^{(1)} - p_{-1}^{(1)}$	
				$p_1^{(2)} - p_0^{(2)}$	
				$p_0^{(2)} - p_{-1}^{(2)}$	
				$p_1^{(3)} - p_0^{(3)}$	
				$p_0^{(3)} - p_{-1}^{(3)}$	
	Rhythmic Contour	Ratio (positive real)	6	$d_1^{(1)} / d_0^{(1)}$	$]0, +\infty[$
				$d_0^{(1)} / d_{-1}^{(1)}$	
				$d_1^{(2)} / d_0^{(2)}$	
				$d_0^{(2)} / d_{-1}^{(2)}$	
				$d_1^{(3)} / d_0^{(3)}$	
				$d_0^{(3)} / d_{-1}^{(3)}$	
	Narmour Group	Label	3	$\text{Narmour}(p_{-1}^{(1)}, p_0^{(1)}, p_1^{(1)})$	$\{\text{'P'}, \text{'D'}, \text{'ID'}, \text{'IP'}, \text{'VP'}, \text{'R'}, \text{'IR'}, \text{'VR'}, \text{'P-'}, \text{'ID-'}, \text{'IP-'}, \text{'VP-'}, \text{'R-'}, \text{'IR-'}, \text{'VR-'}, \text{'none'}\}$
				$\text{Narmour}(p_{-1}^{(2)}, p_0^{(2)}, p_1^{(2)})$	
				$\text{Narmour}(p_{-1}^{(3)}, p_0^{(3)}, p_1^{(3)})$	
	Rest	Boolean	6	$r_1^{(1)}$	true/false
				$r_{-1}^{(1)}$	
				$r_1^{(2)}$	
$r_{-1}^{(2)}$					
$r_1^{(3)}$					
$r_{-1}^{(3)}$					

nation is not fixed to specific intervals but rather continuous. We computed the harmonic charge (Friberg, 1995) on all the notes active within the beat b_0 of the reference note. To compute the harmonic charge, the pitches of notes active within the beat are preliminarily collected in a list. By using the implementation from the open project music21¹ we constructed a chord with the list of pitches and then estimated its root note. We then computed the melodic charge of each note in the chord with respect to the root note. Lastly, we calculated the mean of the obtained melodic charges resulting in the value of harmonic charge $\text{HarmChAt}(b_0)$.

Other harmonic descriptors were provided by the *relative melodic charge*. We computed the difference in melodic charge of every concurrent note with the reference note. We stored the minimum and the maximum value of those differences (minRMC and maxRMC, see Table 4.2). The maxRMC descriptors has been used in the past to define *note salience* in the vertical context. In particular, Sundberg et al. (1989) extracted a *sync-line* from the polyphonic score by keeping only the most salient note at each instant. The most salient note was defined as the one note active at each moment for which $\text{maxRMC} \leq 0$.

4.2 Performance Analysis

The raw data acquired in the recording sessions can be analyzed in various ways. In this dissertation we extract the characteristics of each performed note. For this reason score-performance alignment is a necessary front-end procedure that we had to carry out on all of the analyzed recordings.

We begin the performance analysis with an analysis of the articulations acquired in the recordings S.1-S.5, which are used as our reference template dictionary of bowing articulations. We then present a score-alignment algorithm motivated by observations on the extracted note profiles. We optimize and evaluate the aligner algorithm on a subset of our dataset. Lastly, with the support of the semi-automatic score aligning procedure, we define a set of note expression parameters which can be extracted from audio, gestures and timing. We will study how musicians employed such expression parameters in the next chapter in the various pieces/conditions which were recorded in our dataset.

¹<http://mit.edu/music21/>

4.2.1 Note profiles

As a first approach to studying gestures employed by musicians, we performed a note profile analysis of the datasets S.1-S.5. We manually aligned musicians' performance for all the recordings in S.1-S.5, leading to a precise ground truth note onset/offset time of each note. This manual alignment was obtained marking on screen the onset and offset of each note on top of the sound waveform and a set of audio/gesture descriptors². Once the manual alignment was completed, we looked at the intra-note evolution of descriptors both in the audio and in the gestural domain. To have an homogeneous representation we re-sampled each note trajectory to 25 samples (this number was enough for our analysis but we could have used more); this allowed us to compare notes that would otherwise have different length in samples.

We group notes by musician, by articulation, by dynamics, and by duration. This results in a totality of 180 categories because we have four musicians (first violin, second violin, viola and cello), five articulations (staccato, legato, martelé, staccato on string and détaché), three dynamics ranges (piano, mezzo forte, fortissimo), and three durations (quarter notes, eighth notes and sixteenth notes). For each of the 180 categories (cat) we have recorded 32 notes, each represented by a 25-dimensional vector $v_i^{(\text{feat}, \text{cat})}$ for $i = 1, \dots, 32$ and for each feature (feat). The 32 notes within each category (cat) were performed with similar durations (in milliseconds). For this reason, the re-sampling resulted in negligible stretching factors among notes within each category. We thus define the *note profile* of every given category as the mean trajectory of descriptors. This consisted, in our case, in computing the mathematical center of mass of the 32 vectors in each category $v^{(\text{feat}, \text{cat})} = \frac{1}{32} \sum_{i=1}^{32} v_i^{(\text{feat}, \text{cat})}$.

We used the note profiles derived in each category to study evolution of intra-note performance descriptors in qualitative terms. Such analysis is far from a thorough modeling of bowing articulations because by averaging trajectories we remove many individual note details focusing only on the coarse characteristics of each category. For a thorough study of articulations in bowed instruments we refer the reader to Maestre's (2009) Ph.D. Thesis. The purpose of our qualitative analysis is motivating the score-alignment heuristics method that we will introduce later in this section. We are thus interested in understanding how each prototypical note is realized in terms

²The GNU General Public License software Sonic Visualiser was used for this purpose. Visit <http://www.sonicvisualiser.org> for more information.

of bowing gestures and audio descriptors. We first present note profiles in the velocity-force space, and then we consider features in the audio domain.

The velocity-force space

We focused on force and velocity as they are the most important bowing parameters defining note attack and sustain (Askenfelt, 1989). Figure 4.4a shows bow velocity and bow force profiles of an eighth note played mezzo forte staccato. In Figure 4.4b the same two functions in time are plotted as a 2-dimensional trajectory in a velocity-force space. This last representation is particularly convenient since the two parameters are notoriously linked for the production of sound in bowed instruments. For example, with zero force no string excitation is produced regardless of bow velocity; neither any string excitation can be produced with positive force but zero velocity (the result in this case being a stopping of any vibration of the string). Sound can thus be produced when the 2D trajectory travels across the inside of one quadrant and away from the axes in the velocity-force space. Indeed, in this case the bow transmits energy to the string.

Figures 4.5 to 4.7 show note profiles resulting from the performance of first violinist legato, martelé and staccato respectively. Similar trajectories are obtained for the other musicians and thus we omit here the complete set of plots. In the presented graphs the bow velocity is always positive because, before computing the center of mass, we changed sign to bow velocity trajectories for all the notes where bow velocity was prevalently negative (i.e. when the median bow velocity value was negative).

The difference between the three articulations is self-evident from the plots. The legato articulation is (in average) realized with constant force and velocity so that it appears as a steady point in the velocity-force space (see Figure 4.5). Notice that in the legato profiles any eventual force-velocity transition at note start/end was canceled out by the averaging. In the resulting note profiles, the amounts of force and velocity are modulated depending on note duration and dynamics. The martelé articulation is instead realized with clock-wise oval shapes starting at zero velocity and positive force. The ovals become bigger in louder notes and rounder in shorter notes (see Figure 4.6). The staccato articulation is similar to the martelé but the trajectories are sharper for short notes (see Figure 4.7) and force is zero in the last part of the note.

The purpose of note profile analysis was deriving general heuristics to be applied to a score alignment algorithm based on gestures. As it can be seen

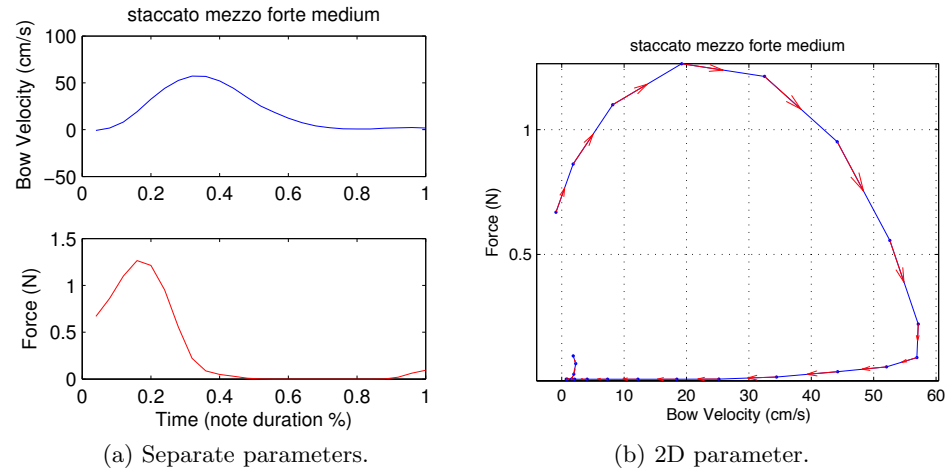


Figure 4.4: Intranote trajectory of bow velocity and force for a mezzoforte staccato note. The same trajectory in Figure (a) can be represented as the 2-dimensional trajectory in (b), which forms a clock-wise closed shape.

in Figures 4.5 to 4.7 note offsets are not well defined, but there are some general principles that we can apply to detect onsets by looking at velocity-force space trajectories. In the profiles of legato notes it is difficult to locate the onsets because we are looking at an average trajectory. However, bow direction changes (where the trajectory crosses the y-axis in the velocity-force space) represent very clear signs of onsets. In staccato and martelé onsets generally occur concurrently with a sudden rise of velocity and/or force. We have thus identified three cases where onsets are more likely to happen: force attack, direction change, and velocity attack. The three cases are graphically described in Figure 4.8 where each case corresponds to a drawing in the velocity-force space.

We implemented a simple detection mechanism of the above onset cases which is explained in details in Algorithm 1. In a first stage, the algorithm detects *onset candidates* based on velocity-force trajectory. Consequently, it defines the onset log-likelihood $g_{att}(i)$ at each frame i depending on the distance in time from the closest onset candidates. The g_{att} likelihood is a Gaussian mixture distribution over time with means at onset candidate instants. This code assumes a sampling frequency of 240Hz and consequently uses a window size of 4 frames (17 ms); however, the pseudo-code can easily be adapted to other sampling frequencies. In Figure 4.9 we show an example of the onset likelihood computed on an excerpt recording. In such

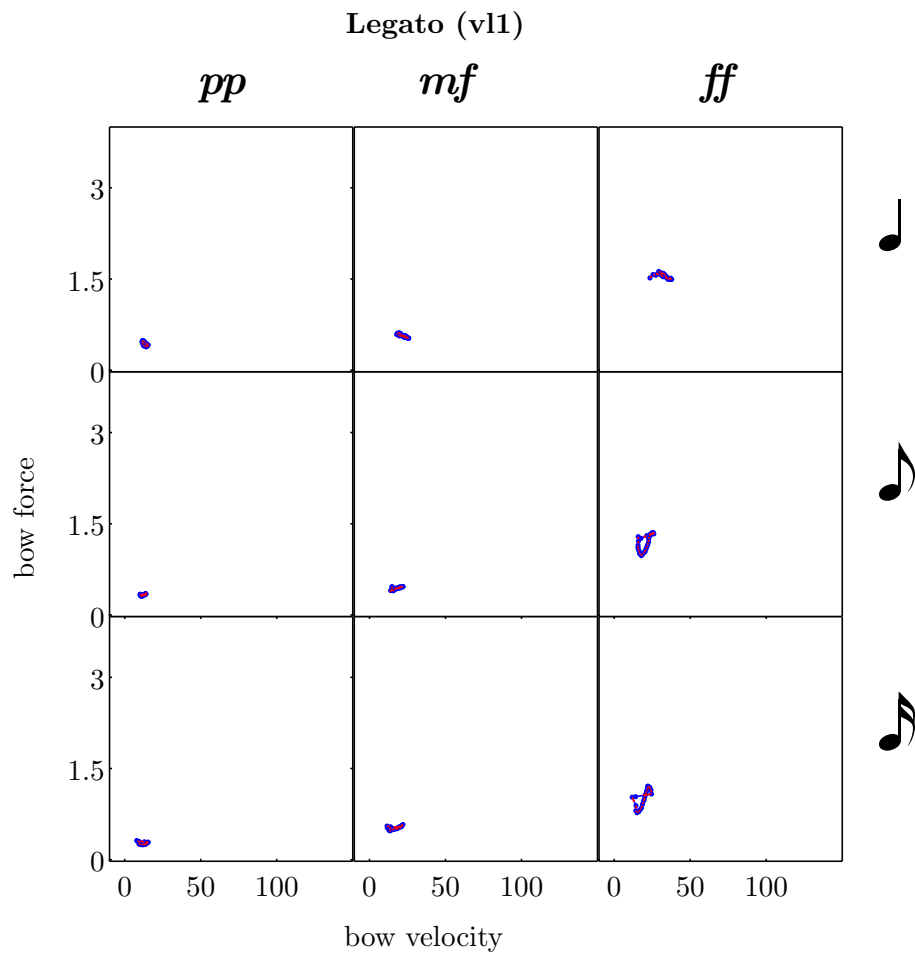


Figure 4.5: Legato mean note trajectories, violin 1.

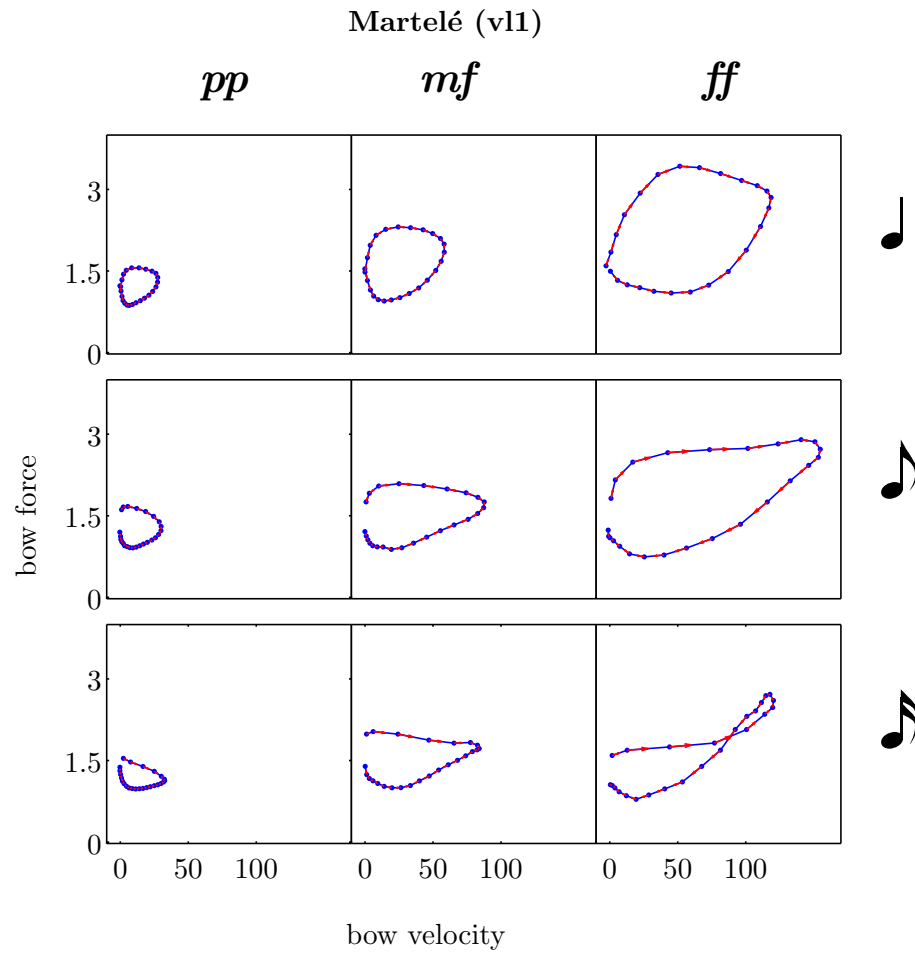


Figure 4.6: Martelé mean note trajectories, violin 1.

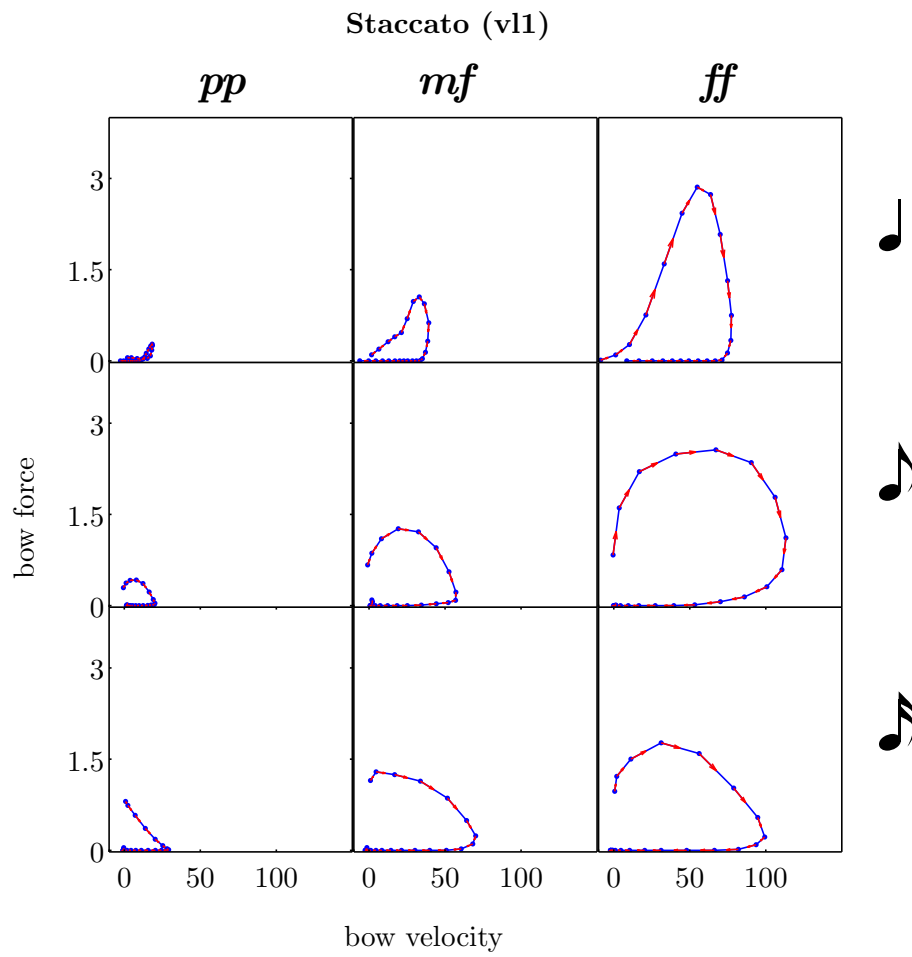


Figure 4.7: Staccato mean note trajectories, violin 1.

plot, the g_{att} likelihood has peaks near onset positions, which correspond to visible onsets in the sound waveform.

From our experience, a good aspect of the defined likelihood g_{att} is that within the onset candidates there are few false positives. False negatives are more common, for example in the case that notes are played legato, and thus two or more notes are played within one bowing direction. However, by mixing the g_{att} with other complementary information from the audio the algorithm is able to recover the missing onsets. Furthermore, each onset candidate is assigned with a different peak value in g_{att} depending on its type (force attack, velocity attack or direction change attack), which tells us how confident we can be that the onset candidate corresponds to an actual performed onset. Lastly, our heuristic formulation of likelihood could still be improved in various ways; for example by adapting the variance of each Gaussian around the candidate onsets depending on further bowing parameters such as bow acceleration.

Intra-note evolution of audio features

Whilst for bow motion it makes sense to compute profiles and show results on different articulations, for audio features we adopt another approach. Computing averages in this domain might, in fact, be counter-productive due to the higher variation of ranges (different pitches and sound level gains). Nonetheless, looking at the recording we could spot out some general and obvious principles. We explain here what we expect from chroma features and sound level at the beginning of a note and in its sustain part.

Not all the sound signal within the duration of a note is considered equally important. We only take into account a sustain period s of $s = 350$ ms between the frame t_k (the note beginning) and t_k^+ (the end of the sustain) or the beginning of the following note t_{k+1} (in the case that $\text{IOI} < s$).

Chroma Figure 4.11 shows five seconds of violin performance recorded with the pickup microphone. At the top, the sound waveform is shown with vertical lines marking the segmentation into notes (as obtained manually).

In Figure 4.11 we can compare HPCP output (see Section 3.3) with a template chroma generated from score pitch information. The template chroma $\text{ChrTempl}(p)$ for a set of pitches $p = \{h_1, h_2, \dots, h_{\tilde{n}}\}$ is a 120-dimensional vector defined as follows. Start from the zero vector v and then for each pitch height h_j (for $j = 1, \dots, \tilde{n}$) add to v a Gaussian shaped distribution (positive only within half semitone) around the bin location corresponding

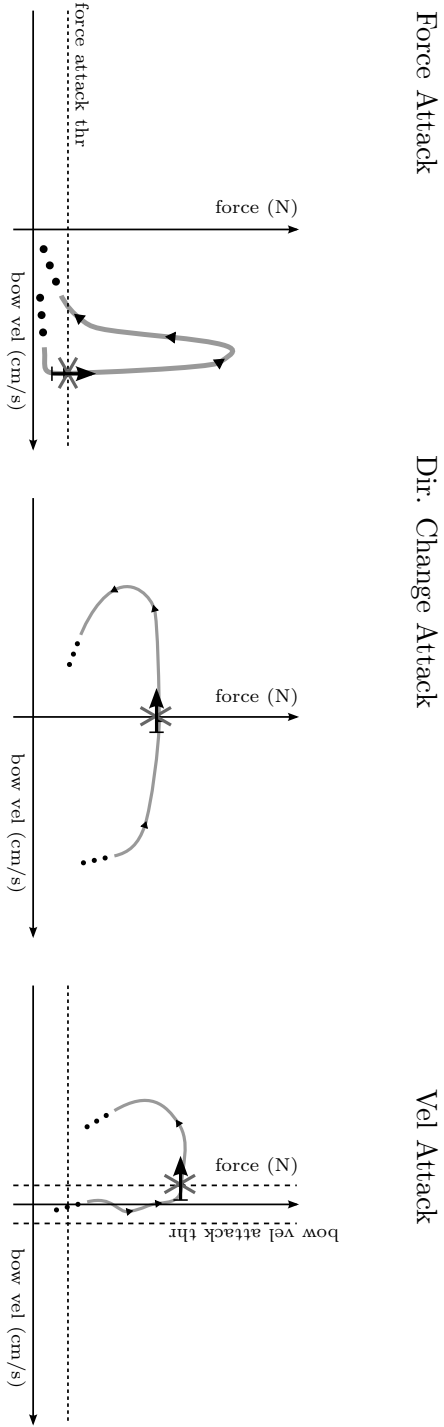


Figure 4.8: Detection of onsets using the gestural trajectory in the velocity-force space. The figure shows the three types of onsets defined and their relative thresholds for automatic onset detection.

Algorithm 1 Algorithm to extract the log-likelihood of onsets from bow velocity and force.

Require: $EBF(i)$ = Force at frame i

Require: $BV(i)$ = Velocity at frame i

for each frame i **do**

if $(EBF(i-2) = EBF(i-1) = 0) \wedge (EBF(i) > 0) \wedge (EBF(i+1) > 0) \wedge (|BV(i)| > velThr)$ **then**

 IsForceAttack \leftarrow true

else

 IsForceAttack \leftarrow false

end if

if $(\max(BV(i-2), BV(i-1)) < 0) \wedge (\min(BV(i), BV(i+1)) \geq 0)$ (or reversed inequalities and min-max signs) and $(|BV(i+1)| > velThr)$

then

 IsDirectionChange \leftarrow true

else

 IsDirectionChange \leftarrow false

end if

if $(\max(|BV(i-2)|, |BV(i-1)|) < velThr) \wedge (\min(|BV(i)|, |BV(i+1)|) \geq velThr)$ **then**

 IsVelocityAttack \leftarrow true

else

 IsVelocityAttack \leftarrow false

end if

if IsForceAttack **then**

$a \leftarrow \ln\text{Logistic}(|BV(i)|; \mu = 100, \sigma = 5)$

for each frame k **do**

$L_k^{(1)} \leftarrow L_k^{(1)} + a + \ln\text{Gauss}(k; \mu = i, \sigma = 1.85)$

end for

else if IsDirectionChange **then**

$a \leftarrow \ln\text{Logistic}(EBF(i); \mu = 0.1, \sigma = 0.005)$

for each frame k **do**

$L_k^{(2)} \leftarrow L_k^{(2)} + a + \ln\text{Gauss}(k; \mu = i, \sigma = 1.85)$

end for

else if IsVelocityAttack **then**

$a \leftarrow \ln\text{Logistic}(EBF(i); \mu = 0.1, \sigma = 0.005)$

for each frame k **do**

$L_k^{(3)} \leftarrow L_k^{(3)} + a + \ln\text{Gauss}(k; \mu = i, \sigma = 1.85)$

end for

end if

end for

for each frame i **do**

$g_{att}(i) \leftarrow \max(-20, L_i^{(1)}, L_i^{(2)}, L_i^{(3)})$ #

Minimum probability is -20 since there might be onsets without bowing attacks events.

end for

Extract from Detache' Violin 1

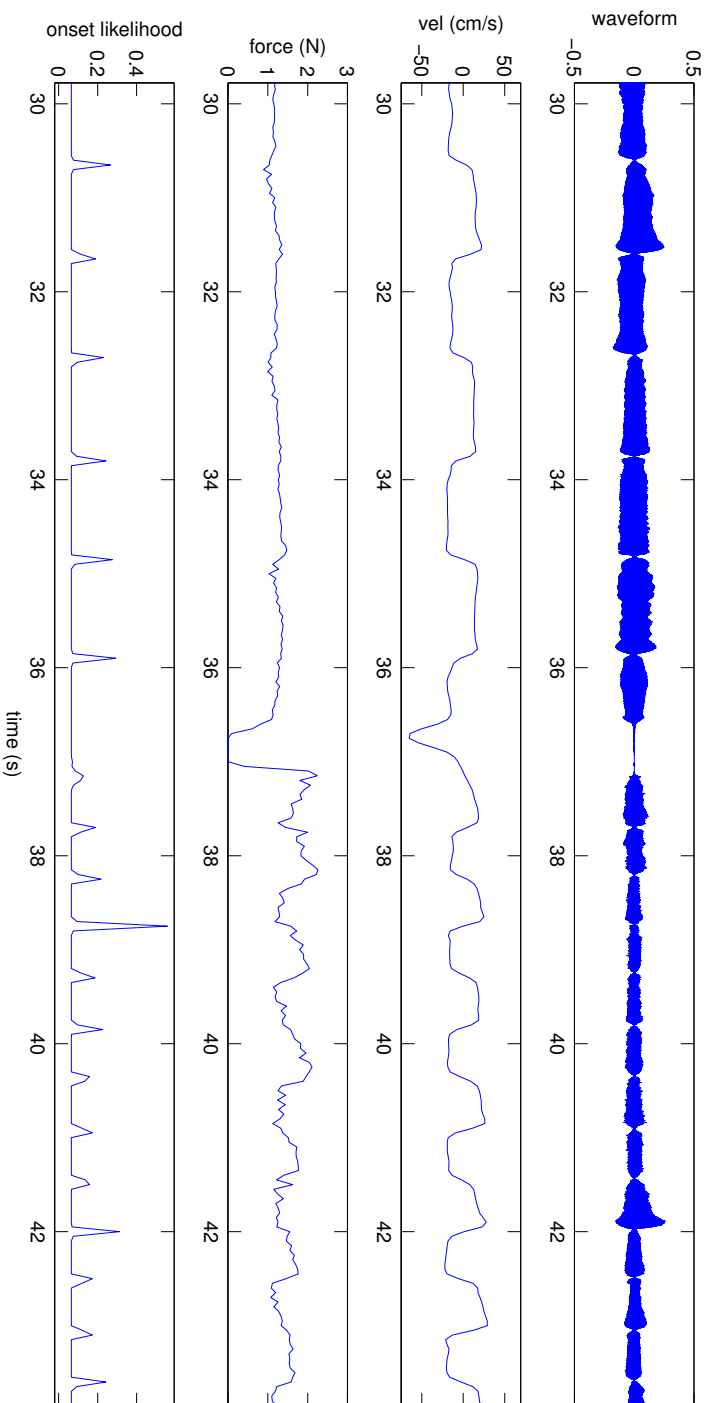


Figure 4.9: An example from the database showing bow velocity, force, and gesture onset likelihood along with the sound waveform. The gestures are here sub-sampled at 20Hz. The range for the likelihood is between zero and one since we do not apply here the logarithmic scale.

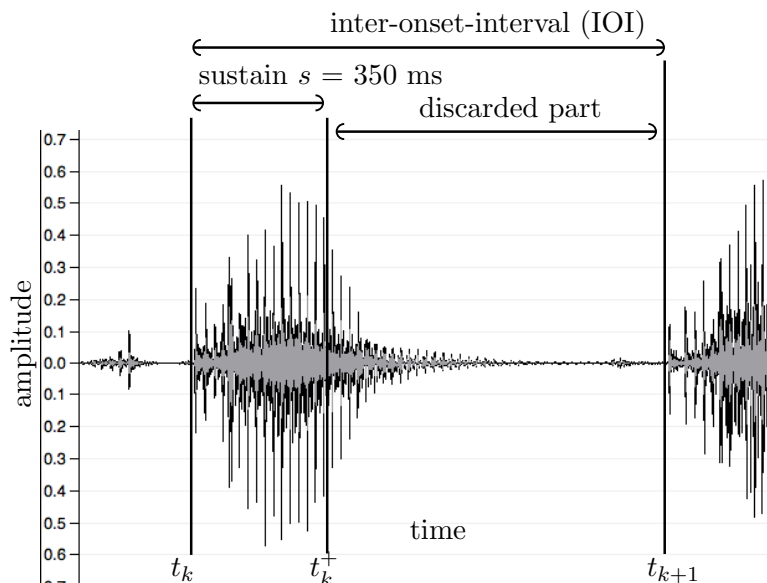


Figure 4.10: Sound waveform of a note with marked sustain period considered and the time instants t_k, t_k^+ and t_{k+1} .

to h_j (the vector v of pitch classes is circular so that if we arrive to the end position we have to continue to the first position). After adding the gaussian shaped bins, normalize v so that it sums to one, the result is $\text{ChrTempl}(p)$. To better illustrate the procedure, Figure 4.12 we represent the chroma template obtained for the C major chord $p = \{C, E, G\}$. At each frame i we can now check how well the set of pitches p fits the sound by computing the *Chroma Activation* of p at frame i as the dot product between measured chroma and template chroma:

$$\text{ChrAct}(i, p) = \langle \text{Chr}(i), \text{ChrTempl}(p) \rangle = \sum_{k=1}^{120} \text{Chr}(i)[k] \text{ChrTempl}(p)[k]$$

The chroma activation $\text{ChrAct}(i, p)$ is a value in the range $[0, 1]$ since both $\text{Chr}(i)$ and $\text{ChrTempl}(p)$ contain positive weight components and they each sum to one. The highest values are obtained if $\text{Chr}(i)$ and $\text{ChrTempl}(p)$ have activations of pitches at the same location of the octave and in similar proportions. In a good performance segmentation the template chroma of each note in the score should match with the corresponding audio chroma within the boundaries of the notes. However, a high matching is sometimes only measured in the *sustain period* after the note onset and not in the last

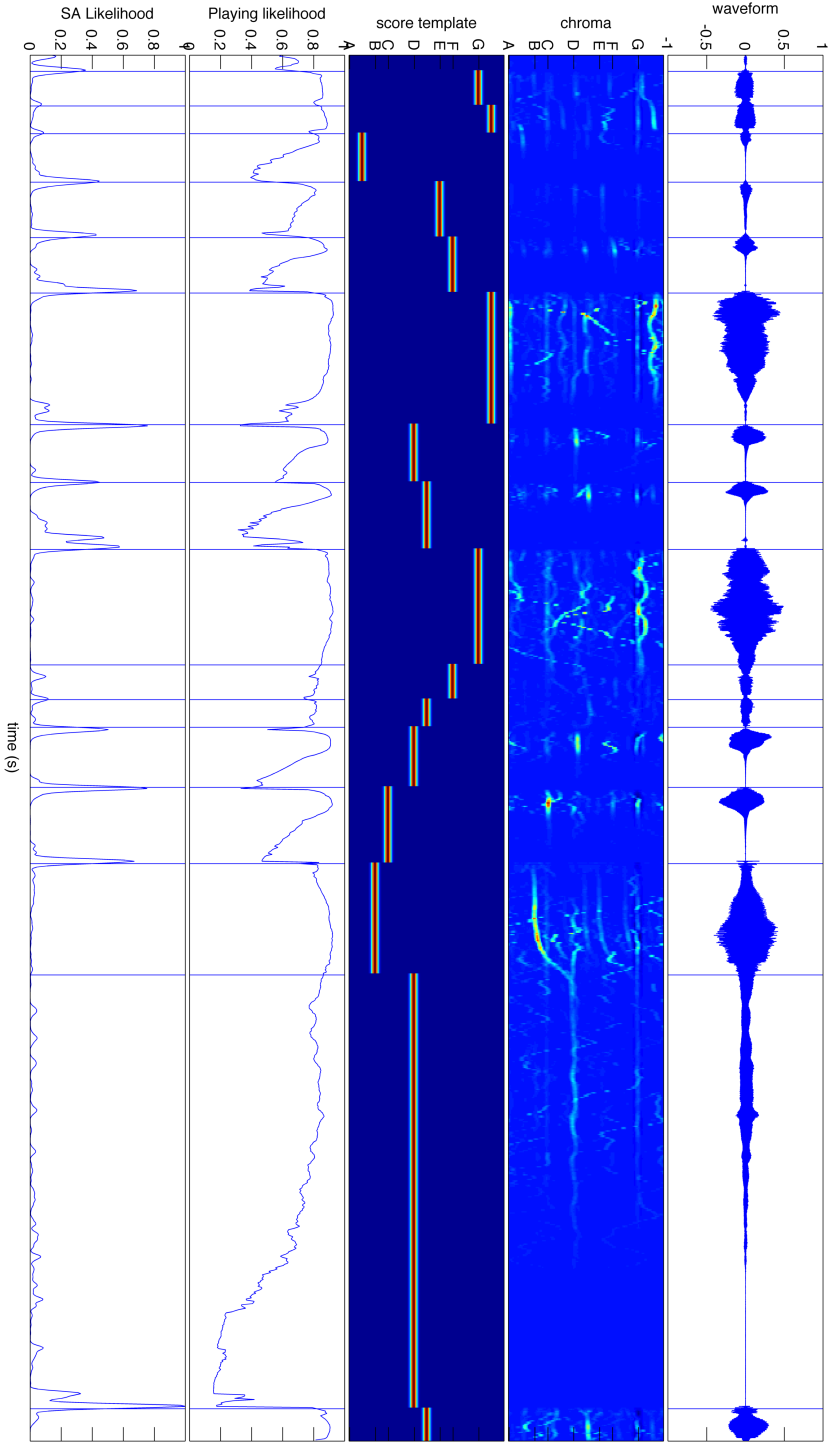


Figure 4.11: Some of the derived audio features and the score template for the chroma.

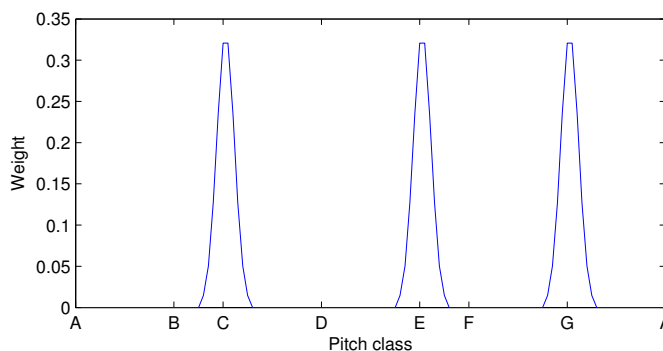


Figure 4.12: Template chroma vector $\text{ChrTempl}(p)$ for the C major chord, $p = \{C, E, G\}$.

part of the note. This is generally due to transitions to the next note which might be realized with a ‘fade away’ in loudness or a ‘glide’ in frequency. For example, in Figure 4.11 (on the second plot from the top) one can notice how the next-to-last note contains a glissando of one third when going from note B to note D. We set the duration of the sustain period to $s = 350\text{ms}$ for our alignment, corresponding to a duration of $s_f = 84$ frames (due to our fixed sampling rate of 240Hz). We expect the highest match between the template and the sound to happen during this period of time. We therefore define the *note chroma matching* $\text{ChrMatch}(p, t_k, t_{k+1})$ between the pitch class p and the audio as an average of the matching between the template chroma and the audio chroma during the sustain period. We calculate the end of the sustain period as $t_k^+ := \min(t_{k+1} - 1, t_k + s_f)$. For a note with pitch p starting at frame t_k ending at frame t_{k+1} the *note alignment chroma matching* is therefore defined as:

$$\text{ChrMatch}(p, t_k, t_{k+1}) = \frac{1}{t_k^+ - t_k + 1} \sum_{i=t_k}^{t_k^+} \text{ChrAct}(i, p)$$

Sound and silence Distinguishing between silence and sound is a first step in order to understand when a note is played or rather there is a pause or a transition. Ideally, we could compress the sound level value to just the two states *on* (playing) or *off* (not playing) since we do not care (at this stage) about the loudness of each note. However, such compression would result in discontinuities in the jumps from one state to the other which we

want to avoid. For this reason we define a *playing likelihood* signal s_{sus} as:

$$s_{\text{sus}}(i) = \text{lnLogistic}(\text{SL}(i); \mu = \text{silenceThr}, \sigma = \text{compressionFactor})$$

where we define the $\text{lnLogistic}(x; \mu = m, \sigma = s)$ as the logarithm of the sigmoid function with parameters m and s :

$$\text{lnLogistic}(x; \mu = m, \sigma = s) = \ln \left(\frac{1}{1 + \exp\left(-\frac{(x-m)}{s}\right)} \right)$$

in which the silence threshold is set manually depending on the instrument and the pickup (usually we set it to -55 dB under audio clipping level), and the compression factor controls how wide the border region between playing and not playing is (we usually set it to 15 dB). An example value of s_{sus} is shown in the fourth plot of Figure 4.11. For reasons analogous to the case of chroma, during the sustain period we expect the value of s_{sus} to stay near one (i.e. there is a sustained sound). Mathematically we can express this condition by stating that the value of $\text{SustainMatch}(t_k, t_{k+1})$ is globally maximized for each note, where:

$$\text{SustainMatch}(t_k, t_{k+1}) = \frac{1}{t_k^+ - t_k + 1} \sum_{i=t_k}^{t_k^+} s_{\text{sus}}(i).$$

We expect the SustainMatch to be close to one for pitched notes. In the case of pauses we instead expect the mean of $s_{\text{sus}}(i)$ within the whole pause boundaries to be close to zero:

$$\text{PauseMatch}(t_k, t_{k+1}) = \frac{1}{t_{k+1} - t_k} \sum_{i=t_k}^{t_{k+1}-1} s_{\text{sus}}(i).$$

A further observation on s_{sus} is that notes usually start on a rise of sound level which can be detected using the derivative of s_{sus} . This consideration is well known in the literature of the most basic onset detector algorithms. By taking the positive part of the derivative of s_{sus} and then convolving it with a Gaussian window we obtain the signal displayed in the last sub-plot of Figure 4.11 and which we refer to as δs_{sus} .

4.2.2 Score-performance alignment via dynamic programming

For the analysis of expressive performance it is fundamental to have a precise score alignment. Here we describe a score-performance alignment algorithm

using audio and motion capture data together with score symbolic data to automatically obtain note-by-note segmentation of performance. All of the corpora studied in this dissertation were score-aligned using this algorithm. Manual inspection was performed on each aligned performance, and wrongly aligned notes were corrected manually for the analysis of Chapter 5.

In this dissertation we employ a dynamic programming algorithm inspired by Maestre (2009), although the complete procedure was revisited here. We provide in the following sections a description of the algorithm in mathematical formalism and an evaluation on real recordings. We show that introducing bow motion data improves the alignment respect to a baseline approach of only audio features.

General overview of the score-alignment algorithm

The problem of score-alignment can be formalized as follows. The recorded performance is a list of sampled features $(X_t)_{t=1,\dots,m}$ where $X_t \in \mathcal{X}$, being \mathcal{X} the set of all possible feature states (e.g. each X_t can be a vector containing sound level, chroma, and g_{att} at time t or more features). The performance is sampled at $f = 240\text{Hz}$ and the duration of the recording is thus $\frac{m}{f}$ in seconds³. The score is a list of n note events, $(t_i, t_{i+1}, p_i)_{i=1,\dots,n}$ where $t_i > t_{i-1}$ are the note onsets and offsets as obtained from the duration of the notes in beats converted to frames at average tempo (the average tempo is the tempo value such that the starting note onset and the ending note offset are correctly aligned while the remaining notes in the middle respect score durations proportionally in time). Each p_i is a list of pitches: empty for pauses, with a singleton for monophonic notes, and with more than one pitch for chords.

Our system of score-performance alignment does not need to address the problem of polyphony since we align each musician independently of the others. In many classical scores for string quartet, each individual part is *homophonic*; in other words, it is either monophonic or notes belonging to each chord are rhythmically dependent (they all start and end simultaneously). This assumption was always verified in our corpora, except for small passages with tied arpeggios⁴. Another important assumption is that

³Here we assume for simplicity that all the features are acquired at 240Hz. Sound was acquired at 44100Hz, but for the alignment we only used the derived audio features which have a much lower sampling frequency.

⁴In a tied arpeggio the notes of a chord are played sequentially. Each note is held to the next so that when the last note is hit the full chord is sounding. This can be realized

musicians play the *exact same* notes dictated by the score. In the context of classical music this is usually the case unless musicians make mistakes or there is an error in the digital score. In our case we had to correct the score several times before it would be exactly as the one performed. Most of the errors were arising due to problems of digitalization of the score (e.g. missing sharp or flat marks) rather than musicians' mistakes. In very few cases we found different played pitches respect to the ones dictated by the score although musicians used different pitches which had musical sense. In those few cases we manually "corrected" the score pitch so as to fit with the musicians performance, otherwise the automatic alignment would produce undesired mis-aligned segments.

The problem of finding the score alignment is formalized as a problem of maximizing the likelihood of a candidate alignment given the recorded observations. The likelihood of an alignment $(t_i^*)_{i=1,\dots,n+1}$, given the observations $(X_t)_{t=1,\dots,m}$, is the product of the likelihoods of each individual note alignment:

$$P((t_i^*)_{i=1,\dots,n+1} | (X_t)_{t=1,\dots,m}) = \prod_{i=1}^n P(t_i = t_i^*, t_{i+1} = t_{i+1}^* | (X_t)_{t=1,\dots,m}) \quad (4.1)$$

where

$$\log(P(t_i = t_i^*, t_{i+1} = t_{i+1}^* | (X_t)_{t=1,\dots,m})) = L_i(t_i^*, t_{i+1}^*) \quad (4.2)$$

The most likely score-alignment $(t_i^*)_{i=1,\dots,n+1}$ can thus be found by maximizing the likelihood in Equation (4.1). The maximum can be found through a Viterbi algorithm. In a forward stage, the algorithm proceeds sequentially frame-by-frame computing the likelihood of each possible note and each possible duration⁵ (see Algorithm 2). In this way it writes two matrices M and D storing the best accumulated likelihood and the corresponding duration, respectively. In Figure 4.13 we plot one part of the matrix M along with the final detected alignment obtained from a real recording. In a second stage, the two matrices are used to find the most likely alignment through a backtracking procedure (see Algorithm 3). The described Viterbi algorithm has a complexity of the order of $O(nm^2)$, although if we know that the deviation of the alignment to the score is bounded by a constant ($\|t_i - t_i^*\| < \maxdev \forall i$) we can reduce the complexity to $O(n)$ since we do not need to compute all of the possible durations and onset positions.

on bowed instruments using a different string for each note.

⁵The notation $\lfloor x \rfloor$ (in Algorithm 2) refers to the *round down* (or floor) integer of x . Analogously, $\lceil x \rceil$ refers to the *round up* (or ceil) integer of x .

Algorithm 2 Forward step algorithm to accumulate the likelihood sequentially.

Require: d_i the score duration (in frames) of each note i

Require: $L_i(x, y)$ the likelihood of the i -th note alignment as a function of start-time x and end-time y as defined in Equation (4.5).

Initialize the matrices M and D

$M(k, i) = -\infty$ for each frame k , each note i

$D(k, i) = 0$ for each frame k , each note i

for each frame $k = 1, \dots, m$ **do**

for each note $i = 1, \dots, n$ **do**

$d_{\min} \leftarrow \lfloor 0.4 \times d_i \rfloor$

$d_{\max} \leftarrow \lceil 1.7 \times d_i \rceil$

avoid negative onset times

$d_{\min} \leftarrow \min(d_{\min}, k - 1)$

$d_{\max} \leftarrow \min(d_{\max}, k - 1)$

initiate best likelihood

$bestL \leftarrow -\infty$

for each duration $d = d_{\min}, \dots, d_{\max}$ **do**

$t \leftarrow k - d$

if $i = 1$ **and** $t \neq 1$ **then**

continue

else

likelihood of i -th note starting at t and ending at k

$l \leftarrow L_i(t, k)$

$accumL \leftarrow l + M(t, i - 1)$

if $accumL > bestL$ **then**

$bestL \leftarrow accumL$

$bestD \leftarrow d$

end if

end if

end for

$M(t, i) \leftarrow bestL$

$D(t, i) \leftarrow bestD$

end for

end for

return M, D

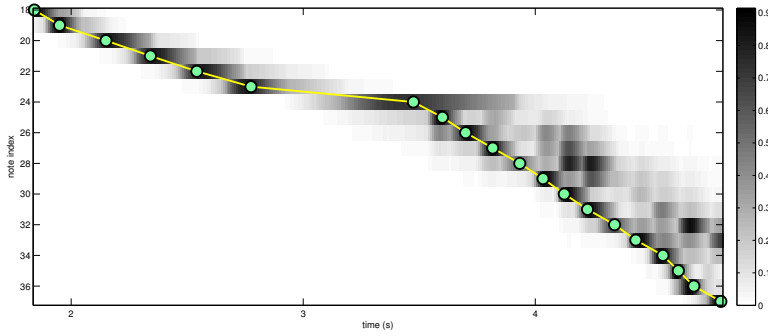


Figure 4.13: Visualization of the matrix of accumulated likelihood M of Algorithm 2. The dots connected by lines show the most likely alignment obtained in the backtracking process of Algorithm 3.

Algorithm 3 Backtracking step to find the most likely alignment.

Require: M, D as returned from Algorithm 2.

```

 $t_{n+1}^* \leftarrow m$ 
for  $i = n, n - 1, \dots, 2, 1$  do
     $d = D(t_{i+1}^*, i)$ 
     $t_i^* = t_{i+1}^* - d$ 
end for

```

The Viterbi algorithm computes the *log-likelihood* of each note L_i (i.e. the term appearing in the right-hand-side product-sequence of Equation (4.1)) depending on candidate note onset and offset positions x and y . Thus L_i needs to be defined for each note and each possible onset and offset. To define L_i we used an heuristic approach motivated by the previous observations on note profiles. Firstly, we assume that the likelihood of having the i -th note starting at frame x and ending at frame y (where x and y are integers satisfying $1 \leq x < y \leq m$) only depends on the observed descriptors recorded in the temporal range $x, x + 1, \dots, y$. This formally translates into the following identity:

$$L_i(x, y) := \log P(t_i = x, t_{i+1} = y | (X_t)_{t=1, \dots, m}) \quad (4.3)$$

$$= \log P(t_i = x, t_{i+1} = y | (X_t)_{t=x, \dots, y}) \quad (4.4)$$

We thus have to define the last term of Section 4.2.2, i.e the probability that the i -th note coincides with the recorded segment $(X_t)_{t=x,\dots,y}$. To define the likelihood of each note, we heuristically combined aspects that increase the possibility of having a specific note event. We consider such aspects as independent variables indicating that the data contain a particular note event at a particular time. The considered aspects are detailed in Table 4.4. Most of the aspects only apply to pitched note events while some other apply to pauses.

Consequently, we define the final log-likelihood of each note as the following weighted list of descriptors:

$$L_i(x, y) = \sum_{r=1}^9 w_r l_r(x, y) \quad (4.5)$$

In Table 4.4 we defined a set of heuristic terms quantifying the likelihood of each aligned note based on a set of audio and gesture descriptors. We have manually defined some parameters such as the μ and σ of each distribution. Such parameters depend on the particular range of the descriptors used; however, it is less clear how to assign values to the set of weights w_r that defines the balance among factors in Equation (4.5). In the following section we detail the procedure to obtain such weights w_r thorough an optimization procedure.

Likelihood weights optimization

The balance among factors w_r of the previous section is very important for achieving a stable and reliable score alignment. We provide here an example of how such factors work together. In case too much importance is given to factor w_1 , then the algorithm would seek solutions where note onsets are aligned to corresponding to the peaks of δs_{sus} . However, whereas some notes barely display any loudness attack at all, some peaks of δs_{sus} do not correspond to note onsets. This bias towards δs_{sus} peaks are likely to produce irregular timing respect to score. In this sense the weights w_5 and w_6 help preventing this issue by penalizing irregular tempi. On the contrary, if too much importance is given to tempi weights, notes would get assigned same durations as score regardless of any observation of loudness, chroma and gesture. The difficulty here is finding the optimal balance of weights w_r , which we empirically seek on real data as explained below.

We run some preliminary tests guided by our intuitions to define a set of meaningful weights. We therefore run an iterative optimization procedure

Table 4.4: Heuristic formulation of the terms in Equation (4.5)

Term	Applies to	Type	Observed variable x	Log-likelihood	Explanation
l_1	Note	Audio	$x = \delta_{\text{Sust}}(t_k)$	$\text{InLogistic}(x; \mu = 10.0, \sigma = 3.0)$	Rise in sound level at note onset time.
l_2	Note	Audio	$x = \text{SustainMatch}(t_k, t_{k+1})$	$\text{InLogistic}(x; \mu = 0.4, \sigma = 0.01)$	Sound sustained within sustain period of pitched note.
l_3	Note	Audio	$x = \text{ChrMatch}(t_k, t_{k+1})$	$\text{InLogistic}(x; \mu = 0.15, \sigma = 0.05)$	Chroma matches the score in sustain period.
l_4	Pause	Audio	$x = \text{PauseMatch}(t_k, t_{k+1})$	$\text{InGauss}(x; \mu = 0.0, \sigma = 0.05)$	Silence within pause boundaries.
l_5	Note/Pause	Timing	$x = r_i = \frac{t_k - t_{k+1}}{t_{i+1} - t_i}$	$\text{InLogNormal}(x; \mu = 0.0, \sigma = 0.15)$	Performed duration is close to score duration.
l_6	Note	Timing	$x = \frac{r_k}{r_{i-1}}$	$\text{InLogNormal}(x; \mu = 0.0, \sigma = 0.15)$	Duration of previous note close to current note if notes have the same durations. In other cases the ratio should be close to the one prescribed by the score.
l_7	Note	Gesture	$x = \text{gatt}(t_k)$	x	Gesture attack log likelihood at onset position
l_8	Note	Gesture	$x = \frac{1}{t_k^+ - t_k^- - 1} \sum_{i=t_k^-}^{t_k^+} \text{gatt}(i + 1)$	$-x$	No gesture attack event in the sustain period

starting from the initial weights. The procedure is based on *coordinate descend*; the algorithm modifies one weight at the time and verifies how this impacts on the mean alignment error (the root mean squared deviation):

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n+1} (t_i^* - t_i)^2}$$

After having tested the aligner with a set of neighboring weight values, the best one is kept and the procedure is repeated on the second weight. When all the weights have been processed, the algorithm starts again from the first. If no improvement is found on a complete cycle around all the weights the step-size is reduced.

We run the optimization for the first violinist jointly on the databases S.1-S.5. Since this algorithm needs to execute the aligner many times we had to make it fast enough for completing the calculation in a reasonable amount of time. For this reason, we under-sampled all of the descriptors to a common sampling frequency of 20 Hz after applying a Gaussian convolution to each descriptor to remove high frequency components higher than 10 Hz (the corresponding Nyquist frequency).

We repeated the optimization twice: once using the whole set of weights w_r for $r = 1, \dots, 8$, and once using only audio and timing descriptors, thus fixing to zero the weights w_7 and w_8 . In Figure 4.14 the RMSD value is showed as a function of the iteration number. The algorithm improved the results both with or without gestures; however, using audio and gesture was always better than audio only.

Evaluation on piece extracts

We tested the final weights derived by the optimization algorithm on another dataset derived from the real pieces. We selected three short excerpts with heterogeneous characteristics from the Beethoven' piece (Bars 1–16, 32–79 and 79–95). We manually aligned the corresponding performance of first violin in the recordings I.1 (mechanical) and I.3 (exaggerated) and used them as a test set. We run the aligner on this test set using the weights obtained in the previous optimization procedure, we also tested other sampling rates since in this case we did not have the same restrictions of computational time (i.e. we run the algorithm only once).

In Figure 4.15 we compare the results of the two aligning methods obtained at different sampling rates. Note how using gesture descriptors improves

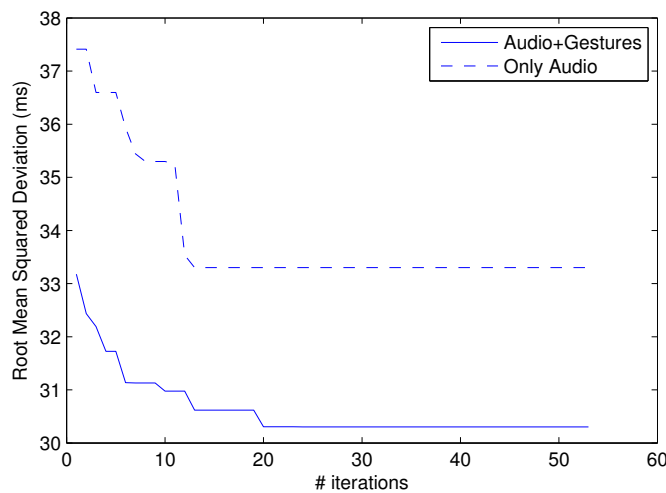


Figure 4.14: Root mean squared deviation for the performance sampling database a number of iterations has been executed. We compare the results obtained using only audio with the results obtained by also employing gestural data.

the alignment at all the sample rates. Additionally we can observe that the same weights we have derived using a sample rate of 20Hz consistently produce better results at higher sampling rates. This means that we have defined the problem in a way that it can work at different sampling rates. At 120Hz, although the result is still better than with 20Hz, we have a decrease in performance respect to 80Hz. There might be two reasons explaining such a performance decrease. The first is that the algorithm might still require a minor adjustment of the parameters depending on the sampling rate. The second reason is that down-sampling was preceded by a Gaussian smoothing and this might improve the results because of an improved blend in time among the different likelihood components.

This completes the discussion on the semi-automatic alignment procedure used for building our datasets. Once the score-performance alignment procedure was completed we extracted several expression parameters for each aligned note. We explain such parameters in the following section.

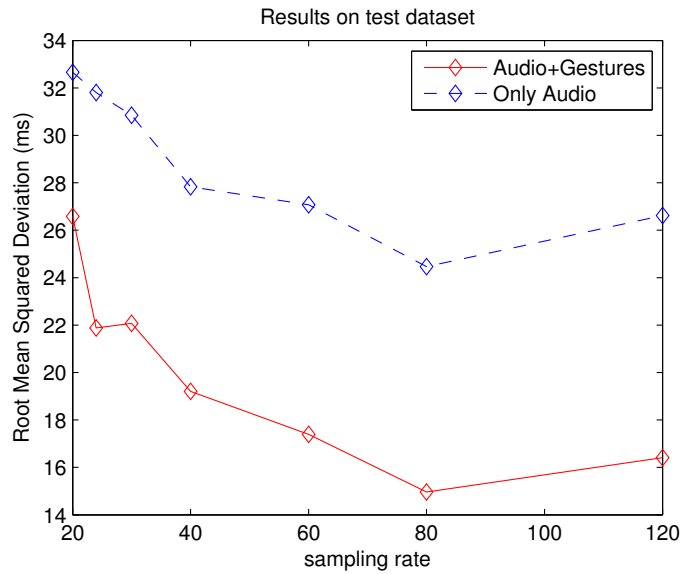


Figure 4.15: Root mean squared deviation obtained by running the aligner on the test dataset. We used the weight parameters obtained from the two optimizations on the performance sampling dataset: both audio only and audio plus gestures. Keeping the same weight parameters we perform several alignments using different sampling rates.

4.2.3 Performed expression parameters

In Chapter 3 we have shown several low level features extracted from audio and motion data. Additionally, in the previous section, we have shown how to use the extracted features to derive automatically a score performance alignment. Once the score alignment procedure has been completed and the alignment has been manually verified we extract the performed expression parameters. Such parameters estimate higher level music descriptors about deviations that musicians introduced in the performance, and are thus the core subject of study in this dissertation.

In the context of ensemble performance, we can make a distinction between *individual* and *collective* expression parameters. For example, as will be explained below, tempo curve can be computed on the performance of one specific musician (individual) or on the performance of the whole quartet (collective).

The possibility of computing both individual and collective expression parameters allows the further step of computing *participatory discrepancy* parameters. Participatory discrepancy parameters have been studied in the literature (Keil, 1987; Prögler, 1995). One of the most popular example of such parameter is swing factor, which is usually computed with respect to an external reference (e.g. a metronome) rather than from an estimated collective expression parameter. As will be explained later we also distinguish between a collective *beat asynchrony* (the amount of asynchrony among musicians on common note attack) and an individual *participatory asynchrony* (amount of anticipation/deferral of note onset time respect to collective tempo, this is similar to the mentioned swing factor).

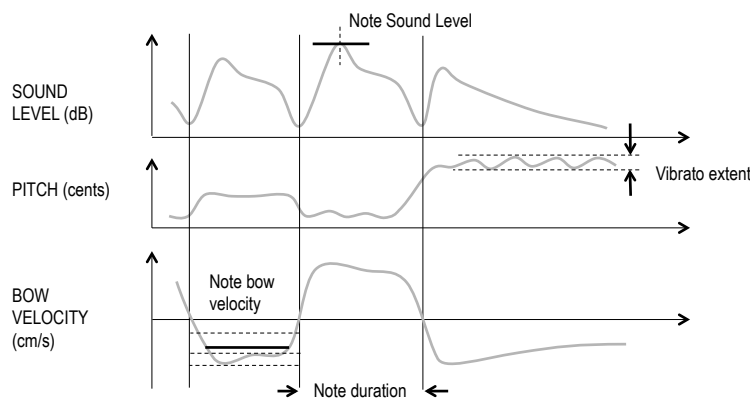


Figure 4.16: An example of the Sound Level, Pitch and Bow Velocity time series along with the descriptors extracted at the note level.

There is no sole way of extracting expression parameters, nor there is yet a set of parameters that can be considered complete. In this thesis dissertation we focused on few expression parameters related to timing, dynamics and articulations. We present here the extracted parameters divided into timing-based, audio-based and gesture-based.

Timing-based expression parameters

Timing-based expression parameters results from note onset-offset time positions of each note as obtained in the alignment procedure. The score dictates note nominal durations in beats. In performance time the beat duration can change during the course of the performance resulting in perceived *tempo changes*. Also abrupt changes of note duration can be mea-

sured in the performance and are instead perceived as *phase changes* since they do not accumulate in the long term to produce a change in tempo (Repp, 2005). There are many ways in which this data can be employed to derive meaningful expression parameters. We describe here *note lengthening*, *tempo curve* and *asynchrony*.

Note lengthening We consider note lengthening as being relative to the local tempo of the ensemble. We define *note lengthening* as the ratio between the actual performed duration and the nominal duration of a note (defined in second by applying the local tempo of the ensemble). We chose to consider duration as being such in an attempt to remove the effect of broad tempo modulations and focus only on residual timing deviations, as it is often done in studies of expressive performance (Widmer and Tobudic, 2003).

A necessary pre-processing step is estimating ensemble tempo for each note in a local context, thereby making it possible to express a nominal duration in seconds. A first approach to estimating the tempo would be to divide the total number of beats recorded by the total performance duration in seconds to derive an average tempo. However, this approach totally neglects possible changes of tempo, which might occur during the performance. For this reason we divided each piece into phrases⁶ with the help of a professional musicologist and looked for possible changes of tempo. By looking at the distribution of note durations in each performance it became evident that abrupt changes of tempo appear at section changes. On the other hand, tempo stays nearly constant within sections (see next section and Figure 4.17 for an example). For these reasons, we chose the average tempo of the section as the reference tempo for each note.

The note lengthening is computed via the formula:

$$NL_k = \frac{\text{notePerformedDuration}}{\text{noteScoreDuration}} \times \text{sectionBPS}$$

where `notePerformedDuration` is the performed duration in seconds, `noteScoreDuration` is the nominal duration of the note in beat units and `sectionBPS` is the average beats per second of the section in which the note is played. Where the note was followed by a pause we included the duration

⁶This procedure led to 128 phrases from the four excerpts (P.1-P.4, see Section 3.1) jointly and 53 phrases for each of the expressive intention piece (I.1-I.3). The phrases were hierarchically linked so as to merge in higher time spans, up to whole sections of the piece.

of the pause in the note enabling us to use the inter onset interval (IOI) to estimate `notePerformedDuration`.

It is worth mentioning that in the case of *ensemble* recordings we used the same `sectionBPS` of the four musicians. On the other hand, for the *solo* recordings we use a different value of `sectionBPS` for each musician since they each employed slightly different tempi.

Tempo curve Tempo curve is a complementary way to look at timing phenomena that transcends single notes. If, instead of looking at the duration of each note, we sum durations on groups of two, three or more notes we can compute a stabler beat-per-minute (BPM) tempo curve. There is not one sole way of computing tempo curve and in the literature several different techniques have been used. Many of the studies on tempo curves are based on beat tracking (either manual or automatic) of performances or tempogram methods (Cemgil et al., 2000; Chuan and Chew, 2007; Grosche et al., 2010).

While in most studies beat tracking is still more convenient, this approach is not enough to address some problems typical of ensemble performances. The first issue is how to compute the collective tempo curve based on the overall ensemble performance. This means that all the musicians contribute to collective tempo curve in an equal manner. The second and last issue is how to define the tempo curve in a way in which we can systematically study discrepancies among musicians respect to a common reference.

We give two operative definitions of tempo curves which we will use for the analysis of macro-timing in Chapter 5, each addressing one of the problems above mentioned. The first definition is based on averaging tempo in larger time-spans than a single note. It regards tempo as a smooth continuous phenomenon by providing a definition of instantaneous tempo, at each instant of the performance. The second definition is based on the idea by Todd (1992) (also known as phrase-arc rule) of a tempo curve modeled by a parabola per phrase. It regards tempo as a continuous phenomenon within each phrase but allows a discontinuity at phrase boundaries. We refer to the first kind of tempo curve as *Smooth tempo*, and to the second as *Parabola tempo*.

Smooth tempo Given the a score onset sequence expressed in beats b_k and with corresponding performance onset times t_k , $k = 1, \dots, n$. If we assume the tempo to be constant within each note, we can compute an

instantaneous tempo as the following piecewise constant function:

$$\text{bpm}_0(b) = 60 \frac{b_{k+1} - b_k}{t_{k+1} - t_k} \quad \forall b \in [b_k, b_{k+1}[.$$

This curve has discontinuities at each note onset and is noisy due to the differences in duration of consecutive notes. In order to remove high frequency content, we convolve it with a Gaussian curve of variance $\sigma > 0$.

$$\text{bpm}_\sigma(b) = (\text{bpm}_0 \star \phi_\sigma)(b) = \frac{\int_{b_0}^{b_n} \text{bpm}_0(x) \phi_\sigma(x - b) dx}{\int_{b_0}^{b_n} \phi_\sigma(x - b) dx}$$

where $\phi_\sigma(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ is a Gaussian centered in zero of variance σ^2 . By changing the parameter σ we can control the time span used to compute tempo and thus view how tempo changes at different levels (see Figure 4.17).

Parabolic tempo and time curve Parabolic tempo is based on the assumption that there exists a continuous function $\tau : \mathbb{R} \rightarrow \mathbb{R}$, which we call *time curve*, that at each instant t in the performance time associates the corresponding beat position b in the score. The parabolic tempo curve is thus the time curve derivative function $\tau' = \frac{d\tau}{dt}$. In order to have a tempo curve with a parabola for each phrase $k = 1, \dots, \rho$, we impose that τ is a continuous piecewise cubic function:

$$\tau(t) = \begin{cases} c_{1,0} + c_{1,1}t + c_{1,2}t^2 + c_{1,3}t^3 & : p_0 \leq x < p_1 \\ c_{2,0} + c_{2,1}t + c_{2,2}t^2 + c_{2,3}t^3 & : p_1 \leq x < p_2 \\ \vdots & \vdots \\ c_{\rho,0} + c_{\rho,1}t + c_{\rho,2}t^2 + c_{\rho,3}t^3 & : p_{\rho-1} \leq x < p_\rho \end{cases}$$

where p_0, \dots, p_ρ are the phrase boundaries time positions. We also impose the following time curve the continuity constraints at phrase boundaries:

$$\tau(p_r) = \lim_{t \rightarrow p_r^-} \tau(t) \quad r = 1, \dots, \rho$$

We optimize the coefficients $c_{p,d}$, using a least square fitting procedure so as minimize the least square sum:

$$\sum_{k=1}^n (\tau(t_k) - b_k)^2$$

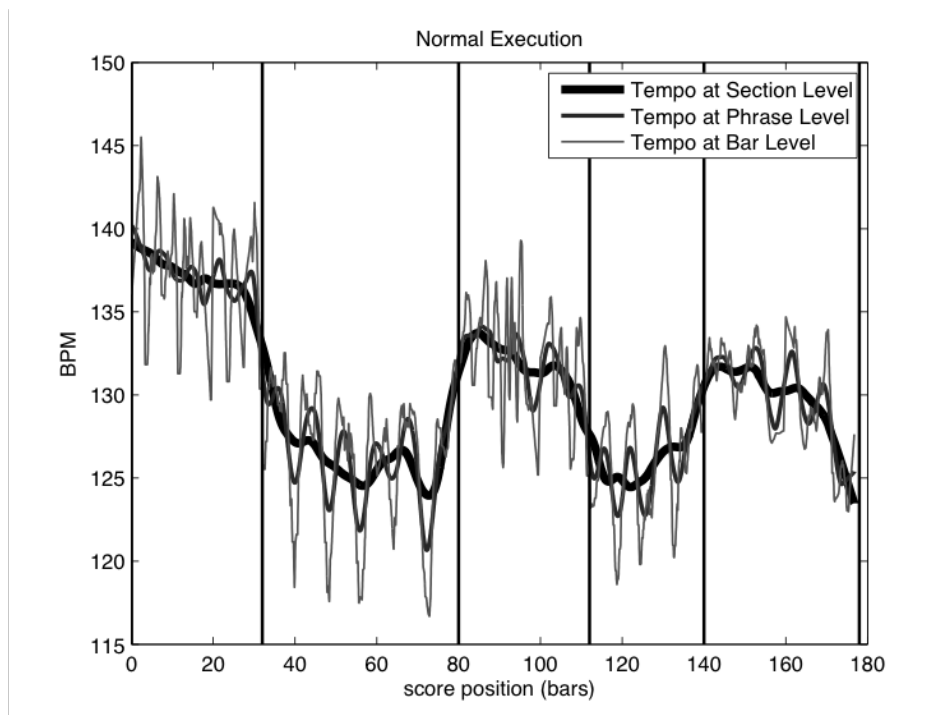


Figure 4.17: Collective tempo curve computed on the performance of recording I.2 with three different smoothing factors: at the level of the section, phrase, and bar. The vertical bars mark boundaries of the sections in the rondó structure.

where here the t_i and b_i in the sum are drawn from the performed notes by *all* the musicians (since we compute the collective time curve).

Several studies have shown how variations in tempo tend to follow similar rules as bodies subject to gravity (Todd, 1992), therefore resulting in parabolic tempo curves. In our study, we have chosen this representation since (1) it is flexible enough to fit the time changes adopted by the musicians, (2) it depends on just three parameters per phrase, and thus prevents over-fitting (3) it allows discontinuities of tempo at the boundaries of phrases. In Figure 4.18 we show the onset times by different musicians playing together four phrases from the I.1 performance, the fitted time curve is drawn on top of the raw onset data. The example of the plot shows an extreme case where musicians performed two fermate; nonetheless the

cubic interpolation manages this situation correctly and the discontinuities at phrase boundaries correspond quite closely to musicians timing.

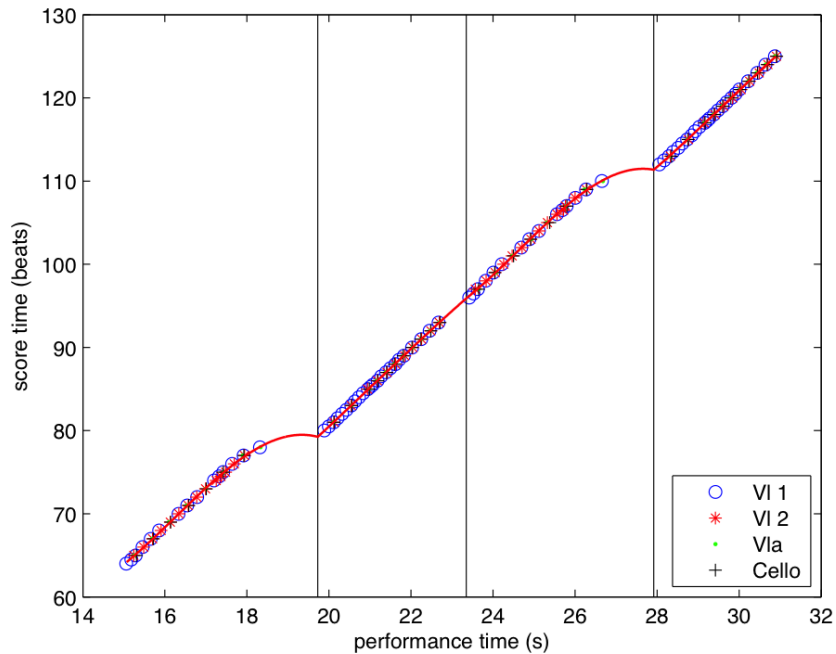


Figure 4.18: Representation of the onset times of the notes in performance time versus score time. The vertical lines mark phrase boundaries and the red line marks the collective time which was obtained with a locally cubic regression with imposed continuity at phrase boundaries. In the plot, two fermatas are evident at time 20, and at time 28 seconds.

Asynchrony and note participatory asynchrony We make a clear distinction between *beat asynchrony* and *note participatory asynchrony*. Beat asynchrony does not require any common time reference and it is only defined for group of notes having common score onset time. The amount of asynchrony of the group is then intuitively proportional to the amount of being spread out in performance time. We measure the beat asynchrony in each group of (at least 2) notes as the standard deviation of their perfor-

mance onset times t_1, \dots, t_n :

$$\text{BeatAsynch} = \sqrt{\sum_{i=1}^n (t_i - \bar{t})^2} \quad \text{where} \quad \bar{t} = \frac{1}{n} \sum_{i=1}^n t_i.$$

In Figure 4.19 we visually represent the idea of beat asynchrony, each group of notes played by different musicians at the same score position can be more or less spread out in performance time. The figure also represents with a red line the collective time as derived from the cubic interpolation although this line is irrelevant for the computation of beat asynchrony.

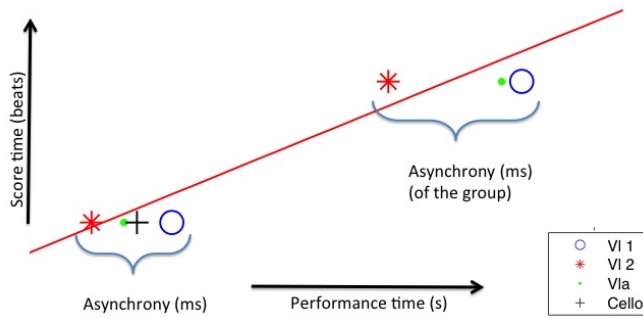


Figure 4.19: Visualization of asynchrony of each group notes with same score onset position. In red the collective time is displayed as obtained through the regression procedure.

Note participatory asynchrony, differently to previous descriptor, is defined for every note in the score with respect to the collective time $\tau(t)$. A note N with onsets at b^* in score beats and at t^* in performance has the following participatory asynchrony:

$$\text{ParticAsync}(N) = \tau(t^*) - b^*$$

The value is measured in beats and it can be seen also as a measurement of swing factor. The value can be either positive if note was anticipated respect to collective time, or either negative if note was delayed (see Figure 4.20).

Audio-based expression parameters

Note Sound Level Through the use of pickup microphones and the consequential acquisitions of each musician's sound in separate audio tracks we

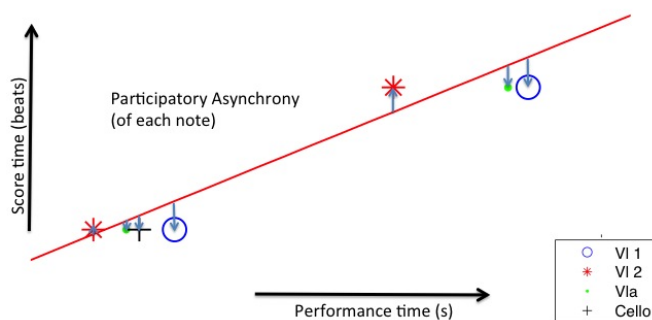


Figure 4.20: Visualization of participatory asynchrony note in the score. The participatory asynchrony is positive if the arrow points downward and negative if the arrow point upward. Its absolute value is the difference in milliseconds from the collective time which is displayed by the red line.

were able to define *note sound level* (NSL) in the simplest possible way. The NSL expression parameter for each note N_i was computed as the maximum sound level value $SL(k)$ (as defined in Chapter 3) within the note boundaries (see Figure 4.16):

$$NSL = \max_i \max_{t_i \leq k < t_{i+1}} SL(k)$$

The obtained NSL_i value for each note is an approximation of note loudness. NSL is related to loudness, also referred to as peak sound-pressure level, and it has also been used as a reference to calibrate key-pressed velocity in expressive performance acquisition of piano (Goebel and Bresin, 2003).

Vibrato Vibrato is an expressive manipulation of pitch corresponding to a frequency modulation of F_0 (fundamental frequency) characterized by its *rate* and *extent* (Prame, 1997). We considered only the extent of the vibrato since the rate is only defined for notes with vibrato and thus plays a secondary role on expression. However, the procedure we implemented to extract the vibrato estimates both rate and extent at the same time.

We estimated the vibrato of each note from a filtered F_0 time series within the note boundaries. The procedure is based on finding the amplitude of the sinusoid that correlates best with the filtered F_0 trajectory. Such amplitude is used as the estimated vibrato extent of the note. Modulations of vibrato

extent within the note boundaries were thus implicitly discarded favoring the selection of the highest vibrato extent reached within the note.

The first step was to filter from the F0 trajectory the transitions from one note to the next characterized by low aperiodicity. Suppose that note k has n_k frames within its onset and offset. Each of such frames $j = 1, \dots, n_k$ is described by the triplet $(t_j^{(k)}, f_j^{(k)}, a_j^{(k)})$ with, respectively, the frame time, the fundamental frequency (F0) and the aperiodicity as obtained with the autocorrelation-based F0 estimation algorithm (De Cheveigné and Kawahara, 2002). We use the threshold A to assign weights $w_j^{(k)}$ to each frame in the following manner:

$$w_j^{(k)} = \begin{cases} 1, & \text{if } a_j^{(k)} \leq A \\ 0, & \text{if } a_j^{(k)} > A \end{cases}$$

We then computed for each note the weight center of mass (in time) $\mu^{(k)}$ and the standard deviation $\sigma^{(k)}$ for each note using the formulae:

$$W^{(k)} = \sum_{j=1}^{n_k} w_j^{(k)}, \quad \mu^{(k)} = \frac{1}{W^{(k)}} \sum_{j=1}^{n_k} w_j^{(k)} t_j^{(k)}, \quad \sigma^{(k)} = \sqrt{\frac{1}{W^{(k)}} \sum_{j=1}^{n_k} w_j^{(k)} (t_j^{(k)} - \mu^{(k)})^2}$$

The select part of note k is then obtained by removing all the frames j such that:

$$|t_j^{(k)} - \mu^{(k)}| \geq B\sigma^{(k)}$$

where B is an opportune positive threshold.

Comparing the plots of F0 and aperiodicity we noticed that the F0 is generally reliable when the aperiodicity is lower than $A = 0.2$. We then tested several values for B and selected $B = 1.8$ as a good compromise between discarding too much data and getting the most reliable part of the F0. This value is such that most times the whole note is selected, except for very long sustained notes where some boundaries at the end are discarded. Furthermore, in order to avoid octave jumps (due to octave error) we mapped the obtained F0 values to a common octave.

Once the F0 time series has been filtered we proceeded to estimate the vibrato extent on the selected part of F0. We employed frequency-domain analysis following a similar procedure as the one described in (Herrera and

Bonada, 1998). We performed a Fourier analysis of the filtered F0 and looked for peaks in the spectrum ranging between 4 and 8 Hz (we employ parabolic interpolation of the peaks in the spectrum). If one or more peaks were found in this range we select the one corresponding to the sinusoidal component with the highest amplitude. The amplitude of such sinusoid (in pitch cents) is our estimation of vibrato extent whereas its frequency is the vibrato rate. Where no salient peak was detected or the select part of F0 was shorter than every possible vibrato period (125 ms), the vibrato extent was set at zero. In Figure 4.21 we show an example of vibrato detection from real data.

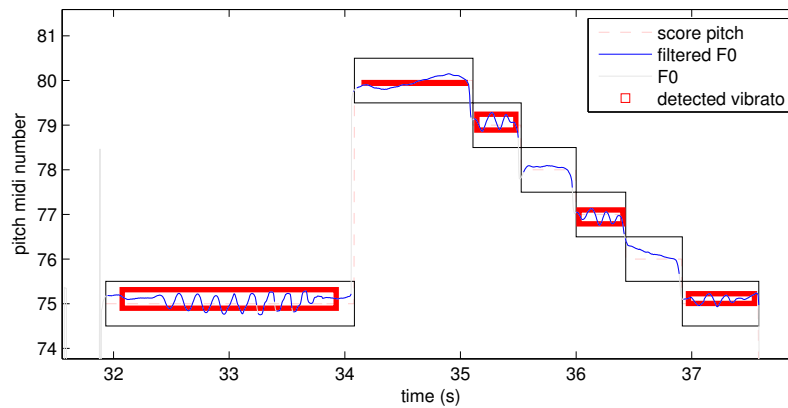


Figure 4.21: Estimated vibrato extent for a short sequence of note from violin 1. The red rectangles mark detected vibratos and their height is proportional to the (detected) vibrato extent of each note.

Gesture-based expression parameters

Bow velocity Bow transversal velocity is one of the main control parameters available for the string player to convey expression, affecting timbre and loudness (Askenfelt, 1989). We measured bow transversal velocity from each frame of motion capture data as described in (Maestre, 2009), leading to a bow velocity time series expressed in cm/s and sampled at 240 Hz. Bow transversal velocity can be positive (down-bow) or negative (up-bow), but since we were interested in the speed of bowing we first computed the absolute value of bow velocity samples across each note. Then, in order to extract a representative bow velocity value for each note, we computed the

interquartile mean on the absolute values of each note segment.

Figure 4.16 shows an example of the time series that are extracted from the raw data as well as the descriptors computed for each note segment.

4.3 Conclusions

In this chapter we have discussed on how to extract score contextual information and expression parameters. For both types of descriptors we have introduced a classification in which we distinguish between *horizontal* and *vertical* score context, as well as between *individual* and *collective* expression parameters.

The extraction of score contextual information involved projecting a multi-dimensional musical surface to several score contextual descriptor attributes. This required in some case the comparison of notes or groups of notes. In the case of expression parameters, we had to go through an alignment procedure in order to segment the performed notes. We described a dynamic programming algorithm that takes into account the acquired bowing gestures. After describing the score-performance alignment algorithm, we showed how to use the basic features introduced on Chapter 3 for defining note level expression parameters.

Whereas a complete list of performance descriptors was beyond the scope of this dissertation, we described a set of important performance descriptors that we will use in Chapter 5 to study expressive performance. In Chapter 5 we will study the relations between the introduced contextual variables and performance variables in with a statistical approach and a machine-learning approach.

Analysis of Ensemble Expressive Performance

5.1 Introduction

In this chapter we present some analyses of ensemble expressive performance combining various approaches. The aim, as in all studies of expressive performance, is to study the relationships between contextual descriptors and expression parameters.

In the previous chapter we have shown how to analyze both score and performance to extract score context and expression parameters. We expect the introduced descriptors to contribute to the understanding of ensemble expressive performance (EEP) and we propose a distinction between two different types of context related to different theoretic performing behaviors.

The first distinction is between *horizontal* and *vertical* score descriptors, which has been already introduced in the previous chapter. Of course when playing expressively musicians take into account horizontal features as it has been showed by most studies of expressive music performance. We discuss what it means to play considering the vertical context or not. This discussion rises several research questions that we later attempt to address in our experiments and analyses.

The second distinction is introduced in Section 5.1.2, and makes the difference between two theoretical modes of performing music: *auto-regressive* and *memoryless*. The term auto-regressive is taken from statistics literature on time-varying processes. In statistics, both linear and non-linear auto-

regressive models apply to the case when a particular outcome at a certain time depends on its own previous values. In *nonlinear auto-regressive exogenous model*, the outcome of a output variable depend on its own previous values and also on the concurrent and past values of a driving exogenous series (Billings, 2013). In this dissertation we model the expressive performance of each musician in a similar way, where the exogenous driving time series is given by the performance of all the other musicians. We make the distinction between *auto-regressive* and *memoryless* behaviors to explain how this translates in different modes of performing. However, this distinction is theoretic as it is hard to imagine a human musician playing in a purely auto-regressive mode or in a purely memoryless mode. Although we can think of particular cases where we can assume that one mode is dominant respect to another, in the majority of cases we believe that musicians employ both modes simultaneously (we explain in Section 5.1.3 how we combine both).

Musicians might perform based only on score context, but that means that they could have already decided *off-line* all the details (including the exact amount of each expression parameter) of the performance before going on stage. This is of course an over-simplification since the performance also evolves continuously respect to itself in an *on-line* fashion. Many studies of expressive performance do not need to make explicitly this distinction because either they focus only on one behavior, or they already include this aspect in their definition of expression parameters. In the case of ensemble performance, making this distinction is very important since the dependencies with past values might also be “spread” across different musicians.

After making the distinctions between different performing behaviors, we consider some related research questions and attempt to empirically answer them through analyzing performance data by following two different approaches. The first approach, which we will refer to as *statistical analysis*, involves the formulation of specific hypotheses and their verification by means of controlled experiments.

The second approach is to use machine-learning modeling of performance. This second approach has the advantage that we can relate a high number of variables, and study performance contexts by comparing the overall properties of the trained models. Employing this approach becomes natural when datasets start growing in size and there are many variables to control. In this context, the distinctions we make between different modes of performing become handy since we can control the way models are built to test to

which degree musicians consider vertical, horizontal, score or auto-regressive variables.

5.1.1 Horizontal and Vertical score context

We have showed in the Chapter 4 how analyzing ensemble scores naturally leads to the concept of horizontal and vertical context. A second step is to understand whether this distinction has any importance at all to the performance. In other words, does the performance of the same melody change if the accompanying voices in the score are changed or removed? Or, equivalently, can we improve the expressive performance model of each musician by feeding it with knowledge about other musicians' voices, namely the score vertical context? When performers play their part solo (without the accompaniment of the others), will they play it in the same way?

The idea here is to test whether some of the introduced vertical descriptors correlate with expression parameters. We do this in the two experiments presented in Section 5.2 using the statistical approach. Then, we aim at understanding how these two types of contexts (horizontal and vertical) can in combination produce models of ensemble expression in different playing conditions. We do this in Section 5.3 where we adopt several strategies to test the importance of horizontal context and vertical context in different performing contexts.

5.1.2 Auto-regressive behavior in expressive performance

Some studies of music performance have neglected completely the score context to focus on specific musical skills employed by musicians (Moore and Chen, 2010). Indeed, there are some cases where the score context does not play any important role at all. As an example, you can easily imagine the case of a score consisting of a single note repeated endlessly. Assume also that there is no meter indication and thus the local context of each note (at a moment sufficiently after the start) is exactly the same as the context of its previous note (since each note is also at the same metrical position). In such case any deterministic model of expressive performance based on local score context would have no option but assigning the same expression parameter to each note.

However, even in this extreme scenario, a musician would very likely introduce expressive deviations. One performance strategy could be for example accentuating one every two subsequent notes (each musician might have a

different preferred grouping tendency, e.g. one every three or four). We can easily predict if a note will be accentuated by checking if the musician accentuated the previous note. We say that a performance strategy has an *auto-regressive behavior* if expression parameters depend on the expression parameters employed on previous notes. When a performance has such behavior, little errors on some notes may affect following notes accumulating and producing a different performance each time.

To the best of our knowledge, the term “auto-regressive behavior” has not been used in the expressive music performance literature although this issue is addressed implicitly by most (if not all) studies. Some researchers define the expression parameters as discrete classes defined when relating a note expression parameter with respect to previous notes. For example, Widmer (2002) divided the notes into two classes “louder” and “softer” based on whether they are louder than the previous note events and the average loudness of the piece. This means, that the derived model is already taking into account variations respect to previous performed notes.

In this dissertation we use the term Auto-Regressive (AR) context to refer to a set of previously played expression parameters, which can help in predicting the expressiveness of the current note. Note that this has nothing to do with the term *performance context* appearing in Figure 2.1 which in our case refers to the various playing conditions: *solo/ensemble*, *mechanical*, *normal*, *exaggerated*. In this chapter, AR descriptors are implicitly included in the statistical analysis (Section 5.2) and explicitly in the machine-learning analysis (Section 5.3).

5.1.3 Verifying hypotheses and analyzing EEP models

The set of contextual descriptors to be used for the prediction of expression parameters lays in four main categories, which are given by all possible combinations of vertical vs. horizontal descriptors and auto-regressive vs. memoryless descriptors. Table 5.1 gives examples of each category of descriptors.

The first aspect we verified was whether the introduced vertical descriptors play a role in ensemble performance. In Section 5.2 besides showing how different playing conditions affect the expression parameters we also demonstrate how some vertical features correlate with the studied expression parameters. We first arrange performed notes into several context groups depending on their score context. We should observe different distributions of expression parameters in each context group. The analysis

Table 5.1: Categorization of different types of contexts for the prediction of expression parameters. Some examples are included in each category.

	Horizontal	Vertical
Memoryless	<ul style="list-style-type: none"> ○ Own-voice note nominal pitch ○ Own-voice Narmour group 	<ul style="list-style-type: none"> ○ Other-voice concurrent note nominal pitch ○ Other-voice Narmour group
Auto-Regressive	<ul style="list-style-type: none"> ○ Performed <i>Note Sound Level</i> of preceding note in own-voice ○ Performed <i>Note Vibrato Extent</i> of preceding note in own-voice 	<ul style="list-style-type: none"> ○ <i>Note Sound Level</i> of other-voice concurrent note as performed by the respective musician ○ Performed <i>Note Vibrato Extent</i> of other-voice concurrent note as performed by the respective musician

revealed simple relations between different context groups and expression parameters (see Section 5.2), justifying to a certain extent the choice of some vertical descriptors. However, the fact that some context descriptors do not show any effect on the performance does not imply we have to remove them. Not all of the descriptors might be important at the same time, for a certain score and performance context. For this reason we limited our analysis to comparing the performance in different playing conditions. We never compare expression parameter distribution recorded from the performance of two different scores.

Simple relations are not enough when analyzing complete pieces, so the machine-learning approach proved to be useful for extending our analysis. As in the example presented in previous subsection, expression parameters might sometimes be explained only in relation to the expression parameters of previous notes. In other moments the score might be the only essential information modeled. Analogously, horizontal context might sometimes prevail on vertical context or the contrary. In general it is through a com-

bination of memoryless/AR contexts and horizontal/vertical contexts that we can achieve best prediction of the outcome. The key question here is how to combine them to produce good prediction. In which case do we expect to need more of one type or more of another type of descriptor? We mainly focused on horizontal and vertical descriptors and tried to answer this question using statistical and machine-learning approaches.

5.2 Statistical analysis

We present two experiments where we applied statistical techniques to analyze how context affects expressive performance. In the first experiment we analyzed the timing of performances of a short exercise, comparing *solo* and *ensemble* performances. In the second experiment we analyzed timing and dynamics in performances of Beethoven's *Allegro-Prestissimo movement* (I.1-I.3 in Table 3.1) at three different expressive intentions (*mechanical*, *normal*, and *exaggerated*) and we show how the distribution of expressive performance descriptors changed with the expressive intention. In this case we also include an exploratory analysis of the synchronization among the musicians suggesting different entrainment behaviors in different sections of the piece.

5.2.1 Experiment I: timing in *solo* vs *ensemble*

We present an experimental framework through which we assign the musicians of a string quartet the task of playing specifically chosen exercises after a brief rehearsal period. In this context we have shown a set of preliminary results on timing synchronization phenomena observing the differences between musicians playing alone or in ensemble.

Material

This experiment is based on the exercise EX.1 of Table 3.1. As already explained in Section 3.1, the exercise (see Figure 3.2) was conceived by Mogens Heimann for training the ability of musicians to coordinate timing and dynamics. The requirement of the exercise is that it should be played “as one instrument”, which is a challenging task as musicians need to keep the tempo together and respect time continuity when switching from one group of semi-quavers (sixteenth-notes) to the next. The musicians repeated the exercise four times in a row. Score-performance alignment was obtained using the method described in Section 4.2.2. We cross-checked and ad-

justed manually the automatic alignment by visual/aural inspection of the individual sound waveforms, and of the audio/gesture descriptors (see Section 4.2.2) to obtain very accurate note onsets positions.

Method

In this experiment we focus on timing, comparing the solo and ensemble performances, dividing the analysis into *macro-timing* and *micro-timing* analysis. Whereas macro-timing can be related to global properties of the performance such as phrases or repetition patterns, micro-timing is usually related to local characteristics of the score within the bar such as metrical position and/or note duration.

For the macro-timing analysis, we extract the Gaussian tempo curve bpm_σ described in Section 4.2.3. We considered several values for σ and selected the value $\sigma = 1.67$ beats. This value was preferred to others since it is such that 78% of the distribution falls within half of the repetition (4 quarter notes), and 98% of the distribution encloses the full repetition (8 quarter notes). The tempo value which this σ produces is thus local enough to measure variations of tempo within each repetition and large enough to ignore irrelevant local fluctuations of note duration. In the ensemble performance we also compute the *joint tempo curve* (see Section 4.2.3) by averaging onset times for notes with same beat position. In this way since the score has only two voices at any time (see score in Figure 3.2) the middle between each of two corresponding onset times is selected. This procedure provides an middle onset time for each chord in the scale, which we then use for extracting the joint tempo curve (see Section 4.2.3) as it is done for each individual voice. The sequence of chord onsets is here referred to as *joint ensemble performance*. As a further inspection of macro-timing variations we include statistics of quarter note durations. At this macro-timing level, we compare performed durations of pauses with performed duration of each block of sixteenth notes.

For the micro-timing analysis, we focus on sixteenth notes, the smallest note value in the score. We compute the ratio between the performed duration of the first note in each block with the corresponding performed duration of the second note. This ratio is referred to as Semi-Quaver Duration Ratio (SQDR). Following the discussion in Section 5.1.2, the SQDR is a performance descriptor that by its definition presupposes an auto-regressive behavior.

Results

Macro-timing From the tempo curves derived in the solo and ensemble cases respectively we find that not only the mean tempo of the four musicians becomes the same, but also the variance of the tempo curve gets significantly smaller in the ensemble case. We can interpret this result both as an indicator that the freedom of the musicians gets restricted and as a result of the collaborative way in which the tempo is jointly shaped. Figure 5.1 shows tempo derived with $\sigma = 1.67$ for the solo and ensemble case. As it is clear from the plots, the individual tempo curves contract to the same tempo when the musicians play together.

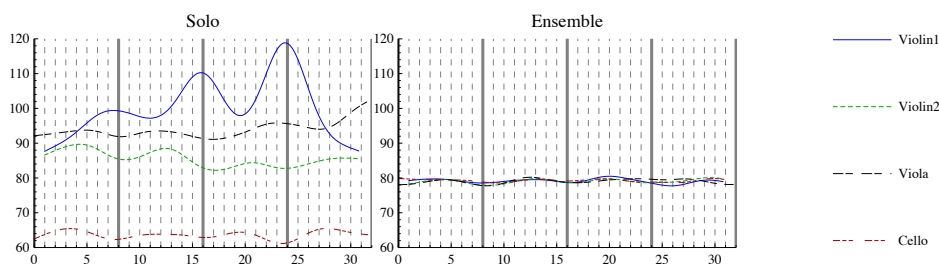


Figure 5.1: Individual tempo curve of the four instruments for the solo (left) and the ensemble case (right). Vertical grid lines mark the boundaries of the repetitions (full line) and the beat start time (dashed line).

Figure 5.2 shows the joint tempo curve. The most evident feature of this

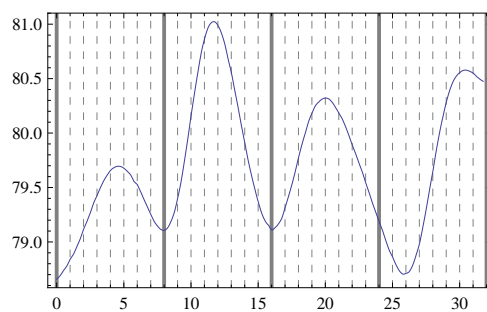


Figure 5.2: Tempo curve of the joint ensemble performance. Vertical grid lines mark the boundaries of the repetitions (full line) and the beat start time (dashed line).

tempo curve is that it correlates with the repetition structure of the exercise. In fact, while remaining relatively constant (just slightly increasing through the performance) the tempo was indeed oscillating by speeding up in the center of the repetition and slowing down towards its boundaries. In this exercise, the structure of the repetition is mirrored by the up-down profile of the pitch sequence. For this reason, we use the pitch sequence¹ as a reference for the repetition structure. In Table 5.2 the correlation between the tempo curve and the pitch sequence is shown for all the cases. In the joint ensemble performance the correlation coefficient is 0.56. This confirms the overall tendency of the performance to speed up at higher pitches. This is predicted by the well-known phrase-arch rule of Friberg et al. It is thus probably unrelated to pitch, and occurs only because the high pitches are in the middle of the phrase. However the performance we are analyzing here, far from being expressive, is just an exercise scale.

Table 5.2: Correlation between pitch and joint ensemble tempo curve.

	Solo		Ensemble		Joint Ensemble	
	Corr	Cov	Corr	Cov	Corr	Cov
Violin 1	-0.54	-33.65	0.59	2.87	0.56	2.78
Violin 2	0.5	8.5	0.75	3.34		
Viola	0.05	0.92	0.59	3.01		
Cello	0.72	6.1	0.35	1.01		

Despite the fact that we only have recoded few repetitions, the correlation coefficient value is highly improbable to arise by chance. To quantify the significance we have used an empirical (Monte-Carlo) method. We generated surrogate performances by perturbing the score onset times with Gaussian noise of standard deviation σ . For each surrogate performance we have carried out the same macro-timing analysis as the one carried out for the real performance. We generated five groups of 2000 surrogate performances with standard deviations σ of 0.1, 2.5, 5, 10 and 25 ms respectively. From each σ value we calculate the 2000 surrogate correlation coefficient values and use their distribution to derive one-tailed p-values. Each empirical Monte-Carlo

¹The pitch sequence has values in number of semitones and has been constructed by taking the higher pitched note of each chord in the score of Figure 3.2.

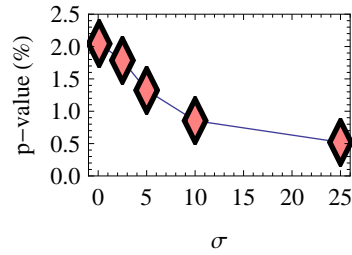


Figure 5.3: The plot shows p-values of the observed 0.56 correlation coefficient (in the joint ensemble) for increasing values of σ . The standard deviation σ by affecting the variance of the Gaussian noise introduced to the score onset times for the empirical significance test.

estimation of the p-values was obtained as

$$p = \frac{\#\{\text{surrogate correlation coefficient values } c \mid c \geq 0.56\}}{\#\{\text{surrogate correlation coefficient values } c\}}$$

The resulting p-values, as shown in Figure 5.3 are bounded by 2.5%. Remarkably, an increase of σ yields a decrease of p-values and not the other way around. This means that, it is even less likely to get large correlation coefficients by increasing the noise. We can, in conclusion, assume a confidence level of 97.5% for the observed correlation coefficient of 0.56.

Note that the excursion of the tempo curve is below the just-noticeable-difference (JND). Recent studies (Thomas, 2007) confirmed the Weber's Law in the perception of tempo change with a JND threshold of 6-8%. In our experiment the fluctuation was of around 2%, as is shown in Figure 5.2 (tempo spanned the range of 79-81 BPM). This means that the musicians were not aware of this fluctuations of tempo. Moreover, we can not distinguish if this mechanism is directly related to repetition structure, pitch or to some more complex underlying mechanism governing the performance.

The duration of each block of sixteenth notes and the pauses revealed larger variations in the solo performance than in the ensemble performance. In the solo case, tempo was kept differently by the musicians in the case of pauses than in the case of semi-quavers blocks. In the ensemble case, the discrepancy between pause and semi-quavers block duration gets smaller because of the interdependence among musicians. Box-and-whisker diagrams of the analysis for the first violin are shown in Figure 5.4. It is evident that the difference between the two cases disappears in the ensemble

case in which the musician has to count tempo with the other musicians. Running t-tests revealed a significant difference between pauses and semi-quavers duration in the cases of solo violin 1 and solo viola ($p < 0.05$).

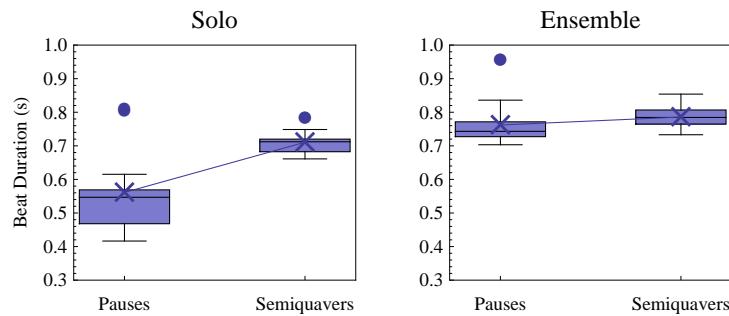


Figure 5.4: Duration of pauses and semi-quavers blocks for violin one: solo vs ensemble.

Micro-timing At the micro-timing scale, we found a correlation between the duration of each semi-quaver and its position within the block of four semi-quavers. An ANOVA test confirmed at a significance level below 1% the effect of metrical position on the joint-performance.

Differences have been found also when comparing the solo performance with the ensemble performance. Since the general tendency is to play the first note of the block longer than the second we have focused only on the first semi-quaver duration ratio (SQDR) of each semi-quavers block. Since the ratio between consecutive durations is not directly dependent on tempo, we were able to compare executions at different tempi. In particular we could compare the solo performance with the ensemble performance.

Remarkably, we could prove at a significance level lower than 2% the effect of the two scenarios (solo/ensemble) to the SQDR for first Violin, Viola and Cello. We can thus report an overall tendency to exaggerate timing accent of consecutive strong-weak semi-quaver couples in the ensemble case respect to the solo. Whereas the second violin keeps a positive SQDR of 1.19 in both cases, the first violinist and the cello increase theirs from 1.07 to 1.27 and from 1.0 to 1.24 respectively. A different behavior was measured for the viola, for which the SQDR decreased from 1.29 in the solo to 1.02 in the ensemble. Box-and-whisker diagrams of SQDR values are shown for both solo and ensemble in Figure 5.5.

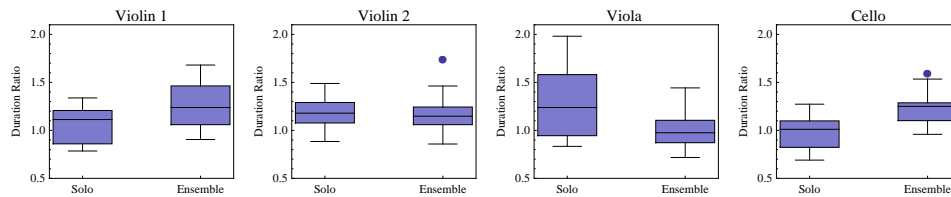


Figure 5.5: SQDR of solo vs ensemble.

A further analysis of the precedence of the onsets seems to explain the different micro-timing results of the musicians in the ensemble case. The onsets of the cello were preceding the ones of the viola by a mean of 8 ms, and the onsets of first violin were preceding the onsets of second violin by 13 ms.

Discussion

In the macro-timing and we have observed broad reduction of the mean tempo and its total excursion in each single instrument. This confirms the hypothesis that, when playing in ensemble, the synchronization among musicians favors a steadier macro-timing. Furthermore, the residual oscillations of ensemble macro-timing significantly correlated to the phrase structure of the repetition.

On the contrary, results on micro-timing showed relatively larger the contrasts of contiguous note durations in the ensemble condition than in the solo condition. By also looking at onset asynchronies between musicians we have formulated the hypothesis that a bigger contrast between contiguous short notes might be the result of coordinated actions constituting an entraining mechanism. In fact, musicians employing a higher SQDR are also anticipating their fellow on the average. This suggests that the use of contrast in successive notes could be the result of a synchronization mechanism.

5.2.2 Experiment II: timing and dynamics with increasing expressiveness

Material

We present a statistical analysis of I.1-I.3 in Table 3.1. We recorded a professional string quartet performing the last movement of Beethoven's quartet

n. 4 in C minor (opus 19 n. 4, allegro-prestissimo movement). After the quartet had played their first version (“normal”) we asked a “mechanical” and an “exaggerated” execution. The three executions were each around 5 minutes long.

The piece is in the classical *Rondó* form and thus its sections follow a structure *ABACABA*. With the assistance of a professional musicologist, sections were further segmented into phrases, leading to phrases of around 4 bars.

Method

Analogously to the previous experiment we carry out an analysis of the data at two complementary time-scales. The *micro-scale analysis* includes Statistics over individual note properties whereas the *macro-scale analysis* includes Statistics over time spans of the size of a phrase.

We carry out macro-scale analysis by analyzing the excursion of note sound level (see Section 4.2.3) and tempo curve (see Section 4.2.3) in the three executions I.1-I.3. The excursion values are computed on moving windows of four bars (the average length of a phrase) with a hop size of 1 quarter value. We thus collect one excursion value for each beat. We compute compare statistics over the excursion values in different expressive intentions.

At a micro-scale level, our analysis focused on note sound level, note lengthening, and participatory asynchrony (see Sections 3.3, 4.2.3 and 4.2.3). Since we want to exclude eventual relations between note value and expression parameter, we restricted statistics to the sub-sample set of eighth notes (the most common in this particular score) discarding any shorter or longer note. This restriction left us with 2946 notes (495 for violin 1, 861 for violin 2, 861 for viola and 729 for cello) counting the three expressive intentions together.

For the statistical analysis of the micro-scale, we use n-way ANOVA to study how note sound level, note lengthening, and participatory asynchrony were affected by the following six note contextual factors: metrical strength, melodic charge, Narmour group and two harmonic descriptors (see Section 4.1.3). In this analysis we quantized all the considered factors to avoid strongly unbalanced levels (some levels highly populated and others with few notes). This also resulted in a reduction of the degrees of freedom of the ANOVA model. The quantization of the contextual descriptors was obtained in the following way.

1. Metrical strength (Factor 1) was quantized to 3 levels: “strongest” and “strong” as defined in Section 4.1, whereas all the weak metrical positions were merged together into a third level.
2. Melodic charge (Factor 2) was not quantized, and thus consisted of 6 levels.
3. Narmour group (Factor 3) was reduced to the 11 classes appearing in the score (see Figure 5.6 to see how the notes were distributed in the datasets).
4. The two harmonic descriptors of Table 4.2 were used to define harmonic factors. The first one (Factor 4) is the harmonic charge of the chord quantized into two levels depending on the results of the test: $\text{HarmChAt}(b_0) \leq 1.75$. Such threshold was set so that notes were evenly distributed between both levels.
5. The second harmonic factor (Factor 5) was not quantized since it already consists of only the two levels *true* and *false* resulting from the formula $c_0 \geq \max(c_0^{(1)}, c_0^{(2)}, c_0^{(3)})$. We refer to Factor 5 as “vertical melodic charge” since it describes the ranking of melodic charge across musicians at a certain instant.
6. Finally the last factor of ANOVA (Factor 6) was the musician who played the note; there were thus 4 levels: first violin, second violin, viola, cello. We are not interested in how the expressive descriptors correlate with the *musician* since some differences might appear due to the calibration (for example due to the calibration of individual sound level gains). We introduced this last factor to guarantee that the observed differences in distributions do not arise because of musicians differences.

We complement the analysis by looking at the participatory asynchrony, and musicians asynchronies (see Section 4.2.3). To compute the participatory asynchrony we used the musicologist segmentation of the piece into phrases to define the boundaries of tempo curve parabolas. We study mean musician asynchronies per section and the correlation of expressive intention to participatory asynchrony per section.

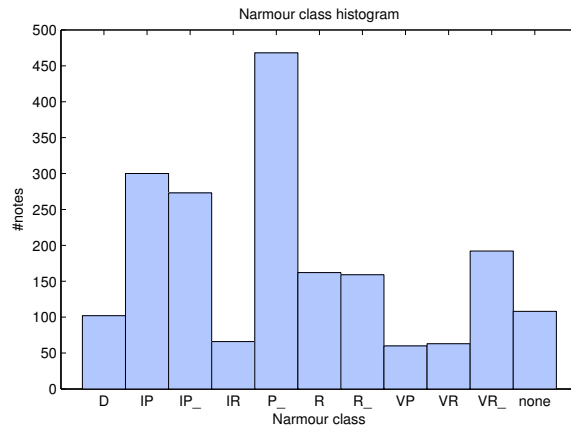


Figure 5.6: Narmour class histogram of eighth notes in the datasets I.1-I.3.

Results

Macro-scale The tempo curves for the first 90 bars of the Beethoven's piece for the three expressive intentions can be seen in Figure 5.7. The tempo oscillates at each phrase between a high and a low value. The excursion between the high and low values increases from mechanical, to normal, to exaggerated execution.

We compute the excursion of sound level in dB whereas the excursion of tempo in percentage respect to the section mean tempo. The distribution of excursion values is represented in Figure 5.8 with box-wiskers-plots. We can see that our previous observation on tempo is statistically valid on average throughout the whole piece. Additionally, we can see that the same happens for the excursion in sound level. A one-way ANOVA on the three expressive cases revealed that the difference is significant ($p < 0.05$) both for tempo and sound level.

Micro-scale The resulting p-values of the ANOVA analysis are shown in Table 5.3. Besides showing that some vertical descriptors affect the considered expression parameters, the results of ANOVA show the statistical significance of vertical context. Note that Factor 4 (harmonic charge) and Factor 5 (vertical melodic charge) are derived from the melodic charge descriptor, in the case of Factor 5 the derivation is straightforward (see Section 4.1 and Table 4.2). In very broad terms, the derived factors are con-

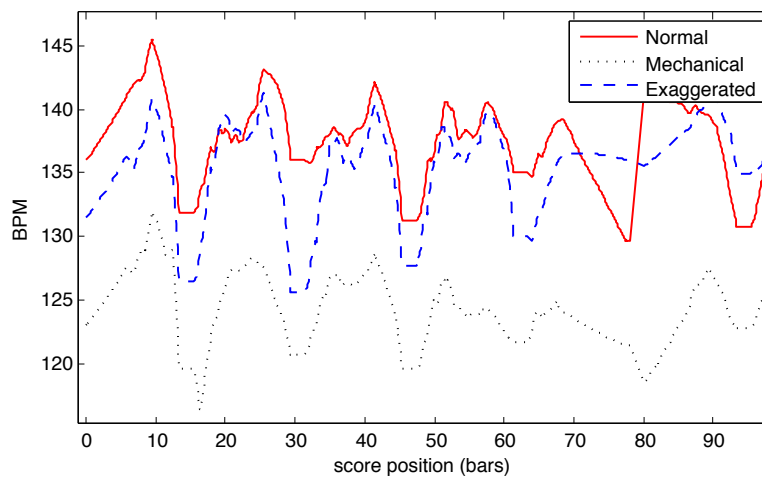
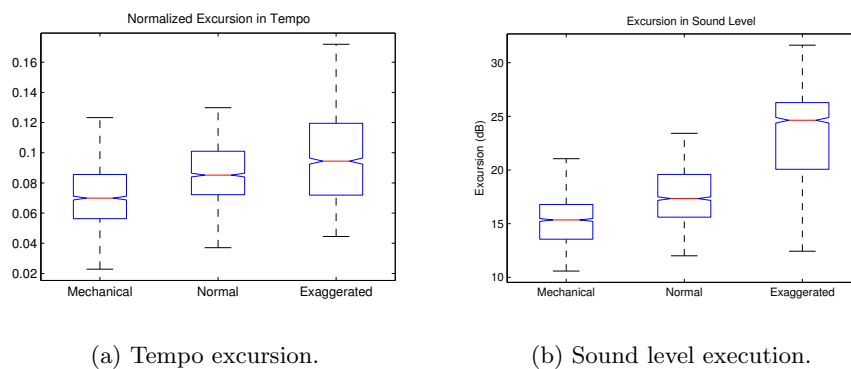


Figure 5.7: Tempo curves for the beginning of the piece in the three expressive intentions.



(a) Tempo excursion.

(b) Sound level execution.

Figure 5.8: The two box plots show the distribution of tempo ((a)) and sound level ((b)) on each expressive intention performance.

structured combining the note melodic charges with concurrent notes from the voices of other musicians. Now notice (from Table 5.3) how Factor 5 affects both note lengthening and note sound level ($p\text{-value} < 0.05$) whereas the simple melodic charge alone (Factor 2) does not affect note lengthening. This result suggests that certain expression parameters do not depend on the melodic charge value of the note played but rather on if such value is

salient in the vertical context (Factor 5). In other words, this confirms our hypothesis that the vertical context plays a role in the choice of expression parameters. Additionally this analysis confirms that the introduced deviations are not noise but are rather related to contextual descriptors.

Table 5.3: P-values of the 6-way ANOVA analysis of the eighth notes in dataset I.1-I-3. We highlight in bold the p-values lower than $p=0.05$ below which the effect of each factor is considered statistically significant.

	Note Lengthening	Note Sound Level	Participatory Asynchrony
Metrical strength (Factor 1)	<0.001	0.001	0.024
Melodic charge (Factor 2)	0.235	<0.001	0.225
Narmour group (Factor 3)	<0.001	<0.001	0.001
Harm. charge (Factor 4)	<0.001	0.064	0.261
Vert. melodic charge (Factor 5)	<0.001	0.014	0.429
Musician (Factor 6)	<0.001	<0.001	0.006

Discussion

Expressive intentions were realized by modulating tempo and dynamics with different amounts of excursion. For both tempo and dynamics the excursion increased towards more expressive executions. This confirmed that the difference between mechanical, normal and exaggerated performance was interpreted clearly by the four musicians. In particular they performed the two extra versions without rehearsing them or discussing them in details. Furthermore, the result suggests that the extracted performance characteristics are related to expressiveness of the performance.

At a micro-scale, the effect of a number of score contextual descriptors was found to be statistically significant. In particular, we found that in the case of note lengthening, the melodic charge was significant only when compared with the other voices (vertical melodic charge). This partially confirms

from the data (with a very high significance) the intuition by Sundberg et al. (1989) that comparing melodic charge of simultaneous note is useful to define the synchronization voice.

5.2.3 Limitations of statistical methods in the analysis of EEP

The statistical methods employed in the above two experiments are a tool for verifying specific hypotheses on the nature of the performance. However, this approach requires controlled experiments in order to obtain balanced levels across all factors. In Experiment I, where repetitions and simplicity of the score favored the analysis, such statistical analysis was meaningful for studying timing. The application of the statistical analysis to Experiment II was instead more difficult, where we had to quantize some descriptors to avoid unbalanced observations and discard some notes.

Another limitation of the statistical approach is that we can check few factors at a time and only on specific hypotheses. Indeed, the more hypotheses one tests, the more likely it is to incur in false positives. In case many hypotheses are checked, a Bonferoni correction should be applied, which means that the significance level has to be divided by the number of hypotheses tested. For example, if 5 hypotheses are tested, then to guarantee a significance level of 5% the p-values have to be smaller than 0.01.

There is no way to apply the complete set of descriptors defined in Chapter 4 to a statistical analysis of our recordings. Nevertheless, different approaches such as the one of expressive performance modeling can be applied. This change in the approach requires a shift in the way the data is analyzed and interpreted. We need to model the quartet as a group of four agents that output expression parameters depending on a multitude of attributes. The results of such investigation also have to be interpreted differently. We introduce in the next section our last approach to modeling expressive performance, which is based on machine-learning algorithms.

5.3 Machine-learning approach

We present here a different approach from the one of verifying statistical hypotheses. This approach has been used already in the past for example by Widmer (2002); Ramirez and Hazan (2006, 2005). The aim is to build a model of expressive performance without any a priori hypothesis on the nature of the dependency between context and expression parameters. This

approach becomes more natural when a large number of variables need to be tested at the same time. It is based on building a model of expression for each musician and then comparing the overall properties of the models.

In this section we build models of ensemble expressive performance assuming that each musician is an independent entity from the other musicians but that can be influenced by the other voices and the performance of such accompanying voices. We build computational models of ensemble performance that are also used as a mean to investigate how humans play music expressively in an ensemble. Additionally we aim at measuring quantitatively the difference between different performing conditions such as solo and ensemble. The results of this section are tightly based on Marchini et al.'s (2014) article, which represents the core contribution of this dissertation.

First we present the learning task given to the machine-learning algorithms. We iterate the same task on each musician and corpus using various feature sets and machine-learning algorithms. On each case we obtain a function that predicts the expression parameters based on the context attributes. We then introduce the machine-learning algorithms that we have used. Thereafter, we introduce the different feature sets and the feature selection procedure. We then present describe the datasets on which we run the experiments and how we constructed them.

We evaluate the individual parts of the method by comparing the predictive power of the models in various conditions and with different settings. We divide the result section in various parts. After discussing the methodological details we present the results on the expressive intention. We then present the analysis of the differences between solo and ensemble models in two steps. First we quantify the predictive power of the two models, then we discuss on the results of feature selection in the two cases.

5.3.1 Learning Task

Data-driven performance models predict note-level expressive transformations given a set of contextual descriptors of the note. Here, we are interested in learning a function f of the form

$$f(\text{Note}) \rightarrow (\text{SL}, \text{BV}, \text{VE}, \text{NL})$$

where Note is a note characterized by the set of descriptors described in Section 4.1, and SL, BV, VE, and NL are, respectively, the predicted sound level, bow velocity, vibrato extent and note lengthening. Each expressive

manipulation of the note is here a different learning task on which we perform a regression.

Each regression is performed by a machine-learning algorithm using a training set. The machine-learning algorithms focus on one form of f and attempt to minimize the error between the prediction and the target. We present in the next section the algorithms that have been used in this analysis. We used various machine-learning algorithms in order to experiment different forms of relating the feature set to the expression parameters.

5.3.2 Algorithms

The machine-learning techniques considered in this approach are the following:

- *K-Nearest Neighbor* (kNN). This is one of the simplest non-parametric machine-learning algorithms. It predicts the value of a test instance by assigning to it an average value of the k closest instances in the training set. The distance between instances is based on the Euclidean distance in the feature space. In this approach we use it with $k=1$.
- *Model Trees*. A decision tree classifier recursively constructs a tree by selecting the most relevant attribute at each node. This process gradually splits up the training set into subsets until all instances at a node have the same classification. The selection of the most relevant attribute at each node is based on the *information gain* associated with each node of the tree (and corresponding set of instances). A model tree, instead of predicting a class, takes into account all instances in a leaf of the decision tree and generates a linear regression in order to compute a real value as prediction. We have applied the model tree algorithm implemented in the Weka data mining software (Hall et al., 2009).
- *Support Vector Machines* (SVM). SVM (Cristianini and Shawe-Taylor, 2000) take great advantage of using a non-linear attribute mapping which enables them to be able to predict non-linear models (though they remain linear in a higher dimension space). Thus, they provide a flexible prediction, but with a higher computational cost necessary to perform all the calculations in a higher dimensional space. The classification accuracy of SVM largely depends on the choice of the kernel evaluation function and the parameters which control the

amount to which deviations are tolerated (denoted by epsilon). In this thesis we have used SVM with a polynomial kernel.

We evaluated the above algorithms by means of a cross-fold validation procedure obtaining an average value of correlation coefficient. The folds were created using phrases of around 16 beats (as provided by the musicological analysis of Section 4.2.3) in order to avoid phrases spanning both test and training set. In order to further guarantee the validity of the results, when creating each fold we removed all repetitions of notes, due to repeated sections (notated with and/or without repeat signs), in the test set from the training set. Doing so left 27 folds for each of the expressive intention corpus and 25 folds for each of the excerpt corpus.

We used the correlation coefficient to measure how well the algorithm was capable of generating a meaningful prediction. In Section 5.3.6 we compare correlation coefficients on different performing scenarios.

5.3.3 Feature Sets

In a preliminary study (Marchini et al., 2013) we tested different combinations of score contextual features considering, among other aspects, the size of the temporal window. We considered a range of window sizes spanning from one to five neighboring notes. This was repeated either for a set of only horizontal score features or a set of both horizontal and vertical score features. We found no evidence that window sizes larger than two improved the prediction. For this reason we limited the extended study to a fixed context length of two neighboring notes. The introduced restriction allowed us to focus mainly on the type of attribute by limiting the number of attributes to a reasonable number.

We selected five different *Feature Sets* (FS) containing an increasing number of contextual attributes of different types. In the first three feature sets we start with information of one's individual voice and progressively add attributes referring to vertical context:

- **FS1**) Horizontal score context of the target voice (Table 4.1)
- **FS2**) FS1 + vertical score context of concurrent notes (Table 4.2)
- **FS3**) FS2 + horizontal context of concurrent notes (Table 4.3)

In addition to score context we also incorporate the expression parameter values employed in the two previous notes in this performance of the score. We include first the expression parameters of the target musician, and then add concurrent expression parameters of the other musicians:

1. **FS4)** FS3 + expression parameter values in the two previous notes of the target voice
2. **FS5)** FS4 + expression parameter values in the concurrent and the previous note of the other voices.

The defined feature sets contain respectively 19, 38, 59, 61 and 67 attributes.

5.3.4 Feature Selection

The number of attributes is reduced when training the system by running a feature selection algorithm. After building each fold following the procedure explained in Section 3.2, we randomly split the training set into two halves. In the first half, we perform a feature selection algorithm in order to reduce the amount of attributes. We then train the regression of expression parameter on the second half of instances using the attributes selected in the first half. Lastly, we used the regression to compute the predictions on the test set.

The employed feature selection algorithm attempts to find a subset of features f_1, \dots, f_k maximizing the Correlation Feature Selection (CFS):

$$\frac{r_{cf1} + r_{cf2} + \dots + r_{cfk}}{\sqrt{k + 2(r_{f_1f_2} + \dots + r_{f_1f_j} + \dots + r_{f_kf_{(k-1)}})}}$$

where each r_{cfi} is the correlation between each feature and the classified attribute and r_{fifj} are correlations between features (Hall, 1999). The idea behind CFS is that a good set of features is made out of attributes poorly correlated with each other while highly correlated with the predicted variable.

In order to find a set of features maximizing the CFS in a reasonable amount of time we reduced the number of combinations to test by employing a greedy step search. The algorithm starts with an empty set and adds one feature at a time. The feature that maximizes the CFS is selected at each

step. The algorithm stops adding features when no further improvement of CFS is possible.

5.3.5 Building the Datasets

Before applying any machine-learning algorithm we needed to construct our dataset. The dataset is a collection of note instances, each characterized by a set of descriptor attributes and a target value. We created each dataset by choosing a musician, a corpus and a feature set (FS) to which one of the four corresponding expression parameters was assigned as target value.

From the recorded excerpts and pieces we create 5 **corpora**:

- *Normal Piece*: I.1 in Table 3.1
- *Mechanical Piece*: I.2 in Table 3.1
- *Exaggerated Piece*: I.3 in Table 3.1
- *Ensemble Excerpts*, where all the note instances of the four excerpts played in *ensemble* are merged in one dataset. It includes all ensemble takes from P.1-P.4 in Table 3.1.
- *Solo Excerpts*, where all the note instances of the four excerpts played *solo* are merged in one dataset. It includes all solo takes from P.1-P.4 in Table 3.1.

For each of the five corpora we create 80 individual datasets with all the combinations given by:

1. **learning tasks**: sound level / bow velocity / vibrato extent / note lengthening
2. **feature sets**: FS1 / FS2 / FS3 / FS4 / FS5
3. **target musicians**: violin 1 (vl1) / violin 2 (vl2) / viola (vla) / cello (cello)

In the case of the first three feature sets (FS1-FS3), datasets differing only by *learning task* (thus having been built on the same corpus, learning task, feature set and target musician) also contain the exact same attribute values. However, in the case of the last two feature sets (FS4-FS5) the type

of expression parameter on the previous notes is conventionally the same as the one of the target (and thus the learning task). As a consequence, in the case of FS4-FS5 the models predicting each expression parameter access to different expressive performance values. For example, where *vibrato extent* was the target expression parameter then the past expression parameter values were drawn from previous note *vibrato extent* values, whereas if *sound level* was the target expression parameter then the past expression parameter values were drawn from previous note *sound level* values.

Another important aspect to notice here is that in the case of the *solo excerpts*, the concurrent expressive performance values included in the feature set FS5 were not heard by the performers. However, we take it from the other musicians' *solo* performance when building FS5. This aspect will be important in our final discussion, since we want to investigate in which sense the ensemble performance was different from the solo performance.

Each dataset was then used for executing one experiment using the previously described cross-validation per phrase resulting in 400 experiments. Each experiment yielded an individual correlation coefficient we collected. As a final approach, we used the *model trees* algorithm, removing repetitions from the test set and using cross-fold evaluation per phrase (as explained in Section 5.3.2) and applying feature selection (as explained in Section 5.3.4). We used the correlation coefficients obtained on each experiment as an estimation of the predictive power of the model. Our results in Section 5.3.6 consists of a statistical analysis of the relations between the factors just introduced (corpus, learning task, feature set and target musician) and the correlation coefficients achieved on the corresponding experiments.

Additionally, we tested another two machine-learning algorithms (SVM and kNN), another type of validation (10-fold cross validation), the effect of feature selection and removing repetition. In total we performed $3 \times 2 \times 2 \times 2 \times 400 = 9600$ individual experiments. We discuss how these additional factors affect the correlation coefficient in Section 5.3.6.

5.3.6 Results

We computed correlation coefficients (using cross-validation) by running the machine-learning algorithms once for each feature set (FS), each corpus, each musician and each learning task. At stage one we compared the performance of each of the collected setups by looking at how each factor considered affects it. In Figure 5.9 we show an example of the collected predictions along with the performed value of the four instruments.

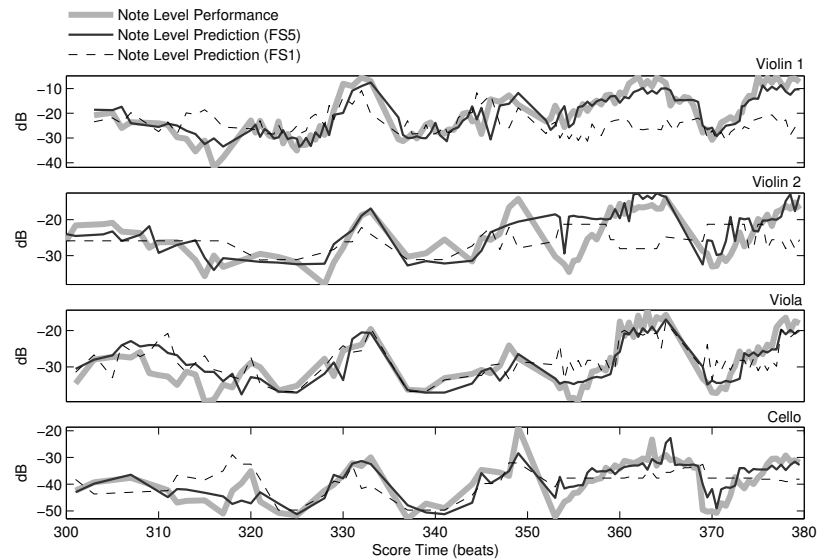


Figure 5.9: We represent the note sound level of the four musicians along with the predicted value (using model trees with the FS1 and FS5). The example is taken from Beethoven’s String Quartet No. 4 (Opus 18) 4th movement. We show the exaggerated expressive execution.

As already mentioned in Section 5.3.5 we use the correlation coefficient as an estimation of the *predictive power* of each model. The sample correlation coefficient however is not normally distributed since its variance becomes smaller for higher values of correlation. We employed the Fisher z-transformation to correct each correlation coefficient r , which distributes the variance equally across all levels of correlation. The Fisher transformation is given by the following formula:

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

We performed an Analysis of Variance (ANOVA with interaction) of the z coefficients (after applying the Fisher z-transform in order to distribute the variance equally across all levels of correlations) obtained from all the considered factors (musician, FS, learning task and corpus). The learning task was responsible for the largest variance encountered. This means that

on average some expression parameters are easier to predict than others. Taking into account the variance, the learning task is followed by the feature set, the dataset and the musician, in that order. For these reasons, the results are more intelligible when each dataset is presented by feature sets, learning tasks and musician.

We present our results with model trees for the expressive intentions dataset (Section 4.1) and then for the performing conditions dataset (Section 4.2). We make some observations about how learning task, musician and feature set affect the result.

Furthermore, we show that our observations are consistent with other regression classifiers (SVR and kNN) and how other corrections we introduced in the evaluation procedure (removing repetitions and using cross fold validation per phrase) affected the result (see Section 4.3).

Lastly, we show that the feature selection procedure, despite removing some of the attributes from the training set, does not worsen the prediction with respect to training with the full set of attributes. Hence, we report that the feature selection procedure tends to select an appropriate subset of features for each of the setups. We, then, present the results of feature selection on the excerpts datasets (without cross-validation) using the largest feature set FS5 for the *solo* vs. *ensemble* scenario. We demonstrate how the selected attributes differ depending on the condition (*solo* or *ensemble*). In particular, the amount of horizontal features is predominant in the *solo* case whereas the vertical features are predominant in the *ensemble* case (Section 4.4).

Expressive intentions

For each expressive intention, we report the best correlation coefficients achieved for each musician and target in Table 5.4. The correlation coefficient of the first violin was higher than that of the other musicians most times. Another interesting phenomenon is that the correlation increased with the degree of expressiveness (increasing from mechanical to exaggerated) for sound level, vibrato and bow velocity. In the case of note lengthening we did not find this trend, which suggests some relation to different tempi (the mechanical was played slower than the other two cases).

If we look at the results more closely by also considering the difference among feature sets we can draw more interesting observations: firstly, we notice that sound level and bow velocity improve on subsequent FS whereas

Table 5.4: Correlation coefficient on each expressive intention. Each cell contains the best correlation coefficient across feature sets for the specific musician and task. The highest value of each column is shown in bold.

	Mechanical				Normal				Exaggerated			
	SLev	Vib	BVel	Dur	SLev	Vib	BVel	Dur	SLev	Vib	BVel	Dur
vl1	0.70	0.48	0.66	0.29	0.76	0.65	0.73	0.25	0.81	0.68	0.84	0.18
vl2	0.46	0.28	0.5	0.26	0.46	0.70	0.53	0.19	0.6	0.65	0.67	0.14
vla	0.67	0.42	0.49	0.26	0.65	0.63	0.47	0.48	0.68	0.54	0.58	0.32
cello	0.44	0.17	0.49	0.14	0.69	0.63	0.61	0.31	0.76	0.54	0.80	0.20

note lengthening and vibrato stay nearly constant (see Figure 5.10). Secondly, we observe that the improvement, when present, follows different profiles depending on the musician and the learning task. An example of this is shown in the left plot (Sound Level) of Figure 5.10 (Exaggerated). The profile of the first violin distinguishes itself from the rest because it features a leap between the FS1 and FS4 and then mildly increases from FS4 to FS5. Other musicians distribute more evenly the improvement across different datasets. The profiles of each instrument are consistent across expressive intentions, and the first violin exhibits the largest improvement in correlation coefficients across feature sets.

An ANOVA on the z coefficient restricted to the results of this dataset confirms our observations. We used an ANOVA model that considers the interaction effects between variables.

The factors showing a significant effect ($p < 0.01$) in order of explained variance (ss: mean sum of squares) are the following: learning task (1.60 ss), expressive intention (0.73 ss), FS (0.59 ss), FS-learning task (0.25 ss), musician (0.23 ss), learning task-intention (0.22 ss), musician-learning task (0.18 ss), musician-intention (0.12 ss), FS-musician (0.03 ss). No significant interaction between the FS and intention was found.

Solo Vs Ensemble

Results for the *solo* vs *ensemble* scenario are comparable with those obtained in the previous scenario. The main factors (in terms of explained variance) are the learning task (0.69 ss), the interaction between musician and learning task (0.16 ss) and the FS (0.15 ss). The ANOVA analysis in

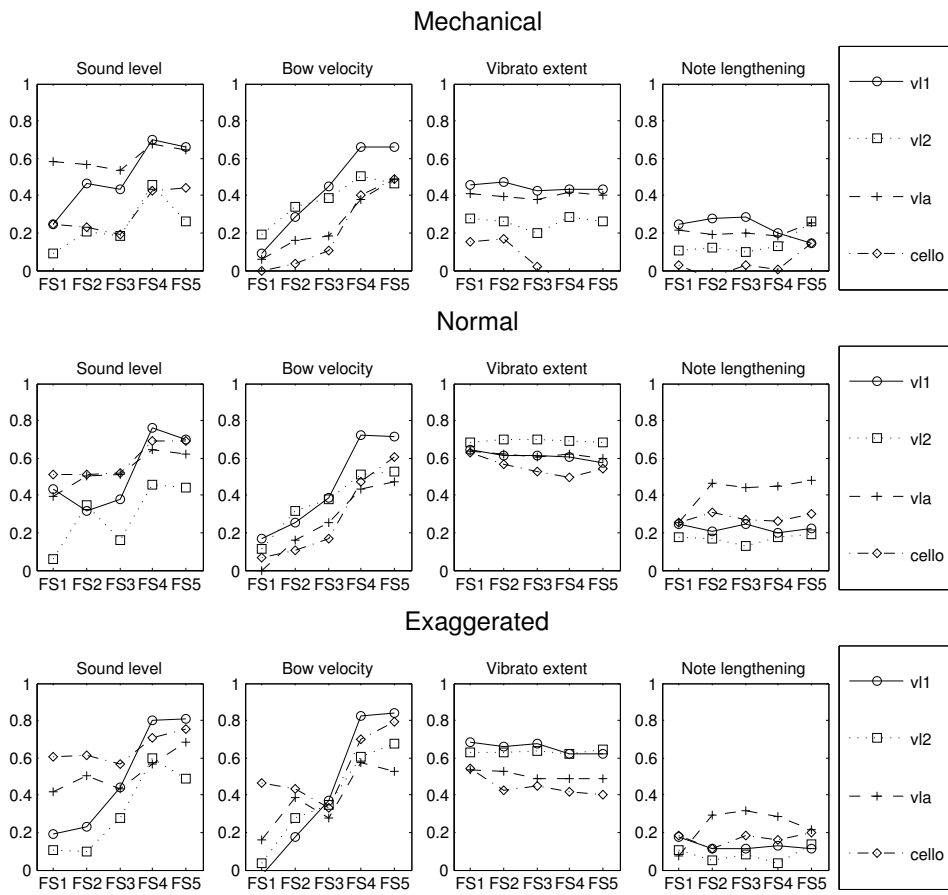


Figure 5.10: Resulting correlation coefficient for each expressive intentions, targets and musicians.

this case revealed an interaction of the condition (*solo/ensemble*) with the learning task (0.06 ss) and the musician (0.04 ss). There is no evidence of any interaction between the condition and the FS ($p=0.86$). We can, in fact, observe similar profiles of improvements in the *solo* and in the *ensemble* case (see Figure 5.11) for every given musician and target.

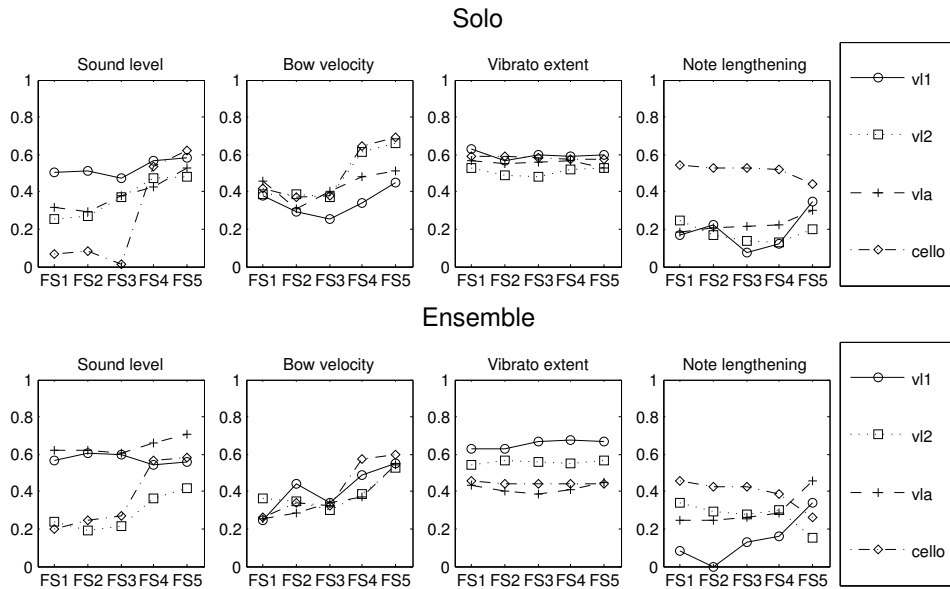


Figure 5.11: Resulting correlation coefficient for different learning tasks and musicians in the “*solo vs. ensemble*” scenario.

Surprisingly, on average, when adding the concurrent expression parameters of the other musicians (in FS5) there is an improvement on the prediction both in the *ensemble* and in *solo* performance (see Figure 5.11). The similarity between the two conditions can be explained as an effect of being accustomed to performing that material together; since the recorded excerpts were part of the ensemble’s repertoire, the musicians were already accustomed to performing those excerpts together and could anticipate the expressive actions of the rest of the ensemble even if they were not present. Nonetheless, we have to consider that the correlation profiles display no information about the underlying process of feature selection through which they were obtained. We comment on the results of feature selection in Section 4.4.

Effect of Different ML Algorithms and Evaluation Method

The results presented in Section 4.1 and 4.2 were obtained by removing notes in the training set which were repetitions of notes in the test set. In order to further validate our results we considered three additional factors. The first is the type of machine-learning algorithm; in order to account for this factor we repeated all of the tests using the SVM and kNN instead of model trees. The second factor is the influence of repetitions in the dataset arising from refrain marks in the score. We wanted to discover what happens if we do not remove the repetitions, and so repeated all the tests without this correction. The third factor was to use a 10-fold cross validation instead of using phrases as folds. In a 10-fold cross validation, instead of using the phrases to build the folds we randomly split the dataset into 10 subsets, which are used, one at a time, as test sets.

We ran an ANOVA on all z values, considering the classifier, the removal of repeated notes and the evaluation method among the factors. All three effects were, alone, found significant (with $p < 0.03$). In particular, the cross-validation per phrase yielded lower values of correlation coefficient with respect to a simple 10-fold cross-validation. Regarding the effect of the classifier, the model trees provided best results, followed, in order, by SVM and kNN. Removing the repetitions also significantly lowered the correlation coefficients, which was expected from the resulting reduction in size of the training sets. We found interaction between classifier and the effect of removing repetitions ($p < 0.03$). Indeed, the effect of removing repetitions was significant (the mean correlation coefficient dropped by 0.10) in the case of kNN whereas both SVM and model trees were less affected (for both the mean correlation coefficient dropped by 0.04). This confirms that removing repetitions was important for guaranteeing that there is no over fitting, especially for the case of kNN. Additionally, we can conclude that model trees and SVM are not reliant on those few repeated instances as much as kNN is and are thus more reliable for building a robust model of expression.

Analysis of Feature Selection on the *Solo* vs. *Ensemble* scenario

On each of the setups we computed the difference between the correlation coefficient after and before applying feature selection. The mean of such differences across all the tests was 0.03 although this value became higher in the models employing larger feature sets. This means that the more features there are, the better the improvement we can achieve by applying feature selection. By considering the difference in terms of z coefficients

we also performed a t-test, which showed that this average difference, is significantly positive ($p < 0.01$). We can deduce that the feature selection procedure improves the prediction.

Table 5.5: Results of feature selection. We report the mean percentage of horizontal features (PHF) across musicians for the *solo* and the *ensemble excerpts* dataset after the feature selection. Notice how PHF is higher in the *solo* than in the *ensemble* case (except for bow velocity). We report the one-tailed p-values for each percentage.

	Solo		Ensemble	
	PHF	p-value	PHF	p-value
Sound Level	54.06%	<0.001	45.26%	0.02
Bow Velocity	35.58%	0.22	38.92%	0.059
Vibrato Extent	60.27%	<0.001	42.98%	0.079
Note Lengthening	52.38%	0.007	23.96%	0.173

The totality of 67 features is divided into 21 horizontal features (all the attributes of Table 4.1 plus 2 expressive performance attributes) and 46 vertical features (all the attributes of Tables 2 and 3 plus 6 previous notes performance attributes). The *Percentage of Horizontal Features* (PHF) was therefore 31.3% in the FS5. We analyzed how this percentage changed after the feature selection in the *solo* vs. *ensemble* scenario.

Table 5.5 reports the mean across musicians of the PHF for each learning task and condition. Overall, we can observe how feature selection tends to select horizontal and vertical features more uniformly with respect to the 31.3% of the starting feature set. The horizontal features were in fact relatively more numerous after feature selection with respect to what they were in the full FS5. We computed p-values for each percentage based on the null-hypothesis that feature selection is modeled by an urn extraction without replacement. This statistic is thus based on a sample percentage mean computed over the results of independent draws with hyper-geometric distributions. We report one-tailed p-values for each mean percentage. Small p-values (< 0.05) indicate that the obtained percentage cannot be attributed to chance. PHF is generally higher in the *solo* than in the *ensemble*. In the

case of *solo* percentages were higher than 50% (except for bow velocity), which is very unlikely, given that by pure chance the expected percentage should be close to the original 31.3%. In the case of sound level, vibrato extent and note lengthening the PHFs are large enough to reject the null hypothesis that they arise from chance (see second column of *solo* in Table 5.5). In the case of *ensemble*, the percentages are all lower than 50% and thus more vertical than horizontal features were selected.

5.3.7 Discussion

In this section we have built models of expressive performance using a machine-learning approach. We quantified the predictive power of various subsets of the descriptors combining horizontal/vertical, and score/auto-regressive descriptors. We also discussed the main differences obtained using three different machine-learning algorithms, two types of cross validation and the effect of refrained notes. Furthermore, we presented the distribution of features after the process of feature selection both on the *solo* and the *ensemble* case.

The analysis of feature selection exhibited a clear tendency of the models to prefer horizontal features (individual voice context) in the solo case, and to prefer vertical features (inter-voice relation context) in the ensemble case. This further confirms the validity of the introduced inter-voice features in the context of ensemble expressive performance. In light of the obtained results, we offer some thoughts on how they can be explained from a musical standpoint involving the musicians and the listeners. The average difference in predictive power among expression parameters suggests different management of each nuance. Sound level and bow velocity are related to expressive amplitude variations. The increase in predictive power towards models with representation of inter-voice context implies that the inter-voice features helped the models. The fact that this improvement can be measured both in solo and ensemble recordings suggests that even when a players are playing solo, they have in their minds the imagined parts of the other members of the quartet. Also, for the listener, expressive use of amplitude variation will simply not be heard if masked by other instruments playing at the same time, so this is highly dependent on simultaneous notes. In the case of vibrato, we believe that, since musicians know from experience that it is more difficult to perform a vibrato in short notes, they might reserve it for notes that are longer. As a result, vibrato extent might be already well predictable based solely on the individual voice. In the

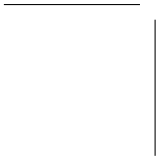
case of note lengthening the results were, instead, not sufficient to prove a substantial advantage of using neither horizontal nor vertical descriptors. This suggests that, in the case of timing, the excess of timing unrelated features corrupted the performance of the models and, thus, simpler models based on time keeping correction mechanisms (Repp, 2005) and phrasing (Widmer and Tobudic, 2003) might still be more suited. However, it is still possible that if trained on larger datasets this method might autonomously discover such synchronization mechanisms and achieve better predictions. We noticed that the predictive power of the models tends to improve when musicians play more expressively. Differences in predictive power across musicians tend to favor the first violin as the musician whose expression is easier to predict in terms of sound level, bow velocity and vibrato extent. We believe that this tendency is a good sign, since it means that the model is capturing the intentional expressive deviations employed by the musicians, and not some perceptually irrelevant byproduct of playing an instrument.

5.4 Conclusion

In this chapter we introduced an approach to analyzing expressive performance in string quartets. In the context of ensemble expressive performance, the models take advantage of the correlations between a set of input variables and the output expression parameters. We characterized the models into horizontal, vertical and memoryless and auto-regressive, depending on the type of input variables that the model have access to.

The approach is based on the introduction of a set of ensemble-specific contextual descriptors. These contextual descriptors are built by relating each note to neighboring notes belonging either to its part or the part of others. The statistical approach provided a great tool for showing how the introduced descriptors behaved in various experiments. In this context we also focused timing and dynamics in macro-scale time spans. This exploratory study highlights the potentials of using this method not only on modeling expressive performance but also in understanding the importance of inter-voice relations in ensemble performance. With the statistical approach we could point out how some descriptors that were individually irrelevant (i.e. the *melodic charge*) assumed an important role when considered in relative terms (i.e. *vertical melodic charge*).

By employing machine-learning techniques, we abandoned the idea of proving hypotheses specific of the individual descriptors approaching an holistic



view over the data. We introduced an approach to modeling expressive performance in string quartets allowing for polyphonic expression and interdependence among voices. We recorded string quartet performances and extracted a set of expression parameters related to sound properties of each performed note by exploiting multi-track audio data and bowing motion data. We assigned an expressive model to each musician and trained it on the collected data using supervised machine-learning algorithms (i.e. regression algorithms in this case). We gradually instilled into the models the ability to represent score and performance inter-voice relations by incorporating into the feature set of the models a number of ensemble-specific contextual descriptors, thus creating *ensemble models* of expressive performance. We evaluated how well models built on different feature sets were fit for predicting the expressive performance data by quantifying their predictive power by means of a cross validation scheme. The type of feature set was found crucial for determining the accuracy of the induced expressive model. As expected, the predictive power tended to increase when introducing inter-voice-relation attributes and recent performed history.

The ensemble models were able to adapt to various degrees of expressiveness and to solo and ensemble performances, proving to be flexible for various styles. Furthermore, similar results were obtained with three different machine-learning algorithms tested (although the nearest neighbor algorithm proved less robust than support vector machines and rule trees). We concluded that the results were mostly dependent on the data being fed to the models. We questioned the difference between models trained on solo performance and models trained on ensemble performance. Both models achieve similar results in predicting the data and, even in the case of solo performance data, ensemble models resulted better than solo models. We concluded that the solo performance data might have a bias from the fact that the musicians were accustomed to play this pieces together, which creates correlations among the individual solo performances. However a more detailed look at the models build in the two cases revealed a significant disparity between the selected horizontal and vertical features in the two cases. The models trained on solo performances selected a majority of horizontal features, and models trained on ensemble selected a majority of vertical features. We concluded that although some residual correlation among solo performances was inevitable, the models trained on ensemble performance could entirely discard some information about the individual voice and replace it with the vertical features.

Conclusion

6.1 Introduction

In this dissertation we defined the hypothesis that expressive performance analysis can be applied to ensemble performance and used as a tool for studying inter-musicians interaction. We proposed a methodology for modeling ensemble expressive performance (EEP) that accounts for the following aspects:

- Polyphonic expression: the simultaneous expressive deviation on each voice might differ with one another.
- Inter-dependence among musicians: the expressive deviations of each musician are shaped partially in accordance with the score context given by simultaneous voices and partially as a reaction to the expressive deviations introduced by other members of the group on stage.

We dealt with polyphonic expression by building independent models of expressive performance for each musician. We allowed inter-dependence among musicians by feeding ensemble local context into each model. Within this framework we defined the further hypothesis that each musician incorporates the expression of other voices in their expressive deviations. Based on these hypothesis, we set ourselves three main goals: to acquire a corpus of string quartet performances containing detailed information about each musician performance both in solo and in ensemble playing, to extract a set of meaningful vertical score context attributes, and to show differences among the models trained in various contexts.

Within the scope of the dissertation we have met such main goals. In Chapter 3, we introduced a corpus of pieces and excerpts that we recorded, and we advanced the state of the art in bowing force estimation in order to reduce the complexity of the multi-modal recording setup. In Chapter 4 we introduced a number of ensemble-specific score contextual descriptors. In Chapter 5 we showed that some of the introduced descriptors correlated significantly with the ensemble performance. Furthermore, we built machine-learning models of EEP using the acquired corpus. The analysis of the models suggested a number of implications over the nature of the ensemble interaction. Among such implications we found that the models trained on solo performance made use of a majority of horizontal context features whereas the models trained on ensemble performance made use a majority of vertical context features.

6.2 Summary of contributions and key results

The contribution of this dissertation were introduced in Section 1.4. We go over them again, referencing to the chapters where each contribution was realized and organizing them into their fields of contribution.

6.2.1 Multi-modal data acquisition

Recording setup In Chapter 3, we proposed a recording setup for acquiring sound, motion and video of string quartet where each musician's performance is independently recorded as a set of multi-modal time series. We discussed the main technical issues arising in this complex recording system and proposed solutions for each of them: the synchronization between device clocks for multi-modal signal acquisition, the extraction of a bowing gesture parameters with minimal intrusiveness, and the selection of music corpus.

Bowing force estimation model With the aim of reducing both the complexity and the intrusiveness of the recording setup, we developed (in Section 3.5) a mathematical model of bow deflection allowing for a non-intrusive estimation of the force applied by the musician on the bow. The model was constructed based on the physics of hair ribbon and depends on the size and the tension of the hair ribbon. We described a methodology for calibrating these parameters with an optimization procedure that was applied to each recording. Furthermore, the resulting force estimation improved the score performance alignment.

Ensemble expressive performance dataset The string quartet recordings represent a valuable resource for both research in ensemble expressive performance as well as research in music information retrieval. We compiled the subset of recordings analyzed in this dissertation and we made it available online¹ for research purposes. The dataset contains the acquired audio and gesture data, as well as the MusicXML scores and the relative score-alignment segmentations. We plan in the future to release the full set of recording that are, at the moment of writing this dissertation, an undergoing work.

6.2.2 Automatic music transcription

Score-performance alignment In Section 4.2, we introduced an algorithm for score-performance alignment which takes advantage of both gestural and audio data. The algorithm is an extension on the heuristic-based dynamic programming described by Maestre (2009). We extended the previous algorithm by handling force and velocity as two linked parameters to define bowing (gesture) attacks events. In Section 4.2.2 we also provided, for the first time, a systematic procedure for optimizing the parameters of the algorithm. We showed that the integration of bowing gestures improved the results with respect to the audio-only baseline algorithm.

6.2.3 Music Representation

Extraction of contextual descriptors in multi-voiced music score In Section 4.1 we introduced a methodology for extracting contextual descriptors in multi-voiced music score. Besides some early work by Sundberg et al. (1989), we could not find in the literature a systematic approach for characterizing a multi-voices score in terms of inter-voice context. We believe that the introduced ensemble contextual descriptors provide a relatively easy computational tool that can find further applications in the study of ensemble music performance.

6.2.4 Ensemble expressive performance modeling

Definition and extraction of expression parameters We gave a clear definition of ensemble expression parameters extracted from timing, audio and bowing gestures. We distinguished between *individual* and *collective* expression parameters. We showed how it is possible to define *participatory*

¹www.mmarchini.com/EEPdataset

discrepancy parameters as individual deviations from collective expression parameters. Following this distinction we gave a few operative definitions of tempo curves and asynchronies in music ensemble performances.

Statistical analysis of ensemble performance In Section 5.2 we analyzed several performance descriptors in two experiments. First, we analyzed the expression parameters employed by each musician across repeated executions of the same voice in various conditions. We showed how the detected differences related to the various performance conditions: solo, ensemble, mechanical, interpreted, and exaggerated. We noticed how it was fairly easy to spot out macroscopic differences among the conditions in terms of dynamics and tempo excursion.

At a micro-level, local differences were hard to find and interpret from a basic statistical analysis standpoint. One of the main reasons was an explosion of the number of hypothesis we could test due to the large possible combinations. This analysis, by showing the possible pitfalls of this method, might be useful for other researchers willing to design experiments in ensemble performance to test very precise hypothesis. Using the basic statistic approach, we could also inspect differences across groups of notes within each part, thanks to the introduced ensemble contextual descriptors. ANOVA analysis provided a valuable tool for detecting correlations between score context and expression parameters. Although we abandoned the idea of testing all the possible hypotheses, we found examples of *vertical* score context descriptors correlating with the expression parameters, which supported the introduction of ensemble score contexts.

Towards a methodology for modeling ensemble expressive performance In Chapter 5 we aimed at discovering the mapping between contextual features and expression parameters. We thereby defined some theoretical modes of playing: horizontal, vertical, auto-regressive, memoryless. We could avoid the imposition of a particular playing mode by using machine-learning algorithms to train the expressive models. Nevertheless, we showed how we can further investigate on the derived models to test whether they are prevalently of one or another mode. In particular, we showed that the models built on ensemble performance make a relative higher use of vertical context descriptors than models built on solo performance, and thus reflect a “prevalently vertical” mode of playing.

6.3 Critique

In this dissertation, we focused on the first problems that a data-driven research on EEP must address: acquiring a dataset of ensemble performances, extracting ensemble specific context features, training EEP models and interpreting the results. In approaching each of those problems, we had to make choices that might be legitimately criticized. We provide here some seeds that might sparkle the debate in the future and might lead to improvements in future works.

Dataset size One main limitation of this work was given by the absence of databases with annotated ensemble recordings and this is why one of our contributions was to carry out multi-modal ensemble recording of string quartets. We believe that on larger datasets we could investigate in-depth information about musicians, roles and music interaction in ensembles allowing a comparison among different music styles and different quartets.

Orthogonality and completeness of contextual features The set of score contextual descriptors of Section 4.1 have, in principle, no redundancy (e.g. pitch and rhythm can be assumed uncorrelated). However, in practice, cross correlations among contextual features can appear depending on the piece structure, recurrent motifs, recurrent counterpoint relations, and recurrent harmonic sequences. In probabilistic terms, we could say that distinct feature sets are not *orthogonal*. If this happens the boundary between horizontal and vertical score context might become blurry. Our results on the percentage of features seem to partially exclude this case (see Section 5.3.6). However, a thorough analysis of the correlations among different score descriptors might help clarify this aspect.

We could say that a set of contextual features is *complete* if it contains enough information in order to reconstruct the score. This was never the case in this dissertation, since we, at least, discarded dynamic marks, articulation marks and ornaments. These information might have been helpful in improving the results and should be considered in the future. Whether it will be possible to derive a complete feature set (to be used on any score and style) is still questionable.

It is because of a possible score bias that we could not compare models built on different corpora. In the case of statistical analysis, the main issue was balancing the levels of the ANOVA since in different music scores the contextual variables might need to be quantized in different ways (see

Section 5.2.3). Analogously, the music score can bias results in ensemble model analysis. The accuracy of each model might be affected by the score, making unsuitable any comparison between performances of different scores. It is questionable whether it would be possible to derive universal features sets as the cross correlations might depend on music genre. However it might be feasible to investigate whether such set exist in a particular music style/genre. Such analysis might have interesting implications not only in EEP but also in the field of music representation and music generation.

6.4 Outlook

We regard this work as the very preliminary step towards expressive performance analysis of ensemble performance. We can foresee a range of applications and extensions of this work. On one side there are direct applications which could reproduce our method to analyze additional performances. On the other side, we can imagine ways in which the method could be extended to other type of data and for additional purposes.

Study of rehearsals One interesting aspect that could be studied is the evolution of musical interpretation during rehearsals. From a psychology standpoint, some studies have studied team roles in ensemble rehearsals (King, 2006). By using the proposed recording setup, we could monitor the evolution of the interpretation and the coordination during rehearsals of a string quartets. Such an analysis could complement the interviews with the musicians that are currently one of the main methodological tools of psychologists. Such a research could have applications in the didactics of ensemble playing.

Another interesting aspect of ensemble performance is how the different musical characters of each member of the ensemble blend with each other to constitute a joint interpretation of a piece. The study of rehearsals could give some insights over how this blend happens. Another objective could be to model the dynamics of the negotiations (implicit and explicit) happening during the rehearsals among musicians. This would be particularly interesting from a social science standpoint. Additionally, it could lead to the possibility of predicting how several individual solo models merge into an ensemble performance model.

Study of other types of ensembles A direct application could be applying the same score contexts and machine-learning tools to other types

of ensembles. In such a study, an important preliminary step would be to understand the expressive capabilities of the music instruments composing the ensemble. It would be particularly interesting to see if it is possible to derive similar results on other types of ensemble. In the case of improvised music such in the case of Jazz, the analysis might considerably differ because of the lack of a pre-established music score. In that case the goal of the musicians would be intrinsically different requiring thus to adapt the contextual features, or even to change completely the approach.

Non-verbal communication With the addition of other sensors, and motion capture data the study could be extended for including non-verbal communication elements. In particular the body gesture of fellow musicians could be integrated in each model as a set of features. Embodiment in music performance is an aspect that has been studied in the recent past, finding correlations between body gestures and sound produced during the performance (e.g. Wanderley et al.'s (2005) work). Such an extension of our method would allow studying the importance of non-verbal communication in comparison with auditory cues in different ensemble contexts.

Model extension The study of rehearsals could lead to models of each performer that do not only react to other musicians, but that also *negotiate with* other musicians. This would result in models that are able to address conflict-resolution problems typical of game theory and multi-criteria optimization. We believe this could be a direction of research where a major leap could be achieved.

Large ensembles It is unclear how our method would scale to large ensembles such as orchestras. A grow in the number of musicians would produce exponentially increasing number of features. In that case some approach of automatic dimensionality reductions might be considered, but could still be insufficient. It might be needed to make further assumptions on the interaction among musicians, by dividing the group in hierarchical sections of musicians and focusing on the prediction of collective performance features. It would be interesting if the hierarchy could emerge as a result of the analysis of the individuals' behaviors thus justifying the approach.

Expressive performance generation Expressive performance generation is also another important future application of our approach. In or-

der to implement a system that generates string quartet expressive performances, some additional steps will be required. Firstly, there is the need to find a proper synthesizer of string instruments that is able to correctly reproduce different intensities and vibratos. One straight-forward candidate system could be the spectral synthesis model developed in the Ph.D. thesis by Pérez (2009). The system works by directly synthesizing the violin sound from the bowing gestures trajectories; the problem of generating such trajectories from the score and expression parameters has already been considered by Maestre (2009). However, respect to these works there is the additional requirement to synthesize vibrato trajectories and, since we did not predict it, the force. Other options for the synthesis of string quartets sounds are commercial software based on sound banks that provide relatively realistic results albeit the reduced flexibility. Secondly, once a synthesizer has been selected, there is the need to map the predicted expression parameters to synthesizer parameters, thus calibrating loudness levels to avoid perceptual bias. Thirdly, each note in the score should be predicted sequentially, eventually switching from one model to another so that the auto-regressive features on the past events are fed to the models correctly. Lastly, the note durations should be adjusted to prevent from accumulating time lags among voices. The approach by Hashida et al. (2007) might be a good candidate (see Section 2.3.3) to perform such post-processing adjustment of durations.

Music accompaniment task In the last years we have increasingly assisted to the appearance of music accompanying systems for training and didactic purposes. Band-in-a-Box² is a MIDI music arranger software that allows to generate an accompaniment for a given sequence of chords and a given style. Such system provides an easy way to overcome the need of having a band to practice a given performance. However the result accompaniment is generally robotic and does not adapt to the performance of the musician in terms of loudness and/or tempo. Using the method developed in this dissertation it would be possible to build ensemble models for each instrument from the data acquired from human bands performances. We could use the models to render the output of the software more human and adaptive to the performance.

²<http://www.pgmusic.com>

Bibliography

Each reference indicates the pages where it appears.

Anders Askenfelt. Measurement of bow motion and bow force in violin playing. *The Journal of the Acoustical Society of America*, 80(4):1007–1015, October 1986. doi: 10.1121/1.393841. 25

Anders Askenfelt. Measurement of the bowing parameters in violin playing. ii: Bow-bridge distance, dynamic range, and limits of bow force. *The Journal of the Acoustical Society of America*, 86(2), 1989. 68, 99

Ingmar Bengtsson and Alf Gabrielsson. *Rhythm research in Uppsala*. Musical Academ, 1977. 15

Ingmar Bengtsson and Alf Gabrielsson. Analysis and synthesis of musical rhythm. *Studies of music performance*, 39:27–60, 1983. 15

Stephen A Billings. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013. 102

AS Bregman. Auditory scene analysis: The perceptual organization of sound. 1990, 1990. 7

Roberto Bresin. Artificial neural networks based models for automatic performance of musical scores. *Journal of New Music Research*, 27(3):239–270, 1998. 21

Ali Taylan Cemgil, Bert Kappen, Peter Desain, and Henkjan Honing. On tempo tracking: Tempogram representation and kalman filtering. *Journal of New Music Research*, 29(4):259–273, 2000. 92

Ching-Hua Chuan and Elaine Chew. A dynamic programming approach to

- the extraction of phrase boundaries from tempo variations in expressive performances. In *Proceedings of the International Conference on Music Information Retrieval*, pages 305–308. Citeseer, 2007. 92
- Herbert H Clark. Managing problems in speaking. *Speech communication*, 15(3):243–250, 1994. 16
- Lothar Cremer. *Physics of the Violin*. The MIT Press, Cambridge, Massachusetts, USA, November 1984. 25
- Renzo Cresti. *La Vita della Musica, ipertesto di Storia della musica*. Feeria, 1970. 2
- Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000. 120
- Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002. 40, 98
- Ramon Lopez De Mantaras and Josep Lluís Arcos. Ai and music: From composition to expressive performance. *AI magazine*, 23(3):43, 2002. 21
- M. Demoucron and R. Caussé. Sound synthesis of bowed string instruments using a gesture based control of a physical model. In *Proceedings of the 2007 International Symposium on Musical Acoustics*, Barcelona, 2007. 26
- M. Demoucron, A. Askenfelt, and R. Causse. Measuring bow force in bowed string performance: Theory and implementation of a bow force sensor. *Acta Acustica united with Acustica*, 95(4):718–732, 2009. 25
- Anders Friberg. *A quantitative rule system for musical performance*. PhD thesis, KTH, Sweden, 1995. 19, 66
- Alf Gabrielsson. The performance of music. *The psychology of music*, 2: 501–602, 1999. 14
- Alf Gabrielsson. Music performance research at the millennium. *Psychology of music*, 31(3):221–272, 2003. 1, 14, 15
- Alf Gabrielsson and Patrik N Juslin. Emotional expression in music performance: Between the performer’s intention and the listener’s experience. *Psychology of music*, 24(1):68–91, 1996. 22
- Werner Goebel and Roberto Bresin. Measurement and reproduction accuracy of computer-controlled grand pianos. *The Journal of the Acoustical Society of America*, 114(4):2273–2283, 2003. 97
- Werner Goebel and Caroline Palmer. Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary*

- Journal*, 26(5):427–438, 2009. 17
- Emilia Gómez. *Tonal description of music audio signals*. PhD thesis, PhD thesis, UPF Barcelona, 2006. 40
- Peter Grosche, M Muller, and Frank Kurth. Cyclic tempogram—a mid-level tempo representation for musicsignals. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5522–5525. IEEE, 2010. 92
- E. Guaus, J. Bonada, E. Maestre, A. Perez, and M. Blaauw. Calibration method to measure accurate bow force for real violin performances. In G. Scavone, editor, *International Computer Music Conference*, pages 251–254, Montreal, Canada, 16/08/2009 2009. The International Computer Music Association, The International Computer Music Association. URL <http://mtg.upf.edu/files/publications/eguaus-ICMC-09.pdf>. 25, 26, 43
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. 120
- Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. 122
- Mitsuyo Hashida, Noriko Nagata, and Haruhiro Katayose. jpop-e: an assistant system for performance rendering of ensemble music. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 313–316. ACM, 2007. 23, 142
- Perfecto Herrera and Jordi Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998. 98
- Ray S Jackendoff and Fred Lerdahl. *A generative theory of tonal music*. Cambridge, Mass.: MIT Press, 1983. 15
- Margaret L Johnson. An expert system for the articulation of bach fugue melodies. *Readings in computer-generated music*, pages 41–51, 1992. 20
- Patrik N Juslin, Anders Friberg, and Roberto Bresin. Toward a computational model of expression in music performance: The germ model. *Musicae Scientiae*, 5(1 suppl):63–122, 2002. 22
- Charles Keil. Participatory discrepancies and the power of music. *Cultural Anthropology*, 2(3):275–283, 1987. ISSN 1548-1360. doi: 10.1525/can.1987.2.3.02a00010. URL <http://dx.doi.org/10.1525/can.1987.2.3.02a00010>. 90
- Peter E Keller. Joint action in music performance. *Emerging Communica-*

- tion, 10:205, 2008. 17
- Roger A Kendall and Edward C Carterette. The communication of musical expression. *Music perception*, pages 129–163, 1990. 22, 35
- Tae Hun Kim, Satoru Fukayama, Takuya Nishimoto, and Shigeki Sagayama. Polyhymnia: An automatic piano performance system with statistical modeling of polyphonic expression and musical symbol interpretation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 96–99, 2011. 23
- Elaine C King. The roles of student musicians in quartet rehearsals. *Psychology of music*, 34(2):262–282, 2006. 140
- Hiroshi Kinoshita and Satoshi Obata. Left hand finger force in violin playing: Tempo, loudness, and finger differences. *The Journal of the Acoustical Society of America*, 126(1):388–395, 2009. 25
- Alexis Kirke and Eduardo R Miranda. An overview of computer systems for expressive music performance. In *Guide to Computing for Expressive Music Performance*, pages 1–47. Springer, 2013. 18, 22
- J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, 1999. ISSN 1052-6234. 50
- E. Maestre, J. Bonada, M. Blaauw, A. Perez, and E. Guaus. Acquisition of violin instrumental gestures using a commercial emf device. In *International Computer Music Conference*, Copenhagen, Denmark, 27/08/2007 2007. URL files/publications/bd052f-ICMC07-Maestre-Bonada-Blaauw-Perez-Guaus.pdf. 25, 43
- Esteban Maestre. *Modeling instrumental gestures: an analysis/synthesis framework for violin bowing*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, Novembre 2009. 10, 25, 38, 41, 42, 67, 81, 99, 137, 142
- Marco Marchini, Panos Papiotis, Esteban Maestre, and Alfonso Pérez. A hair ribbon deflection model for low-intrusiveness measurement of bow force in violin performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Oslo, Norway, 2011. URL <http://www.nime2011.org/proceedings/papers/M17-Marchini.pdf>. 10, 43
- Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban Maestre. Inducing rules of ensemble music performance: a machine learning approach. In *3rd international conference on Music & Emotion, Jyväskylä*, 2013. 121
- Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban Maestre.

- The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43(3):303–317, 2014. doi: 10.1080/09298215.2014.922999. URL <http://www.tandfonline.com/doi/abs/10.1080/09298215.2014.922999>. 119
- George P Moore and Jessie Chen. Timings and interactions of skilled musicians. *Biological cybernetics*, 103(5):401–414, 2010. 17, 103
- Eugene Narmour. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992. 15, 60
- Elisabeth Pacherie. The phenomenology of joint action: self-agency vs. joint-agency. *Joint attention: new developments*, pages 343–389, 2012. 16
- Caroline Palmer. Mapping musical thought to musical performance. *Journal of experimental psychology: human perception and performance*, 15(2):331, 1989. 15
- Caroline Palmer. Anatomy of a performance: Sources of musical expression. *Music Perception*, pages 433–453, 1996a. 16, 23
- Caroline Palmer. On the assignment of structure in music performance. *Music Perception*, pages 23–56, 1996b. 16, 23
- Caroline Palmer. Music performance. *Annual review of psychology*, 48(1):115–138, 1997. 1, 13, 15
- Panos Papiotis. *Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data*. PhD thesis, Universitat Pompeu Fabra, in preparation. 34
- Panos Papiotis, Marco Marchini, Alfonso Perez-Carrillo, and Esteban Maestre. Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. *Frontiers in Psychology*, 5(963), 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00963. URL http://www.frontiersin.org/cognitive_science/10.3389/fpsyg.2014.00963/abstract. 34
- J. A. Paradiso and N. A. Gershenfeld. Musical applications of electric field sensing. *Computer Music Journal*, 21(2):69–89, 1997. 25
- Alfonso Pérez. *Enhancing spectral synthesis techniques with performance gestures using the violin as a case study*. PhD thesis, Ph. D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2009. 25, 38, 142
- Giovanni De Poli. Methodologies for expressiveness modelling of and for music performance. *Journal of New Music Research*, 33(3):189–202, 2004. 22

- Eric Prame. Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America*, 102:616, 1997. 25, 97
- J. A. Prögler. Searching for swing: Participatory discrepancies in the jazz rhythm section. *Ethnomusicology*, 39(1):pp. 21–54, 1995. ISSN 00141836. URL <http://www.jstor.org/stable/852199>. 90
- R. Ramirez and A. Hazan. Discovering expressive transformation rules from saxophone jazz performances. *Journal of New Music Research*, 34(4):319–330, 2005. 118
- Rafael Ramirez and Amaury Hazan. A tool for generating and explaining expressive music performances of monophonic jazz melodies. *International Journal on Artificial Intelligence Tools*, 15(04):673–691, 2006. 21, 118
- Nicolas Rasamimanana. Gesture analysis of bow strokes using an augmented violin. Master’s thesis, IRCAM, Paris, France, 2003. 26
- Bruno Repp. Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review*, 12(6):969–992, 2005. ISSN 1069-9384. 17, 91, 133
- Bruno H Repp. Diversity and commonality in music performance: An analysis of timing microstructure in schumann’s “träumerei”. *The Journal of the Acoustical Society of America*, 92:2546, 1992. 16
- Bruno H Repp. The dynamics of expressive piano performance: Schumann’s “träumerei” revisited. *The Journal of the Acoustical Society of America*, 100(1):641–650, 1996. 16
- E. Schoonderwaldt, N. Rasamimanana, and F. Bevilacqua. Combining accelerometer and video camera: Reconstruction of bow velocity profiles. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Paris, France, jun 2006. IRCAM. 25
- Erwin Schoonderwaldt. *Mechanics and acoustics of violin bowing*. PhD thesis, Stockholm Royal Institute of Technology, Stockholm, Sweden, 2009a. 24, 25, 43
- Erwin Schoonderwaldt. The violinist’s sound palette: Spectral centroid, pith flattening and anomalous low frequencies. *Acta Acustica united with Acustica*, 95(5):901–914, 2009b. 25
- Carl Emil Seashore. *The psychology of music*. Courier Dover Publications, 1938. 8, 14, 24, 25, 26
- J. Sundberg, L FrydÈn, and A. Askenfelt. What tells you the player is musical? an analysis-by-synthesis study of music performance. In *Quarterly*

- Progress and Status Report - KTH*. Royal Swedish Academy of Music, 1983. 19
- Johan Sundberg, Anders Friberg, and Lars Frydén. Rules for automated performance of ensemble music. *Contemporary Music Review*, 3:89–109, 1989. doi: 10.1080/07494468900640071. 17, 66, 118, 137
- Kim Thomas. Just noticeable difference and tempo change. *Journal of Scientific Psychology*, pages 14–20, 2007. 110
- Neil P McAngus Todd. The dynamics of dynamics: A model of musical expression. *The Journal of the Acoustical Society of America*, 91(6):3540–3550, 1992. 20, 92, 94
- Leroy Ninde Vernon. *Synchronization of chords in artistic piano music*. PhD thesis, note on label mounted at head of title. University of Iowa, 1937. 22
- Marcelo M Wanderley, Bradley W Vines, Neil Middleton, Cory McKay, and Wesley Hatch. The musical significance of clarinetists' ancillary gestures: an exploration of the field. *Journal of New Music Research*, 34(1):97–113, 2005. 141
- G. Widmer and W. Goebel. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3): 203–216, 2004. 21
- Gerhard Widmer. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50, 2002. 21, 22, 104, 118
- Gerhard Widmer and Asmir Tobudic. Playing Mozart by Analogy: Learning Multilevel Timing and Dynamics Strategies. *Journal of New Music Research*, 32:259–268, 2003. doi: 10.1076/jnmr.32.3.259.16860. 91, 133
- Alan M. Wing, Satoshi Endo, Adrian Bradbury, and Dirk Vorberg. Optimal feedback correction in string quartet synchronization. *Journal of The Royal Society Interface*, 11(93), 2014. doi: 10.1098/rsif.2013.1125. URL <http://rsif.royalsocietypublishing.org/content/11/93/20131125.abstract>. 17
- J. Woodhouse and P. M. Galluzzo. The bowed string as we know it today. *Acta Acustica united with Acustica*, 90(4):579–589, 2004. 24
- Diana S. Young. Wireless sensor system for measurement of violin bowing parameters. In *Proceedings of the Stockholm Music Acoustics Conference*, Stockholm, Sweden, 2003. 25, 26



Publications by the author

Peer-reviewed journals

Srikanth Cherla, Hendrik Purwins, and Marco Marchini. Automatic phrase continuation from guitar and bass guitar melodies. *Computer Music Journal*, 37(3):68–81, 2013.

Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban Maestre. The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43(3):303–317, 2014. 119

Panos Papiotis, Marco Marchini, Alfonso Perez-Carrillo, and Esteban Maestre. Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. *Frontiers in Psychology*, 5(963), 2014. ISSN 1664-1078. 34

Full-article contributions to peer-reviewed conferences

A.R. Addressi, L. Ferrari, D. Benghi, F. Regazzi, M. Marchini, and F. Pachet. The children and the continuator. In *4th Conference of the European Network of Music Educators and Researchers of Young Children*, pages 597–608, Bologna, 22/07/2009 2009. Bononia University Press, Bononia University Press.

Katerina Kosta, Marco Marchini, and P. Purwins. Unsupervised chord-sequence generation from an audio example. In *International Society for Music Information Retrieval Conference (ISMIR 2012)*, 2012.

Marco Marchini and P. Purwins. Unsupervised generation of percussion sound sequences from a sound example. In *Sound and Music Computing Conference, Barcelona*, 2010a.

Marco Marchini and P. Purwins. An unsupervised system for the synthesis of variations from audio percussion patterns. In *7th International Symposium on Computer Music Modeling and Retrieval (CMMR), Malaga*, pages 277–278, 2010b.

Marco Marchini, Panos Papiotis, Esteban Maestre, and Alfonso Pérez. A hair ribbon deflection model for low-intrusiveness measurement of bow force in violin performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Oslo, Norway, 2011. 10, 43

Marco Marchini, Panos Papiotis, and Esteban Maestre. Timing synchronization in string quartet performance: a preliminary study. In *International Workshop on Computer Music Modeling and Retrieval (CMMR12)*, pages 117–185, 2012.

Marco Marchini, Panos Papiotis, and Esteban Maestre. Investigating the relationship between expressivity and synchronization in ensemble performance: an exploratory study. In *International Symposium on Performance Science, Vienna*, 2013a.

Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban Maestre. Inducing rules of ensemble music performance: a machine learning approach. In *3rd international conference on Music & Emotion, Jyväskylä*, 2013b.

O. Mayor, Quim Llimona, M. Marchini, Panos Papiotis, and Esteban Maestre. repovizz: a framework for remote storage, browsing, annotation, and exchange of multi-modal data. In *ACM International Conference on Multimedia (MM'13)*, Barcelona, 2013.

P. Papiotis, Esteban Maestre, Marco Marchini, and Alfonso Pérez. Synchronization of intonation adjustments in violin duets: towards an objective evaluation of musical interaction. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, Paris, 19/09/2011 2011a.

P. Papiotis, Marco Marchini, Esteban Maestre, and Alfonso Pérez. Measuring ensemble synchrony through violin performance parameters: a preliminary progress report. In *2nd Workshop of Social Behavior in Music (SBM-2011/INTETAIN-2011)*, Genova, Italy, 27/05/2011 2011b.

P. Papiotis, Marco Marchini, and Esteban Maestre. Computational analysis of solo versus ensemble performance in string quartets: Dynamics and intonation. In *12th International Conference on Music Perception and Cognition*, Thessaloniki, Greece, 23/07/2012 2012.

Panos Papiotis, Perfecto Herrera, Marco Marchini, and Esteban Maestre. Aural-based detection and assessment of real versus artificially synchronized string quartet performance. In *3rd international conference on Music & Emotion, Jyväskylä*, 2013a.

Panos Papiotis, Marco Marchini, and Esteban Maestre. Multidimensional analysis of interdependence in a string quartet. In *International Symposium on Performance Science, Vienna*, 2013b.

Book chapters

Marco Marchini and Hendrik Purwins. Unsupervised analysis and generation of audio percussion sequences. In Sølvi Ystad, Mitsuko Aramaki, Richard Kronland-Martinet, and Kristoffer Jensen, editors, *Exploring Music Contents*, volume 6684 of *Lecture Notes in Computer Science*, pages 205–218. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23125-4.

Theses

Marco Marchini. Unsupervised generation of percussion sequences from a sound example. Master's thesis, Universitat Pompeu Fabra, 2010.

