Université de Lille

École Graduée MADIS-631

COMPUTATIONAL ANALYSIS OF EXPRESSIVE TIMING IN SOLO FLUTE PERFORMANCES

Analyse computationnelle de l'expressivité temporelle dans les interprétations de flûte solo

THÈSE

pour obtenir le grade de :

Docteur de l'Université de Lille

dans la spécialité Informatique

par

Patrice THIBAUD

Thése à soutenir le 23 Novembre 2026 devant le jury composé de:

Elaine CHEW	Professeure, King's College, London	Rapportrice
Jean-Louis GIAVITTO	Directeur de recherche CNRS, STMS Lab, IRCAM	Rapporteur
Juliette CHABASSIER	Chargée de Recherche INRIA, Modartt	Examinatrice
Anja VOLK	Professeure, Utretch University	Examinatrice
Clarisse DHAENENS	Professeure des Universités, Université de Lille, CRIStAL	Présidente
Mathieu GIRAUD	Directeur de Recherche, CNRS, CRIStAL	Directeur
Louis BIGO	Professeur des Universités, Bordeaux INP, Université de Bordeaux, LaBRI, SCRIME	Co-directeur
Yann TEYTAUT	Maître de Conférences, Université de Lille, CRIStAL	Co-encadrant

Résumé

Aucune interprétation d'une même œuvre musicale n'est exactement identique à une autre. Même lorsqu'ils suivent une même partition, les interprètes façonnent en permanence le rythme, l'articulation, les nuances et le phrasé afin d'exprimer leur propre vision musicale. Ces variations sont au cœur de l'expression musicale. Comprendre et modéliser ces nuances constitue aujourd'hui un enjeu central en extraction d'information musicale et, plus largement, en analyse computationnelle de l'interprétation musicale.

Cette thèse s'intéresse plus particulièrement à la micro-temporalité expressive, c'est-à-dire aux variations temporelles de très faible amplitude qui contribuent à l'expressivité musicale, ainsi qu'aux défis spécifiques posés par le répertoire pour flûte seule.

Premièrement, nous présentons deux corpus annotés originaux : les *Douze Fantaisies* de Telemann pour flûte solo, représentatives du répertoire baroque, ainsi qu'un corpus consacré à deux œuvres majeures du début du XX^{e} siècle, *Syrinx* de Debussy et *Density 21.5* de Varèse. Ces jeux de données intègrent des annotations multi-niveaux décrivant la structure tonale, les voix implicites, les motifs musicaux, ainsi que les positions temporelles précises des notes.

Nous développons ensuite plusieurs stratégies d'alignement musical. D'une part, nous introduisons une méthode d'alignement partition-audio reposant sur des modèles harmoniques adaptés aux spécificités de la flûte traversière ; d'autre part, nous comparons cette approche à différentes représentations audio, notamment les chromagrammes et les estimations de hauteur multiples issues de réseaux de neurones. Nous montrons que les erreurs d'alignement présentent des biais temporels spécifiques aux représentations utilisées et qu'un post-traitement simple permet d'en améliorer significativement la précision. Nous proposons également une nouvelle stratégie de fusion tardive pour l'alignement audio–audio, reposant sur la cohérence de transferts multiples, qui renforce la robustesse des alignements.

Enfin, l'analyse des données alignées à l'aide de nouvelles métriques basées sur les intervalles entre attaques met en évidence des stratégies d'interprétation liées à la structure métrique, aux groupements rythmiques et à la polyphonie implicite. Les interprètes soulignent les voix entrelacées et les motifs musicaux par des asymétries temporelles systématiques. Ces travaux ouvrent des perspectives vers une analyse multimodale de l'interprétation intégrant notamment la capture du souffle et les techniques instrumentales étendues.

Mots-clés : alignement, micro-temporalité, flûte, analyse d'interprétation, MIR, Debussy, Telemann, Varèse.

Abstract

No two performances of the same musical work are ever exactly alike. Even when performers follow the same score, they continuously shape timing, articulation, dynamics, and phrasing to convey their own musical interpretation. These variations are at the heart of musical expression. Understanding and modelling these nuances has become a central topic in Music Information Retrieval and, more broadly, in computational performance analysis.

This thesis specifically investigates expressive microtiming, i.e., very small-scale temporal variations that contribute to musical expressivity, as well as the specific challenges posed by the solo flute repertoire.

First, we introduce two original annotated corpora: Telemann's *Twelve Fantasias* for solo flute, representative of the Baroque repertoire, and a corpus dedicated to two landmark works of the early twentieth century, Debussy's *Syrinx* and Varèse's *Density 21.5*. These datasets include multi-level annotations describing tonal structure, implicit voices, musical motives, and precise note onset times.

Next, we develop several music alignment strategies. On the one hand, we introduce a score-to-audio alignment method based on harmonic models adapted to the specific characteristics of the flute; on the other hand, we compare this approach with different audio representations, including chromagrams and neural network-based multi-pitch estimations. We show that alignment errors exhibit temporal biases specific to the representations used and that a simple post-processing step effectively compensates for these biases, significantly improving alignment accuracy. We also propose a new late-fusion strategy for audio-to-audio alignment, based on the consistency of multiple transfer paths, which improves alignment robustness.

Finally, using newly proposed metrics based on inter-onset intervals, we analyse aligned data and reveal interpretive strategies related to metric structure, rhythmic grouping, and implicit polyphony. Performers emphasise interleaved voices and musical patterns through systematic temporal asymmetries. These findings open new perspectives for multimodal performance analysis, notably through the integration of breath capture and extended instrumental techniques.

Keywords: alignment, microtiming, flute, performance analysis, MIR, Debussy, Telemann, Varèse.

Contents

Résumé **2**

Abstract **3**

Acronyms **8**

Remerciements/Acknowledgments **10**

1 Introduction **12**
1.1 Motivations 13
1.2 The Flute: a Brief Introduction 14
1.3 Solo Flute Music 15
1.4 Outline of this Thesis 19
1.5 Publications and Communications 21

I Theoretical Framework and Tools **23**

2 Processing Music: From Representation to Perception **24**
2.1 Symbolic Representations 25
2.1.1 Score Based Representations 27
2.1.2 Event-Based and Derived Representations 28
2.2 Audio Representations and Feature Extraction 30
2.2.1 From Time-Domain Signals to Time–Frequency Representations 31
2.2.2 Pitch Features 33
2.2.3 Chroma Representations 34
2.2.4 Cepstral Representations 38
2.2.5 Learned Features 40
2.2.6 Onset Activation Functions 40

2.3 Performance . . . 44
2.3.1 Parameters of Performance and Representations . . . 45
2.3.2 Focus on Timing . . . 46
2.4 Perceptual and Musicological Models . . . 48
2.4.1 Perception of Realisation . . . 48
2.4.2 Implied Voices . . . 50

3 Musical Alignment: Tasks, Methods and Tools 54
3.1 General Framework of Musical Alignment . . . 55
3.2 Alignment Tasks and Scenarios . . . 56
3.2.1 Alignment Modalities . . . 56
3.2.2 Offline vs. Online Alignment . . . 58
3.3 Alignment Algorithms: Focus on Dynamic Time Warping . . . 61
3.3.1 Overview of Alignment Algorithms . . . 61
3.3.2 Dynamic Time Warping (DTW) . . . 63
3.3.3 Multi-Scale and Multi-Resolution DTW . . . 67
3.4 Frameworks . . . 71
3.4.1 Overview of Existing Frameworks for Offline Alignments . . . 71
3.4.2 Focus on Synctoolbox . . . 72
3.5 Annotations . . . 73
3.5.1 Annotations Transfer . . . 73
3.5.2 Triple Transfer and Late-Fusion . . . 74
3.6 Datasets . . . 76

II Contributions : Data, Methods and Performance Analysis in Solo Flute Music 79

4 Telemann, Debussy, Varèse: Two Solo Flute Datasets 80
4.1 Fantasias Dataset . . . 82
4.1.1 Musical Annotations . . . 85
4.1.2 Recordings and Unfolded Scores . . . 91
4.2 Early 20th Dataset . . . 94
4.2.1 Scores and Recordings . . . 96
4.2.2 Onset Annotations . . . 96
4.2.3 Musical Patterns . . . 98
4.3 Data Availability . . . 101
4.3.1 Dataset Fantasias . . . 102
4.3.2 Dataset Early 20th . . . 102

5 Improving Alignment Methods for Solo Flute **104**
5.1 Spectral Templates in Score-to-Audio Alignments . . . 106
5.1.1 Fantasias Alignments . . . 106
5.1.2 Alignment Evaluation . . . 108
5.2 Comparison of Audio Features in Score-to-Audio Alignment . . . 113
5.2.1 Dataset of Monophonic Music Performances . . . 113
5.2.2 Alignment Score-to-Audio . . . 116
5.2.3 Audio Features Evaluation . . . 118
5.2.4 Optimal Shift Investigation . . . 120
5.2.5 Annotator Agreement and Robustness of Alignment Shift . . . 122
5.3 Soft-Late Fusion Strategy for Audio-to-Audio Alignments . . . 126
5.3.1 Audio-to-Audio Alignment . . . 126
5.3.2 Evaluation of Alignment Quality . . . 129

6 Performance Analysis of Expressive Timing **137**
6.1 Levels of Temporal Analysis . . . 139
6.2 From Notated Rhythm to Performed Timing in Fantasias . . . 140
6.2.1 Rhythmic Deviation Metrics . . . 140
6.2.2 Fantasias Across Performances . . . 142
6.3 Timing Across Patterns in 20th Solo Flute Music . . . 149
6.3.1 Musical Patterns Selection . . . 149
6.3.2 Timing Descriptors . . . 152
6.3.3 Inter-Performer Analysis of Patterns in Syrinx . . . 153
6.3.4 Timing Analysis of Patterns in Density 21.5 . . . 156

7 Conclusion and Perspectives **161**
7.1 Contributions and Discussion . . . 162
7.2 Perspectives . . . 163
7.2.1 Tonguings . . . 163
7.2.2 Extended Techniques . . . 164
7.2.3 Performance Capture . . . 165
7.2.4 Respiratory Control and Sound Production . . . 166
7.3 Toward Richer Representations of Musical Performance . . . 168

A Interleaved Voices Detection **169**

Interleaved Voices Detection **169**
A.1 Algorithm for Voice Change Detection . . . 170
A.1.1 Feature Design and Scoring Function . . . 170
A.1.2 Algorithm . . . 171

A.2 Voice Change Detection in the Fantasias 172
A.2.1 Reference Annotations 172
A.2.2 Evaluation Protocol 172
A.2.3 Results 173
A.3 Study Case: Oboe 3 Parts Fugue 175
A.3.1 Evaluation of the Voice-Fission Algorithm 176

List of Figures **178**

List of Tables **185**

Bibliography **188**

Acronyms

CENS Chroma Energy Normalized Statistics. 37, 68, 72

CNN Convolutional Neural Network. 57

CQT Constant-Q Transform. 35, 36

DAW Digital Audio Workstation. 29, 178

DLNCO Decaying Locally adaptive Normalized Chroma-based Onset. 38, 72, 117, 127

DTW Dynamic Time Warping. 61–63, 65–68, 70–72, 107, 113, 117, 126

HMM Hidden Markov Model. 33, 62, 65

IIR Infinite Impulse Response Filter Bank. 117, 127

IMSLP International Music Score Library Project. 85

IOI Inter Onset Interval. 20, 46, 98, 138, 148, 152, 158, 160, 186

IQR Interquartile Range. 154, 155, 157, 183

ISMIR International Society for Music Information Retrieval. 76

JSON JavaScript Object Notation. 152

MEI Music Encoding Initiative. 28, 46

MFCC Mel-Frequency Cepstral Coefficients. 38, 39

MIDI Musical Instrument Digital Interface. 28–30, 33, 57, 166, 167

MIR Music Information Retrieval. 13, 14, 20, 25, 30, 42, 55, 56, 61, 62, 71, 75, 76, 78, 81, 98, 163, 165, 168, 180

MPE Multi-Pitch Estimation. 40, 113

MrMsDTW Memory-Restricted Multiscale Dynamic Time Warping. 68–70, 72, 117, 118, 127, 180

MsDTW Multiscale Dynamic Time Warping. 67, 70

PAT Perceptual Attack Time. 97

PFCC Perceptual Frequency Cepstral Coefficients. 38, 39

PSD Power Spectral Density. 57

QPM Quarter notes per minutes. 149, 183

RMS root-mean-square amplitude. 45

STFT Short-Time Fourier Transform. 31, 32, 72, 106, 117

WSD Wagner Ring Dataset. 59, 180

XML Extensible Markup Language. 27, 28

Remerciements/Acknowledgments

First of all, I would like to thank my reviewers, *Elaine Chew* and *Jean-Louis Giavitto*, for the time they spent thoroughly reading this thesis. I would also like to thank *Juliette Chabassier*, *Clarisse Dhaenens* and *Anja Volk* who have kindly agreed to be part of my jury. Thank you to *Philippe Rigaux* for agreeing to serve on my *CSI* (individual monitoring committee).

When I embarked on this adventure, I was a PRAG in the Computer Science Department at the University of Lille. After several years spent teaching Physics and Chemistry in middle and high schools, I decided to take my professional career in a new direction by devoting myself to Computer Science. This manuscript, in a way, marks the end of a chapter in this journey, but it also opens up new paths ahead. I would like to take this opportunity to express my gratitude to the many people who have been present at different stages of this journey and who have contributed to it in various ways.

Tout d'abord, un grand merci à *Mathieu*, qui a eu la gentillesse de m'accueillir dans l'équipe pour échanger autour de l'informatique et de la musique, un an avant le démarrage de ce projet de thèse. Ma première participation aux JIM en 2023 m'a définitivement convaincu que je ne m'étais pas trompé en choisissant cette voie. Tu as été ensuite un soutien précieux tout au long de ces trois ans, et j'ai énormément appris à tes côtés.

Je tiens également à remercier chaleureusement mes co-directeurs, *Louis* et *Yann*, pour la richesse de nos échanges, leurs conseils avisés et leurs précieuses remarques tout au long de l'élaboration de ce manuscrit. La complémentarité de leurs points de vue et de leurs approches a joué un rôle essentiel dans la construction de ce travail.

Un immense merci à *Manu* pour tout le temps consacré à Dezrann (et pour avoir supporté mes nombreuses sollicitations). Merci également à *Ken* pour nos échanges et ses conseils à différents moments de ces trois années. Je remercie aussi *Florence* pour ses retours, notamment avant ma première soumission à ISMIR.

Je pense bien évidemment à l'ensemble des collègues, étudiants et étudiantes, post-doctorants et stagiaires (dont certains étaient d'anciens étudiants) que j'ai eu la chance de côtoyer au sein de l'équipe Algomus : *Ajda*, *Aurélien*, *Alexandre*, *Baptiste*, *Charles*, *Claudio*, *David*, *Francesco*, *Jérémie*, *Louis*, *Marc-Antoine*, *Pascal*, *Pouya*, *Quentin*, *Richard*, *Romain*,

Rui, *Sebastian*, *Théo*, *Tom*, *Valentin*, *Vanessa*, *VietToan*, *Xiaobo*, *Yanis*, *Zakaria*, *Ziyun*... (la liste est certainement incomplète, et je prie celles et ceux que j'aurais oubliés de bien vouloir m'en excuser). Je vous remercie toutes et tous pour les discussions et les échanges qui ont accompagné ces trois années de recherche.

During this PhD, I also had the great opportunity to spend one month in Würzburg, Germany, within *Christof Weiß*'s group, where I had the pleasure of interacting with a fantastic team: *Benjamin*, *Lele*, *Leonie*, *Lucas*, *Natalia*, *Yaozhong*, *Yannik*, and *Yiwei*.

A huge thank you, *Christof*, for all the rigour and scientific discipline you shared with me during this period, as well as afterwards while preparing our submissions. I am deeply grateful for your guidance, your advice, and the trust you placed in me. Your high standards and dedication have been a great source of inspiration and have significantly contributed to my growth as a researcher.

Je souhaite également remercier l'ensemble du département informatique, au sein duquel je m'épanouis pleinement en tant qu'enseignant. Je suis particulièrement reconnaissant pour l'ambiance chaleureuse qui y règne, et pour les nombreuses personnes qui contribuent chaque jour à faire de ce département un lieu où il fait bon travailler. Je tiens tout particulièrement à remercier *Jean-Christophe*, qui m'a permis d'effectuer mes premières vacations au sein du département. Je tiens également à remercier *Jean-Stéphane*, qui était directeur au moment de mon recrutement et qui m'a accompagné dans la préparation de ma demande de thèse.

Je remercie également l'Université de Lille pour m'avoir accordé une décharge d'enseignement, un dispositif qui témoigne de l'importance accordée au développement de la recherche et qui m'a permis de mener à bien ce projet de thèse tout en conciliant mes activités de recherche et mes responsabilités d'enseignement.

Je souhaite également remercier mes parents, qui m'ont donné l'opportunité de faire de la musique, passion qui m'accompagne encore aujourd'hui. Pour terminer, je voudrais remercier Marie de m'avoir accompagné durant ces trois années, au cours desquelles j'ai certainement été parfois moins disponible. Merci pour ta patience, ton soutien et ta compréhension. Enfin, j'espère que Salomé et Gabriel seront fiers de leur papa.

Chapter 1

Introduction

Contents of this chapter

1.1 Motivations 13
1.2 The Flute: a Brief Introduction 14
1.3 Solo Flute Music 15
1.4 Outline of this Thesis 19
1.5 Publications and Communications 21
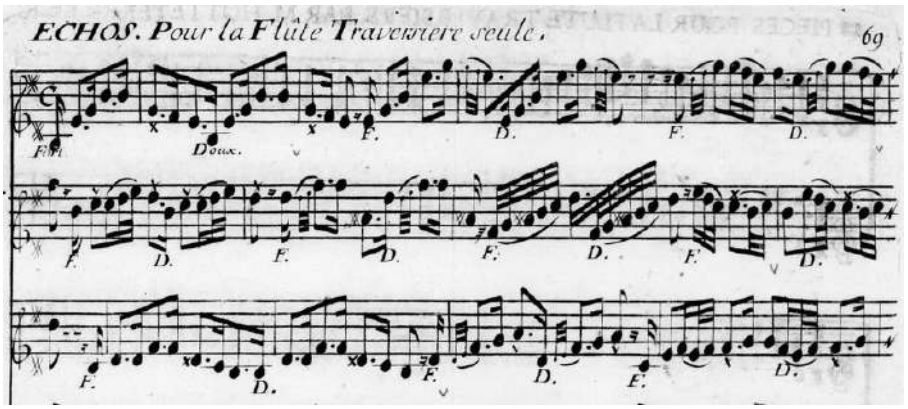


Figure 1.1: First measures of "ECHOS Pour la Flûte Traversiere seule" by Jacques Hotteterre (1708) [Hotteterre (1708)], the earliest known score for solo flute. You can listen a performance by Guillermo Peñalver at traverso (https://patthmus.github.io/phd-manuscript-links/chap1.html).

1.1 Motivations

No two performances of the same musical work are ever exactly alike. Music, as we experience it daily, can exist in different forms. It can be improvised, leading by nature to unique realisations at each performance. It can also be notated, for example through scores written in common Western notation. However, even when a musical work is fixed in a written form, every performance of this work will exhibit distinctive characteristics.

When interpreting the same score, performers continuously transform the music through subtle variations of timing, dynamics, articulation, timbre, phrasing, and breathing. They choose to explore these different facets offered by their instrument, drawing on their intention and intuition to produce a unique living musical interpretation. For listeners, these differences, even if they are immediately perceptible, can be difficult to describe precisely. They give each performance its own identity and invite us to reflect on what it means to interpret music.

Understanding this transformation from notation to sound has long been a central concern of musicology and performance studies, which have emphasised performance as a primary site of musical meaning [Cook (2013)]. Traditionally, such questions have been addressed through close listening and detailed analysis of a limited number of recordings. The increasing availability of large collections of digital scores and recordings now opens new perspectives [Orio (2006)]. It becomes possible to compare interpretations across performers, instruments, and historical periods. This evolution is part of a broader movement known as *Digital Humanities* [Schreibman et al. (2004)], in which computational approaches participate in the study of cultural and artistic artifacts. However, this shift not only expands the scale of analysis. Music, which offers a rich language and a great expressive dimension, brings into focus a set of more fundamental questions. What is a pulse? What defines a musical pattern? What is a musical voice: a notated part, a perceived melodic line?

Within this perspective, Music Information Retrieval (MIR) is a research field that emerged in the late 1990s and early 2000s [Downie (2003)]. It aims to design computational methods for analysing and understanding musical data. One of its central challenges is to establish a correspondence between different representations of a musical work (e.g., audio, score, or other symbolic representations). This problem is commonly referred to as music synchronisation or music alignment [Müller (2021)]. More specifically, it addresses the question of how symbolic and acoustic representations of music can be related. Such a correspondence provides a structural framework for studying musical interpretation, as it enables the comparison of performances at a fine temporal scale.

Most existing approaches and datasets have primarily focused on piano and orchestral music [Emerson and Harrison (2025)]. By contrast, woodwinds and especially the flute have received considerably less attention, even though it occupies an important place in Western music. The monophonic nature of the flute removes many of the harmonic cues exploited by alignment methods designed for polyphonic music. At the same time, it introduces its own

challenges, such as onset detection, central to the alignment process. These limitations motivate the work presented in this thesis. We propose new datasets for solo flute performance, adapt and improve existing alignment methods to better account for its specificities. We also propose a new methodology to study expressive timing in performance, thereby contributing both to MIR and to a broader understanding of musical interpretation.

1.2 The Flute: a Brief Introduction

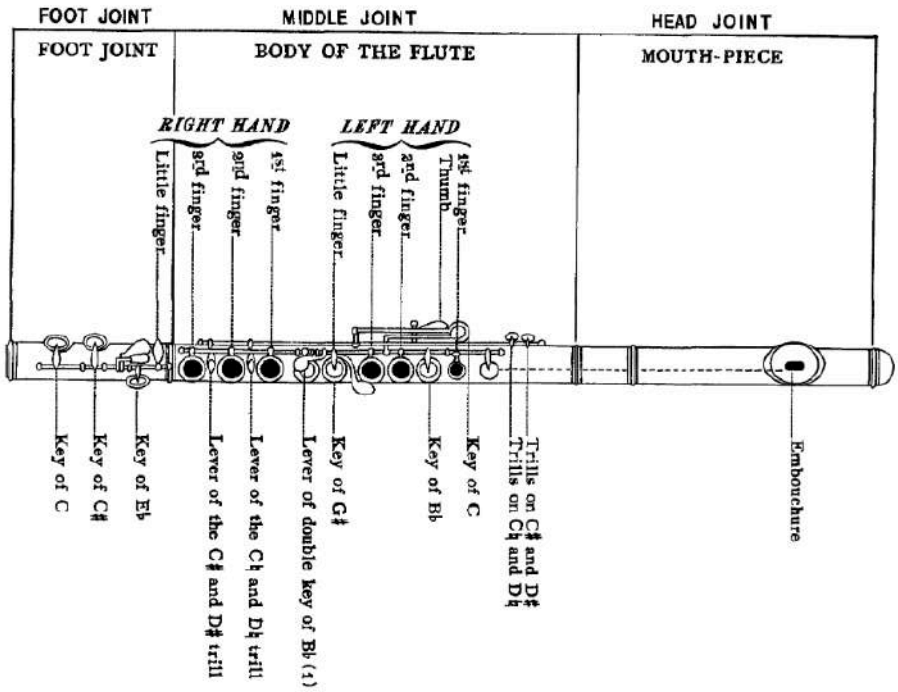


Figure 1.2: Drawing and description of the Böhm flute in an english translated edition of the french *"Famous Complete Flute Method"* ALTES (https://imslp.org/wiki/Special:ReverseLookup/280140).

The flute belongs to the family of *aerophones*, i.e., musical instruments whose sound production is primarily based on the vibration of a column of air. The excitation mechanism of the flute is based on an air jet directed against a sharp edge (the *embouchure*), which creates oscillations in the air column contained within the tube. The resulting sound is determined by the interaction between the excitation mechanism, the geometry of the instrument, and the acoustic resonances of the air column [Fletcher and Rossing (2008)].

From an acoustical perspective, the flute can be considered as a resonant tube in which the player's fingering modifies the effective length of the air column and therefore the fundamental frequency of the produced note. The resulting sound is not a pure sinusoidal tone but a complex waveform composed of a fundamental frequency and a series of harmonic partials [Chaigne and Kergomard (2008)].

Beyond its acoustical properties, the flute has undergone significant organological transformations throughout its history. Early flutes were commonly made of wood and produced through relatively simple mechanisms, while modern concert flutes are typically manufactured from metal alloys and rely on the Böhm key system, which optimised fingering, intonation, and tonal consistency (Fig. 1.2).

Today, the modern concert flute occupies an important position within the orchestra, where it is primarily associated with the upper register and is characterised by its agility, wide dynamic range, and timbral flexibility, making it a particularly versatile instrument for both traditional and contemporary musical practices [Powell (2002)].

1.3 Solo Flute Music

Earliest surviving works for flute date from the late 16th century [Montagu et al. (2014)]: *Ricercate, passagi et cadentie* (Giovanni Bassano, 1585) and *Il dolcimento, Book 2* (Aurelio Virgiliano, 1600). In fact, these pieces could be performed on a variety of treble instruments, including violin, lute, and flute.

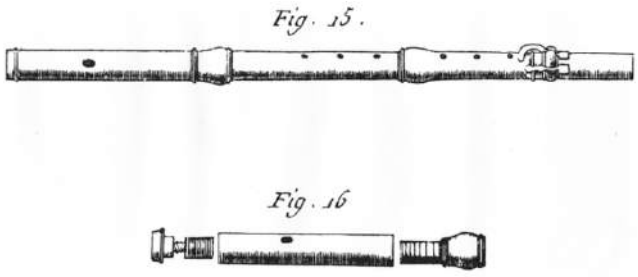


Figure 1.3: Quantz's double keywork from the Supplement to the Diderot's Encyclopédie (http://ark.bnf.fr/ark:/12148/cb30340683w.public).

The Baroque Flute The Baroque transverse flute (also known as the *traverso*, or *one-key flute*) was developed by the Hotteterre family, a prominent family of French instrument makers, in the 1660s–1670s. This instrument was typically pitched around $A \approx 415$ Hz, corresponding to the common Baroque pitch standard, although lower pitch levels (around $A \approx 392$ Hz) were also in use depending on regional practices. It was also characterised by a single key (Fig. 1.3)

Work	Composer	Year
Baroque		
ECHOS, Op.2	J.M. Hotteterre	1708
Partita in A minor, BWV 1013	J.S. Bach	1722–1723
12 Fantasias for Solo Flute, TWV 40:2–13	G.P. Telemann	1732–1733 (sometimes 1722–1723)
Sonata in A minor, Wq 132 H 562	C.P.E. Bach	1747
Early 20th Century		
Syrinx, L.129	C. Debussy	1913
Danse de la chèvre, H.39	A. Honegger	1921
Trois pièces pour flûte	P.O. Ferroud	1922
Oiseaux tendres	J. Rivier	1935
Density 21.5	E. Varèse	1936
Image, Op. 38	E.J. Bozza	1936
Mid 20th Century		
Stèle funéraire, Op. 224	C. Koechlin	1950
Sequenza I (for flute)	L. Berio	1958

Table 1.1: Significant Solo Flute Works: Baroque to the Mid-20th Century

and a fingering system that relied heavily on cross-fingering (fingering configurations in which an open tone hole is surrounded by closed ones). Chromatic notes were produced either using the key or through these cross-fingerings—lifting a finger between two stopped holes—which often resulted in slight intonation irregularities, particularly for pitches such as `Bb` and `G#`.

The instrument's practical range extends approximately from `D4` to `A6`, although notes above `E6` are generally considered difficult to produce and less stable in tone, as noted by Jacques-Martin Hotteterre in his 1722 treatise [Hotteterre (1722)]. The notation used at the time often employed the so-called *French violin clef*, equivalent to `F4` octave notation used for violin in 17th–18th century French music (Fig. 1.1).

Early repertoire for the instrument reflects both its expressive capabilities and its technical constraints. Among the first significant works, Michel de Labarre published *Pièces pour la flûte traversière* (1702), the earliest known collection for transverse flute with basso continuo. But the earliest surviving solo flute piece is *ECHOS for the Transverse Flute, Premier livre de pièces pour la flûte-traversière, Op.2* composed in 1708 by Jacques-Martin Hotteterre, who was also known as "le Romain" and served as a flautist at the court of Louis XIV (Fig. 1.1).

More broadly, the Baroque flute repertoire remains relatively limited in comparison to later periods. This can be partly explained by the instrument's technical limitations. Certain

keys—such as *Bb minor*, *C minor*, or *F minor*—were particularly challenging due to the use of cross-fingerings, leading to uneven tone quality and intonation issues across registers. These constraints are extensively discussed in the treatise of Quantz, *Essai d'une méthode pour apprendre à jouer de la flûte traversière* (1752) [Quantz, Jean Joachim (1752)], who also emphasises differences in sound production depending on register and fingering. Such limitations likely contributed to the relatively small number of works composed for solo flute during this period.

Despite this, several major works stand out. The masterpieces of Johann Sebastian Bach, Georg Philipp Telemann, and Carl Philipp Emanuel Bach are widely regarded as the most significant contributions to the unaccompanied flute repertoire before the twentieth century [Zohn (2008)]. Table 1.1 presents selected significant solo flute works from the early Baroque period to the mid-20th century.

From Classical to Romantic Flute Repertoire While unaccompanied flute repertoire remains scarce during the late eighteenth and nineteenth centuries, the Classical and Romantic eras nevertheless witnessed a substantial expansion of flute writing in concerto, chamber, and pedagogical contexts. Composers such as François Devienne [1] contributed to the concert repertoire. Pedagogical figures such as Paul Taffanel, founder of the French flute school and known for the development of pedagogical materials, including methods, études, and technical studies, played a central role in flute training during this period [Powell (2002)]. Although relatively few works for solo unaccompanied flute were composed compared to later twentieth-century developments, this period was crucial for the evolution of the flute itself, culminating in Theobald Böhm's nineteenth-century redesign.

Modern flute – the Böhm Flute The limitations of the Baroque flute—particularly in terms of fingering complexity, tonal homogeneity, and intonation across keys—motivated significant organological developments in the nineteenth century.

Building on the innovations introduced by Charles Nicholson in the early nineteenth century—most notably larger finger holes and a wider embouchure hole— Theobald Böhm developed a radically new flute design and fingering system, first patenting it in 1847 (Fig. 1.4). The Böhm system introduced a new key mechanism and acoustically optimised hole placement, greatly improving intonation, technical facility, and consistency of tone across registers. With only minor modifications, the Böhm system remains the standard mechanism used on modern concert flutes today.

Consequently, the solo repertoire expanded significantly. Major works include early twentieth-century compositions such *Syrinx* (1913) by Claude Debussy, *Danse de la chèvre* (1921) by Arthur Honegger, and *Density 21.5* (1936) by Edgard Varèse. A decisive turning

[1] Catalogue of Devienne works involving the flute (BnF): https://data.bnf.fr/fr/see_all_activities/13893235/page1?role=220&searchletters=fl%C3%BBte



Figure 1.4: Personal photograph of Dr. Heinz Prager's flute collection exposed at the Deutsches Museum (Munich, Germany). This collection was compiled by Heinz Prager together with Ludwig Böhm's (great great-grandsom of Theobal Böhm). We can see the evolution of the models, from a conical Flute (T. Böhm 1838–1839) at the top to a cylindrical flute, Böhm system No. 67 (T. Böhm 1852) at the bottom.

point occurs with *Sequenza I* (1958) by Luciano Berio, after which composers begin to explore extended techniques such as multiphonics, microtonality, percussive articulation, and the integration of electronics, considerably broadening the expressive and sonic possibilities of the instrument.

This thesis is based on two solo flute corpora drawn from contrasting musical periods. The first corpus is rooted in the Baroque tradition and consists of Telemann's *Twelve Fantasias for solo flute*. The second corpus represents the early twentieth-century repertoire and includes Debussy's *Syrinx* and Varèse's *Density 21.5*.

1.4 Outline of this Thesis

This thesis is organised into two parts (Fig. 1.5).

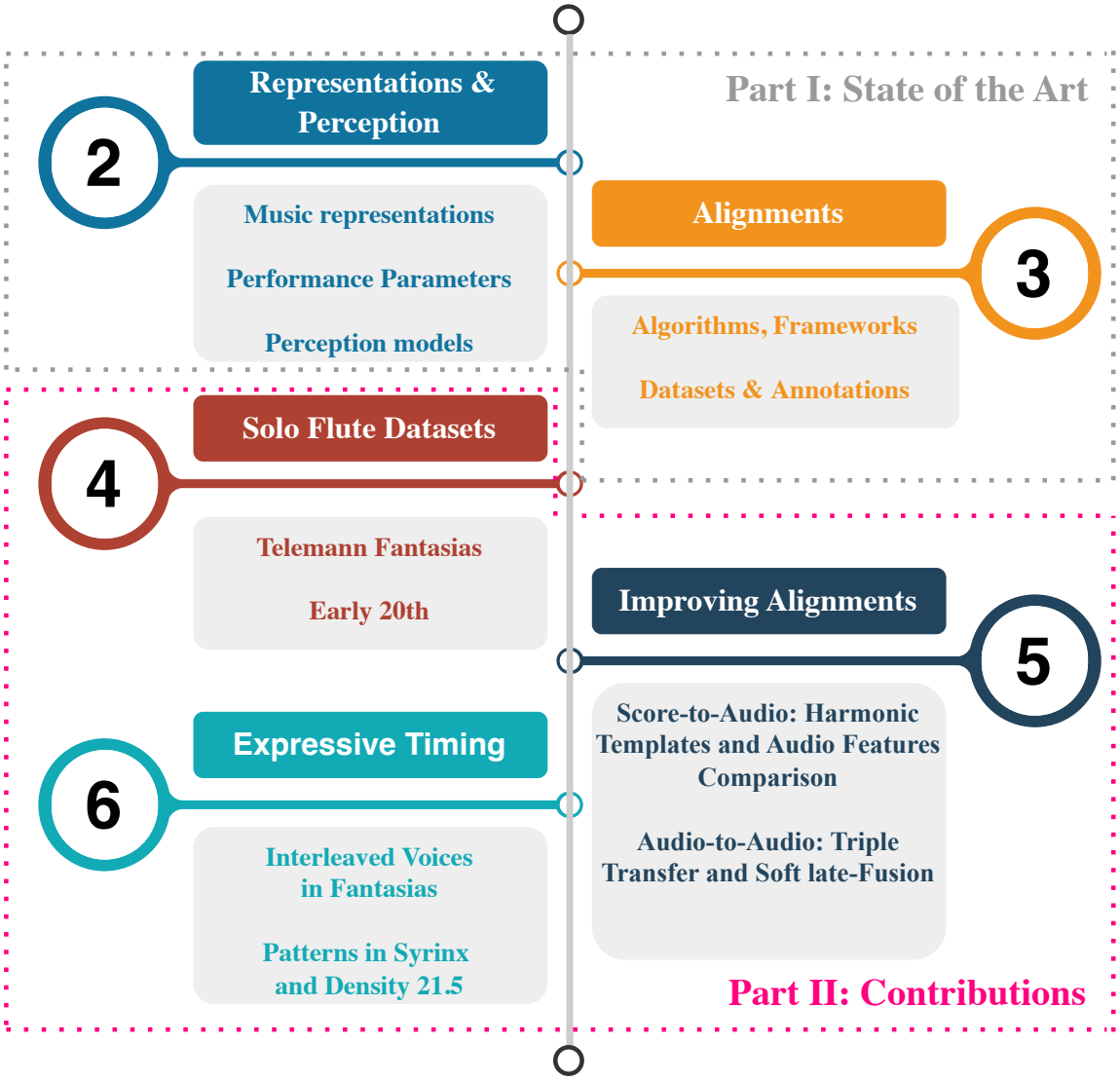


Figure 1.5: Thesis Outline.

Part I presents the state of the art that underpins this work.

Chapter 2 reviews symbolic representations formalising the score, and audio representations capturing the sound itself both abundantly used in MIR. The acoustic realisation of a performance can then be analysed through specific related parameters; we focus here specifically on timing. Finally, perceptual models contributing in understanding of the melodic structure are presented, including aspects such as implied polyphony in monophonic compositions.

Chapter 3 presents the main concepts and resources underlying musical alignment. It introduced alignment modalities and application scenarios, the general alignment framework, representative software tools, annotation transfer strategies, and the role of annotated datasets as benchmarks.

Part II presents the original technical contributions of this thesis

Chapter 4 introduces two new annotated corpora for solo flute repertoire: the first on the *Twelve fantasias for solo flute* by G.P. Telemann, the other dating from the early 20th century, including Debussy's *Syrinx* and Varese's *Density 21.5*. We discuss the selection of the musical material, the annotation strategy, and the methodological choices made to represent the repertoire at several levels (structure, harmony, voices, and onsets for evaluation purposes).

Chapter 5 describes three alignment strategies for solo flute and related woodwind performances. We first develop a score-to-audio framework based on harmonic templates. Then, we investigate the choice of audio representation in this strategy extending to woodwinds performances. Finally, we address audio-to-audio alignment for a highly expressive repertoire using a multi-version framework based on triple-transfer consistency. We propose a soft late-fusion strategy that improves robustness by combining multiple references, reducing extreme alignment errors in challenging passages.

Chapter 6 introduces a methodological framework for studying expressive timing, with a focus on local microtiming deviations. Applied to the two datasets used in this manuscript, this framework distinguishes global tempo variations from fine-grained temporal variations and provides measures based on the normalisation of Inter Onset Interval (IOI) and onset deviations to enable systematic comparison across performances and musical contexts. Results reveal consistent expressive strategies linked to metrical structure and rhythmic grouping, including timing asymmetries within beats, effects related to interleaved voices, and pattern-dependent timing behaviours.

Chapter 7 concludes outlining several directions for future research.

Appendix A presents a complementary study on the symbolic detection of implied voices in two different monophonic works.

1.5 Publications and Communications

The research undertaken in the course of this thesis has led to the following publications, workshops, and seminars.

In particular, two publications have been accepted at ISMIR. A third paper, initially submitted to ISMIR, has undergone further revisions and is being prepared for a new submission. Furthermore, this research has contributed to a joint publication with members of the research team in TISMIR.

Journal

- [Ballester et al. (2025)] Charles Ballester, Baptiste Bacot, Louis Bigo, Vanessa Nina Borsan, Louis Couturier, Ken Déguernel, Quentin Dinel, Laurent Feisthauer, Klaus Frieler, Mark Gotham, Richard Groult, Johannes Hentschel, Alexandre d'Hooge, Dinh-Viet-Toan Le, Florence Levé, Francesco Maccarini, Ivana Maričić, Gianluca Micchi, Meinard Müller, Alexandros Stamatiadis, Tom Taffin, Patrice Thibaud, Christof Weiß, Rui Yang, Emmanuel Leguy, and Mathieu Giraud.

  Interacting with Annotated and Synchronized Music Corpora on the Dezrann Web Platform.

  Transactions of the International Society for Music Information Retrieval (TISMIR), 2025, 8(1), pp.121-139. ⟨10.5334/tismir.212⟩. ⟨hal-04956003⟩

International conferences

- [Thibaud et al. (2025)] Patrice Thibaud, Mathieu Giraud, Yann Teytaut.

  When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute.

  In Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR), Sep 2025, Daejeon, South Korea. ⟨hal-05102415⟩

- Patrice Thibaud, Louis Bigo, Mathieu Giraud, Yann Teytaut.

  Soft Late-Fusion Alignment and Pattern-Based Performance Analysis of Early 20th-Century Solo Flute Music.

  Accepted at International Society for Music Information Retrieval Conference (ISMIR), 2026.⟨hal-05690528⟩

National Conferences, Seminars, Workshops

- [Thibaud et al. (2024)] Patrice Thibaud, Mathieu Giraud, Yann Teytaut.
  Towards Detecting Interleaved Voices in Telemann Flute Fantasias.
  DMRN+19: Digital Music Research Network One-day Workshop, 2024. ⟨hal-04923955⟩

- Seminar: *Timing analysis in solo flute performance from Telemann to Varèse*
  LaBRI, Bordeaux (France), May 2026.
  https://scrime.u-bordeaux.fr/scrime/evenements/presentation-par-patrice-thibault-timing-analysis-solo-flute-performance-telemann-varese

- [Thibaud et al. (2025)] Patrice Thibaud, Mathieu Giraud, Yann Teytaut.
  When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute.
  Presentation at Journées d'Informatique Musicale (JIM), June 2026, Lille.
  https://jim2026.sciencesconf.org/

In preparation

- Patrice Thibaud, Yannik Venohr, Christof Weiß.
  Score-to-Audio Alignment for Solo Woodwind Instrument Recordings.

Part I

Theoretical Framework and Tools

Chapter 2

Processing Music: From Representation to Perception

Contents of this chapter

2.1 Symbolic Representations . . . 25
2.1.1 Score Based Representations . . . 27
2.1.2 Event-Based and Derived Representations . . . 28
2.2 Audio Representations and Feature Extraction . . . 30
2.2.1 From Time-Domain Signals to Time–Frequency Representations . . . 31
2.2.2 Pitch Features . . . 33
2.2.3 Chroma Representations . . . 34
2.2.4 Cepstral Representations . . . 38
2.2.5 Learned Features . . . 40
2.2.6 Onset Activation Functions . . . 40
2.3 Performance . . . 44
2.3.1 Parameters of Performance and Representations . . . 45
2.3.2 Focus on Timing . . . 46
2.4 Perceptual and Musicological Models . . . 48
2.4.1 Perception of Realisation . . . 48
2.4.2 Implied Voices . . . 50

In Western classical music, the score remains the primary medium for orchestral instruments. The staff provides a framework for musical signs that allow the performer to identify, among other things, the pitches and durations of the sounds to be played.
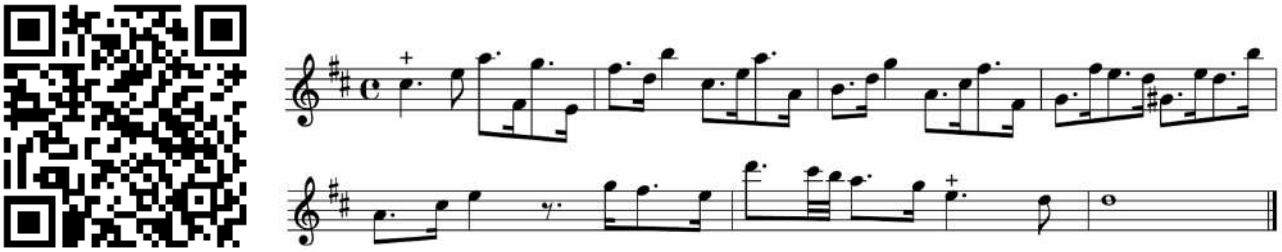


Figure 2.1: Score and interpretation examples by Rampal and Leleux (https://patthmus.github.io/phd-manuscript-links/chap2.html) of Fantasia No. 7 by G.P. Telemann (TWV 40:8) in *D Major*, mm. 89–95.

Is the score alone sufficient for analysing or understanding music?

Consider two short excerpts from the end of the first movement of Telemann's Fantasia No. 7 (Fig. 2.1), first performed by J.P. Rampal (flute) and then by F. Leleux (oboe). The timbre of each instrument is distinct, and the overall performance speed differs, even though the score does not indicate tempo variations. Some notes are more or less emphasised, while others—ornaments—are added by the performer.

To analyse such performance differences, we turn to the field of MIR. This chapter first reviews the main symbolic and audio representations used in MIR. Then, performance-related parameters, with a particular focus on timing, are discussed. Finally, within the context of monophonic melodies, perceptual models related to implied polyphony are examined

2.1 Symbolic Representations

The history of musical representation is marked by a wide diversity of notational systems, shaped by cultural traditions, instrumental practices, and historical contexts. Early forms of notation include neumatic systems used in Gregorian chant, as well as tablatures and grid-based notations developed for instruments such as the lute and early keyboard instruments. These heterogeneous practices reflect the multiplicity of ways in which musical ideas have been externalised, transmitted, and preserved.

Western staff notation gradually emerged between the Middle Ages and the seventeenth century, building upon earlier neumatic traditions and the foundational contributions of figures such as Guido of Arezzo, who introduced around 1028 systematic principles of pitch organisation and solmization [Hugues and Gerson-Kiwi (2001)]. Alternative symbolic systems were also developed, such as the numerical notation initiated by Jean-Jacques Rousseau and subsequently expanded by Nanine Paris and others, leading to the Galin–Paris–Chevé method [Bullen (1877);
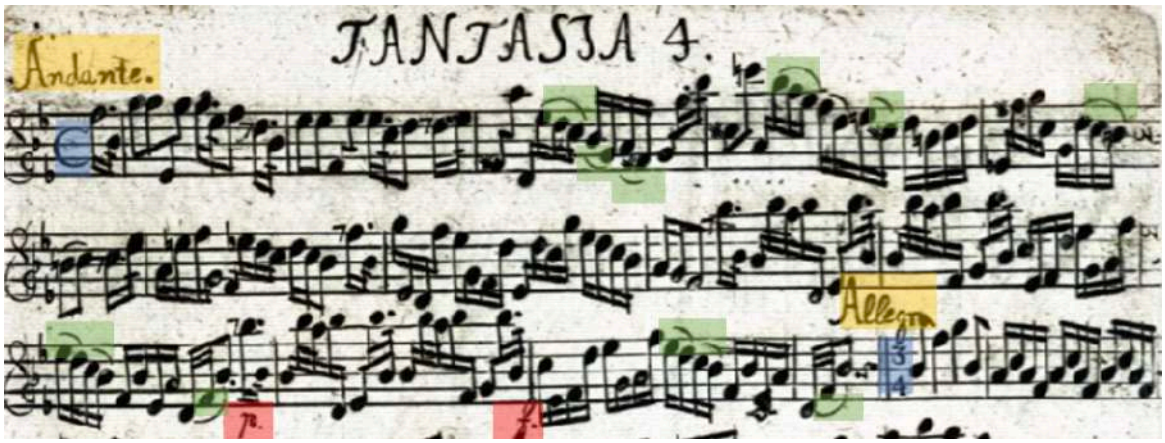


Figure 2.2: Fantasia for solo flute No. 4 (TWV 40:5) by G. P. Telemann, mm. 1–16, autograph score.

Galin (1835)], which remains widely used in East Asian musical pedagogy and practice [Yang et al. (2025)]. These historical and cultural variations illustrate that musical notation does not constitute a universal and immutable language; rather, it is a flexible and evolving medium that mediates between the composer's conceptual intentions and the performer's embodied realisation.

The advent of music printing marked a decisive turning point in the dissemination and standardisation of musical notation. As early as 1501, Ottaviano Petrucci published the *Harmonice Musices Odhecaton*, using movable type to produce staff scores, thereby facilitating large-scale circulation [Boorman et al. (2001)].

Musical scores (Fig. 2.2) convey a wide range of information essential for performance. They primarily encode discrete pitches and their associated durations, which correspond to the rhythmic values of notes and rests. Scores further include bar segmentation governed by a metre, which may evolve throughout a piece, as well as a tempo indication, often specified only approximately. Additional layers of musical meaning are introduced through articulations, dynamics, and key signatures, all of which guide the performer's expressive interpretation while interacting with the symbolic notation.

Despite its richness, traditional notation remains an abstraction of sound, omitting many continuous and context-dependent aspects of musical performance, such as micro-timing, timbral nuance, and expressive variability. This gap becomes particularly salient in computational contexts, where explicit and machine-readable encodings are required.

In the following, we introduce two categories of symbolic representations: score-based and event-based. All the illustrations are derived from the short two-measure excerpt shown in Fig. 2.3.

2.1.1 Score Based Representations
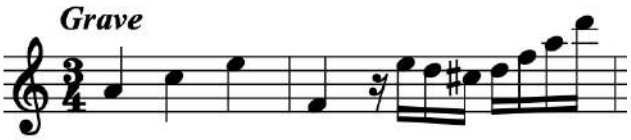


Figure 2.3: Fantasia No. 2 (TWV 40:3), mm 1–2.

```
<note default-x="80.21" default-y="-25">
  <pitch>
    <step>A</step>
    <octave>4</octave>
    </pitch>
  <duration>4</duration>
  <voice>1</voice>
  <type>quarter</type>
  <stem>up</stem>
  </note>
```

(a) MusicXML encoding of the the first note.

```
<measure xml:id="mx2wvat" n="1">
  <staff xml:id="m1s1" n="1">
    <layer xml:id="m1s1l1" n="1">
      <note xml:id="nxgjzkt" dur="4" pname="a" oct="4" />
      <note xml:id="n84egs8" dur="4" pname="c" oct="5" />
      <note xml:id="np1ykjd" dur="4" pname="e" oct="5" />
    </layer>
  </staff>
  <tempo xml:id="t1qkow5c" type="mscore-infer-from-text"
    midi.bpm="45" staff="1" tstamp="1">Grave</tempo>
</measure>
```

(b) MEI encoding of the first measure.

Figure 2.4: Tree-based score representations of Figure 2.3 exerpts, in MusicXML and MEI.

Score-based formats such as MEI and MusicXML encode music as structured symbolic documents. They preserve not only pitch and rhythm, but also hierarchical and semantic relationships such as voices, measures, articulations, and layout information.

In particular, these formats can be represented as hierarchical structures:

$$\mathcal{T} = (\mathcal{N}, \mathcal{E}), \tag{2.1}$$

where nodes $\mathcal{N}$ represent musical entities (notes, measures, voices, articulations, dynamics), and edges $\mathcal{E}$ capture their structural relationships. This representation explicitly models the logical and semantic organisation of the score.

MusicXML

The open MusicXML format, introduced in the early 2000s [Good (2001)], has become widely adopted in notation software and research applications. It represents musical content as a structured Extensible Markup Language (XML) document, where each musical symbol is explicitly tagged. This includes pitch information (step, octave, accidentals), rhythmic values, time signatures, articulations, dynamics, ornaments, lyrics, and layout instructions. These elements make MusicXML particularly suitable for tasks such as symbolic analysis and computational musicology.

MEI

The Music Encoding Initiative (MEI) is a community-driven, open-source XML standard for the semantic encoding of musical notation [Hankinson et al. (2011)]. MEI encodes the structural and semantic content of musical documents (pitches, rhythms, articulations, editorial annotations), while delegating rendering decisions to downstream software. Tools such as *Verovio* [Pugin et al. (2014)] illustrate this separation by converting MEI files into high-quality graphical notation.

2.1.2 Event-Based and Derived Representations

This section introduces computational representations of symbolic music that operate at different levels of abstraction. Starting from an event-based formulation, we derive two widely used computational encodings: the MIDI event stream and the piano-roll representation. Together, these formats constitute a bridge between symbolic score descriptions and time–frequency representations used in downstream MIR tasks.

Event-based Representation

In the case of monophonic musical lines extracted from single-staff scores, such material can be represented as an ordered sequence of discrete musical events:

$$\mathcal{S} = (e_i)_{i=1}^{N}, \tag{2.2}$$

where each event is defined as:

$$e_i = (p_i, d_i, t_i^{\text{on}}, t_i^{\text{off}}, v_i), \tag{2.3}$$

with pitch p_i, symbolic duration d_i expressed in quarter notes, onset and offset times t_i^{on} and t_i^{off}, and optional expressive parameters v_i, such as dynamics or articulation.

This event-based representation provides a unified temporal abstraction that can be derived from symbolic score formats such as MEI or MusicXML. Importantly, onset and offset times may deviate from strict symbolic durations due to expressive timing in performance data.

From the event-based representation $\mathcal{S}$, two widely used computational formats can be derived: the MIDI event stream and the piano-roll representation.

MIDI Representation

The Musical Instrument Digital Interface (MIDI) standard, introduced in the early 1980s, remains one of the most widely used symbolic formats in electronic music production and digital instruments [Rothstein (1995)].

A MIDI file represents music as a sequence of time-stamped events describing discrete musical actions. These include *note-on* and *note-off* messages, which specify the onset and offset of notes identified by integer MIDI pitch numbers, as well as velocity values encoding attack intensity. In the MIDI specification, pitch is encoded on a linear integer scale from 0 to 127, where each value corresponds to a semitone in the equal-tempered scale (e.g., 69 corresponds to `A4` = 440 Hz). Temporal information is encoded using discrete ticks or PPQ (pulses per quarter note), which determine the achievable timing resolution.

While MIDI offers a compact and efficient representation, it exhibits significant limitations for detailed musical analysis, including limited temporal resolution, a restricted representation of expressive parameters (e.g., dynamics and articulation), and the absence of explicit notational information such as beams, slurs, or score layout.

Piano-roll Representation

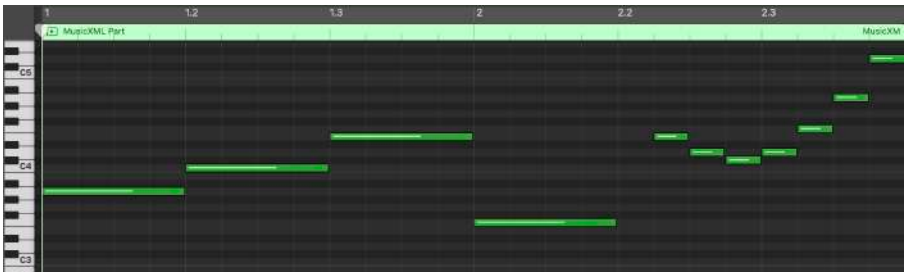


Figure 2.5: Piano-roll representation corresponding to the Figure 2.3, displayed in GarageBand (Digital Audio Workstation (DAW)).

One of the most intuitive computational representations is the *piano-roll* (Fig. 2.5), historically associated with mechanical reproduction systems dating back to the late nineteenth century, and now widely employed in digital audio workstations as a visual interface for editing symbolic note data.

The piano-roll is a matrix-based symbolic representation in which pitch is mapped along the vertical axis and time along the horizontal axis, making rhythmic structures explicit. Time can be defined either in absolute units or using a tempo-normalised symbolic grid. Matrix entries indicate note activity weighted by velocity. Formally, the piano-roll can be described as a discrete-time pitch function

$$\mathcal{P}(t, p) \in \mathbb{R}^{+}, \tag{2.4}$$

indicating the activation level of pitch p at time t. After discretising time into uniform frames, this representation yields a matrix

$$\mathbf{P} \in \mathbb{R}^{P \times T}, \tag{2.5}$$

where P denotes the number of discrete pitch values and T the number of temporal frames.

Typical temporal resolutions vary depending on the application. In signal-oriented MIR tasks, frame sizes between 10–50 ms are commonly used, depending on the analysis window and hop size. In contrast, symbolic and score-informed representations often rely on tempo-normalised grids ranging from values such as 24 PPQ in early MIDI implementations to 960 PPQ in modern digital audio sequencers, depending on the required temporal precision.

Overall, symbolic representations provide structured abstractions of musical sound, capturing pitches, rhythms, articulations, and dynamics, while inevitably leaving room for interpretative freedom. The resulting discrepancy between discrete symbolic encoding and continuous musical performance is a central challenge in audio–symbolic alignment and computational music analysis.

Section summary – Symbolic Representations

Symbolic representations explicitly encode the discrete structural aspects of music—pitch, duration, metre, articulation—but remain incomplete formalisations of performed sound. Audio representations, derived from digital sampling and time–frequency analysis, capture the acoustic and perceptual richness of the signal, yet require transformation to become computationally meaningful.

2.2 Audio Representations and Feature Extraction

Audio representations provide a complementary perspective by directly operating on the acoustic signal, allowing fine-grained temporal and timbral information to be retained.

An audio signal can be represented as a waveform: a curve showing how the electrical amplitude—corresponding to variations in air pressure—changes over time. It contains a rich set of information from which descriptors can be extracted to establish a correspondence with the musical information encoded in the score, including the performed pitches, their durations, and, consequently, the precise perceptual beginnings of musical events, which we call *onsets*.

Digital Audio Sampling

In digital audio processing, the continuous acoustic pressure signal is converted into a discrete-time representation through uniform sampling in time, a process known as *digital audio sampling*. The sampling rate, expressed in Hertz (Hz), indicates the number of samples recorded per second.

Let $s(t)$ denote a continuous-time acoustic signal, $s : \mathbb{R} \to \mathbb{R}$. Uniform sampling at rate f_s

maps $s(t)$ to a discrete-time signal $s[n] : \mathbb{Z} \to \mathbb{R}$ defined as

$$s[n] = s\left(\frac{n}{f_s}\right), \quad n = 0, \ldots, N-1. \tag{2.6}$$

where N denotes the total number of samples, determined by the duration of the signal and the sampling rate.

According to the Nyquist–Shannon sampling theorem [Shannon (1949)], all frequency components below $f_s/2$ can be perfectly reconstructed from $s[n]$. In practice, standard sampling rates such as 44.1 kHz or 48 kHz provide sufficient bandwidth to capture the full range of human hearing. This discretised waveform constitutes the fundamental representation from which all subsequent audio descriptors are derived.

2.2.1 From Time-Domain Signals to Time–Frequency Representations

While the waveform provides a complete representation of the audio signal in the time domain, many musical properties—such as pitch content, harmonic structure, and note onsets—are more easily analysed in a joint time–frequency representation (Fig. 2.6).

Spectrograms

A first well-known representation is the *magnitude spectrum*, obtained by applying the discrete Fourier transform to a discrete-time signal, which describes the distribution of spectral amplitudes across frequencies.

To account for temporal variations, the Short-Time Fourier Transform (STFT) $S(m, k)$ computes such spectra over successive overlapping frames, yielding a time–frequency representation known as a *spectrogram*. The magnitudes are typically expressed on a logarithmic scale (e.g., in decibels, *dB*) to better reflect the characteristics of human loudness perception.

The STFT, originally introduced by Dennis Gabor [Gabor (1946)], analyses signals locally in both time and frequency by applying a Fourier transform to short-time segments of the signal, called *frames*, obtained by sliding a window function along the time axis. The hop size determines the temporal step between successive frames:

$$S(m, k) = \sum_{n=0}^{N-1} s[n + mH]\ \overline{w}[n]\ e^{-j2\pi kn/N} \tag{2.7}$$

where $w[n]$ is a discrete window function of length N, H is the hop size, m indexes the time frames, and k the frequency bins. Each frame consists of N consecutive samples of the discrete-time signal. The Fourier transform applied to each frame yields a set of complex-valued

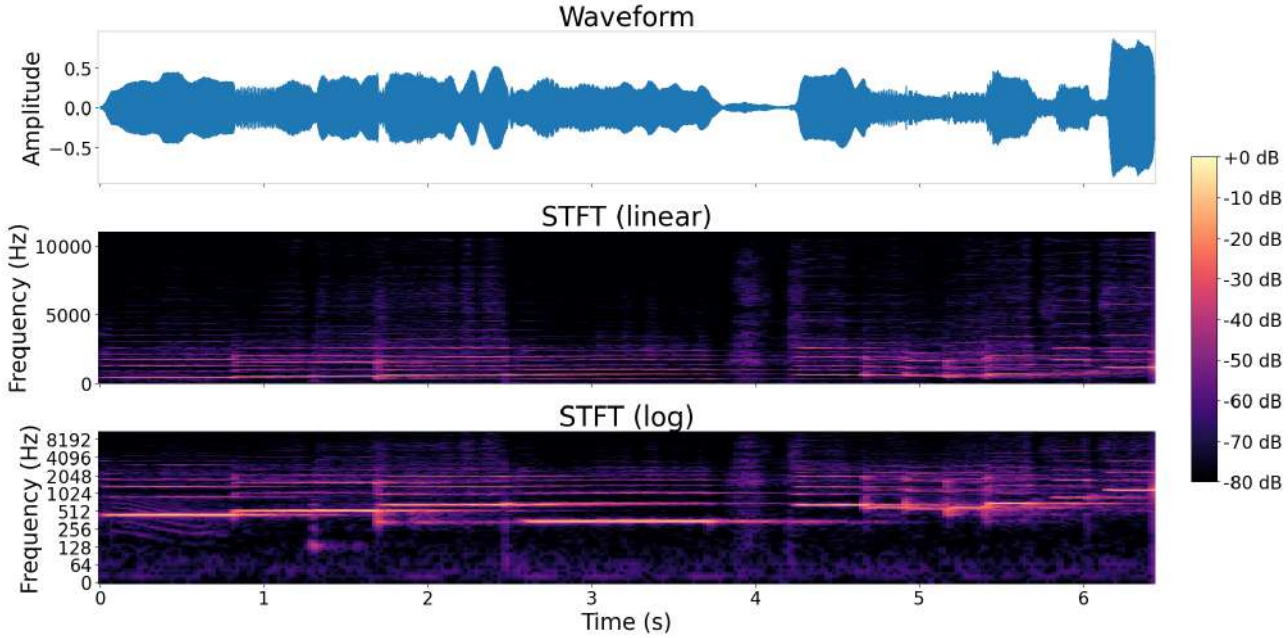


Figure 2.6: Various audio representations of the first 6 seconds of Fantasia No. 2 (corresponding score Fig. 2.3). Recorded by the author on a Yamaha YFL-587 modern Flute (Würzburg, Germany, April 2025). From top to bottom: waveform, STFT, STFT (logarithmic frequency axis). The following audio representations (Fig. 2.7, 2.8, 2.9, 2.10, 2.11) are based on the same audio signal and the same magnitude scale and correspond to a performance derived from the score excerpt presented in Fig. 2.3.

coefficients indexed by k, referred to as frequency bins, each representing the signal's content within a narrow frequency band.

The hop size H defines the number of samples between the starting points of successive frames. It controls the temporal resolution and the amount of overlap between frames: a smaller H increases the overlap and provides finer time resolution at the cost of a higher computational load, while a larger H reduces overlap and computational cost but decreases temporal precision.

In its standard form, the frequency axis of the STFT spectrogram is linearly spaced. However, it is often beneficial to represent the frequency axis on a logarithmic scale, which better reflects perceived pitch relationships and harmonic structures in Western music.

2.2.2 Pitch Features

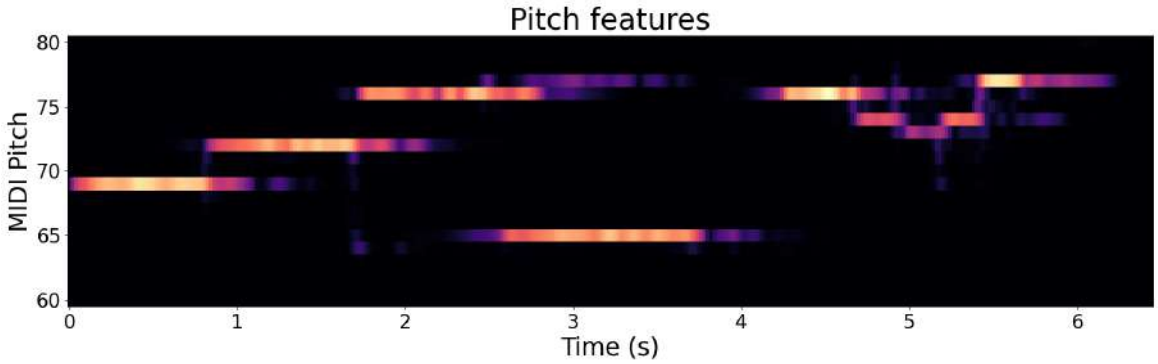


Figure 2.7: Pitch features computed via an IIR filterbank aggregated as short-time mean-square power.

To move from low-level spectral representations toward musically meaningful descriptors, many approaches rely on pitch-based features that explicitly encode the distribution of spectral energy across pitch-related frequency bands (Fig. 2.7).

Pitch-Related Features

Pitch-related features aim at extracting musically meaningful information related to note identity and temporal events from the audio signal. These representations are particularly useful for establishing a correspondence between acoustic observations and symbolic music information, such as performed pitches and note events. Pitch-based onset detection aims at identifying the temporal positions of note beginnings and other musical events in an audio signal. A common strategy consists of decomposing the signal into multiple subbands, each capturing complementary spectral information, and computing a novelty function independently in each band before combining them to estimate event locations [Müller (2021); Müller and Kurth (2006)]. In practice, these subbands can be defined in a musically meaningful way, for instance by aligning them with MIDI pitch bins[1].

Beyond onset detection, more detailed pitch-related information can be obtained through fundamental frequency (F_0) estimation. Methods such as pYIN [Mauch and Dixon (2014)] formulate pitch tracking as a probabilistic inference problem. For each analysis frame, multiple pitch candidates are extracted and assigned probabilities based on a thresholded likelihood model. These frame-wise observations are then integrated over time using an Hidden Markov Model (HMM), where Viterbi decoding is used to infer the most likely continuous pitch

[1] SyncToolbox documentation.

trajectory. This temporal smoothing reduces local estimation errors such as octave jumps and yields a more stable pitch track, particularly for monophonic signals such as singing voice.

Template–based representations

Template–based representations encode the harmonic structure of musical pitches by comparing the magnitude spectrum to a set of predefined harmonic templates. To ensure robustness against varying timbres, the input spectrum is often subjected to spectral whitening to flatten the spectral envelope before template matching [Klapuri (2006); Mauch and Dixon (2010)].

Each template $T_p(k)$ models the expected amplitudes of the partials for a given pitch p. These templates are constructed either as Gaussian peaks centred on the fundamental frequency and its harmonics (typically the first ten), or as idealised note profiles where overtone amplitudes follow a geometric decline. In some models, the importance of each harmonic is further refined by learned weight functions that prioritise specific partials to maximise pitch salience.

At each time frame m, a similarity score $y_p(m)$ is computed.

While this can be expressed as a simple weighted sum of harmonic amplitudes to determine pitch strength, real-time alignment systems frequently rely on the Kullback-Leibler (KL) divergence. This probabilistic measure calculates the distance between the observed frequency distribution and the reference template distribution, allowing for a more precise likelihood estimation in dynamic environments.

In more complex polyphonic scenarios, the spectrum is modelled as a linear combination of these templates, where the presence of notes is estimated using Non-Negative Least Squares (NNLS) [Mauch and Dixon (2010)].

These scores are then transformed into a probabilistic pitch distribution using a *softmax* function or an exponential mapping [Cont (2010)] :

$$P(p \mid m) = \frac{\exp(\beta, y_p(m))}{\sum_{p'} \exp(\beta, y_{p'}(m))}, \tag{2.8}$$

where β controls the sharpness of the distribution. This normalisation enhances pitch selectivity, enforces competition between candidates, and yields a sparse, noise-robust time–pitch representation. While effective for monophonic signals, these methods are particularly powerful for polyphonic tasks and provide a robust foundation for automatic transcription, chord recognition, and real-time score-to-audio alignment.

2.2.3 Chroma Representations

In Western music, pitch is commonly organised according to the *twelve-tone equal temperament*, in which the octave is divided into 12 equal semitones. A *pitch class* groups all notes that are

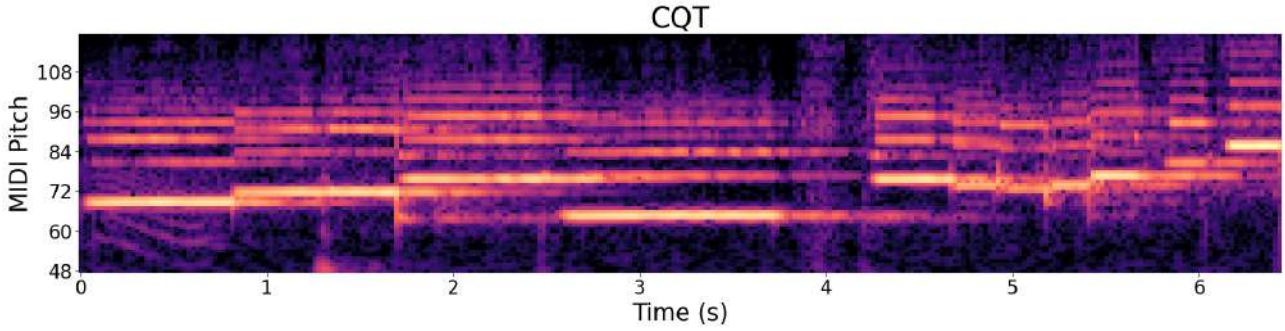


Figure 2.8: Constant-Q transform (CQT) of the audio signal, showing the time–pitch representation with the vertical axis expressed in MIDI pitch values.

equivalent modulo octaves; for example, all `C` notes (`C2, C3, C4`, ...) belong to the same pitch class `"C"`.

The centre frequency $f_{\text{pitch}}(p)$ of a MIDI pitch $p \in [0 : 127]$ can be computed from the standard reference $A_4 = 440$ Hz as

$$f_{\text{pitch}}(p) = 440 \cdot 2^{(p-69)/12} \tag{2.9}$$

which allows one to convert pitch indices to frequencies.

CQT

A more musically motivated time–frequency representation is the Constant-Q Transform (CQT) [Brown (1991)], which employs a bank of band-pass filters whose centre frequencies are geometrically spaced on a logarithmic scale. The nominal centre frequencies are given by

$$f_k = f_{\min} \cdot 2^{k/B}, \tag{2.10}$$

where B denotes the number of bins per octave and k is the frequency-bin index.

Each CQT bin is associated with a band-pass filter of bandwidth Δf_k, which determines the frequency resolution of the analysis. The corresponding bandwidths satisfy the constant-Q property

$$Q = \frac{f_k}{\Delta f_k}, \tag{2.11}$$

ensuring that the relative bandwidth remains constant across all frequency bins.

The CQT naturally provides a logarithmic frequency axis that aligns with musical pitch perception (Fig. 2.8). It can be used to construct a *pitch-based log-frequency spectrogram* $\mathcal{Y}_{\text{LF}}(n, p)$ by assigning each frequency bin k to the pitch $p \in [0 : 127]$ whose nominal frequency

is closest:

$$\mathcal{Y}_{\mathrm{LF}}(n,p) = \sum_{k \in \mathcal{P}(p)} |S(n,k)|^2, \tag{2.12}$$

where

$$\mathcal{P}(p) = \{k : f_{\mathrm{pitch}}(p-0.5) \leq f_k < f_{\mathrm{pitch}}(p+0.5)\} \tag{2.13}$$

denotes the set of frequency bins associated with pitch p. The bandwidth of pitch p is then defined as

$$\mathrm{BW}(p) = f_{\mathrm{pitch}}(p+0.5) - f_{\mathrm{pitch}}(p-0.5), \tag{2.14}$$

so that each pitch covers a frequency interval proportional to its position on the logarithmic scale.

A key property of the CQT is that the relative positions of the harmonics of a given note are independent of the fundamental frequency when plotted on a logarithmic axis. Specifically, the distance between successive harmonics remains constant in log-frequency, forming a characteristic harmonic pattern for each note.

Chroma

To summarise harmonic content in a compact, musically meaningful form, one can compute *chroma vectors*, which are 12-element representations with one element per pitch class (`C`, `C#`, `D`, `...`, `B`). Given a pitch-based log-frequency spectrogram $\mathcal{Y}_{\mathrm{LF}}(n,p)$, chroma features are obtained by aggregating all pitch coefficients that belong to the same pitch class:

$$C(n,c) = \sum_{\substack{p \in [0:127] \\ p \bmod 12 = c}} \mathcal{Y}_{\mathrm{LF}}(n,p), \tag{2.15}$$

for $c \in [0:11]$. This results in a compact representation that is invariant to octave transpositions, while retaining the harmonic information relevant to Western tonal music (Fig. 2.9).

A chromagram computed from a musical excerpt shows how energy is distributed across pitch classes over time. Each element measures the energy of the corresponding pitch class in the audio signal, indicating how strongly that note is present, regardless of the octave. By assuming an equal-tempered scale, any frequency in the spectrum can be mapped to one of the twelve pitch classes. Chroma-based representations are widely used in piano or orchestral music [Bartsch and Wakefield (2005)], highlighting the harmonic and tonal structure of a musical excerpt.

While rich in information, these chromagrams exhibit several inherent limitations: they often span a very large dynamic range, are sensitive to short-term fluctuations, and can vary with overall loudness or recording conditions. To overcome these challenges, a range of post-processing steps can be applied.

- Logarithmic amplitude compression [Müller (2021)] reduces amplitude disparities, enhances low-energy components, and produces representations that better reflect human auditory perception.

$$\tilde{C}(n,c) = \log\big(1 + \gamma\, C(n,c)\big), \tag{2.16}$$

  where γ controls the compression strength;

- Temporal smoothing reduces short-term fluctuations and emphasises stable harmonic content [Bartsch and Wakefield (2005); Ewert (2011); Müller (2021)]. By averaging or low-pass filtering the feature vectors over a window of several frames, transient variations caused by rapid amplitude changes, expressive articulation, or noise can be attenuated;

- Normalisation is often applied to reduce sensitivity to global loudness variations. A general approach is to normalise each chroma vector $C(n) \in \mathbb{R}^{12}$ using a vector norm:

$$\hat{C}(n) = \frac{C(n)}{\|C(n)\|_k}, \tag{2.17}$$

  where $k \geq 1$ specifies the type of norm, and $\|v\|_k$ for a vector $v = [v_1, \ldots, v_{12}]$ is defined as

$$\|v\|_k = \Big(\sum_{i=1}^{12} |v_i|^k\Big)^{1/k}. \tag{2.18}$$

  For $k = 1$, this corresponds to *L1 normalisation*, which scales the vector such that the sum of the absolute values of its components equals one. For $k = 2$, this corresponds to the standard *L2 (Euclidean) normalisation*, which scales the vector to unit Euclidean norm.

Together, logarithmic compression and normalisation provide a perceptually motivated foundation for robust extraction of various chroma-based features described in the literature.

CENS

The Chroma Energy Normalized Statistics (CENS) representation [Ewert (2011); Kurth and Muller (2008); Müller et al. (2005)] is a robust extension of classical chromagrams, developed to minimise the effects of dynamics, timbre, and transient events. These features provide a weighted statistical summary of energy across consecutive chroma vectors. By aggregating statistics over relatively large temporal windows, CENS reduces the impact of small timing variations, such as those caused by articulation, and ensures that even performances with substantial differences in playing style or instrumentation produce comparable sequences. This stability makes CENS particularly well suited for applications such as score-to-audio alignment, cover song detection, musical structure analysis, and tuning estimation.

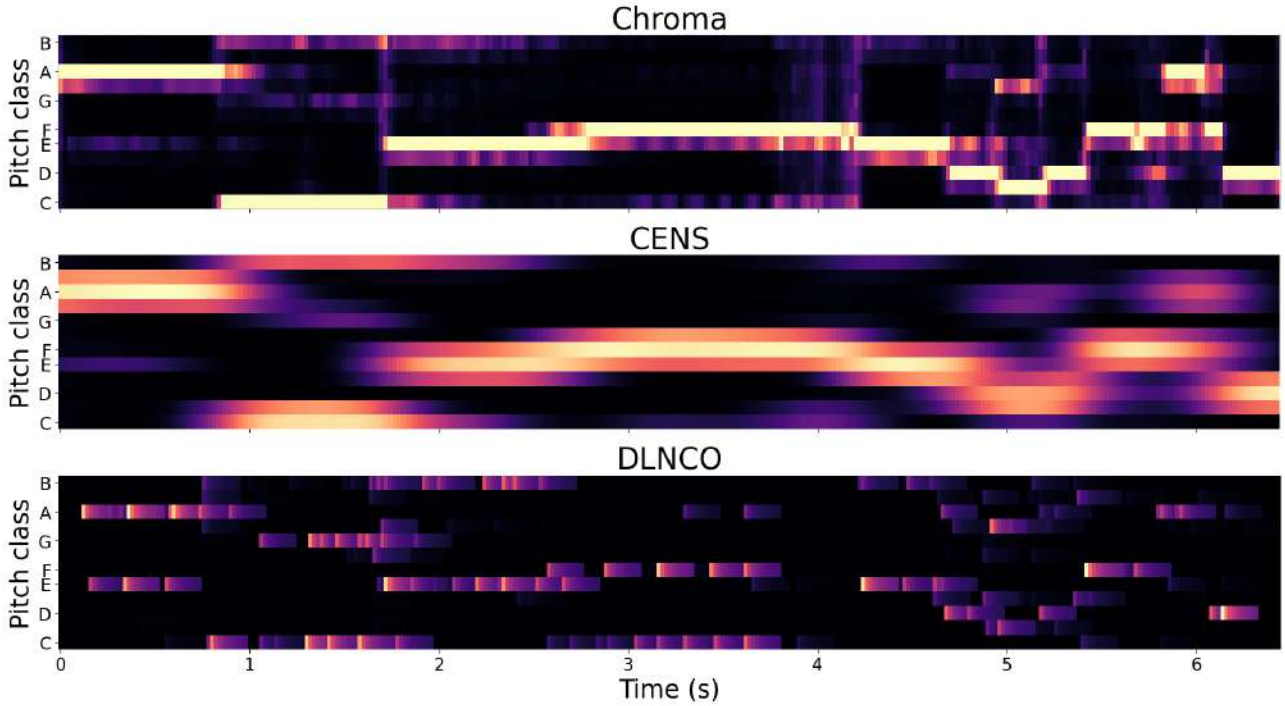


Figure 2.9: Various chroma-based representations. From top to bottom: chroma, CENS, and DLNCO.

DNLCO

Low-energy onsets are less perceptually relevant in high-energy passages. Decaying Locally adaptive Normalized Chroma-based Onset (DLNCO) features address this with locally adaptive normalisation, remaining robust to dynamics while preserving precise onset information. Traditionally, chroma-based features are used for their timbral robustness, often combined with DLNCO to improve temporal precision [Ewert et al. (2009); Grosche et al. (2009); Müller (2021)]. Their octave-folded, logarithmically compressed, normalised representation emphasises transients while reducing the influence of overall dynamics, making them well-suited for polyphonic instruments like the piano.

2.2.4 Cepstral Representations

Cepstral representations offer a compact, perceptually motivated encoding of the spectral envelope, effectively separating harmonic structure from overall timbre. Among the most widely used are the Mel-Frequency Cepstral Coefficients (MFCC) and the Perceptual Frequency Cepstral Coefficients (PFCC).

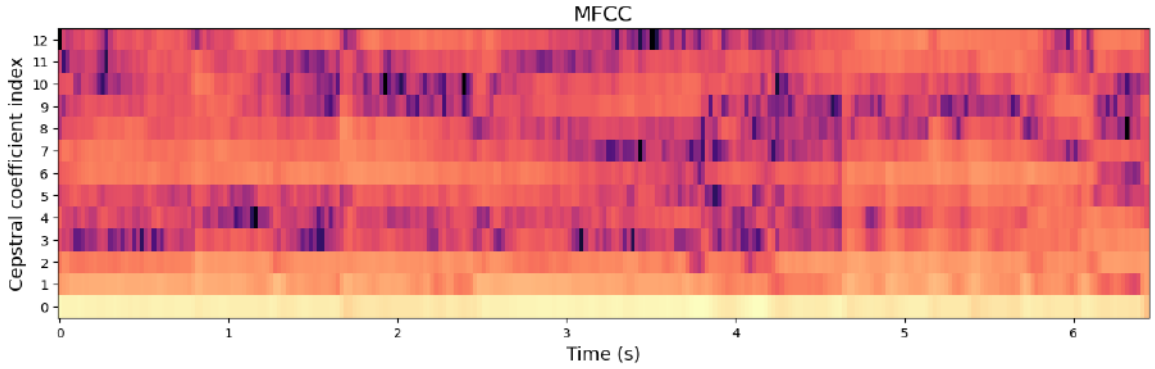


Figure 2.10: Mel-frequency cepstral coefficient (MFCC) representation of the audio signal, with the vertical axis corresponding to cepstral coefficient indices.

MFCC

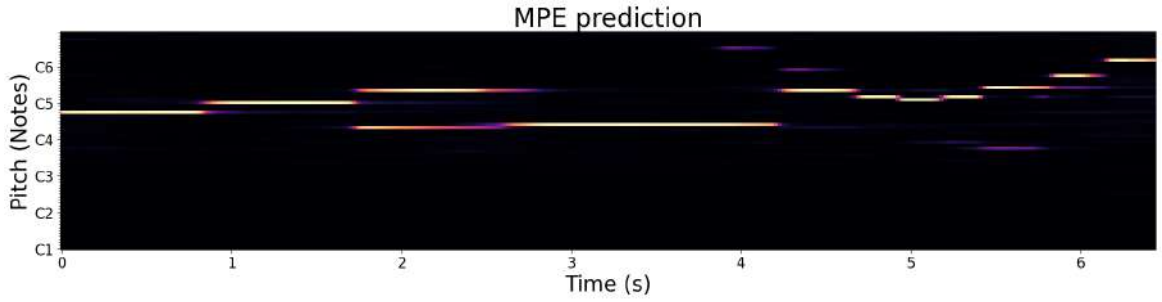
MFCCs [Logan and Ma (2000); Rabiner and Juang (1993)] are computed by applying a Mel scale filterbank to the magnitude spectrum, taking the logarithm, and performing a discrete cosine transform. In practice, only the first few coefficients are typically retained, as they capture the coarse spectral envelope of the signal.

They capture the overall timbral shape of the signal and are widely employed in music and speech analysis for tasks such as timbre characterisation, onset detection, and audio alignment. Musically, MFCCs are useful because they primarily encode the spectral envelope rather than fine harmonic structure, making them largely invariant to changes in fundamental frequency and therefore suitable for comparing timbral characteristics across different instruments, registers, and articulations. Furthermore, by applying the discrete cosine transform to a logarithmically scaled Mel spectrum, MFCCs compress spectral information along perceptually meaningful frequency bands, which better approximate the human auditory system than linearly spaced frequency bins [Dias and Hadjileontiadis (2021)]. This frequency warping enables a more compact and perceptually relevant representation of sound, useful not only for musical analysis but also for applications such as instrument identification [Nagawade and Ratnaparkhe (2017)], genre classification [Rajesh and Bhalke (2016)], or music emotion recognition [Nalini and Palanivel (2016)].

PFCC

PFCCs [Muller and Ewert (2010)] extend this approach by incorporating additional perceptual weighting to better reflect human auditory sensitivity. Both MFCC and PFCC representations can also serve as inputs to learning-based models, providing robustness to timbre variations and recording conditions.

2.2.5 Learned Features



a

Figure 2.11: Multi-pitch estimation representation of the audio signal, with estimated pitches shown over time [Weiß et al. (2021b)].

More recently, learned features derived from neural networks have gained prominence due to their ability to capture musically meaningful information directly from audio data. In particular, models trained for Multi-Pitch Estimation (MPE) can provide frame-level pitch activations that encode harmonic and melodic content in a compact representation.

Compared with handcrafted signal-processing descriptors, learned representations can capture complex spectral patterns and adapt to variations in timbre, instrumentation, and recording conditions. Earlier approaches incorporated transcription-informed features by combining several symbolic cues, while more recent methods exploit continuous neural predictions as audio representations for downstream tasks [Maman and Bermano (2022); Zeitler et al. (2024a)].

These learned representations have demonstrated strong performance in tasks such as automatic transcription and audio-to-score alignment, where they provide a data-driven alternative to manually designed acoustic features.

While the representations described above capture complementary aspects of the audio signal, including spectral content, pitch, harmony, timbre, and learned musical descriptors, they do not directly provide the temporal positions of musical events. For tasks requiring precise localisation of note beginnings, these representations are further transformed into onset activation functions, also referred to as novelty functions [Müller and Chiu (2024)]. More recently, learned features derived from neural networks trained for MPE have gained prominence due to their ability to capture musically meaningful pitch and harmonic information.

2.2.6 Onset Activation Functions

An *onset activation function*, also referred to as a *novelty function*, converts an audio representation into a continuous temporal curve whose peaks indicate potential onset locations.

Different activation functions emphasise different properties of the acoustic signal. Spectral approaches measure rapid changes in frequency content, while pitch-based and learned approaches exploit higher-level musical information to improve robustness for harmonic and polyphonic instruments.

Spectral Flux

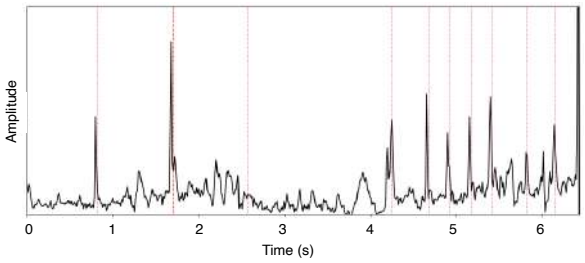


Figure 2.12: Spectral flux of the introduction recording of Fantasia No. 2 (Fig. 2.6), computed using *libfmp* [Müller and Zalkow (2021)]. The vertical lines indicate the positions of the manually annotated reference onsets. Peaks in the spectral flux novelty function highlight potential onset locations.

The *Spectral Flux* (SF) measures temporal changes in a spectrogram by computing the positive difference between successive spectral frames [Müller (2021)].

Given a logarithmically compressed magnitude spectrogram

$$\mathcal{Y} = \log(1 + \gamma|S|) \tag{2.19}$$

A discrete temporal derivative is then computed. The half-wave rectification suppresses decreases in spectral energy and retains only positive changes associated with attacks.

$$SF(m) = \sum_{k} \max(0, \mathcal{Y}(m+1, k) - \mathcal{Y}(m, k)) \tag{2.20}$$

Where $\mathcal{Y}(m, k)$ denotes the (log-compressed) spectral representation at frame m and frequency bin k.

The combination of Spectral Flux and chroma features has been shown to achieve DTW alignment accuracy comparable to that of the DLNCO approach [Özer et al. (2021)].

SuperFlux

The *SuperFlux* (SF^*) method extends spectral flux to improve robustness against frequency modulations such as vibrato [Böck and Widmer (2013)]. Instead of directly comparing two

consecutive spectral frames, it first tracks local spectral trajectories using a maximum filter along the frequency axis. By comparing spectral energy with its locally shifted maximum, SuperFlux reduces false detections caused by moving harmonic components while preserving genuine onset-related increases. This makes it particularly suitable for sustained instruments such as string ensembles.

Learned Onset Activations

Recent approaches increasingly rely on neural networks trained directly for onset detection. These models typically output frame-wise activation values interpreted as the probability of a musical event occurring at each time frame.

Compared with handcrafted novelty functions, learned onset activations can capture complex temporal and spectral patterns directly from data. Convolutional neural networks have achieved state-of-the-art performance in onset detection, further strengthening their role as complementary signals in alignment systems [Schlüter and Böck (2014)]. These learned activations are now widely integrated into automatic transcription and audio-to-score alignment systems (alignment tasks will be discussed in Chapter 3).

To summarise, Table 2.1 provides an overview of audio representations, from low-level spectral features to high-level timbral and harmonic descriptors, highlighting some representative musical applications and their relevance for analysis.

Section summary – Audio Representations and Feature Extraction

Audio representations convert the sampled acoustic signal into descriptors that capture complementary musical information, including pitch, harmony, timbre, and note onsets. They range from classical signal-processing approaches, such as spectrograms, chroma, and cepstral features, to learned neural representations that offer increased robustness. These representations form the basis of many MIR tasks, including transcription, alignment, and musical analysis.

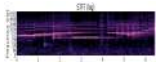
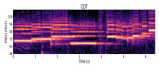
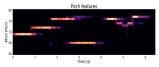
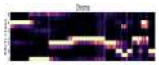
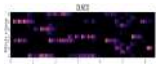
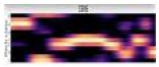
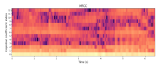
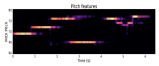
	Representation	Typical Usage
Low-level / Signal-based		
	**Waveform**	Full time-domain signal, contains all audio information; can serve as input to neural networks for tasks such as source separation, MPE, or onset detection.
	**STFT**	Time–frequency analysis, onset detection, transient localisation.
	**Log-STFT**	Perceptually scaled spectral analysis, coarse pitch and timbre cues.
	**CQT**	Log-frequency representation, music transcription robustness, tonal analysis, harmonic analysis, polyphonic transcription.
Mid-level / Pitch-based		
	**Pitch-related features**	Monophonic pitch tracking, note transcription, onset refinement.
High-level / Harmonic / Timbre-based		
	**Chroma (CQT)**	Harmonic content representation and similarity, key estimation, chord recognition.
	**DLNCO**	Onset-enhanced chroma features for polyphonic alignment and structure analysis.
	**CENS**	Smoothed and quantised chroma for robust alignment, large-scale structure retrieval and cover song detection.
	**MFCC / PFCC**	Cepstral representations of the spectral envelope for timbre modelling, instrument recognition, genre classification, audio segmentation, and robust audio characterisation.
	**Learned features(MPE)**	Neural representations learned from data, capturing discriminative harmonic and temporal structures; used for multi-pitch estimation, onset detection, and alignment.

Table 2.1: Typical usages of common audio representations organised by abstraction level, ranging from low-level signal descriptors to higher-level harmonic and timbral features.

2.3 Performance

While the previous sections have focused on signal representations and feature extraction, musical meaning ultimately emerges at the level of performance, where expressive decisions ultimately structure the temporal and spectral structure of the audio signal.

Musical performance is shaped by multiple layers of interaction (Fig. 2.13), where expressive decisions continuously influence the temporal and spectral structure of the audio signal. Performers engage not only with the notated score but also with their instrument, fellow musicians, and the listening audience, negotiating timing, expression, and dynamics throughout. Their interpretation draws on personal experience and broader musical knowledge, including stylistic and social conventions, performance practices, and familiarity with the repertoire. The audience contributes as well, both through direct feedback and through their memory of previous performances, influencing perception and expectation.

The interactions described above give rise to observable manifestations of musical performance, which can be analysed through measurable parameters. In the following, we first introduce general performance parameters before focusing more specifically on timing, which plays a central role in expressive interpretation.

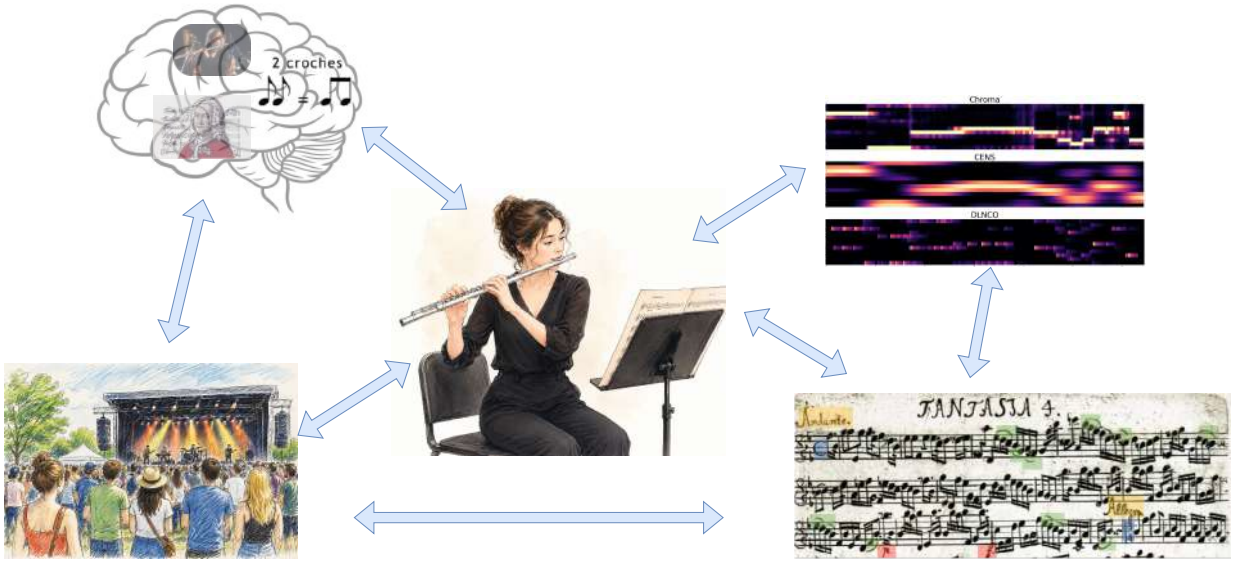


Figure 2.13: Interactions in performance (Images of the flautist and the audience generated using OpenAI DALL·E, 2026).

2.3.1 Parameters of Performance and Representations

Parameters

Musical performance extends beyond the mere reproduction of a notated score. As reviewed by [Lerch et al. (2019)], performers shape music through a combination of core parameters such as tempo, timing, dynamics, pitch, and timbre.

Tempo, "the speed at which a piece of music is performed" [Latham (2011)], has long been conceptualised not just as a single rate of beats per minute but as a continuous temporal profile over the course of an interpretation. The notion of *tempo curves* has been proposed to model how symbolic time (score time) is transformed into physical performance time [Konz et al. (2009)], producing a detailed trajectory of tempo across an interpretation. Such curves reflect expressive timing decisions such as accelerando and ritardando rather than a single averaged tempo, but can be limited in fine-grained analysis [Desain and Honing (1993)]

Timing, in particular, encompasses both macro-level tempo and micro-timing deviations from the notated rhythm that give rise to expressive gestures such as rubato. It will be more detailed in Section 2.3.2.

Dynamics operate at multiple levels, from overall loudness to note-by-note variations. A quantitative measure of loudness over time is given by root-mean-square amplitude (RMS) curves [Narang et al. (2022)], allowing one to visualise dynamic contours throughout a performance. Such RMS curves not only capture the overall shape of crescendos and decrescendos but can also reveal subtler expressive details like emphasis on particular notes or phrases. Dynamics can even be correlated to timing deviations, reflecting how performers may slightly stretch or compress time to highlight dynamic peaks or transitions [Repp (1992)].

Pitch includes fine intonation, vibrato [Seashore (1937)], and ornamentation. Fine pitch control involves not only hitting the nominal notes but also maintaining stability and tonal precision throughout performance, which depends on the coordination of breath, embouchure, and fingerings. Han and Lee [Han and Lee (2014)] present a hierarchical approach to detect common mistakes of beginner flute players, analysing the audio signal to identify errors related to intonation, tone production, and fingering. Their method demonstrates that deviations in pitch and other performance inaccuracies can be systematically detected, providing objective feedback on technical difficulties. This work illustrates that analysing pitch goes beyond simple note accuracy: it is a crucial tool for understanding and modelling both the technical and expressive aspects of musical performance, and it has practical applications in music education and performance assessment.

Timbre, finally, captures both instrument-specific sound qualities and extended techniques such as multiphonics [Riera et al. (2014)], which enrich the expressive palette of a performance.

Beyond these measurable dimensions, higher-level attributes like expressivity and emotional engagement are crucial in shaping musical interpretation [Dias and Hadjileontiadis (2021)]. Computational analysis of performance seeks to make these nuances accessible to systematic

study by encoding them in structured formats and extracting quantitative features.

Representations

Several symbolic music formats provide frameworks for representing notated aspects of performance.

Devaney and Gauvin [Devaney and Léveillé Gauvin (2019)] propose extensions to Humdrum and MEI to better capture performance-relevant information: Humdrum is augmented with *beat-* and *measure-level descriptors*, which allow the encoding of rhythmic and dynamic variations at a coarse-grained level, while MEI incorporates continuous data capable of representing expressive parameters over the course of a performance, such as tempo and dynamic evolution.

In contrast, the *Music Performance Markup* (MPM) proposed by Berndt et al. [Berndt (2021)] offers a model-driven XML-based framework designed specifically to describe performed expressive parameters in a systematic and analysable way, enabling detailed computational studies of timing, dynamics, and articulation.

These representations are complemented by software tools and large-scale datasets that facilitate data-driven analysis. The Python toolkit *pyAMPACT* [Devaney et al. (2026)] allows researchers to extract, visualise, and analyse features related to timing, dynamics, pitch inflexion, and spectral timbre.

Although these tools enable multi-dimensional analyses of musical expression, Section 2.3.2 focuses more specifically on timing, which plays a central role in expressive performance. In particular, we examine temporal deviations relative to the score and their interpretation in terms of expressive timing phenomena such as notes inégales.

2.3.2 Focus on Timing

Timing is a central dimension of musical performance and will be studied in this thesis (Chapter 6). Whereas tempo refers to the rate at which musical notes are played [Dixon (2001)], *timing deviations* refer to the differences in the performed notes' onsets and durations compared to a *score time* performed at a constant tempo.

Such deviations are not random but exhibit systematic and expressive patterns [Gabrielsson, Alf (1999)]. A common low-level descriptor used to characterise temporal structure is the inter-onset interval (IOI), which captures the time elapsed between consecutive note onsets and can be analysed at multiple hierarchical levels, including notes, beats, bars, and phrases [Friberg and Battel (2002)].

Only a few performance studies have explicitly related timing deviations to specific structural or compositional aspects of the music. Repp [Repp (1992, 1998)] examined piano performances of "Träumerei" by Schumann across different interpretations and measured similarities in the relative durations of consecutive grace notes. He also highlighted systematic lengthening of

downbeats as a timing strategy for performing ritardandi at the ends of melodic gestures, for instance in Chopin's "Etude in E Major". Palmer further demonstrated systematic timing patterns within chords and ornaments, such as the "melody lead" effect and structured temporal organisation in trills [Palmer (1996a,b)].

From a perceptual perspective, Clarke [Clarke (1989)] examined listeners' sensitivity to timing variations in both tonal and atonal contexts. The study showed that timing deviations as small as approximately 20 ms can be perceptible depending on note duration and context, and that sensitivity varies with metrical position. Such timing deviations include the baroque practice of playing *notes inégales*, two consecutive notes of the same rhythmical value with different durations [Fuller (2001)].

Notes Inégales
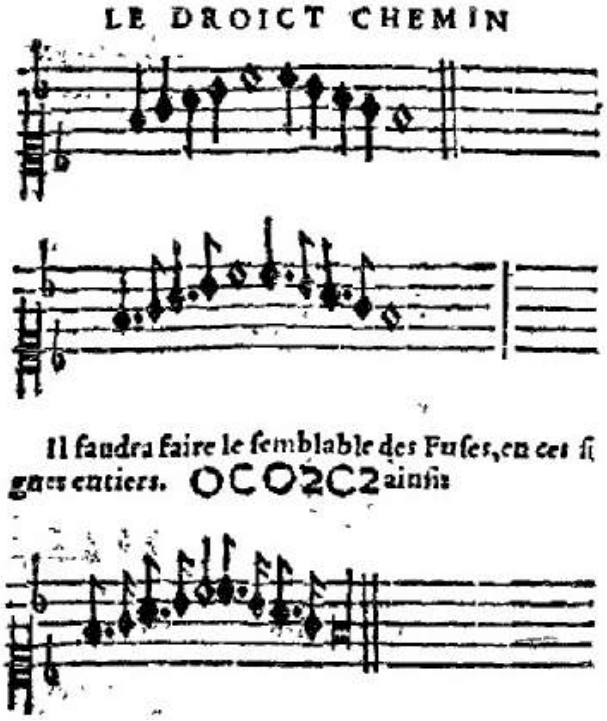


Figure 2.14: Excerpt of Chapter "De entrelaceur, & de la maniere de chanter les demiminimes" from "Le Droict Chemin de musique" by Loys Bourgeois [Bourgeois (1550)].

From its first description by Loys Bourgeois in 1550 (Fig. 2.14) to the late 18th century, this practice was initially restricted to one-quarter of the *tactus* in duple metre [Houle (1987)]. The

strong beats or the first part of each beat were often emphasised compared to the following notes, creating pairs of notes with strong and weak accents. Such an (unequal) performance was generally not mentioned in the scores, but theorised in musical treatises and flute methods from this period [Boland (1998); Fontijn (1995)].

Cyr gathered various baroque sources (Quantz, Mattheson) describing where the inequality occurs in baroque dances [Cyr (2016)]. Moelants studied the *timing ratio* as defined by Étienne Loulié in 1696 [Loulié (1696)], between the longest and the shortest notes when harpsichordists and violinists play *gavotte* [Moelants (2011)]. He highlighted the influence of interval size, beat position, and tempo on the *notes inégales*. Honing showed that this timing ratio decreases when tempo increases [Honing (2006)].

Section summary – Performance

Performance is not a mere reproduction of notation, but the result of expressive choices that become measurable through parameters such as tempo, timing, dynamics, pitch, and timbre.
A particular focus is placed on timing, distinguishing global tempo from micro-timing deviations relative to score time. These deviations—far from random—reflect structural, stylistic, and expressive principles (e.g., rubato or notes inégales) and can be studied through measures such as inter-onset intervals and tempo curves.

2.4 Perceptual and Musicological Models

Having examined measurable aspects of musical performance and expressive timing, we now turn to perceptual and musicological models that describe how listeners organise, interpret, and anticipate musical structures. Beyond their acoustic realisation, melodic and temporal patterns give rise to expectations, grouping processes, and perceptual organisations that shape musical understanding. These models provide a theoretical framework for studying how musical events are cognitively structured, both in the perception of melodic continuation and in the emergence of implied polyphonic voices from monophonic textures.

2.4.1 Perception of Realisation

The *Implication/Realization* (I-R) model, developed by Eugene Narmour [Narmour (1990, 1992)], provides a cognitive model of melodic perception in which listeners continuously form expectations about the continuation of a melody. Each melodic interval, termed the *implicative interval*, generates expectations that are either confirmed or violated by the subsequent *realised*

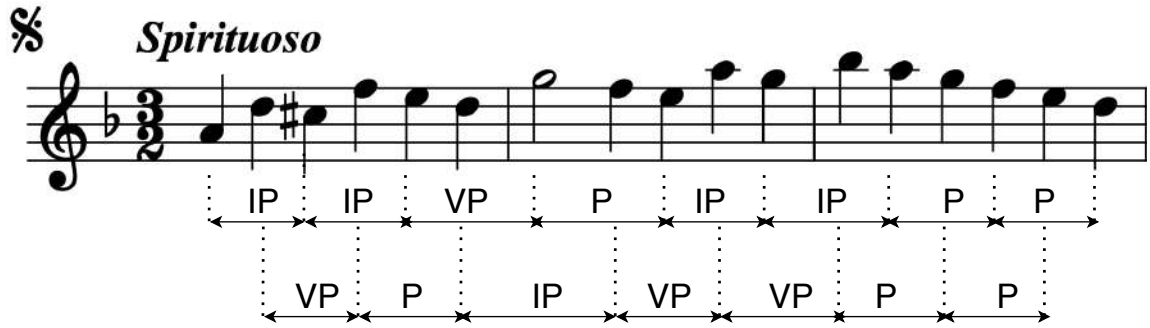


Figure 2.15: Analysis by the author of the first three measures of the last movement, *Rondo*, from Fantasia No. 6 (TWV 40:7) for solo flute by G. P. Telemann, according to Narmour's *Implication–Realization* model. The labels indicate process (`P`), intervallic process (`IP`), and vectorial process (`VP`).

interval. These relations are primarily governed by two parameters: interval size and melodic direction.

I-R model is grounded in principles of Gestalt psychology [Koffka (1935); Wertheimer (1923)]. In melodic perception, listeners tend to favour continuity, proximity, and good form, leading to the grouping of successive notes into perceptually meaningful units. These principles support the idea that melodic understanding emerges not from isolated pitch events but from dynamic relational patterns unfolding over time.

Core Principles of I-R Model

Narmour distinguishes between small and large intervals, with small intervals typically promoting continuation in both size and direction, while large intervals tend to induce a reversal of direction and a subsequent reduction in interval size. Directional continuity, registral return, and proximity thus act as central forces shaping melodic expectancy. Tritone intervals occupy a perceptually ambiguous status, often weakening predictive strength.

From these principles, Narmour derives a set of melodic archetypes describing typical implication–realization patterns that capture common perceptual responses to melodic motion and form the basis of a grammar-like system of melodic expectation.

Figure 2.15 illustrates some of these principles. Several ascending fourths are followed by descending stepwise motions (e.g., `C#5-F5-E5` and `D5-G5-F5`), producing *intervallic process* (`IP`) patterns in which a leap is compensated by a subsequent step in the opposite direction. By contrast, configurations such as `D5-C#5-F5` are interpreted as *vectorial processes* (`VP`), since the ascending leap disrupts the continuation implied by the preceding descending step. The passage concludes with an extended descending *process* (`P`): `B5-A5-G5-F5-E5-D5`, reinforcing

directional continuity and providing a sense of local stabilisation.

Extensions and Simplifications

Beyond pitch, I-R model extends to other musical dimensions, including duration, dynamics, and consonance, suggesting that similar cognitive mechanisms underlie the perception of temporal and expressive structures. Schellenberg [Schellenberg (1997)] proposed a streamlined version of the theory, reducing it to five general principles: registral direction, intervallic difference, proximity, registral return, and closure.

This formulation highlights the interaction between bottom-up perceptual constraints and top-down processes shaped by musical experience and cultural familiarity [Pearce and Wiggins (2006)]. Bottom-up mechanisms arise directly from the acoustic and structural properties of the musical signal, such as proximity or contour continuity, whereas top-down processes depend on learned stylistic conventions, listening habits, memory, and expectations acquired through prior exposure to musical repertoires.

2.4.2 Implied Voices

Beyond melodic expectation, listeners may also perceive higher-level structural organisations within a single melodic stream. In particular, monophonic melodies can sometimes evoke the impression of multiple simultaneous voices, a phenomenon commonly referred to as *implied polyphony*. Although only one note is physically produced at a time, listeners may perceptually segregate the melodic line into distinct virtual voices unfolding in different pitch registers or following different contour trajectories.

The perception of multiple simultaneous voices within a single melodic line has been extensively discussed in both music theory and music cognition. Early theoretical descriptions appear in Piston's *Counterpoint* (1947) [Piston (1947)], particularly in its first Chapter, where he introduces the concept of the *melodic curve* and discusses the notion of the *compound melodic line*, notably in the context of J.S. Bach's works.

In music theory, the term *compound melody* is frequently used within Schenkerian analysis to describe melodies that suggest multiple contrapuntal voices. In contrast, cognitive and perceptual studies often employ the term *melodic fission*, attributed to Dowling (1973) [Dowling (1973)], to describe the perceptual splitting of a single melodic stream into multiple auditory streams. According to this framework, large pitch leaps, contour changes, and other contextual cues can induce the listener to perceive two or more distinct melodic lines emerging from a single sequence of notes. Additional factors, such as silences and abrupt dynamic changes, may further reinforce this perceptual separation.

Perceptual Foundations

Dowling's study on the perception of interleaved melodies demonstrated that listeners can segregate alternating pitch sequences into distinct perceptual streams under specific conditions [Dowling (1973)]. This phenomenon is particularly prominent in Baroque repertoire, where interleaved melodies are used to generate a contrapuntal texture in monophonic instruments. Canonical examples include the solo sonatas and partitas of J.S. Bach, as well as the solo flute fantasias of Telemann. In such contexts, intervals smaller than approximately three and a half semitones are systematically avoided in passages designed to evoke melodic fission, thereby facilitating perceptual voice separation.

More recent examples can also be found in jazz improvisation, notably in the work of John Coltrane, where rapid alternation between registral regions produces the impression of multiple simultaneous melodic strands [Dowling (1973)].

Experimental studies on auditory streaming further support these observations. Classical works by Miller and Heise [Heise and Miller (1951); Miller and Heise (1950)], van Noorden [van Noorden (1975)], Narmour [Narmour (1990)], and others have established the importance of pitch distance, contour, and temporal coherence in auditory stream segregation.

Symbolic Rule-Based Analysis of Implied Polyphony

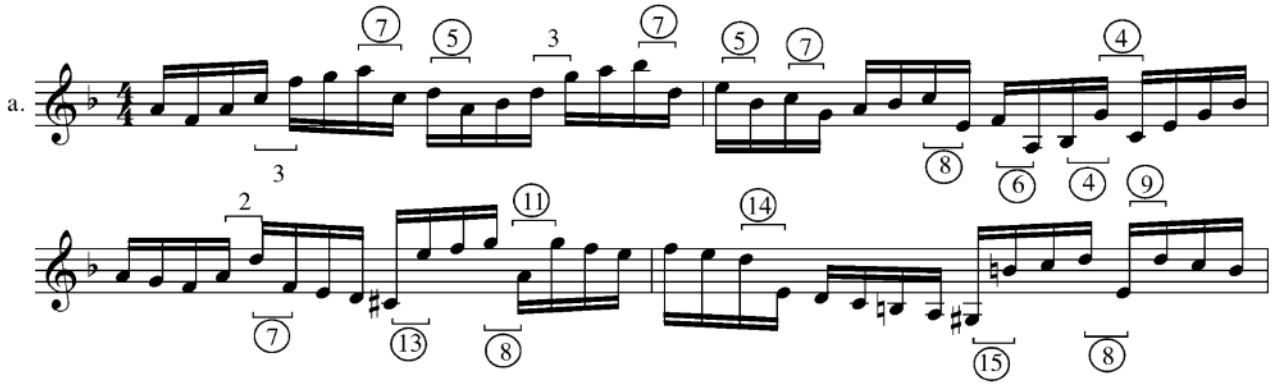


Figure 2.16: From [Davis (2006)]: excerpt of J. S. Bach's *Allemande* from Violin Partita No. 2 in *D minor*, mm. 4–7. Numbers represent point totals generated by Davis's rule system; circles indicate intervals that mark the transition from one implied voice to another.

A systematic computational framework for analysing implied polyphony in monophonic music is proposed by Davis [Davis (2006)] (Fig. 2.16). The method relies on the detection of salient pitch events that may act as perceptual boundaries between implied voices, primarily based on intervallic and contour-related cues. Large melodic intervals are interpreted as

potential indicators of voice separation, with the perfect fourth serving as a perceptual threshold distinguishing small from large intervals. Each additional diatonic step beyond this threshold incrementally strengthens the likelihood of perceptual segmentation.

Contour changes further reinforce this effect: abrupt reversals of melodic direction contribute positively to perceptual separation, reflecting Gestalt principles of good continuation and supported by experimental evidence showing that V-shaped pitch trajectories enhance segmentation [Heise and Miller (1951); Huron and Royal (1996)]. Conversely, conjunct motion, such as unison or stepwise progressions, tends to promote perceptual continuity within a single voice and therefore reduces the likelihood of voice separation [Dowling (1973)]. These cues are combined into a cumulative separation score, to which a perceptual threshold is applied in order to determine the presence of a voice change.

Additional musical factors, such as articulation and phrasing marks, may further reinforce perceptual separation when explicitly notated, although their reliability must be considered carefully due to potential editorial intervention. Notably, rests and silences are not explicitly modelled in this framework.

As a complementary work, we investigate the implementation and improvement of Davis's algorithm in Appendix A.

Taxonomy of Implied Polyphony

Building on this analytical framework, Davis proposes a taxonomy of implied polyphonic textures that captures the diversity of perceptual strategies used by listeners to infer multiple voices from monophonic material. In *linear* textures, voices move primarily by step within distinct pitch registers, preserving stable registral separation. *Motivic* textures rely on the alternation of characteristic figures assigned to different voices, often involving arpeggiated motion and overlapping registers. In *antiphonal* textures, a single motif is exchanged between voices, typically articulated through phrasing and slur patterns.

Other configurations include *pedal* textures, in which one voice sustains or reiterates a pitch while another unfolds melodically, and *sequential* textures, where phrases or motifs are repeated at different pitch levels, creating the impression of multiple simultaneous strands. Together, these categories provide a perceptually grounded framework for describing how monophonic musical textures, particularly in idiomatic writing for solo instruments, can evoke complex polyphonic structures.

Section summary – Perceptual and Musicological Models

Now listeners perceive melodic structure and implied polyphony?
Implication/Realization (I/R) theory models melodic expectation based on interval size and direction, forming archetypal patterns of melodic continuation or reversal.
Implied voices describe how a single line can be heard as multiple streams, driven by large intervals, contour changes, and dynamic cues. Computational frameworks, like Davis's rule-based system, formalise perceptual segmentation.

Chapter summary – Processing Music: From Representation to Perception

In Music Information Retrieval, the analysis of a musical performance—the acoustic realisation of a score—relies on tools and representations from both the symbolic and audio domains. Symbolic representations formalise the score, while audio representations capture the sound itself. The acoustic realisation can then be analysed through specific performance- related parameters. Among them, timing refers to the differences in the performed notes onsets compared to a metronomic realisation. Perceptual models further support the understanding of the melodic structure, including aspects such as implied polyphony in monophonic compositions.

To the question *"Is the score alone sufficient for analysing or understanding music?"*, the answer is no: although the score provides a structured representation of musical intentions, it does not fully describe the acoustic, expressive, and perceptual dimensions of a performance. A complete understanding of music therefore requires combining symbolic analysis with audio information and perceptual models that account for how music is actually performed and experienced by listeners.

Chapter 3

Musical Alignment: Tasks, Methods and Tools

Contents of this chapter

3.1 General Framework of Musical Alignment . . . 55
3.2 Alignment Tasks and Scenarios . . . 56
3.2.1 Alignment Modalities . . . 56
3.2.2 Offline vs. Online Alignment . . . 58
3.3 Alignment Algorithms: Focus on Dynamic Time Warping . . . 61
3.3.1 Overview of Alignment Algorithms . . . 61
3.3.2 Dynamic Time Warping (DTW) . . . 63
3.3.3 Multi-Scale and Multi-Resolution DTW . . . 67
3.4 Frameworks . . . 71
3.4.1 Overview of Existing Frameworks for Offline Alignments . . . 71
3.4.2 Focus on Synctoolbox . . . 72
3.5 Annotations . . . 73
3.5.1 Annotations Transfer . . . 73
3.5.2 Triple Transfer and Late-Fusion . . . 74
3.6 Datasets . . . 76

Musical alignment addresses the problem of establishing meaningful correspondences between multiple renditions of the same musical content, despite variations in timing, expression, and interpretation.

How can such heterogeneous performances be reliably synchronised at a fine-grained musical level?

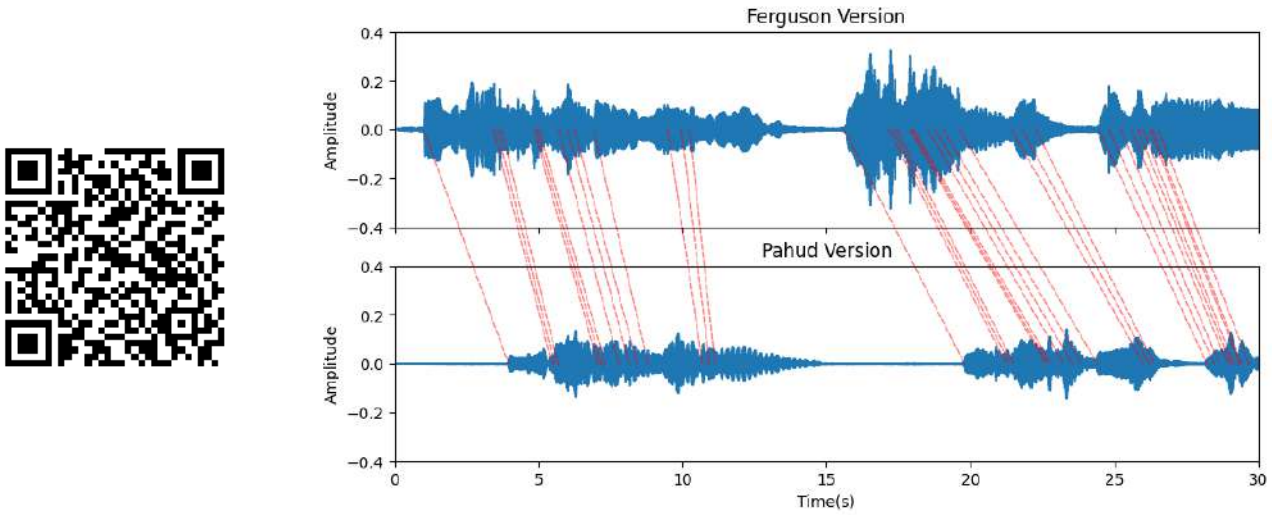


Figure 3.1: Waveforms of the first 30 s of *Syrinx* (L. 129 - CD. 137) by Claude Debussy, as performed by Emi Ferguson (top) and Emmanuel Pahud (bottom). Corresponding score onsets are connected by red lines (https://patthmus.github.io/phd-manuscript-links/chap3.html).

Figure 3.1 illustrates this issue through two performances of *Syrinx* by Claude Debussy aligned at the level of score onsets.

This chapter reviews the main concepts, methods, and tools of musical alignment. After defining the task and its principal application scenarios, we focus on dynamic time warping (DTW) as a core family of alignment algorithms, before presenting widely used software frameworks and discussing the role of annotations and datasets, which are essential for reproducible evaluation.

3.1 General Framework of Musical Alignment

Musical alignment refers to the problem of establishing temporal correspondences between different musical representations. These representations may include symbolic scores, MIDI files or audio recordings of several distinct performances of the same musical piece. The goal is to determine how musical events in one representation relate temporally to those in another.

This problem plays a central role in numerous MIR applications, such as performance analysis, score following or large-scale corpus annotation. Musical alignment is inherently challenging due to the expressive freedom of performers and the complex nature of musical signals. Tempo fluctuations and particularly expressive timing lead to non-linear temporal deformations between performances. By providing a common temporal reference, musical

alignment enables detailed comparisons between these performances and facilitates the integration of heterogeneous musical data.

Let $\mathcal{T} = \{t_n \mid n = 1, \ldots, N\}$ denote a set of onset positions defined over a continuous temporal axis. In the case of audio signals, these positions correspond to physical time (in seconds), whereas for symbolic representations, they are expressed in musical time (such as in quarter notes). Within this framework, we formalise musical alignment as a mapping between onset sets associated with two performances or representations. Given two representations X and Y, the alignment operator $\mathcal{A}^{X \to Y}$ maps each onset in $\mathcal{T}_X$ to a corresponding position in $\mathcal{T}_Y$, thus providing a continuous temporal correspondence between both representations.

Section summary – General Framework of Musical Alignment

Musical alignment establishes temporal correspondences between different representations of the same piece, enabling detailed comparisons and integration of heterogeneous musical data. Formalising it as a mapping between onset sets provides a clear framework for MIR applications.

3.2 Alignment Tasks and Scenarios

Musical alignment encompasses a broad family of tasks that differ according to the nature of the aligned representations, the temporal constraints imposed by the application, and the intended usage of the resulting correspondences. Depending on the considered scenario, alignment may involve synchronising multiple audio recordings or matching audio to symbolic representations.

In this section, we first distinguish the main alignment modalities based on the nature of the representations involved. This categorisation focuses on the methodological aspects of the alignment problem, independently of its intended application. We then discuss the distinction between offline and online alignment scenarios, which primarily reflects different operational contexts and temporal constraints, and has strong implications in terms of algorithmic design and achievable accuracy.

3.2.1 Alignment Modalities

Musical alignment consists of establishing temporal correspondences between two musical representations, which may belong to different modalities. Depending on the nature of the aligned data, two main alignment paradigms can be distinguished: *audio-to-audio* and *score-to-audio*. Each of these modalities raises specific methodological challenges and is associated with distinct application scenarios.

In this work, we focus on cases where a symbolic score representation is available and therefore do not consider sheet music–to–audio alignment, which requires processing image data to extract a symbolic music representation prior to synchronisation [Thomas et al. (2012)].

Score–Audio Alignments

Score–audio alignment consists of mapping score events to their corresponding times in an audio recording [Muller et al. (2004); Soulez et al. (2003); Turetsky and Ellis (2003)].

Let $\mathcal{S} = \{s_n \mid n = 1, \ldots, N\}, s_n \in \mathbb{R}$ the set of onset positions in a symbolic score defined over a continuous temporal axis where s_n is the metric position (e.g., cumulative durations in quarter-note units, beats, ticks, ...) of the n-th event. To convert these positions in the time domain, a rendering function can map symbolic positions to physical time, depending on tempo (or even expressive parameters). This conversion mechanism underlies standard symbolic-to-audio representations such as MIDI, where event positions are specified in ticks and rendered into physical time during playback.

Several methods have been proposed to align a symbolic music representation with a corresponding audio performance. Early approaches, such as [Dannenberg and Hu (1987)], first convert the MIDI file into an audio signal using a synthesiser. Both the synthesised and recorded audio are then transformed into chroma features for comparison. Alternatively, MIDI data can be directly converted into chromagrams, bypassing the intermediate audio synthesis step, which reduces computational cost and potential timbral mismatches.

Other approaches focus on spectral characteristics. For example, Orio and Schwarz [Orio and Schwarz (2001)] use the Power Spectral Density (PSD) of both the performance and the score to compute local distances between frames. In this framework, the score is represented by a simple harmonic spectrum for each note, augmented with models for silences and note attacks to better capture temporal onset information. Specifically, spectral bands corresponding to the harmonics of each note are generated and multiplied by the Fourier magnitude of the corresponding performance frames. After normalisation, the peak structure is computed for both the score and the performance, and local distances are evaluated based on the difference between these PSD peak structures. In this setting, both data streams are assumed to be fully available prior to the alignment, which enables global optimisation strategies to compute the alignment [Müller (2021)].

More recently, deep learning approaches have shifted the paradigm from manually designed acoustic features towards learned representations and end-to-end alignment models. Rather than relying on predefined similarity measures between handcrafted features, these methods learn task-specific representations directly from data. For instance, [Teytaut et al. (2023)] introduced a framework for vocal music alignment based on deep Convolutional Neural Network (CNN). Although developed for singing voice, this approach illustrates the broader trend towards neural alignment models capable of learning robust correspondences between

audio signals and symbolic representations. The framework enables automatic alignment of vocal recordings with lyrics/syllables (voice-to-text alignment) and musical notes (voice-to-note alignment) directly from commercial recordings, reducing the need for isolated vocal tracks and extensive manual pre-processing. These approaches generally rely on large annotated datasets for training, which remain difficult to obtain for specialised instrumental repertoires. This limitation is particularly relevant for expressive solo performances, where variations in timing, articulation, and interpretation introduce additional challenges for learned alignment models.

Although often grouped under the term *score–audio alignment*, some authors distinguish between *score-to-audio* and *audio-to-score* alignment according to the operational scenario. However, this terminology is not used consistently throughout the literature, and the two expressions are often employed interchangeably or replaced by more specific terms such as *score following* for real-time applications (see Section 3.2.2).

Audio-to-Audio Alignments

Audio-to-audio alignment aims at establishing temporal correspondences between multiple audio recordings of the same musical piece. Since all aligned signals belong to the same modality, feature representations can be directly compared, which often leads to robust alignment results.

Most approaches rely on mid-level representations that aim at capturing musically meaningful characteristics of the signal while discarding irrelevant low-level details: chroma-based representations [Bartsch and Wakefield (2005); Müller et al. (2005)] or onset-features [Ewert et al. (2009); Grosche et al. (2009)]. Compared to score-informed alignment, audio-to-audio methods avoid potential mismatches induced by symbolic abstraction, but must cope with expressive timing variations, articulation differences, and acoustic variability.

Recent works exploit the availability of multiple performances to improve alignment accuracy. For instance, [Wang et al. (2014); Zalkow et al. (2017)] propose multi-version alignment frameworks that simultaneously align several recordings, leveraging redundancy to stabilise the estimated correspondences. Progressive and reference-based strategies are effective in reducing local misalignments caused by expressive timing, articulation differences, or recording conditions.

While alignment modalities are primarily determined by the nature of the aligned representations, another fundamental distinction arises from the temporal constraints imposed by the application. In particular, alignment methods can be categorised as offline or online, depending on whether future observations are accessible during the alignment process.

3.2.2 Offline vs. Online Alignment

Musical alignment tasks can be further distinguished according to the temporal constraints imposed by the application, leading to the classical separation between *offline* and *online* alignment scenarios (Fig. 3.2).



Figure 3.2: From left to right: *Dezrann* an open-source web platform to interact with corpora while sharing music and music analysis [Ballester et al. (2025)] https://www.dezrann.net/~/bach-fugues/bwv851?audio=1; *Metronaut* (screenshot of a commercial video): a real-time score-following application using *Antescofo* https://metronautapp.com/; An overview of Wagner Ring Dataset (WSD): a large-scale annotated and audio-to-audio synchronised operas [Weiß et al. (2023)].

Offline Alignment

Offline alignment refers to scenarios in which the complete data streams to be synchronised are available prior to the alignment process. This setting enables the use of global optimisation strategies. As a result, offline approaches generally achieve high temporal accuracy and robustness, making them particularly well-suited for large-scale musicological studies and detailed performance analyses.

In the context of *audio–audio* synchronisation, offline alignment plays a central role in applications such as expressive timing analysis [Liem and Hanjalic (2011)], identification of structural differences between performances [Grachten et al. (2013)], cover song identification [Serra et al. (2008)], and large-scale corpus annotation [Weiß et al. (2021a, 2023)].

For *score–audio* synchronisation, offline alignment is primarily employed for score-informed music analysis, performance studies [Devaney et al. (2026)], and annotation transfer across recordings [Pratzlich and Muller (2016)], enabling systematic large-scale investigations.

Offline synchronisation techniques also constitute a key component of multimodal music navigation systems and interactive visualisation platforms, where reliable cross-modal alignment enables fine-grained exploration of musical content, as exemplified by the Dezrann platform [Ballester et al. (2025)].

Online Alignment

Online score–audio alignment arises in interactive musical applications where temporal correspondences between a live performance and a symbolic score must be established as the performance unfolds. Typical applications include automatic accompaniment systems [Dannenberg (1984)], live electronic performance environments [Cont (2008); Vercoe (1984)],

and music education tools providing real-time feedback to performers [Arzt et al. (2014)].

To support such applications, the system must continuously estimate the performer's current position within a symbolic score as well as the instantaneous tempo from an ongoing audio stream. This task is commonly referred to as *real-time score following* [Cont (2010)]. The simpler term *score following* can, however, be ambiguous, as it has also been used in the literature to describe offline score–audio alignment tasks [Dannenberg and Hu (1987)].

In this setting, the symbolic representation of the piece is assumed to be known in advance, while the audio signal is processed sequentially. In the simplest case of solo performance tracking, at least two tasks are involved: (i) detecting what the performer is currently playing, and (ii) matching this information to the score while maintaining both high tracking accuracy and low latency [Arzt (2016); Cont (2010)]. In automatic accompaniment scenarios, a third task consists of generating and synchronising the accompaniment in real time [Dannenberg (1984)].

Unlike offline alignment approaches, online methods are subject to strict causality constraints, since future observations are unavailable. Consequently, score following is particularly challenging in the presence of expressive tempo variations, performance errors, omissions, insertions, or pitch-estimation inaccuracies, all of which may lead to tracking failures and delayed recovery [Müller (2021); Orio et al. (2003)].

Online score following has played a central role in the development of mixed music repertoire, notably at Ircam (Institut de Recherche et Coordination Acoustique/Musique, Paris, France), where numerous works composed from the late 1980s and 1990s onward relied on real-time score following [Orio et al. (2003)]. Beyond concert performance, such techniques are now widely employed in educational and practice-oriented systems, such as the *Piano Companion* system, which provides real-time score tracking and feedback for piano practice [Arzt et al. (2014)].

In the remainder of this work, we focus exclusively on offline alignment scenarios (see Chapter 5), which enable reliable annotation transfer and support detailed cross-performance analysis.

Section summary – Alignment Tasks and Scenarios

Musical alignment can be classified by data modality (score–to-audio vs. audio-to-audio) and by temporal setting (offline vs. online). Offline alignment uses complete signals and global optimisation, while online methods operate under real-time constraints. This work focuses on offline alignment for analysis and annotation tasks.

3.3 Alignment Algorithms: Focus on Dynamic Time Warping

Given the alignment problem defined in the previous section, a wide range of algorithmic approaches have been proposed to compute temporal correspondences between musical representations. We first provide an overview of key alignment algorithms, and then focus more specifically on Dynamic Time Warping and its variants.

3.3.1 Overview of Alignment Algorithms

Musical alignment methods can generally be categorised into three main families: dynamic programming-based approaches, probabilistic models, and learning-based techniques. These families differ in their underlying mathematical formulations, optimisation strategies, and suitability for offline or real-time applications.

Despite these differences, they can be broadly interpreted as solving a structured optimisation problem over sequences, subject to temporal consistency constraints. In this view, the alignment task consists of estimating a path of correspondences between two temporal sequences that optimises a global or implicit objective function.

While sharing this common principle, the three families differ in how this objective is defined and how the optimal alignment is computed or approximated. In particular, dynamic programming approaches explicitly minimise an accumulated distance over an alignment grid, probabilistic models formulate alignment as inference over latent state sequences, and learning-based methods aim to learn feature representations, similarity measures, or the alignment process itself directly from data. These formulations can be applied in both offline and online settings, depending on whether future observations are available during inference.

Dynamic Programming Approaches

Dynamic Time Warping (DTW) is a classical dynamic programming algorithm that quantifies the similarity between two temporal sequences and computes an optimal alignment, even when the sequences differ in length or exhibit local timing variations.

In the context of MIR, it is particularly employed for tasks like score-to-audio and audio-to-audio alignment. DTW-based methods typically operate on frame-level feature representations, such as chroma features, spectral templates, or pitch-related descriptors. Numerous variants have been proposed to enhance computation; these are discussed in detail in Section 3.3.2.

Beyond music, DTW has been widely applied in diverse fields such as speech processing [Rabiner and Juang (1993); Sakoe and Chiba (1978)], bioinformatics [Aach and Church (2001); Keogh and Ratanamahatana (2005)], chemistry [Bork et al. (2013)], climatology and

environmental sciences [Berndt and Clifford (1994)] or economics [Rakthanmanon et al. (2012)]. These applications date back to early and foundational work in each respective domain.

Probabilistic Models

Probabilistic approaches recast musical alignment as an inference problem over latent variables, where the score position and tempo are treated as hidden states that evolve according to a stochastic model. A widely used class of models in this family is the HMM, which defines a Markovian progression of discrete hidden score states and associates observed audio features with those states via emission probabilities [Duan and Pardo (2011)]. In this framework, the optimal alignment corresponds to the most probable state sequence under the model, which can be efficiently computed using dynamic programming (e.g., the *Viterbi* algorithm [Viterbi (1967)]) given a parametric score–feature relationship.

In addition, *Conditional Random Fields* (CRFs), which model the conditional distribution of score positions given observations without requiring a generative model of the audio, have been explored as discriminative alternatives to HMMs for sequence alignment and related tasks [Joder et al. (2011)]

Probabilistic models naturally handle uncertainty in audio observations and performance variations and serve as the foundation for many online score-following systems.

Learning-Based Methods

Recent advances in machine learning have led to a growing number of learning-based alignment approaches that leverage neural networks to automatically learn both feature representations and similarity measures from data, reducing the need for handcrafted descriptors or manually defined distance functions [Agrawal et al. (2022); Simonetta et al. (2021)].

Some methods aim to *learn the alignment process itself* rather than relying on a fixed matching algorithm. These approaches may predict correspondence matrices directly or incorporate differentiable alignment objectives within the training pipeline. A prominent example is the differentiable variant of Dynamic Time Warping known as *soft-DTW*, which smooths the traditional DTW cost function to make it differentiable and therefore suitable as a training loss in gradient-based optimisation [Cuturi and Blondel (2017)]. *Soft-DTW* enables end-to-end learning of both feature extraction and alignment, making it particularly well suited to neural audio-to-score alignment and music transcription tasks [Zeitler and Müller (2026); Zeitler et al. (2023)].

Among these approaches, DTW plays a central role as it provides a direct and intuitive formulation of alignment as a constrained optimisation problem over a cost matrix. For this reason, and due to its wide use in MIR, we focus in the following sections on DTW and its variants.

3.3.2 Dynamic Time Warping (DTW)

To perform sequence alignment, audio descriptors (such as chroma features) are extracted and compared either to other audio descriptors or to a symbolic representation of the score (mainly in the form of chroma features). A temporal alignment algorithm is then used to estimate an optimal path between the two sequences by minimising an accumulated cost function.

Dynamic Time Warping

Let

$$X = (x_1, x_2, \ldots, x_N), \quad x_n \in \mathbb{R}^d, \tag{3.1}$$

and

$$Y = (y_1, y_2, \ldots, y_M), \quad y_m \in \mathbb{R}^d, \tag{3.2}$$

be two sequences of feature vectors of dimension d and possibly different lengths N and M, respectively. The goal of DTW is to find a path between these two sequences that minimises the cumulative distance under a set of constraints.

Local Cost Matrix. To compare the sequences, a local distance measure is defined between feature vectors. A common choice is the cosine distance:

$$c(x_n, y_m) = 1 - \frac{x_n \cdot y_m}{\|x_n\| \, \|y_m\|}, \tag{3.3}$$

which is robust to variations in signal energy and suitable for spectral representations commonly used in music processing.

These local distances are assembled into a cost matrix $C \in \mathbb{R}^{N \times M}$, with entries

$$C(n, m) = c(x_n, y_m), \quad 1 \leq n \leq N, \; 1 \leq m \leq M. \tag{3.4}$$

In Figure 3.3, two sequences of 12-dimensional chroma vectors corresponding to two audio recordings are compared. The sequences have different lengths (respectively $N = 147$ and $M = 180$ frames). In the cost matrix, darker regions correspond to low cost values, indicating high similarity between pairs of chroma vectors. An alignment path can then be computed to map corresponding frames between the sequences.

Warping Path. The alignment is represented by a warping path of length L:

$$W = (w_l)_{l=1}^{L}, \quad w_l = (n_l, m_l) \in [1 : N] \times [1 : M], \tag{3.5}$$

which defines a sequence of index pairs mapping frames of X to frames of Y. This path must satisfy three fundamental constraints:

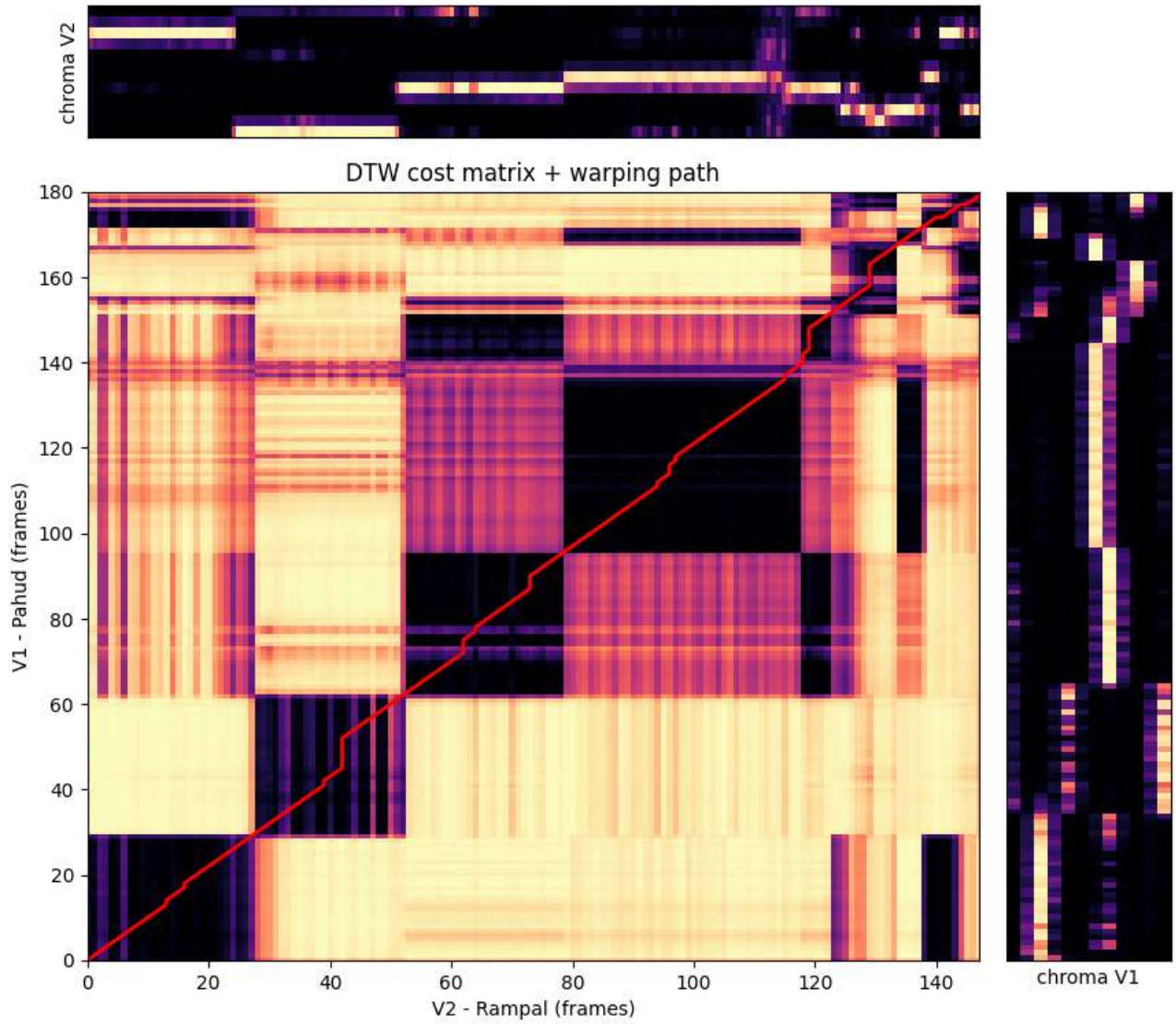


Figure 3.3: Cost matrix, chroma vectors and DTW warping path of the Intro of Flute 2, Grave movement (see Fig. 2.3), performed by J.-P. Rampal and E. Pahud on modern flutes.

- *Boundary conditions:* the first and last elements of each sequence must be aligned:

$$(n_1, m_1) = (1, 1), \quad (n_L, m_L) = (N, M). \tag{3.6}$$

- *Monotonicity:* the order of elements in each sequence must be preserved, ensuring that successive elements follow the temporal order:

$$n_{l+1} \geq n_l, \quad m_{l+1} \geq m_l. \tag{3.7}$$

- *Step size constraint:* allowed transitions are typically

$$(1,0), \quad (0,1), \quad (1,1) \tag{3.8}$$

ensuring that no elements are skipped and preventing excessive temporal jumps.

A warping path defined in this way does not yet incorporate the musical information encoded in the descriptors. Using the cost matrix, an optimal warping path can be computed.

Accumulated Cost and Objective Function. Finding an optimal warping path is a combinatorial optimisation problem under the above constraints. The objective is the cumulative cost along a path:

$$D(W) = \sum_{l=1}^{L} C(n_l, m_l) \tag{3.9}$$

The DTW computation thus corresponds to the following optimisation problem:

$$W^* = \arg\min_{W \in \mathcal{W}} D(W) \tag{3.10}$$

where $\mathcal{W}$ denotes the set of admissible warping paths.

From this definition, one can construct an $(N \times M)$ *accumulated cost matrix* D, where each entry $D(n, m)$ represents the minimal cost of a path starting at $(1, 1)$ and ending at (n, m). In particular, the final alignment cost is

$$D(N, M) = \mathrm{DTW}(X, Y) \tag{3.11}$$

This problem can be efficiently solved using dynamic programming, and is closely related to other sequence alignment problems, such as edit distance or alignment of musical partitions [Mongeau and Sankoff (1990)].

Relation to Probabilistic Formulations. While DTW is formulated as a deterministic optimisation problem over a cost matrix, it is closely related to probabilistic sequence alignment approaches such as the *Viterbi* algorithm for HMM [Rabiner (1989)].

A local distance $c(x_n, y_m)$ can be converted into a log-probability or emission log-likelihood

$$s(x_n, y_m) = \log P(x_n \mid y_m), \tag{3.12}$$

where higher values indicate a better match.

These probabilistic scores can then be used directly as entries in the cost matrix for DTW.

Recursive Formulation. The accumulated cost matrix D is computed recursively as follows:

$$\text{Initialisation:} \quad D(1,1) = C(1,1), \tag{3.13}$$

$$D(n,1) = C(n,1) + D(n-1,1), \quad n \in [2:N], \tag{3.14}$$

$$D(1,m) = C(1,m) + D(1,m-1), \quad m \in [2:M], \tag{3.15}$$

$$\text{Recursion:} \quad D(n,m) = C(n,m) + \min \begin{cases} D(n-1,m-1), \\ D(n-1,m), \\ D(n,m-1) \end{cases}, \quad n \in [2:N],\ m \in [2:M]. \tag{3.16}$$

The first cell of the matrix simply takes the value of the local cost between the first elements of the two sequences. Each cell in the first row adds its local cost to the accumulated cost of the cell immediately to its left, while each cell in the first column adds its local cost to the accumulated cost of the cell directly above. This sets up the starting edges of the matrix.

For all remaining cells, the accumulated cost is calculated by adding the local cost at that cell to the minimum of the accumulated costs from the three neighbouring cells (diagonal, above, or left). Musically, these directions correspond to different types of temporal alignment: the diagonal represents a direct alignment of notes or events between the two sequences, moving in sync; moving from above corresponds to a local prolongation or slowing down in the first sequence, while moving from the left corresponds to a local prolongation in the second sequence. This procedure ensures that the minimal-cost path is found while respecting the DTW constraints: the alignment must move forward in time in both sequences and cannot skip frames. In other words, DTW "stretches" or "compresses" time locally so that musical events correspond, and these three directions represent the fundamental elementary movements that allow this temporal flexibility.

Optimal Warping Path. Resolving this problem allows us to find one of the better values of the optimal solution, e.g. a minimal accumulative cost along an optimal warping path. In practice, the implementation can be bottom-up (iterative), resulting in $\mathcal{O}(NM)$ computational complexity both in time and space. The optimal warping path can thus be reconstructed using the accumulative cost matrix D by backtracking.

While expensive for long sequences, this formulation guarantees an exact solution under the defined constraints. Various optimisations exist to reduce complexity, such as multiscale approximations.

3.3.3 Multi-Scale and Multi-Resolution DTW

DTW is widely used in Music Information Retrieval (MIR) for audio–audio and audio–score alignment. However, its quadratic time and memory complexity, $\mathcal{O}(NM)$, becomes prohibitive for long sequences and high-resolution feature representations. To address this limitation, several strategies have been proposed, including global path constraints, weighted transition models, and multiscale processing schemes.

Multiscale Dynamic Time Warping (MsDTW)

Among early multiscale approaches, Salvador and Chan [Salvador and Chan (2004)] introduced a FastDTW framework based on progressively refining an alignment computed at coarse temporal resolutions. This paradigm was later adapted to audio synchronisation by Müller et al. [Muller et al. (2006)], who demonstrated substantial reductions in computation time while preserving alignment accuracy using a MsDTW

Weighted Transition Model. To control local warping behaviour, a weighted transition scheme is introduced through a weight vector $(w_d, w_h, w_v) \in \mathbb{R}^3$, modifying the DTW recursion as follows:

$$D(n,m) = \min \begin{cases} D(n-1,m-1) + w_d.C(n,m), \\ D(n-1,m) + w_h.C(n,m), \\ D(n,m-1) + w_v.C(n,m), \end{cases} \qquad n \in [2:N],\ m \in [2:M]. \tag{3.17}$$

The standard DTW formulation is recovered for $(w_d, w_h, w_v) = (1,1,1)$. Assigning higher costs to horizontal and vertical transitions biases the warping path towards diagonal moves, which correspond to locally linear time mappings and prevent excessive stretching or compression. In the context of music synchronisation, Müller et al. propose the weighting $(w_d, w_h, w_v) = (2, 1.5, 1.5)$, which empirically improves temporal stability and reduces pathological alignments.

Multiscale Processing. Two complementary mechanisms are employed to reduce the computational burden of DTW:

Global Path Constraints. The admissible warping paths are restricted to a predefined region $R \subset [1:N] \times [1:M]$, such as Sakoe–Chiba bands or Itakura parallelograms [Rabiner and Juang (1993)]. These constraints significantly reduce computational cost while enforcing musically plausible tempo deviations. However, overly tight constraints may degrade alignment performance in the presence of expressive timing variations, and adaptive global constraints are privileged, such as in multiscale DTW.

Coarse-to-fine Temporal Refinement. Feature sequences are recursively downsampled to obtain a multiresolution representation. At the coarsest level, DTW is applied to low-rate features, yielding a rough alignment path. This path is subsequently projected to the next finer resolution, where local refinement is performed within a narrow corridor around the projected path. This hierarchical strategy drastically reduces both time and memory complexity while maintaining high alignment accuracy.

In practice, CENS features (see Section 2.2.3) are used at coarse temporal resolutions to automatically determine the number of levels in the multiscale hierarchy and their respective time scales, by jointly optimising the temporal aggregation of chroma vectors and the downsampling factor.

Memory-Restricted Multiscale Dynamic Time Warping (MrMsDTW)

The Memory-Restricted Multiscale Dynamic Time Warping (MrMsDTW) algorithm proposed by Prätzlich et al. [Pratzlich et al. (2016)] addresses the prohibitive memory requirements of classical DTW by combining multiscale processing with a strict memory constraint.

The core idea consists of constructing a global warping path from a sequence of locally computed alignments, each restricted to a rectangular constraint region. The size of each region is bounded by a memory restriction parameter τ, which represents an upper bound on the number of cost matrix cells that can be stored and processed at any time. Consequently, the memory complexity becomes $\mathcal{O}(\tau)$, independent of the lengths of the input sequences.

Local Rectangular Constraint Regions. Instead of computing the full $N \times M$ cost matrix, MrMsDTW partitions the alignment problem into a sequence of local DTW computations. Each local alignment is carried out within a rectangular constraint region of size at most τ, centred around an estimated alignment position. Within each region, a standard DTW is applied independently, yielding a local warping path segment. These segments are then concatenated to form a continuous global alignment path.

This strategy ensures bounded memory usage while maintaining the optimality of the local alignments. However, the quality of the resulting global path strongly depends on the accuracy of the region placement, which is determined using a multiscale strategy.

Multilevel Coarse-to-Fine Strategy. MrMsDTW employs a hierarchical coarse-to-fine alignment strategy based on multiple temporal resolution levels.

At the coarsest resolution level, the input feature sequences are temporally downsampled, and a global DTW alignment is computed using a multiscale DTW framework. Due to the strongly reduced temporal resolution, this step is computationally efficient and provides a rough estimate of the global warping path. This coarse alignment captures the large-scale temporal structure of the performance, including global tempo trends and major timing deviations.

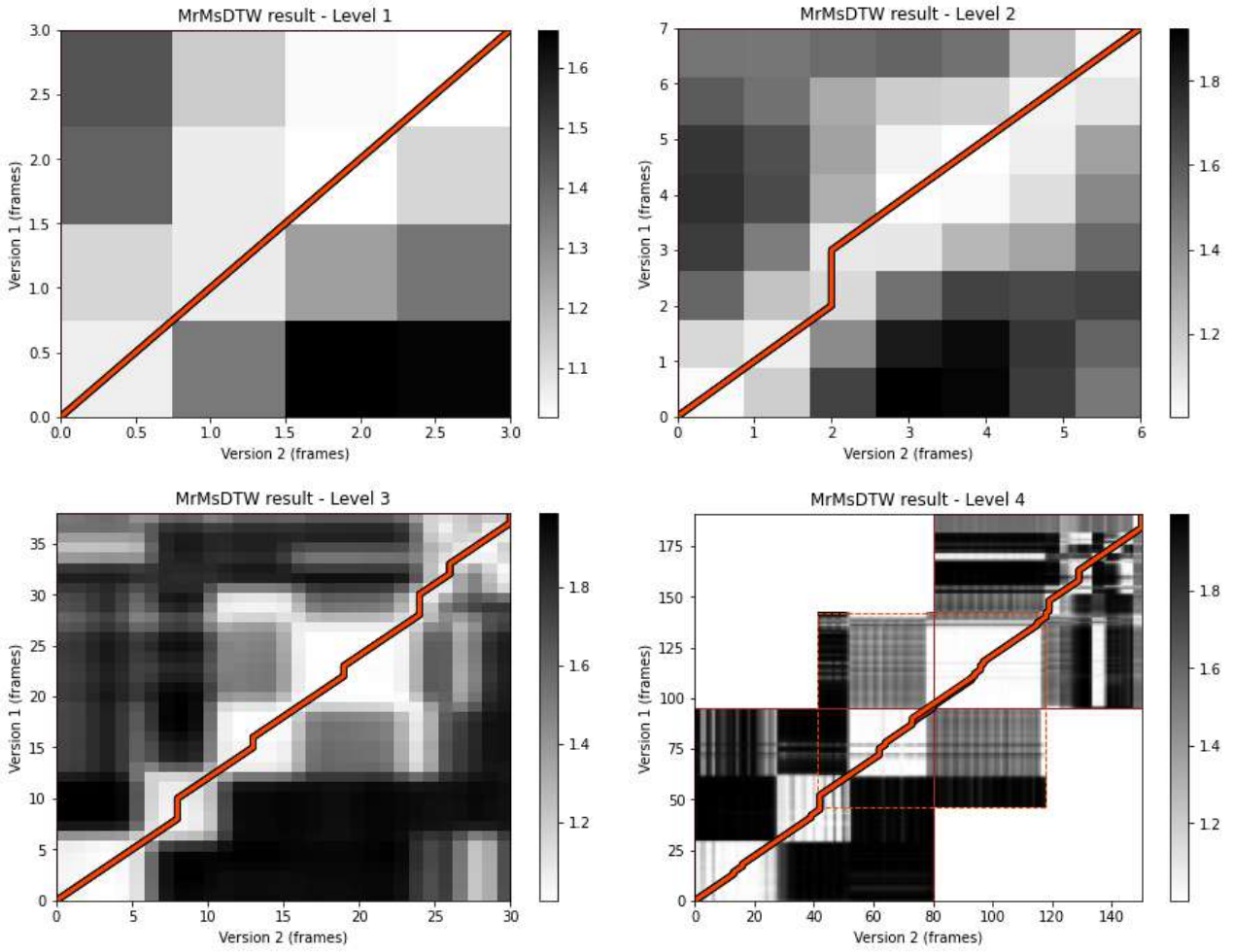


Figure 3.4: Multilevel refinement of the warping path corresponding to Fig. 3.3 using the MrMsDTW algorithm. From coarse (top-left) to fine resolution (bottom-right), the alignment corridor progressively narrows, enabling accurate local synchronisation under strict memory constraints. In our experiments, we used four resolution levels with smoothing windows [201, 101, 21, 1] and downsampling factors [50, 25, 5, 1]. Step weights were set to [1.5, 1.5, 2.0] to favour diagonal transitions, and the memory threshold was set to 10^6. Visualisations are produced by *SynctoolBox* (Section 3.4.2): white regions are absent at intermediate levels due to full cost-matrix coverage, while finer levels use locally constrained regions defined by neighbouring anchors, leading to partial coverage of the alignment cost matrix.

The estimated warping path is then iteratively projected onto finer temporal resolutions. At each refinement level, the projected path defines a sequence of local rectangular constraint

regions of bounded size τ. Within each region, an independent standard DTW is computed, yielding locally accurate alignments while strictly respecting the memory constraint. The concatenation of the resulting local paths produces a continuous global alignment at the original temporal resolution.

Computational Complexity. While classical DTW exhibits quadratic time and memory complexity $\mathcal{O}(NM)$, and multiscale DTW reduces the time complexity to near-linear behaviour, MrMsDTW achieves a constant memory complexity $\mathcal{O}(\tau)$, independent of the sequence lengths. The time complexity becomes linear with respect to the input lengths, making the approach particularly suitable for long audio recordings and large-scale music synchronisation tasks.

Beyond algorithmic developments, musical alignment research depends on practical tools and resources that facilitate reproducibility and reliable evaluation.

The following sections review some frameworks available for reproducible research (Section 3.4), discuss the creation of high-quality annotations (Section 3.5), and present some of the datasets most commonly used in the field (Section 3.6).

Section summary – Alignment Algorithms: Focus on Dynamic Time Warping

Musical alignment algorithms can be classified into three main families: dynamic programming (e.g., DTW), probabilistic models (e.g., HMMs, CRFs), and learning-based approaches (e.g., neural networks, soft-DTW). DTW computes an optimal temporal mapping between sequences of musical features, handling local timing variations and differing sequence lengths. Multiscale variants, such as MsDTW and MrMsDTW, reduce computational and memory costs by combining coarse-to-fine alignment with constrained local regions, enabling efficient and accurate alignment for long audio recordings.

3.4 Frameworks

A variety of software frameworks have been proposed to support audio alignment in MIR, ranging from lightweight research libraries to full synchronisation pipelines, offering reusable implementations of state-of-the-art algorithms. In the context of this thesis, we focus on a subset of existing offline frameworks and place particular emphasis on *Synctoolbox*, which is used as the main alignment tool in our experiments (Chapter 5).

3.4.1 Overview of Existing Frameworks for Offline Alignments

Rather than providing an exhaustive survey, we first present some representative frameworks used for offline alignment.

MATCH

Presented in 2005 and developed in Java, *Music Alignment Tool Chest* (MATCH) [Dixon and Widmer (2005)] is a toolkit for aligning multiple audio recordings of the same musical piece. The system relies on low-level audio features (positive spectral difference vectors) and an incremental DTW [Dixon (2005)].

Beyond alignment, MATCH provides an interactive user interface allowing users to navigate, mark, and compare corresponding musical positions across different versions of the same piece. The system also supports visualisation of alignment results, including cost matrices and forward/backward warping paths, which can be displayed and updated during playback for real-time inspection of the alignment process.

pyAMPACT

Python-based Automatic Music Performance Analysis and Comparison Toolkit (pyAMPACT) is a framework for score–audio alignment and expressive performance analysis [Devaney et al. (2026)], derived from the original MATLAB *AMPACT* toolkit [Devaney et al. (2012)]. It uses DTW and extracts frame-level acoustic descriptors including fundamental frequency, spectral power, harmonic content, and spectral centroid.

These low-level features are then aggregated into note-level descriptors of pitch, dynamics, and timbre, enabling higher-level analysis of expressive performance parameters. Interestingly pyAMPACT can integrate a diversity of annotations formats like Humdrum [Huron (2002)] or Dezrann [Ballester et al. (2025)].

Synctoolbox

[Müller et al. (2021)] is a modular Python toolkit that implements state-of-the-art audio synchronisation pipelines, including multiscale DTW variants and memory-efficient algorithms. It supports feature extraction (chroma, onset-related features, DLNCO), tuning correction, transposition compensation, and multiresolution alignment. The workflow is described in the following (Section 3.4.2).

3.4.2 Focus on Synctoolbox

As we use *Synctoolbox* in several of our experiments (see Sections 5.2.2 and 5.3.1), here, we describe the synchronisation pipeline implementation (see https://meinardmueller.github.io/synctoolbox/build/html/index.html for complete documentation).

The complete synchronisation pipeline is implemented in Python. By default, audio signals are first resampled to a sampling rate $F_s = 22\,050$ Hz and analysed using a STFT with a frame size of $N = 2\,048$ samples and a hop size of $H = 1\,024$ samples, yielding a feature frame rate of approximately 21.5 Hz. All subsequent feature representations are computed at this temporal resolution.

To compensate for possible tuning deviations between recordings, a global tuning offset is estimated independently for each audio signal using a pitch-based tuning estimation method. This tuning correction is applied prior to chroma computation, ensuring pitch consistency across recordings.

Pitch-based chroma features are then extracted using a logarithmic frequency mapping followed by chroma quantisation. In parallel, onset-related pitch features are computed and transformed into DLNCO representations (see Section 2.2.3), which capture note onset information. These two complementary feature streams—quantised chroma and DLNCO—are subsequently combined during the alignment process.

To handle potential global pitch transpositions between performances, an optimal chroma shift is estimated. This is achieved by computing low-resolution CENS features (see Section 2.2.3) at 1 Hz and determining the chroma shift that minimises the DTW alignment cost between the two sequences. The resulting pitch shift is applied to both chroma and onset features of the second recording, ensuring consistent tonal alignment.

Finally, temporal alignment is performed using the MrMsDTW algorithm (see Section 3.3.3), which integrates both chroma and onset features within a memory-restricted multiscale DTW framework. The step weights by default $(1.5, 1.5, 2.0)$ are employed to favour diagonal transitions. A memory threshold $\tau = 10^6$ is used to bound the size of the local cost matrices. Multiscale processing is achieved through successive smoothing and downsampling stages, while local constraint regions are recursively refined from coarser alignment estimates in order to reduce computational complexity. Chroma features are ℓ^2-normalised after each smoothing and downsampling stage, and on the finest scale, chroma- and onset-based cost matrices are

combined according to

$$C = \alpha C_{\text{chroma}} + (1 - \alpha) C_{\text{onset}}$$

where $\alpha = 0.5$. The output of the algorithm is a high-resolution warping path that provides frame-accurate temporal correspondence between the two recordings.

However, the evaluation and comparison of such methods depend on reliable reference data. We therefore now discuss the role of annotations and datasets in musical alignment.

Section summary – Frameworks

Software frameworks such *Synctoolbox* provide reusable implementations of alignment algorithms, ensuring reproducibility and methodological comparability across studies.

3.5 Annotations

Beyond offline alignment frameworks, the study of performance and their cross comparison depend critically on high-quality annotations.

Annotations may describe various musical aspects, including temporal events (e.g., measure positions, beats, or even note onset times) [Wang et al. (2014)], structural markers (sections, phrases) [Pratzlich and Muller (2013)], and higher-level musical descriptors (chords, expressive timing deviations) [Konz et al. (2013); Liem and Hanjalic (2011)]. In this section, we will describe the process of annotation transfer and particularly the triple transfer strategy.

3.5.1 Annotations Transfer

Score-to-audio alignment is generally less accurate than manual annotation [Weiß and Pratzlich (2016)]. Consequently, a central application of audio-to-audio musical alignment is *annotation transfer*, which involves propagating time-aligned symbolic or analytical information from a reference performance to one or more target performances via an alignment function. In other words, *manual annotations* produced by human annotators for a given performance can be synchronised to other performances, yielding *computed annotations* [Weiß and Pratzlich (2016)].

Software tools [Cannam et al. (2010)] and multimodal platforms [Ballester et al. (2025)] capable of handling multiple annotation encoding formats can assist in the production of such musical annotations. Nevertheless, manual annotation remains time-consuming, costly, and often subjective [Koops et al. (2019)]. These limitations hinder the construction of large, high-quality annotated datasets. Annotation transfer, therefore provides a scalable and reproducible alternative, enabling the rapid propagation of annotations across extensive corpora once a small set of reference performances has been carefully annotated.

However, the annotation transfer process is inherently sensitive to alignment errors. Moreover, interpretative variability, ornamentation, articulation differences, and local timbral changes introduce additional uncertainties that further complicate reliable propagation. While several methods have been proposed to improve alignment accuracy, achieving robust and precise annotation transfer across diverse performances remains a challenging problem.

3.5.2 Triple Transfer and Late-Fusion

Triple Transfer

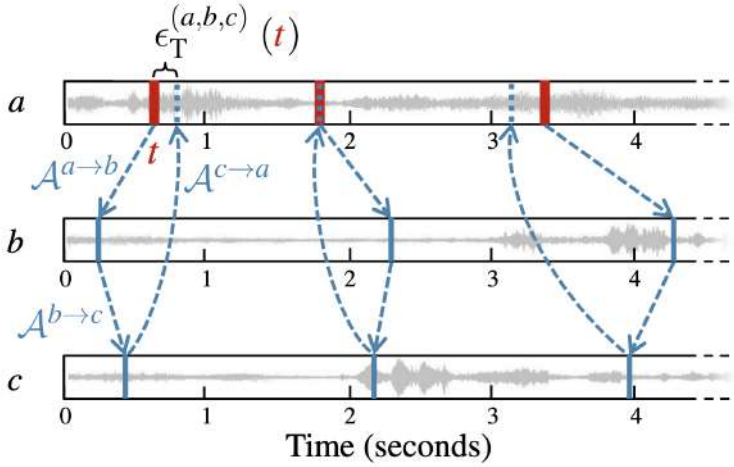


Figure 3.5: Triple transfer principle as proposed by Zalkow et al. (schema from [Zalkow et al. (2017)]).

To improve robustness and quantify transfer reliability, multi-version and triple-transfer strategies have been proposed [Pratzlich and Muller (2016); Zalkow et al. (2017)] providing a principled framework for assessing alignment reliability and propagating annotations across large corpora.

Instead of a simple single transfer, starting from a reference performance a, the measure positions (the onset of the downbeats) are first transferred to a target performance b, then to a third and distinct unannotated version c_j, before being transferred back to the original reference (Fig. 3.5):

$$\mathcal{A}^{c\to a} \circ \mathcal{A}^{b\to c} \circ \mathcal{A}^{a\to b}$$

While direct transfers already provide plausible onset estimates in the target performances, individual alignments may suffer from local mismatches or performance-specific artifacts. Using

multiple other versions of c, these transfer paths serve to assess the self-consistency of the direct transfer $a \to b$ by exploiting redundancy across multiple performances. Indeed, for a given onset time o_n occurring in the reference performance at time t_n , a deviation called *triple error* e_T can be quantified as the absolute difference between the original onset time in a: t_n^a and the one obtained after the triple transfer process $t_n^{(a,b,c)}$:

$$e_T^{(a,b,c)}(o_n) = \left| t_n^{(a,b,c)} - t_n^a \right| \tag{3.18}$$

For each onset o_n, target performance b and reference performance a, the reliability of the direct transfer across all intermediate versions c are summarise using the median triple error:

$$\epsilon_n^{(b,a)} = \text{median}_j \left(e_T^{(a,b,c_j)}(o_n) \right) \tag{3.19}$$

This quantity serves as a confidence measure for the onset estimate obtained from the direct transfer $a \to b$ but is not sufficient to prove the accuracy of the transfer. It is, in fact, used for choosing between multiple reference performances.

Late-Fusion

If we have more than one reference performance, the triple error can be used to choose between two sources of annotations during the transfer. Since two independent reference performances are available, the late-fusion strategy selects, for each onset, the reference performance that yields the smallest median triple error, i.e., the most self-consistent transfer.

The final fused onset is obtained by selecting the direct alignment from the best reference:

$$a_k^* = \underset{a_k \in \{R_1, R_2\}}{\arg\min}\ \epsilon_n^{(b,a_k)}, \qquad \widehat{t}_n^{\,(b)} = t_n^{(b,a_k^*)} \tag{3.20}$$

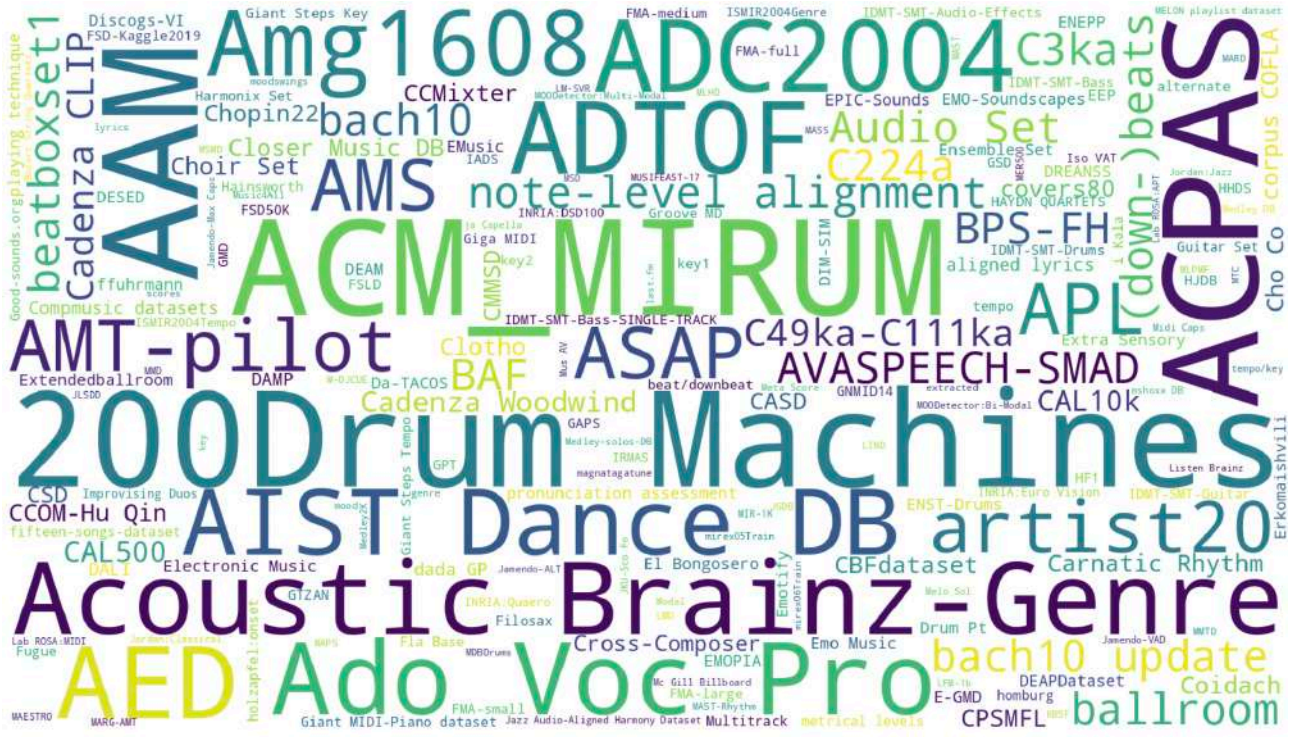
When applied to two reference performances, this strategy resulted in improved accuracy for annotation transfer. We will investigate and improve this strategy in Section 5.2.5.

To produce high-quality annotation transfers that will be useful, the quality and availability of open datasets (annotations, scores, audios) are crucial. This is one of the major research questions in the MIR community, as will be explained in the following section.

Section summary – Annotations

Beyond implementation aspects, annotations have a central role. Particularly in the context of annotation transfer, it enables the scalable propagation of musical information. This propagation remains sensitive to alignment accuracy and performance variability.

3.6 Datasets


Amg1608
ADC2004
AAM
ACM MIRUM
ADTOF
AMS
note-level alignment
Audio Set
C224a
BPS-FH
APL
ACPAS
AMT-pilot
ASAP
C49ka-C111ka
AVASPEECH-SMAD
200Drum Machines
AIST Dance DB
artist20
Acoustic Brainz-Genre
AED
Ado
Voc
Pro
bach10
bach10 update
ballroom
Cadenza Woodwind
pronunciation assessment
CCOM-Hu Qin
CAL500
Electronic Music
CBFdataset
Carnatic Rhythm
Cross-Composer
Emo Music
Filosax
Giant MIDI-Piano dataset
Multitrack
CPSMFL
Choir Set
Chopin22
CCMixter
EPIC-Sounds
EMO-Soundscapes
Closer
Music DB
Giga MIDI
IDMT-SMT-Drums
aligned lyrics
IDMT-SMT-Bass-SINGLE-TRACK
Extendedballroom
DAMP
Da-TACOS
BAF
Clotho
CASD
CAL10k
beat/downbeat
GAPS
IRMAS
Improvising Duos
Drum Pt
El Bongosero
EMOPIA
DEAPDataset
E-GMD
homburg
Coidach
Erkomaishvili
Listen Brainz
ENST-Drums
Guitar Set
Extra Sensory
tempo
Compmusic datasets
Discogs-VI
FSD-Kaggle2019
Giant Steps Key
Harmonix Set
beatboxset1
Cadenza
CLIP
DESED
FSD50K
EMusic
IADS
INRIA:DSD100
Groove MD
DEAM
FSLD
CMMSD
key1
key2
DIM-SIM
ENEPP
C3ka
IDMT-SMT-Bass
EEP
alternate
Ensemble Set
GSD
Iso VAT
DREANSS
covers80
HAYDN QUARTETS
WHDS
corpus COFLA
(down-)beats
Cho Co
HJDB
aligned lyrics
tempo
Da-TACOS
Woodwind
GNMID14
extracted
Midi Caps
tempo/key
Giant Steps Tempo
Medley-solos-DB
magnatagatune
HF1
IDMT-SMT-Guitar
MIR-1K
mirex05Train
Filosax
Fla Base
MDBDrums
Jamendo-VAD
Mc Gill Billboard
FMA-Large
FMA-small
MAST-Rhythm
Jazz Audio-Aligned Harmony Dataset
metrical levels
Fugue
Lab ROSA:MIDI
MAESTRO
MARG-AMT
holzapfel2012
GTZAN
dada GP
DALI
fifteen-songs-dataset
Electronic Music
Filosax
Emotify
Melo Sol
Drum Pt
Coidach
ffuhrmann
scores
Hainsworth
Jamendo-Max Caps
GMD
Good-sounds.org
playing technique
lyrics
woodwinds
LM-SVR
MUSIFEAST-17
MASS
MERSH
DREANSS
Iso VAT
Jordan:Jazz
WHDS
Medley DB
Kala
Lab ROSA:MIDI
WLPMF
MTC
Midi Caps
tempo/key
mshoxx DB
Listen Brainz
HF1
IDMT-SMT-Guitar
Erkomaishvili
MARD
MELON playlist dataset
IDMT-SMT-Audio-Effects
FMA-medium
ISMIR2004Genre
FMA-full
MOODetector:Multi-Modal
MOODetector:Bi-Modal
HAYDN QUARTETS
Cho Co
corpus COFLA
(down-)beats
alternate
Ensemble Set
Iso VAT
DREANSS
covers80
WHDS
Guitar Set
Extra Sensory
tempo/key
Listen Brainz
HF1
IDMT-SMT-Guitar
Erkomaishvili
ENST-Drums
INRIA:Euro Vision
Carnatic Rhythm
CBFdataset
Drum Pt
Coidach
DEAPDataset
E-GMD
homburg
CPSMFL
ballroom
bach10 update
Jamendo-VAD
MATD
Emo Music
Cross-Composer
EMOPIA
Mc Gill Billboard
FMA-Large
FMA-small
MAST-Rhythm
Jazz Audio-Aligned Harmony Dataset
metrical levels
Multitrack
Giant MIDI-Piano dataset
Fugue
Lab ROSA:MIDI
MAESTRO
MARG-AMT
holzapfel2012
GTZAN
dada GP
DALI
fifteen-songs-dataset
Electronic Music
CAL500
CCOM-Hu Qin
CSD
Improvising Duos
pronunciation assessment
Cadenza Woodwind
CASD
CAL10k
beat/downbeat
GAPS
IRMAS
magnatagatune
Medley-solos-DB
Giant Steps Tempo
GPT
Clotho
BAF
Da-TACOS
DAMP
Extendedballroom
AMT-pilot
ASAP
C49ka-C111ka
AVASPEECH-SMAD
APL
BPS-FH
IDMT-SMT-Drums
aligned lyrics
IDMT-SMT-Bass-SINGLE-TRACK
Giga MIDI
Groove MD
INRIA:DSD100
MUSIFEAST-17
LM-SVR
MOODetector:Multi-Modal
EPIC-Sounds
EMO-Soundscapes
EEP
IDMT-SMT-Bass
ENEPP
C3ka
FMA-full
ISMIR2004Genre
FMA-medium
IDMT-SMT-Audio-Effects
Choir Set
Chopin22
CCMixter
Harmonix Set
Giant Steps Key
Discogs-VI
FSD-Kaggle2019
beatboxset1
Cadenza
CLIP
DESED
FSD50K
EMusic
IADS
Closer
Music DB

Existing datasets vary widely along several dimensions: musical texture (isolated notes, monophonic, polyphonic), instrumentation (piano, orchestral, mixed ensembles), recording format (audio only vs. multitrack), and annotation granularity (note-level, beat-level, or higher-level structural annotations). This diversity supports a broad range of MIR tasks, but also introduces strong biases in the evaluation of alignment methods.

In particular, alignment research has largely focused on piano music, benefiting from its well-defined percussive onsets and relatively stable timbre. Datasets such as *(note-)Aligned Scores And Performances* (ASAP) [Foscarin et al. (2020)], *Beethoven Piano Sonata Dataset* (BPSD) [Zeitler et al. (2024b)], and *MIDI and Audio Edited for Synchronous TRacks and Organization* (MAESTRO) [Hawthorne et al. (2019)] provide high-quality beat- or note-level alignments for piano performances and have become standard benchmarks.

Beyond piano, several datasets include a wider range of instruments. For instance, *Schubert Winterreise Dataset* (SWD) [Weiß et al. (2021a)], *TRIOS* [Fritsch and Plumbley (2013)], and *MusicNet* [Thickstun et al. (2017)] incorporate woodwind instruments. However, these datasets typically focus on ensemble or mixture scenarios, where multiple instruments are present simultaneously. As a result, they are less suited for evaluating alignment methods in controlled monophonic conditions.

Some datasets provide isolated recordings within polyphonic contexts. For example, *Bach10* [Zhiyao Duan et al. (2010)] offers both individual instrument tracks and corresponding mixtures, enabling controlled experiments on source separation and multi-pitch estimation. Nevertheless, its primary design remains oriented toward polyphonic music rather than expressive solo performance.

In contrast, datasets specifically dedicated to solo woodwind instruments remain scarce. *ChoraleBricks* [Balke et al. (2025)] consists of modular multitrack recordings of 10 chorales from the German sacred repertoire (including Bach, Telemann, and Vulpius), covering four-part SATB writing. It includes isolated instrumental recordings from a diverse set of wind instruments (woodwinds and brass), which can be freely combined. Nevertheless, its chorale-based, ensemble-oriented structure limits its applicability to expressive solo alignment tasks.

Overall, despite the growing number of datasets, there remains a significant imbalance across instruments. In particular, datasets providing high-quality, time-aligned annotations for solo woodwind performances—especially flute—are extremely limited. This lack of dedicated resources poses a major challenge for the development and evaluation of alignment methods in these contexts, motivating the need for new datasets tailored to monophonic wind performances.

In the following Chapter 4, we present one of the main contributions of this thesis, namely two new corpora dedicated to solo flute performance, designed to address the lack of high-quality, time-aligned datasets for expressive monophonic woodwind analysis.

Section summary – Datasets

Datasets play a central role as shared benchmarks for MIR evaluation, but their distribution across instruments and musical contexts remains highly uneven. In particular, solo woodwind performance data is still underrepresented, limiting progress in alignment methods for these instruments.

Chapter summary – Musical Alignment: Tasks, Methods and Tools

This chapter presents the main concepts and resources underlying musical alignment. It introduced alignment modalities and application scenarios, the general alignment framework, representative software tools, annotation transfer strategies, and the role of annotated datasets as benchmarks. Together, these elements provide the methodological foundation for the alignment approaches developed in the remainder of this thesis, while highlighting the need for dedicated resources for expressive solo flute performance.

To the question *"How can heterogeneous performances be reliably synchronised at a fine-grained musical level?"*, the answer is by combining appropriate alignment models, musical representations, and reliable annotated resources. While existing methods provide effective solutions for many alignment scenarios, their reliability remains dependent on the quality of the representations, annotations, and datasets available, particularly for highly expressive solo performances where timing variations in interpretations are substantial.

Part II

Contributions : Data, Methods and Performance Analysis in Solo Flute Music

Chapter 4

Telemann, Debussy, Varèse: Two Solo Flute Datasets

Contents of this chapter

4.1 Fantasias Dataset 82
4.1.1 Musical Annotations 85
4.1.2 Recordings and Unfolded Scores 91
4.2 Early 20th Dataset 94
4.2.1 Scores and Recordings 96
4.2.2 Onset Annotations 96
4.2.3 Musical Patterns 98
4.3 Data Availability 101
4.3.1 Dataset Fantasias 102
4.3.2 Dataset Early 20th 102
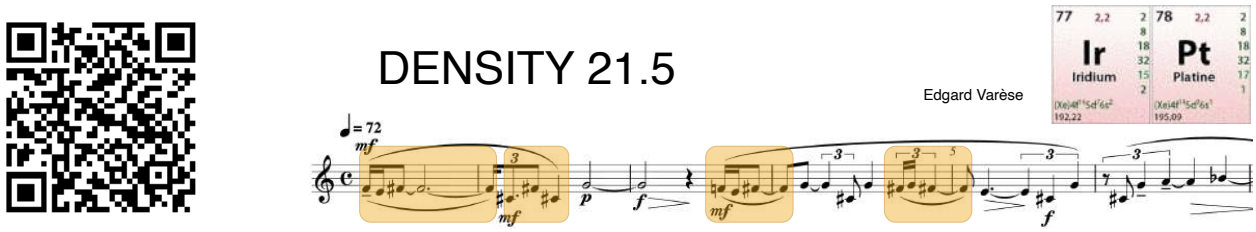


Figure 4.1: Density 21.5 is a solo flute piece by Edgard Varèse, composed in 1936 and revised in 1946. It was written for Georges Barrère's platinum flute made of a 90/10 platinum–iridium alloy with a similar density (∼21.6 g/cm³). Listen to Claire Chase's performance (2019, Cambridge, MA) at https://patthmus.github.io/phd-manuscript-links/chap4.html.

Symbolic annotations and aligned audio recordings constitute essential resources in MIR. As discussed in 3.6, datasets focusing on solo flute repertoire are especially under-represented. While symbolic scores provide a foundational representation of musical content, they remain limited in their ability to capture interpretative information.

How can higher-level musical structures that are not explicitly encoded in the score be systematically represented through enriched annotation layers in monophonic music corpora?

Figure 4.1 illustrates this issue through *Density 21.5*, a solo flute work by Edgard Varèse, whose writing is characterised by recurring motivic cells transformed through rhythm, register, and articulation. Although these relationships strongly contribute to the perception of musical structure, they are not explicitly encoded in the score and therefore require additional symbolic annotations to be formally described and analysed alongside recorded performances.

This chapter presents the construction of two corpora dedicated to the solo flute repertoire, with a particular focus on the *Fantasias* by Georg Philipp Telemann. We briefly complement this with an additional corpus drawn from early 20th-century repertoire characterised by reduced metric structure.

Research outputs associated with this chapter

Towards detecting interleaved voices in Telemann Flute Fantasias

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In DMRN+ 19: Digital Music Research Network One-day Workshop, 2024

When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In 26th International Society for Music Information Retrieval Conference (ISMIR), 2025

Soft Late-Fusion Alignment and Pattern-Based Performance Analysis of Early 20th-Century Solo Flute Music

Patrice Thibaud, Louis Bigo, Mathieu Giraud, Yann Teytaut

Accepted to ISMIR 2026

4.1 Fantasias Dataset

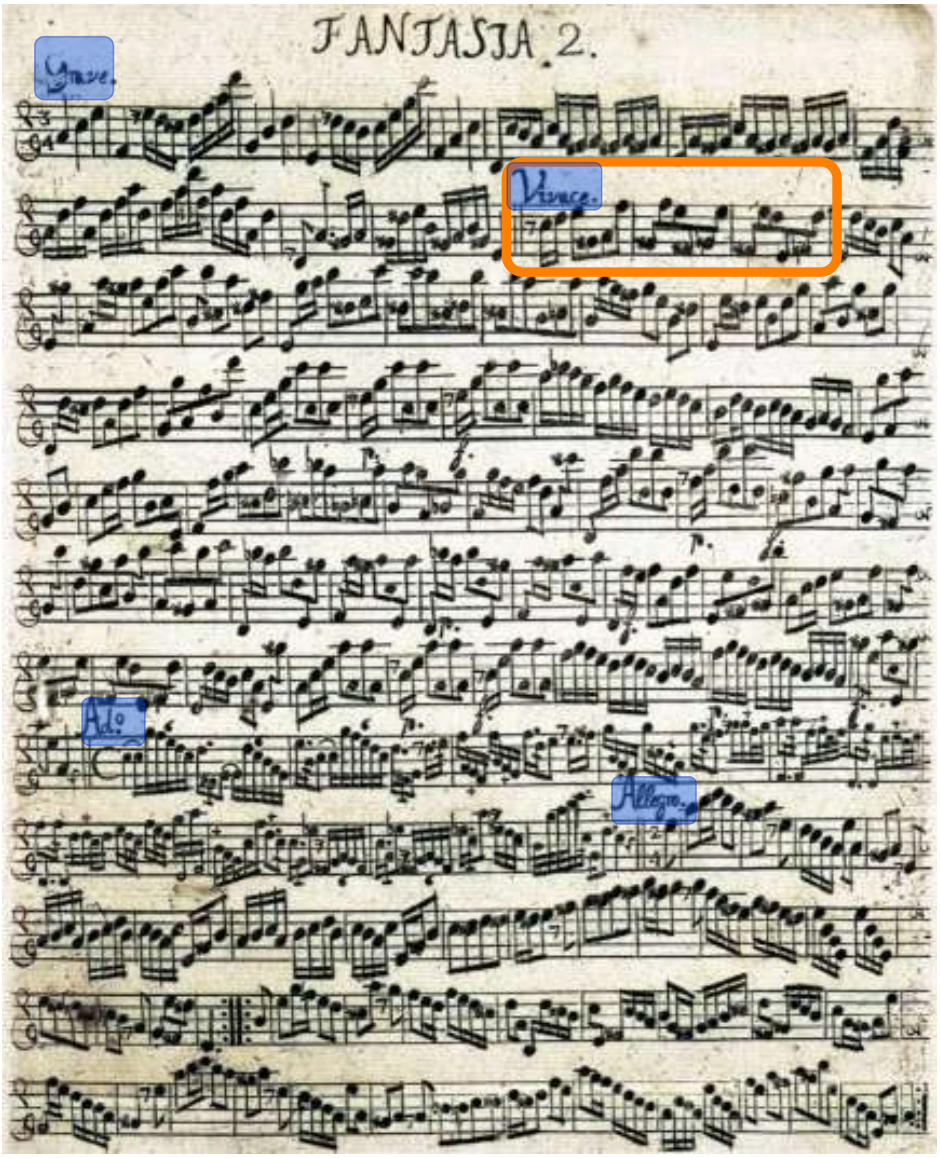


Figure 4.2: Autograph score of Fantasia No. 2 (TWV 40:3) by G.P Telemann. Four Sections are outlined corresponding to different movements. The boxed excerpt is shown in Fig 4.3.



Figure 4.3: In the *Vivace* of Fantasia No. 2, some of the eighth notes in the same beamed group have stems down while the others have stems up. Telemann uses these typographic changes in direction to emphasise the interleaved lines.

As outlined in the introduction of this thesis (Section 1.3), the *twelve fantasias for solo flute* by Georg Philipp Telemann, edited by himself in Hamburg, probably by 1722-1723, are one of the rare "significant works for unaccompanied flute before the twentieth century" [Zohn (2008)]. Telemann printed these fantasias himself in Hambourg and nowadays, the only authenticated copy of the fantasias is preserved at the Royal Library of Brussel (T.5823).

These works are situated within the broader Baroque period, a stylistic era of Western classical music that followed the Renaissance and preceded the Classical period. Baroque music is closely associated with the gradual formalisation of tonality and witnessed important developments in compositional and performance practice. Improvisation remained an essential skill for performers, particularly in the realisation of ornaments and stylistic embellishments.

Within this stylistic context, Telemann's writing frequently combines contrapuntal procedures with emerging tonal thinking. In the *twelve fantasias*, this results in passages where single-line writing nonetheless suggests multiple simultaneous voices. Such passages feature implied polyphony through interleaving of two or more melodic lines, what Piston describes as a "compound melody" [Piston (1947)].

Such voices [Silva and Carlos (2012)] may be notated on the Telemann autograph by a beamed group of notes where some stems point downward and others upward (Fig. 4.3).

This section is organised as follows: we first describe the methodology used for engraving the scores, collecting and encoding the collected musical annotations. Then, we present selected audio performances that constitute our dataset. The methodology chosen for aligning the recordings to the score is developed in Chapter 5.

Source	Movement names	Fugatos description	Voices annotations	Harmony
Telemann [G.P. Telemann (1732)]	only tempi	–	stems indications	–
Beyer [Marion Beyer, Rachel Brown (2013)]	partial	–	–	–
Brown [Brown (2015)]	partial	6	6 (fugato) ▲	–
Byrd [Byrd and Crawford (2002)]	–	–	7 ●	–
DeBree [de Bree (2020)]	–	1, 6	7, 9–12 ●	CP (1,6) ●
Eppinger [Eppinger, S. (1984)]	all	partial	4–5, 7, 10–12 ●	–
Hunt [Hunt (2016)]	4–7	4–7	7 ●	–
Kuijken [Barthol Kuijken (1987)]	partial	–	–	–
Min [Kyoung Joo Min (1998)]	all	all	all fugatos ▲	LK (all) ■ CP (all) ●
Porter [Porter (2008)]	all	5–6	–	–
Silva [Silva and Carlos (2012)]	1–3	1–2	1–3 ■	LK (all) ■ CP (1–3) ■
Zohn [Zohn (2008)]	all	3	–	–

Table 4.1: The musicological sources collected in this dataset analyse some or all of the Telemann's *twelve fantasias* referred by their number. For the movements names some sources propose either full identification by fantasia or just sparse information (partial). For the fugatos movements, in the same way, the fugato can be just identified (partial) or fully described. Concerning voices annotations only three fantasias are exhaustively analysed ■, in the other cases either voices are fully identified on specific movements ▲ or are given in a sparse way ● . Harmony annotations include local keys *(LK)* and chord progressions *(CP)*.

4.1.1 Musical Annotations

Engraving the Scores

We built the corpus from *MusicXML* files transcribed by Árpád Zoltán Szabó[1] and available in Creative Commons Attribution-ShareAlike 4.0 Licence on International Music Score Library Project (IMSLP). We amended these transcriptions by carefully checking the only authenticated copy of the fantasias [G.P. Telemann (1732)].

Each fantasia is a free-form composition written in a different key, with a sequence of themes evoking improvisation. Telemann's music reflects the mixed taste, "a blending of the French, Italian, English, and Polish national styles" [Zohn (2008)], the fantasias reflect these influences.

Collecting upon 12 musicological sources (Tab. 4.1), we gathered annotations about movement names, implied chords, local keys, and voices to produce consistent annotations.

Movements Identification

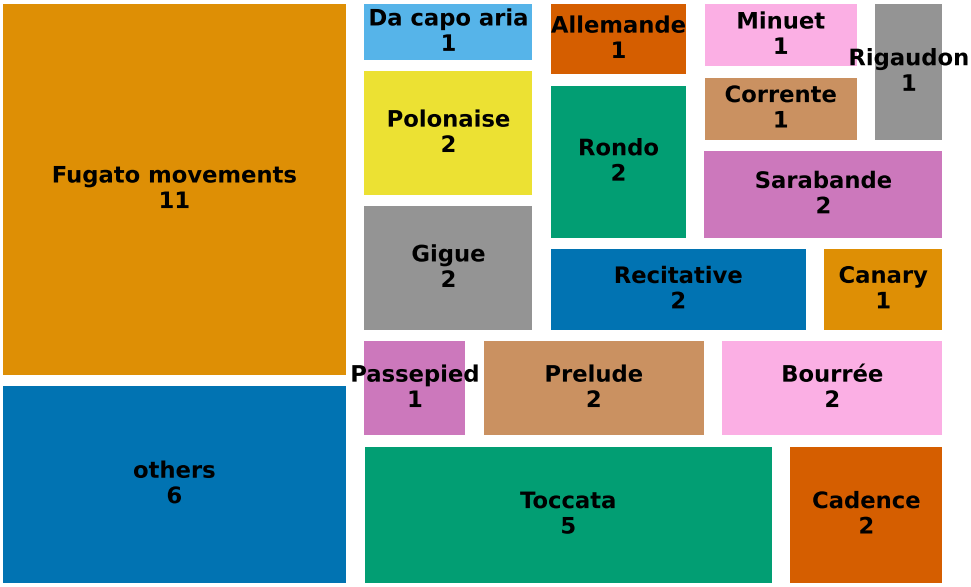


Figure 4.4: Treemaps of the repartition of the movements in the fantasias

[1] https://imslp.org/wiki/Special:ReverseLookup/236786

Each fantasia is divided into at least three short movements and up to seven, with the final one always being a dance. The division and naming of these sections were not provided by Telemann but are derived from consensus or majority agreement among the consulted sources. Across the 12 fantasias, a total of 45 short movements can be identified (Fig. 4.4).

In labelling the movements, we rely primarily on the terminology provided by the sources and on the consensus found in the musicological literature.

In some cases, however, the identification of a specific dance type remains debated. This is particularly the case in the last movement of Fantasia No. 8 in *E minor* (TWV 40:9) in which Telemann gives only the tempo indication *Allegro*, but scholars have proposed different implicit dance types. While Zohn and Eppinger suggest a *Passepied* because of the lively triple meter, others such as Brown and Porter point to features associated with a *Polonaise*. In our analysis, we adopt the latter interpretation, as the movement lacks the hemiola patterns typically associated with the *Passepied* and instead emphasises recurring rhythmic figures resembling the characteristic *Polonaise* pattern, which are further reinforced through sequential repetition in the second section.

Beyond the naming of individual movements, several sources also point to higher-level formal archetypes underlying certain fantasias. For instance, some pieces follow patterns reminiscent of the *Sonata da Chiesa*, traditionally organised as slow–fast–slow–fast as in Fantasia No. 2. Other fantasias evoke stylistic models such as the *French overture* (Fantasia No. 7) associated with Jean-Baptiste Lully. When such broader formal interpretations are explicitly mentioned in the sources and consistent across them, they have also been included alongside the movement-level annotations. In any case, the choices can be discussed et are not an absolute truth.

Focus on Fugatos Movements

Several studies describe certain movements of Telemann's Fantasias as exhibiting a fugato character (Fig. 4.5). Among the 11 movements identified as fugato (Fig. 4.4), five are *Vivace*, three are *Allegro*, two are *Gigue*, and one is *Presto*. This subsection details the specific characteristics of this type of movement and the corresponding annotations.

What is a Fugue ? In the Western contrapuntal tradition, a fugue is commonly described as a polyphonic form based on the systematic imitation of a principal theme (the *subject*), first stated in a single voice and then successively answered by other voices, typically alternating between tonic and dominant entries, and developed through a combination of episodic passages and additional thematic statements [Fux et al. (1971); Giraud et al. (2015)].

Depending on the analytical perspective, fugatos sections may be interpreted across our sources as *fully structured fugues*, *double fugues*, or more loosely as passages employing *fugal imitation*. The distinction is important because many of these movements evoke fugal procedures without necessarily fulfilling all the formal criteria of a complete fugue.

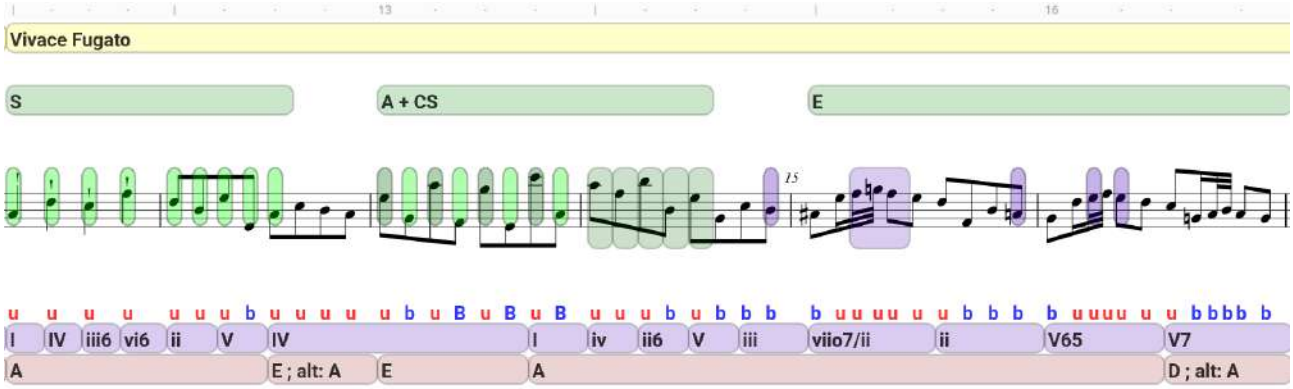


Figure 4.5: The *Vivace fugato* from Fantasia No. 1 (TWV 40:2), mm. 11-16 with harmonic labels and identification of the fugato form available on Dezrann. After exposition of a first subject (`S`), the answer (`A`) is combined with the counter-subject (`CS`). Later an episode (`E`) develops the preceding material through a sequence of harmonic progressions.

Fugatos Sections in the Fantasias Different types of fugato movements can be identified across the sources. In some cases, the labels (subject, answer, countersubject, and episode) reflect an analytical framework derived from later pedagogical traditions of fugue theory [Giraud et al. (2015)], as found in treatises by François-Joseph Fétis [Fétis (1825)] or André Gedalge [Gedalge (1901)], and should therefore be applied cautiously to early eighteenth-century repertories. Nevertheless, these terms provide a convenient framework for describing recurrent contrapuntal patterns. For example, the first movement (*Vivace*) of Fantasia No. 1 (Fig. 4.5) displays a recognisable exposition in which the theme called subject `S` is stated in A major. The answer `A` corresponds to the same subject transposed a fifth upper (*E major*) and interleaved with a counter subject `CS`. A development of these patterns is proposed through a passage called episode `E` in this excerpt.

Other movements have been described as *double fugue*: the *Vivace* of Fantasia No. 2 according to Silva.

In other cases, the music presents what may be described as a *fugal entry* rather than a fully developed fugue. Such passages introduce a subject followed by a response, thereby suggesting the beginning of a fugal exposition, but remain limited in scope before the texture moves toward a less strictly contrapuntal continuation. This situation can be observed in Fantasia No. 6 (*Allegro*), where an initial subject–answer pair is followed by several additional thematic entries that do not receive corresponding answers.

Finally, certain movements may be more accurately described as instances of *fugal imitation*. In these passages, the contrapuntal writing evokes the stylistic procedures of the fugue without adhering to its structural conventions. The imitation is typically shorter and more motivic, without the systematic deployment of subject and answer characteristic of a full exposition.

Examples include Fantasia No. 9 (*Allegro*), which presents multiple thematic entries without a clearly defined answer, and the opening of Fantasia No. 10, where the texture suggests fugal imitation rather than a fully developed fugue.

For each fugato movement, when present, we identified the subjects (`S`), counter-subjects (`CS`), answers (`A`), and episodes (`E`) and detailed the notes belonging to the different subjects for a total of 556 annotations. When contradictions occurred between sources—mainly concerning the end of a subject—we followed the majority view to resolve them. A quarter of the movements of the fantasias are fugatos and are the main source of interleaved voices, as we will see.

Implied Chords and Local Keys

Several analytical approaches have been proposed to explain how polyphony can be suggested in Telemann's solo writing. For instance, Debree discusses different technical devices used to simulate polyphony, drawing on the tradition described by Flux [Fux et al. (1971)]. These include latent counterpoint, where alternating notes imply two voices (often through equal rhythmic values such as successive eighth notes), as well as arpeggiation (see Fig. 4.6) and ornamental figures that outline underlying harmonic structures.

On Tonal Analysis Tonal analysis plays a central role in musicology, as well as in performance and listening practices, by providing a structured framework for understanding harmonic organisation and musical coherence. Historically, Western music gradually evolved from Renaissance modal practice toward the tonal system.

A decisive step was taken in the sixteenth century by theorists such as Gioseffo Zarlino, who emphasised the structural importance of consonant intervals, particularly thirds and fifths [Zarlino (1558)]. Around 1600, the emergence of *basso continuo* and figured bass further transformed musical thinking by reducing complex polyphonic textures to their essential harmonic foundations [Abromont and Montalembert (2001)]. These developments were later systematised by Jean-Philippe Rameau in his *Traité de l'harmonie* [Rameau (1722)], which established a theoretical framework based on chords, harmonic functions, and tonal relationships.

In tonal music, pitches and harmonies are organised around a tonic, which functions as the tonal centre and is perceived by listeners as a point of stability, being typically "sounded with relative frequency and with longer duration" [Krumhansl (2001)].

The notion of *global key* refers to the tonal centre governing an entire work, serving as a primary interpretative reference for both analysis and performance. It is commonly included by the composer within the title of the work, and can be inferred from the beginning and ending sections [Par et al. (2006)]. However, tonal organisation frequently operates at a more local level through *modulations* described by *local keys*.

Finally, a chord is generally defined as a simultaneity of pitches perceived as a single harmonic unit [Abromont and Montalembert (2001)]. In practice, especially in monophonic or solo instrumental writing, such vertical structures are often not explicitly sounded but instead cognitively inferred from melodic context. *Implied chords* refer to harmonic structures that are not directly present as simultaneous sonorities, but are suggested through arpeggiation or voice-leading patterns.

Harmonic Annotations in the Fantasias We encoded harmonic annotations describing local keys and implied chords, comprising 192 key labels and 731 Roman numeral chords.

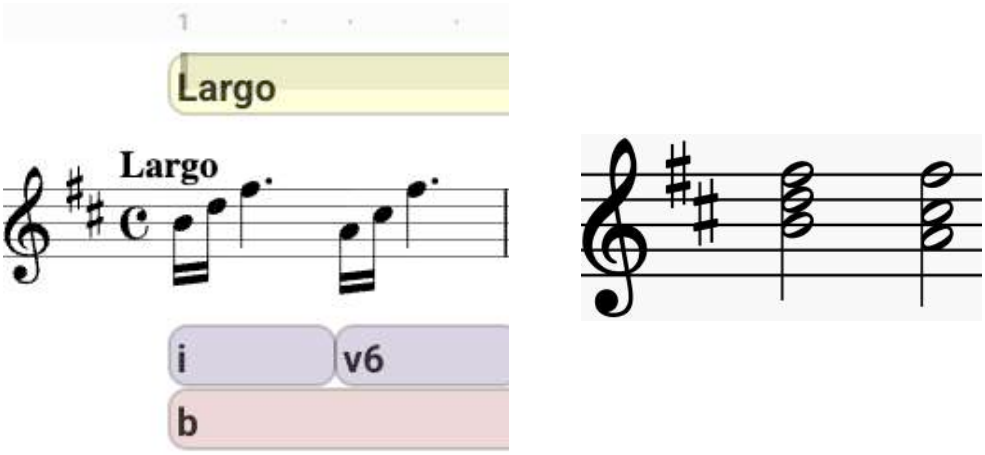


Figure 4.6: The first measure of Fantasia No. 3 (TWV 40:4) implies harmonies, tonic (`I`) then dominant (`v6`). The local key being there the main key *B minor* 🔊.

For the first three fantasias, Silva provides a detailed analysis of local keys. A second important source is Min, who proposes chord progressions across all twelve fantasias. These sources were complemented with additional analyses when necessary. In cases of disagreement between sources—for example, when a modulation is interpreted as beginning one measure earlier, or when a passage can be heard as a full modulation - we retained a main local-key annotation and an alternative interpretation.

These annotations were encoded (Fig. 4.5) using the open-source web platform Dezrann. In the example shown, the last two beats of measure 12 can either be interpreted as the end of a phrase in the local key of *A major* (alternative analysis) or as the beginning of a modulation toward *E major*. Whenever implied chords are specified, they are derived consistently from the main local-key annotation.

We choose not to add our own analysis to keep these annotations within the scope of the work of scholars specialised in this area.

Voices Annotations

The identification of voices are much or less exhaustive across the sources. If Silva detailed these voices completely in the Fantasias No. 1–3, more often only some movements are annotated (Min, Brown), or examples are given at different level (just for an excerpt of several measures, on a complete subject of a fugato, or even on one measure). As for the other annotations, we chose, even if it was tempting, to use only these sources and not our feelings about the voices on which we have no scholarly indications. However, it seems clear to us that the annotations could be a little extended at some point.

The voice annotations account for 2,324 notes (28.2% of the notes of the corpus) and mostly denote two voices. A third voice appears in only three movements: the *Allegro fugato* of Fantasia No. 7 [Hunt (2016)], the *Corrente* of Fantasia No. 10, and the *Vivace fugato* of Fantasia No. 11 [de Bree (2020); Eppinger, S. (1984)]. We produce, annotated files by encoding the voices related to the position in the score by `u` (upper voice, 48% of these annotations), `m` (middle voice, 4.6%), `b` (bass voice, 21.6%), `B` (bass voice, notated with *opposite stem-notes* by Telemann, 25.9%) as in Figure 4.5.

The sources are largely consistent. For example, only six opposite stem-notes in the corpus (0.3%) are interpreted as upper or middle voice by some sources. We kept them as bass voices, as in the autograph. Notes without annotations or rare ambiguous cases are tagged by `x`.

voice sequences	`b u b u`	`B u B u`	`u u u u`	`u b b u`	`u b u b`	`B u b u`
%	**22.5**	**15.3**	10.8	8.1	**6.3**	**4.5**

voice sequences	`b u u b`	`b m m m`	`B u u u`	`b b u u`	`B u m m`	`m m m m`
%	4.5	3.6	3.6	3.6	2.7	2.7

Table 4.2: Distribution of main voice annotations in groups of four sixteenths in duple metre movements of the corpus. Bold values show alternating upper/bass voices.

Focusing on full annotated groups (*e.g.*, groups with annotation on all notes) of four sixteenth notes within a beat in duple metre, 49.6% of the sixteenth are interleaved, alternating between an upper voice and a bass voice (`bubu`, `ubub`, and all combinations with also `B`). In contrast, only 13.5% consist of a single voice, with most being the upper voice and 2.7% the middle voice (Tab. 4.2). These groups of alternating notes are mainly (81%) found in the *fugatos* movements.

In Appendix A, we present an improved version of a change-of-voice detection algorithm and evaluate its application on the entire *Fantasias* corpus.

4.1.2 Recordings and Unfolded Scores

We describe in this section the datasets of recordings and the preparation of the scores for the score-to-audio alignments described in Chapter 5.

Recordings

performer	year	instrument	hyperlinks	**Expressive Timing** (5.1 and 6.2)	**Audio Features** (5.2)
Jean-Pierre Rampal	1973	Flute	YT / MBID	🕓	🎧
Barthold Kuijken	1978	Traverso	YT / MBID	🕓	🎧
Marion Verbruggen	1997	Recorders	YT	–	🎧
François Lazarevitch	2016	Traverso	YT / MBID	🕓	–
François Leleux	1997	Oboe	YT	–	🎧
Amy Porter	2008	Flute	YT	🕓	–
Marek Nahajowski	2014	Alto recorder	YT	–	🎧
Thomas Indermühle	2015	Oboe	YT	–	🎧
Emmanuel Pahud	2018	Flute	YT / MBID	🕓	🎧
Katja Pitelina	2020	Traverso	YT	🕓	🎧

Table 4.3: Complete audio recordings of the 12 Telemann fantasias used in this thesis. Recorded during the COVID-19 lockdown, the Pitelina recording is the only non-commercial one and also contains videos. The hyperlink column reports, when available, YouTube playlists (YT) and MusicBrainz identifiers (MBID). The last columns indicate inclusion in two different subsets (🕓 , 🎧) and their corresponding main applications.

In 2020, Garrison [Garrison (2020)] identified approximately 50 complete sets of recordings of the Fantasias performed on instruments from the flute family, including the modern flute, traverso, and recorder.

In Chapter 6, to investigate interpretative variability in these works—particularly temporal deviations and the interaction between interleaved voices—we select a subset of six performances spanning the period from 1973 to 2020, balancing modern flute and traverso interpretations (🕓 in Tab. 4.3). The corresponding score-to-audio alignments are described in Section 5.1.1. On this subset of performances, we aimed to include both renowned recordings—such as those by Rampal for the modern flute and Kuijken for the traverso—and more recent interpretations, including those by Lazarevitch. All of these recordings are commercially available, with the exception of Pitelina's recording, which was made in a church during the COVID-19 lockdown.

In Section 5.2, to study the influence of audio features on the score-to-audio alignment accuracy, we select a different subset of eight recordings including recorders and oboe (🎧 in Tab. 4.3). With this second subset, we provide manual annotations of onsets at the measure-level, described in 5.2.1. The two subsets overlap in four performers: Rampal, Kuijken, Pahud, and Pitelina.

The diversity of interpretations in our dataset is not limited to timbre, tempo, or articulation choices, but also concerns the formal realisation of the scores themselves. In particular, performers do not always observe written repeats in the same way, which has important consequences for alignment and annotation procedures.

Unfolded Scores

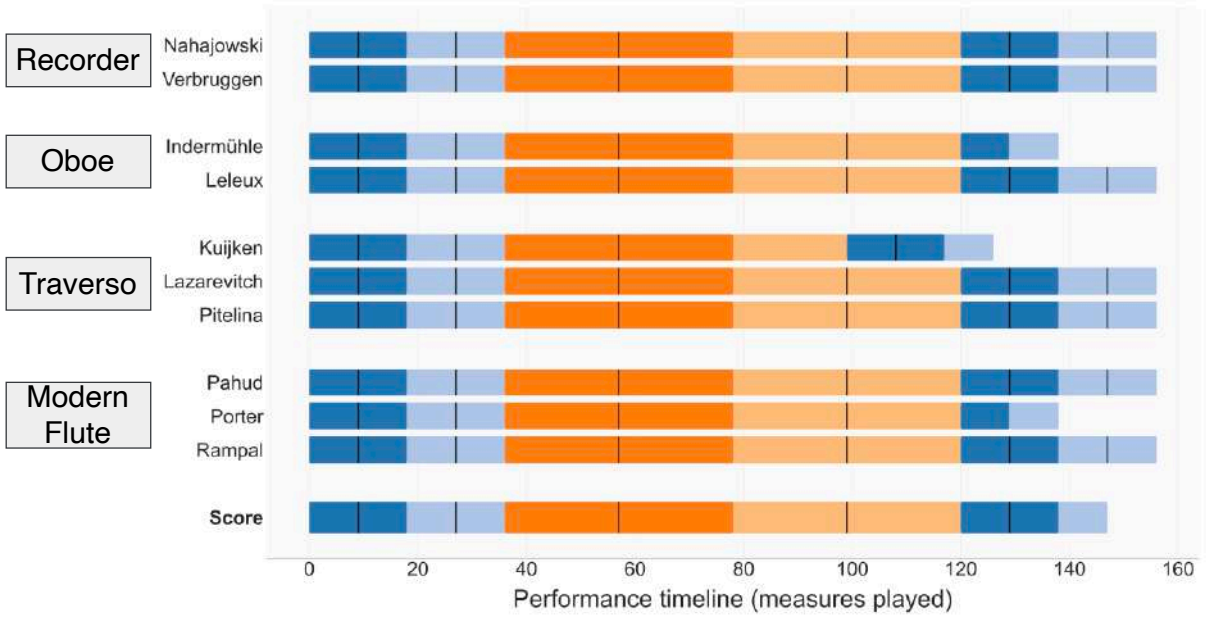


Figure 4.7: Performed repeats structure of Fantasia No. 12 (TWV 40:13) last movement (*Rigaudon*).

fantasia	**1**	2	3	4	5	6	7	**8**	9	10	**11**	12
min	88	129	98	121	133	121	151	96	155	192	113	195
max	88	152	118	149	156	138	231	96	196	225	113	225

Table 4.4: Min/max measures played per fantasia across performances.

Structural differences such as repeated sections are a common challenge in score–to-audio alignment. For example, in the final movement (*Presto* of Fantasia No. 12), the notated structure is `AABBCCDDAAB`, although this pattern is not always realised in performance. Instead, three alternative realisations are observed: `AABBCCDDAAB`, `AABBCCDDAB`, and `AABBCCDDAABB` (Fig. 4.7). While previous approaches such as *JumpDTW* [Fremerey et al. (2010)] have attempted to explicitly handle such structural variations, in our case, we carefully checked the actual repetitions in the score to establish these variants.

This variability in structural realisation is reflected in the duration of the performances, which ranges from 1592 measures (Kuijken) to 1887 measures (Pahud and Rampal). By contrast, only three fantasias (Nos. 1, 8, and 11) are consistently performed with the same structural realisation across our dataset (Tab. 4.4).

Ornamentations

Performing such baroque music often implies adding ornamentation. Telemann notated only 16 grace notes throughout his entire corpus, most of them in Fantasia No. 5, and 64 *tremblements* marked with a "+" sign indicating trills or mordents. Performers may add other ornaments, particularly in slow movements, such as the introductory *Sarabande* of Fantasia No. 9, as played by Lazarevitch. Repeats are often ornamented the second time, as abundantly demonstrated by Rampal (for example in the *Passepied* of Fantasia No. 1). We have chosen not to include the written ornaments in the expanded scores and will discuss why in Section 5.1.2.

Section summary – Fantasias Dataset

This section presents the construction of the annotated corpus built from Telemann's *Twelve Fantasias* for solo flute. Starting from MusicXML transcriptions checked against the authenticated source, we enriched the scores with several layers of musicological annotations. These include movement segmentation and naming, identification of fugato passages, implied harmonic structures, local keys, and voice annotations related to latent polyphony. The corpus also includes six historically and stylistically contrasted recordings performed on modern flute and traverso. Because performers do not always realise written repeats in the same way, unfolded versions of the scores were prepared to match the actual performed structures. Finally, although ornamentation plays an important role in performance practice, additional ornaments introduced by performers were not encoded in the expanded scores, as discussed later in the manuscript.

4.2 Early 20th Dataset


COMŒDIA. — MARDI 2 DECEMBRE 1913

THÉATRE PARTICULIER DE M. MORS

Première représentation de "PSYCHÉ"

Poème dramatique en trois actes de M. Gabriel Mourey

Figure 4.8: Scrapbook of programmes and press articles on Gabriel Mourey's Psyché (Act III, Scene 1), Bibliothèque nationale de France, Performing Arts Department, 8-RF-68061.

In contrast to Telemann's Fantasias, several works from the early twentieth-century solo flute repertoire exhibit a much more flexible relationship to metre and pulse. This is particularly the case for two emblematic solo flute pieces: *Syrinx* (L.129 – CD.137), composed in 1913 by Debussy, and *Density 21.5*, composed in 1936 and revised in 1946 by Varèse. These works emerge in the context of early 20th-century modernism, where tonal stability and regular pulse begin to dissolve. Debussy's *Syrinx*, originally written as incidental music for Act III, Scene 1 of Gabriel Mourey's *Psyché* (Fig. 4.8), reflects a speech-like rhythmic flow [Trezise (2003)].

Varèse's Density 21.5, composed for the inauguration of a platinum flute, prioritises timbre and pitch organisation over traditional metric structure. Together, these pieces contribute to a new conception of musical time in the modern solo flute repertoire. Indeed, they often lack a stable underlying pulse, relying instead on local temporal shaping, gesture, and motivic development [Bernard (1987); Nattiez and Barry (1982)]. As a result, score–performance alignment becomes particularly challenging. To address this issue, we adopt an audio-to-audio alignment strategy that enables robust transfer of onset-level annotations, as described in Chapter 5. This approach relies on manually annotated note onsets as a prerequisite for reliable alignment.

In this section, we present the score encoding, the dataset of recordings, the onset annotation protocol and the motives chosen for expressive timing investigation (Chapter 6).

Syrinx, L.129 (C. Debussy, 1913)

Version	Performer	Year	Youtube ID
R_1	E. Pahud	2000	YG2Ymz7FULs
R_2	E. Ferguson	2023	IZovibvtoYg
E (V_3)	W. Auer	2023	NcWuqs0TeXI
$V_4 \cdots V_{23}$	R. Bourdin	1959	f4sn6Dx7kr4
	J.-P. Rampal	1975	bvXN_9DoMPs
	P.-L. Graf	1977	vFMDh2SzdY8
	J. Galway	1986	M3otKgy4qGs
	W. Schulz	1990	IC3iGyfxfr0
	E. Zukerman	1996	9KWDPdq_F78
	M. Wiesler	1997	qExOd10z8Yc
	P. Gallois	1998	vjD-Ey8oRQE
	A. Marion	1998	amLxouHCKgc
	J. Hurel	2001	PnB8HjHdDGw
	V. Lucas	2008	4wVURxW57oA
	S. Zampetti	2008	plDH3KIPHls
	P. Bernold	2012	JqejwmbRPnQ
	J.-L. Beaumadier	2012	qggGTrm-KFk
	M. Nastasi	2012	BsPguqbyA20
	P. Robinson	2016	JXcWC9ykM8U
	A. Kossenko	2020	c8eQLhJu9jg
	L. Orden	2020	sAmkVnOokHI
	K.-H. Schütz	2023	3AYk7MeMRWE
	M. Calderini	2024	TyD_QcCMUTM

***Density 21.5* (E. Varèse, 1936, rev. 1946)**

Version	Performer	Year	Youtube ID
R_1	E. Pahud	2018	GHjMbgLnkAY
R_2	J. Hurel	2001	mjgpL0Dts9o
E (V_3)	E. Zukerman	1996	kYtIfiG7r5U
$V_4 \cdots V_{23}$	S. Gazzelloni	1958	H5T9UZRy1jM
	L. Beauregard	1984	J025vOlyZOg
	P.-L. Graf	1989	xTYSbe8HCdU
	R. Fabbriciani	1994	ODvaInMr-e4
	P. Pierlot	1996	ac8naviYmmM
	M. Wiesler	1997	BK6IM2w9RG8
	J. Zoon	1998	s6-TBeDLlnA
	C. Marchand	1999	DI-z9dQaqv0
	P. Fried	2005	bl1_fb6l83w
	M. Grochowska	2008	y4AIbgFpTKo
	P. Bernold	2008	vlCxOJqcyxA
	M. Nastasi	2012	JmKRnE9y2N4
	H. Yoon	2013	9nFuvrEhBU8
	C. Chase	2013	J5-nkJz0guk
	P. Bocquillon	2013	JHYVkMX7nUc
	M. Barth	2018	lxdY4hCzrzI
	L. Orden	2020	DfaW__NS2SU
	M. Bellavance	2021	r9P6xHmxZuQ
	K.-H. Schütz	2023	RpVWg3p6nMk
	V. Prats	2025	Mgi6uoPdoEQ

Table 4.5: Among the 23 performances used in this study for each piece, three are manually annotated: R_1 and R_2 are the reference performances and E (also designed as V_3) the evaluative one.

4.2.1 Scores and Recordings

Most interpretations of these pieces are non-metronomic, reflecting expressive timing rather than strict adherence to a beat, which makes them particularly suitable for performance analysis. To address these challenges, we introduce a corpus of 46 solo flute performances of early 20th-century repertoire, evenly distributed between *Density 21.5* and *Syrinx*, enriched with onset-level annotations for fine-grained timing analysis.

We manually transcribed the scores from published editions and processed them using `music21` [Cuthbert and Ariza (2010)]. This processing step focused on encoding and extracting structural and musical descriptors such as symbolic durations, pitches, and time signatures. Tied notes were treated as single sustained events, with only their initial onset retained for representational consistency across the score.

Among the 46 performances collected (Tab. 4.5), 43% were recorded from 2010, and 41% between 1990 and 2009. All the audios can be streamed on Youtube. Three of them are accompanied by video recordings of the performances (Kossenko, Ferguson, and Calderini for *Syrinx*).

In total, the dataset encompasses 156 minutes of solo flute performances (Tab. 4.6), providing a rich foundation for analysing interpretive variation across performers in such important pieces for the contemporary solo flute repertoire.

Close pitches are particularly challenging for alignment tasks, especially in score-to-audio settings. In a piece such as *Syrinx*, 61% of consecutive melodic intervals are smaller than or equal to two semitones, with 36% corresponding to minor seconds (for *Density 21.5*, the proportions are respectively 42% and 29%). This high density of stepwise motion significantly increases the ambiguity of onset matching and motivates the need for precise manual onset annotations.

Piece	N_{Onsets}	$d_{\min}$	$d_{\max}$	d_{total}
Syrinx	259	2'2"	3'14"	58'
Density 21.5	249	3'4"	5'36"	88'

Table 4.6: Number of onsets; minimum, maximum and total duration among the 46 recorded performances used.

4.2.2 Onset Annotations

For each piece, I manually annotated note onsets for three performances, denoted R_1, R_2, and $E = V_3$ (Tab. 4.5), using *Sonic Visualiser*, for a total of 1524 annotations (Fig. 4.9).

The first pass consisted of a full listening to assess the recording's overall phrasing and interpretation. During the second pass, onsets were annotated by stroking a key while listening

at 80% playback speed. In the third pass, the annotated onsets were refined and corrected at 20% playback speed, with zooming on waveform features. A fourth pass involved auditing the annotations through sonification using *librosa* [McFee et al. (2015)].

This process allows to combine auditory judgment with visual inspection of the signal. However, ambiguities may arise due to reverberation, soft onsets, and Perceptual Attack Time (PAT), as the perceived onset can occur later than the "acoustic onset" [Wright (2008)] (and even this notion is subject to debate).

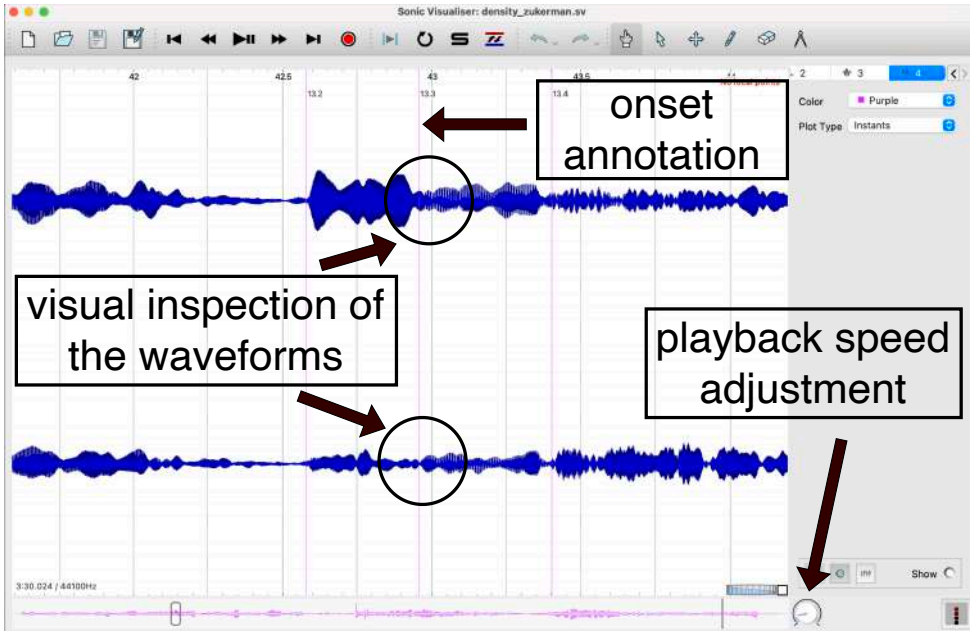


Figure 4.9: *Sonic Visualiser* screenshot during the annotation process.

Statistics derived from these manual annotations reveal systematic differences in temporal variability across both performers and pieces (Tab. 4.7). In particular, *Syrinx* shows higher coefficients of variation, indicating more pronounced local timing fluctuations compared to *Density 21.5*, where timing appears comparatively more stable across interpretations. However, the present analysis remains limited by the small number of manually annotated performances, which restricts the generalisability of these observations.

A larger dataset enabled by annotation transfer (see Chapter 5) will allow for a systematic investigation of expressive timing in these works, which will be further explored in Chapter 6.

Piece	Performer	μ	$\tilde{x}$	σ	CV
Syrinx	Auer	0.63	0.35	0.85	1.36
	Ferguson	0.75	0.39	0.99	1.32
	Pahud	0.67	0.30	1.1	1.6
Density 21.5	Barth	0.90	0.59	0.87	0.96
	Hurel	0.85	0.55	0.84	0.98
	Zukerman	0.81	0.53	0.79	0.98

Table 4.7: Comparison of IOI-based timing statistics across the six manually annotated performances of *Syrinx* and *Density 21.5*. Mean μ, median $\tilde{x}$, and standard deviation σ of IOI are reported in seconds. The coefficient of variation (CV), defined as $CV = \sigma/\mu$, provides a normalised measure of temporal variability across performances.

4.2.3 Musical Patterns

A motif (or pattern) is a short musical idea, melodic, harmonic, rhythmic, or any combination of these three [Drabkin (2001)]. They constitute fundamental building blocks of music and are widely used by composers to create coherence and unity within a work.

The importance of motifs has long been emphasised in music theory and extensively studied by musicologists. Schoenberg defines motives as "intervals and rhythms, combined to produce a memorable or contour which usually implies an inherent harmony" [Schönberg (1967)].

The computational study of musical patterns has also become an important topic in MIR through pattern discovery [Conklin and Bergeron (2010); Lartillot (2009); Meredith et al. (2003); Vantalon et al. (2022)], but despite their importance, publicly available datasets containing expert-annotated musical patterns remain relatively scarce [Ren et al. (2017)].

In this study, rather than defining patterns ourselves, we rely on motivic analyses established in the musicological literature.

Patterns in Syrinx

For *Syrinx*, we collect some patterns proposed by Nattiez [Nattiez (1975)].

We consider four short *cell-level* patterns: `A`, the lead pattern; `B`, consisting of three triplet eighth notes with an appoggiatura on the second note; `C`, an arch-shaped figure appearing both in sixteenth-note and thirty-second-note realisations; and the descending figure `E` (Fig. 4.10). Because pattern `C` occurs with different rhythmic realisations, its occurrences may have different durations while preserving the same melodic contour.

In addition, we include two larger *phrase-level* patterns, `P` and `Q`. Pattern `P` corresponds to the principal thematic statement of the piece, combining two occurrences of motive `A` followed

by a descending sixteenth-note passage.

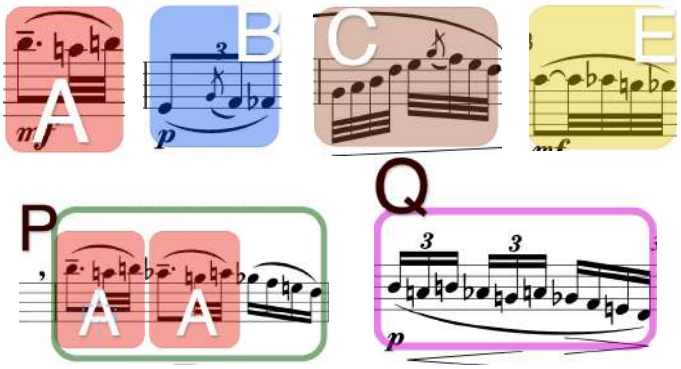


Figure 4.10: Cell-level (`A,B,C,E`) and phrase-level (`P,Q`) patterns in Debussy's *Syrinx*. Full identification of all patterns in the score is provided in Fig. 6.10.

Head Pattern in Density 21.5

For *Density 21.5*, we follow Ohriner [Ohriner (2019)], who identifies a recurrent three-note configuration as a central structural element of the piece. The pattern is defined as:

> "a segment of three notes without intervening rests in which the interval between the first and third notes is a unison or half step, and the ratio of the notated durations is between 1:2 and 2:1".

Unlike the motifs selected for *Syrinx*, this pattern is defined through a set of intervallic and rhythmic constraints that capture a characteristic gesture recurring throughout the piece. All occurrences matching this definition were annotated and included in the dataset (Fig. 4.11). The resulting annotations are used throughout the experiments presented in Chapter 6.

Section summary – Early 20th Dataset

This dataset comprises 46 solo flute performances of *Syrinx* (Debussy) and *Density 21.5* (Varèse), two works characterised by highly flexible temporal structures and weak metrical regularity. To support fine-grained timing analysis, all scores were manually encoded, and 1524 note onsets were annotated across selected performances. It offers a solid basis for annotations transfer and, in a second time, study of expressive timing in early twentieth-century solo flute repertoire.
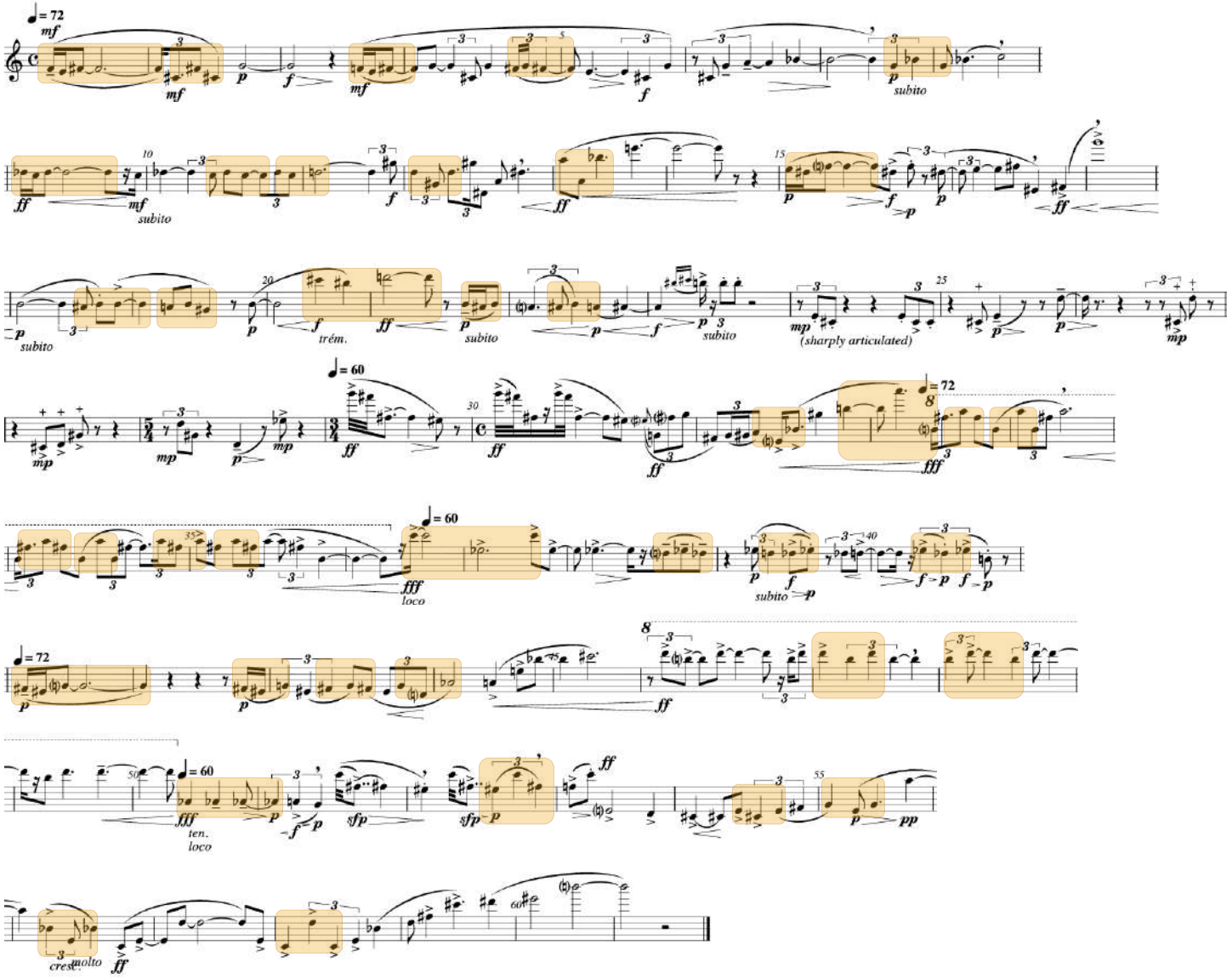


Figure 4.11: *Density 21.5* with 40 pattern α occurrences identified based on Ohriner's definition.

4.3 Data Availability

	 Dataset **Fantasias** Telemann (1732)		 Dataset **Early 20th** Debussy (1913) Varèse (1936)
Symbolic Labels			
Sources	autograph manuscript + 11 musicological studies (Tab. 4.1)		published scores
Scores	12 (Sec. 4.1)		2
Structure	45 movement names 556 fugato labels		–
Harmony	192 local-key 731 Roman-numeral chords		–
Voices	2 324 610 opposite stem notes		–
Patterns	–		7 patterns 77 occurrences (Sec. 4.2.3 and 6.3.1)
	Subset 🕐 ***Expressive Timing***	**Subset** 🎧 ***Audio Features***	
Audio Resources			
Recordings	48 recordings 12 × 6 performances 5.5 h (Tab. 6.1)	96 recordings 12 × 8 performances 7.5 h (Tab. 5.2)	46 recordings 2 × 23 performances 156 min (Tab. 4.5)
Automated Onsets Annotations	67 826 note onsets (Sec. 5.1.1)	80 600 note onsets (Sec. 5.2.2)	12 719 note onsets *on 42 recordings* (Sec. 5.3.1)
Curated Human Onsets Annotations			
	on 1 recording *Fantasia No. 1* 673 note onsets (no ornaments) 759 with ornaments (Sec. 5.1.2)	on 96 recordings *Fantasias No. 1–12* 14 602 measure onsets (+ 704 for Fantasia No. 1) (Sec. 5.2.1 and 5.2.5)	on 6 recordings 1 524 note onsets (Sec. 4.2.2)

Table 4.8: Overview of the datasets: symbolic labels, aligned performances, and manual onset annotations.

Our two datasets are described in Table 4.8 with cross references to the sections of this Thesis where the different elements are described.

The note onsets in the **Audio Resources** are outputs of score-to-audio alignment for the Dataset Fantasias or outputs from audio-to-audio alignments with triple transfer and soft-late fusion strategy for the Early 20th dataset (all alignments are described in Chapter 5).

This thesis also involved a total of 17 589 human **Onsets Annotations**, of which 16 885 were produced by the author. In the following, we describe all the necessary material available online.

4.3.1 Dataset Fantasias

Subset Expressive Timing

We provide, under open licenses, curated scores and all previous described annotations, under the https://recherche.data.gouv.fr long-term archive at https://doi.org/10.57745/MSLVWS.

All the data can also be explored with using *Dezrann* platform https://www.dezrann.net/explore/telemann-flute-fantaisies

Subset Audio Features

We release the dataset, annotations, alignment outputs and code on https://github.com/patthmus/phd-manuscript-links/tree/main/audiofeatures-sec5.2

4.3.2 Dataset Early 20th

Recording links, annotations, alignment data, and code are available under open licenses from https://gitlab.com/algomus.fr/early20thsoloflute.

The synchronised recordings and patterns can also be explored using *Dezrann* platform https://www.dezrann.net/explore/early20thsoloflute and through a long-term archive at https://doi.org/10.57745/IZRHFF.

Chapter summary – Telemann, Debussy, Varèse: Two Solo Flute Datasets

This chapter introduced two annotated corpora for solo flute repertoire, spanning from Baroque implied polyphony writing in Telemann to early 20th-century works in Debussy and Varèse. Together, these datasets provide symbolic representations enriched with higher-level musical annotations and aligned audio recordings, supporting score-to-audio and audio-to-audio alignment tasks.

To the question *"How can higher-level musical structures that are not explicitly encoded in the score be systematically represented through enriched annotation layers in monophonic music corpora?"*, the answer is by extending traditional score representations with enriched higher-level musical annotation layers that describe perceptually and musically meaningful structures, such as patterns, or implied polyphony. These annotations provide resources adapted to the study of expressive timing.

Chapter 5

Improving Alignment Methods for Solo Flute
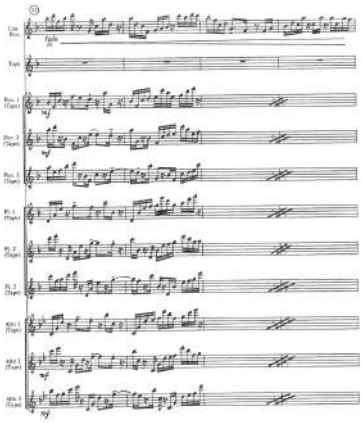


Figure 5.1: Vermont Counterpoint by Steve Reich linked to a performance by flautist Joe LaRocca.

Contents of this chapter

5.1 Spectral Templates in Score-to-Audio Alignments 106
5.1.1 Fantasias Alignments 106
5.1.2 Alignment Evaluation 108
5.2 Comparison of Audio Features in Score-to-Audio Alignment 113
5.2.1 Dataset of Monophonic Music Performances 113
5.2.2 Alignment Score-to-Audio 116
5.2.3 Audio Features Evaluation 118
5.2.4 Optimal Shift Investigation 120
5.2.5 Annotator Agreement and Robustness of Alignment Shift 122
5.3 Soft-Late Fusion Strategy for Audio-to-Audio Alignments 126
5.3.1 Audio-to-Audio Alignment 126
5.3.2 Evaluation of Alignment Quality 129

Linking a musical score to an audio signal, or vice versa, amounts to establishing an alignment between two heterogeneous representations whose temporal coherence is not given but must be reconstructed.

How can onset-level correspondences be robustly inferred across diverse performances ?

This issue becomes particularly salient in situations where temporal structure is either explicitly encoded in a score or only implicitly realised in performance.

It is illustrated in Steve Reich's *Vermont Counterpoint* (Fig. 5.1), where a live flute interacts with multiple pre-recorded flute layers. The perceptual continuity of the piece relies on precise temporal coordination across parallel streams, highlighting the need for reliable alignment between related musical events.

In this perspective, two complementary alignment scenarios are considered in this chapter. In Telemann's *Twelve Fantasias*, alignment can be formulated as a score-to-audio mapping problem. We first present a pipeline using harmonic templates designed specifically for the flute. We then investigate the influence of the choice of audio features on alignment accuracy, extending the evaluation to a broader range of woodwind instruments. Conversely, in early twentieth-century repertoire such as Debussy's *Syrinx* and Varèse's *Density 21.5*, the absence of a stable metrical framework motivates an audio-to-audio strategy, where correspondences are inferred directly between performances through annotation transfer.

Research outputs associated with this chapter

When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In 26th International Society for Music Information Retrieval Conference (ISMIR), 2025

Interacting with Annotated and Synchronized Music Corpora on the Dezrann Web Platform

Ballester et al.

Transactions of the International Society for Music Information Retrieval (TISMIR), 2025

Soft Late-Fusion Alignment and Pattern-Based Performance Analysis of Early 20th-Century Solo Flute Music

Patrice Thibaud, Louis Bigo, Mathieu Giraud, Yann Teytaut

Accepted to ISMIR 2026

5.1 Spectral Templates in Score-to-Audio Alignments

This section presents the alignment framework developed for Telemann's *Twelve Fantasias* for solo flute, focusing on score-to-audio alignment. The selected performances correspond to all traverso and modern flute recordings (🕓 in Tab 4.3)

We address this problem through a spectral template-based approach that bridges symbolic pitch representations and audio signals via a probabilistic pitch model. This framework enables the extraction of onset-level correspondences between score events and performance realisations. Then, we present one evaluation protocol.

5.1.1 Fantasias Alignments

Spectral Templates and Alignment

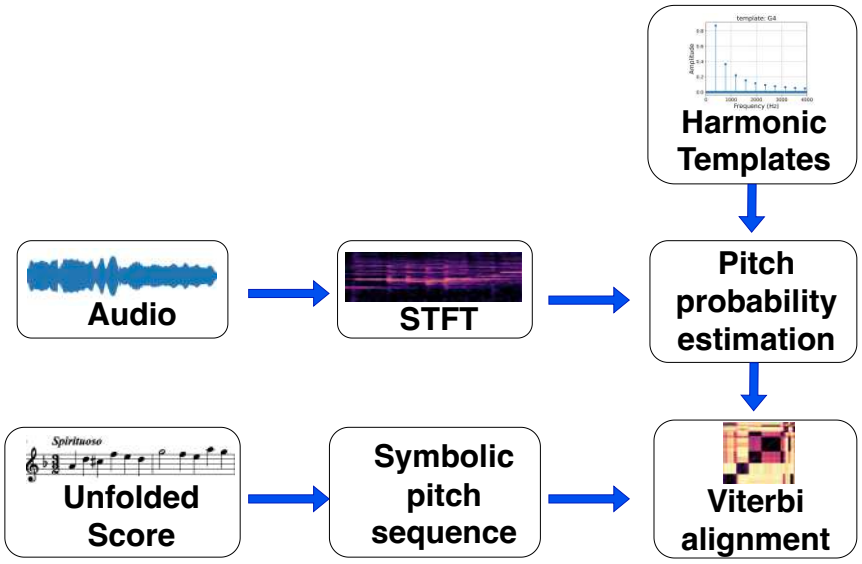


Figure 5.2: Score-to-audio alignment pipeline.

We describe the score-to-audio pipeline (Fig. 5.2) used for aligning performed scores to audio performances. In the following, spectrograms are computed at 22.05 kHz using a 1 024-sample window and a 512-sample hop.

Audio Representation. We compute a linear-frequency magnitude spectrogram via the STFT. A global tuning offset is first estimated from the audio and used to adjust the reference

fundamental frequency, ensuring consistency with performance tuning variations. For each candidate pitch within the instrument range, we construct discrete harmonic templates by placing weighted impulses at integer multiples of the (tuning-corrected) fundamental frequency nf_0 ($n = 1, \ldots, 10$). The amplitude of each partial follows a power-law decay of the form $f^{-\alpha}$, where the decay exponent is fixed to $\alpha = 1.25$ and chosen empirically to provide a balance between fundamental prominence and harmonic structure preservation. The resulting harmonic representation (see Fig. 5.2 "Harmonic Templates") is projected onto the FFT frequency grid and subsequently L2-normalised to ensure scale invariance across templates.

At each time frame t, we compute a time–pitch similarity matrix by evaluating the correspondence between the spectral magnitude vector $X_t \in \mathbb{R}^F$ and each discrete harmonic template $T_p \in \mathbb{R}^F$, defined over the FFT frequency bins. The similarity score is given by:

$$S(p, t) = \langle T_p, X_t \rangle, \tag{5.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard dot product.

This yields a score matrix $S \in \mathbb{R}^{M \times N}$, where M is the number of candidate pitches and N is the number of time frames.

To obtain a probability estimation over pitches, we apply a numerically stabilised *softmax* independently at each time frame:

$$P(p \mid t) = \frac{\exp(S(p, t) - \max_{p'} S(p', t))}{\sum_{p'} \exp(S(p', t) - \max_{p'} S(p', t))} \tag{5.2}$$

This results in a frame-wise posterior distribution over pitches $P(p \mid t)$, which serves as the observation model for subsequent alignment.

Score Representation. As it is described in Chapter 4, the symbolic score is first unfolded to account for repeats and structural variations. The resulting sequence is then encoded into a discrete pitch-index representation aligned with the instrument-specific pitch set, yielding a symbolic sequence $(i_k)_{k=1}^K$ used as the discrete state space for alignment.

Alignment Procedure. Alignment is formulated as a DTW-like monotonic sequence-to-sequence matching problem, solved via dynamic programming. More precisely, it corresponds to a Viterbi-style decoding over a left-to-right state space, where score events define the hidden states and frame-wise pitch probabilities serve as emission likelihoods.

Given a sequence of score note indices $(i_k)_{k=1}^K$ and a probability matrix $P(p \mid t)$ obtained from harmonic template matching and *softmax* normalisation, we compute an accumulated cost matrix using DTW scheme that enforces a monotonic progression through the score. At each time frame m, the alignment can either remain on the current score note i_k or advance to the next one i_{k+1}.

The local alignment score is given directly by the pitch probability $P(i_k \mid t)$, such that the algorithm favours paths that maximise the likelihood of the observed audio under the template-based representation. In this sense, the probabilistic output of the template matching stage serves as an observation model for the alignment.

The optimal alignment path is recovered through a backward pass over the accumulated cost matrix, yielding a correspondence between score events and audio time frames. From this alignment, each note n is assigned a performance onset time $o_p(n)$, from which inter-onset intervals $IOI_p(n)$ are derived for subsequent analysis of expressive timing in Chapter 6.

5.1.2 Alignment Evaluation
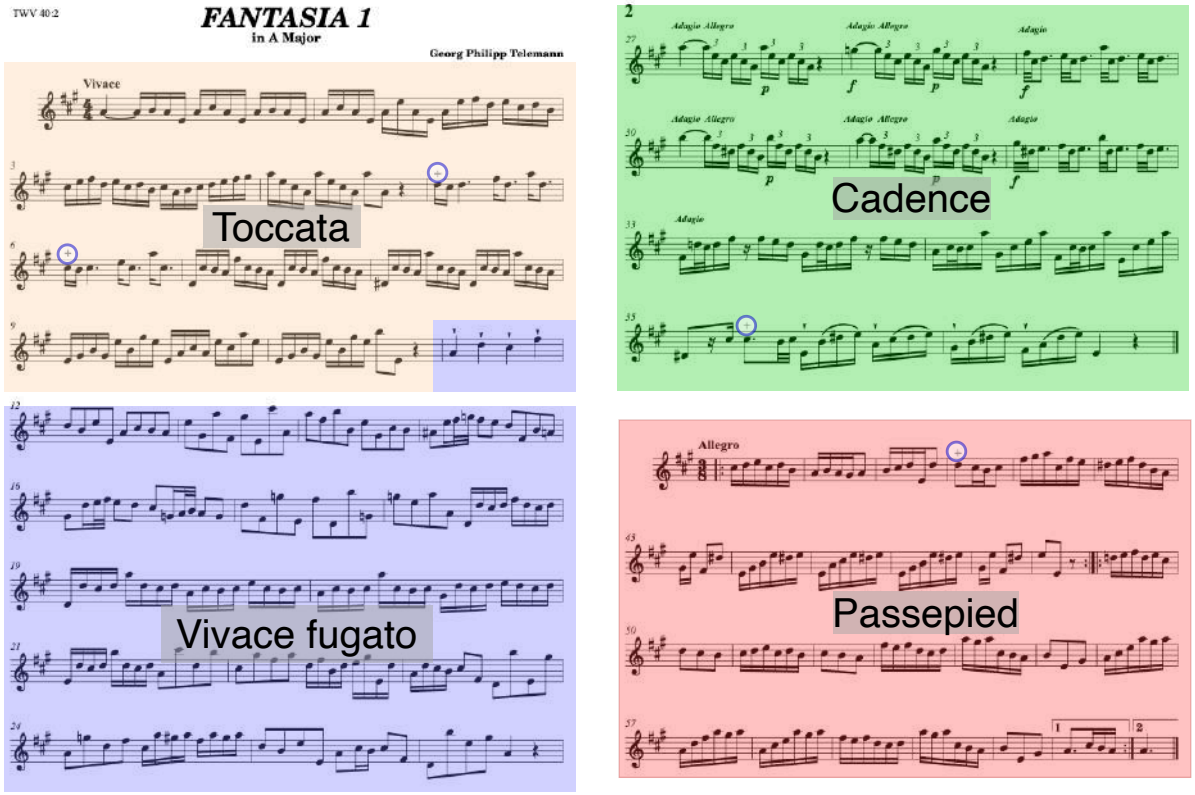


Figure 5.3: Structure of Fantasia No. 1 (TWV 40:2) with movements names. Four *mordents* are indicated by a "+" sign in the score.

A corpus-scale evaluation of harmonic templates, including comparisons with other audio feature representations, is presented in Section 5.2.2. In this section, we focus on a detailed

alignment evaluation for Fantasia No. 1 (TWV 40:2), performed by Lazarevitch on traverso. The performance structure, including repeats, follows the score structure shown in Figure 5.3.

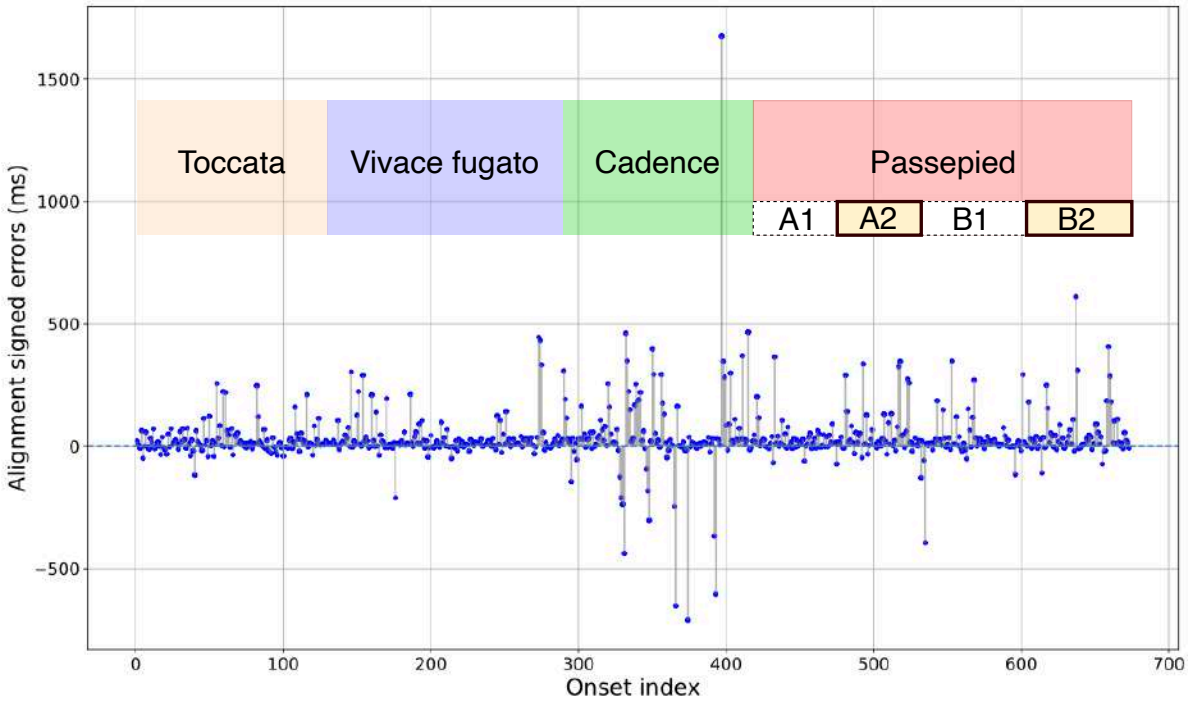


Figure 5.4: Signed alignments errors by onset in Fantasia No. 1 performed by Lazarevitch.

Note onsets were manually annotated by the author, following the procedure described in Section 4.2.2. In a first step, ornamentations were not taken into account, resulting in 673 annotated onsets. For each onset, we compute the signed alignment error, defined as the temporal difference between the onset estimated by the score-to-audio alignment and the corresponding manual annotation. A visualisation of these errors is provided in Figure 5.4.

Across the entire performance, the median absolute error is 21 ms, which is on the same order of magnitude as the analysis frame size. However, several large outliers are observed, reaching up to 1675 ms in the third movement (mm. 35).

Detailed statistics by movement are reported in Table 5.1. The first two movements appear less prone to alignment errors than the *Cadence* and the *Passepied*, especially in the repeated sections of the latter. Overall, 75% of the alignment errors remain below 50 ms.

The *Cadence* presents particular challenges for the harmonic-template representation due to the presence of arpeggiated figures and implied harmonic structures, which produce less stable spectral configurations. In the *Passepied*, ornamentation practices common to Baroque performance further increase the local temporal complexity, especially in repeated sections

where performers frequently introduce additional embellishments. Similar behaviour can be observed in the final movements of several other fantasias.

	Fantasia /Movement	$\bar{e}$	σ	μ_e	onsets	annotated voices
Without ornaments	Fantasia	59	122	21	673	180
	Toccata	**37**	**57**	20	126	6
	Vivace fugato	39	76	**14**	162	162
	Cadence	117*	226*	31*	128	3
	Passepied	53	92	20	257	6
	Passepied (only A2, B2)	66	110	25	127	3
With ornaments	Fantasia	61	104	23	759	180
	Toccata	**40**	**57**	21	136	6
	Vivace fugato	**40**	76	**14**	162	162
	Cadence	100*	161*	33	137	3
	Passepied	65	99	24	324	6
	Passepied (only A2, B2)	80	114	34*	180	3

Table 5.1: Absolute errors mean ($\bar{e}$) , standard deviation (σ), and absolute errors median (μ_e) by movement with and without taking account of ornaments on score-to-audio alignments of Fantasia No. 1 performed by Lazarevitch. We also report the number of onsets and annotated voices.

Do Ornaments Matter?

An important question is whether explicitly modelling ornamentations improves alignment accuracy.

Four of these (Fig. 5.3) occur in Fantasia No. 1: two in the *Toccata*, one in the *Cadence*, and one in the repeated section of the *Passepied*. The largest alignment outlier (1675 ms) corresponds to an ornamented passage in mm. 35, where Lazarevitch performs an extended trill 🔊.

Beyond the written ornaments, many additional performer-added embellishments appear in the final movement. Telemann left no written instructions, but the contemporary musical suggestions of Quantz [Quantz, Jean Joachim (1752)], among others, can help in performing stylistically appropriate ornaments.

To evaluate their influence, we realigned the performance using a manually transcribed score including both written and performed ornaments (Fig. 5.5). This second annotation version

contains 759 annotated onsets, compared to 673 in the simplified representation.

The comparison between the two annotation settings (Tab. 5.1) shows that including ornaments results in a denser symbolic representation which does not improve alignment precision. The mean absolute error increases slightly from 59 ms to 61 ms, while the median absolute error rises from 21 ms to 23 ms. Similarly, alignment accuracy at 25 ms or 50 ms decreases by a few percentage points.

The impact of ornamentation is particularly visible in the fourth movement and in repeated passages, where both the mean and median absolute errors increase more substantially. These sections appear especially sensitive to the additional temporal complexity introduced by ornamental notes. By contrast, the first and second movements remain comparatively stable across both annotation settings.

At the same time, the global standard deviation decreases from 122 ms to 104 ms when ornaments are included, and the maximum alignment error is significantly reduced. This suggests that ornament annotations may help reduce extreme alignment failures, even though they slightly decrease the typical local precision.

Overall, these results indicate that explicit ornament annotations are not strictly necessary for achieving reliable note-level alignment in this flute repertoire. A simplified symbolic representation without ornaments provides very similar median alignment accuracy while requiring fewer annotated events and producing slightly better overall precision metrics. Ornament annotations therefore appear to introduce additional local complexity without yielding a clear improvement in alignment quality, although they may contribute to reducing large temporal deviations in difficult passages.

Since the present work primarily focuses on interleaved voices, which are mainly encountered in fugato movements where relatively few ornaments are performed, ignoring ornamentations appears to be a reasonable methodological choice.

Section summary – Spectral Templates in Score-to-Audio Alignments

This section introduced a score-to-audio alignment framework for Telemann's *Twelve Fantasias*, based on harmonic spectral templates and a probabilistic pitch representation. An evaluation on Fantasia No. 1 shows that the method achieves generally accurate onset-level alignment, with median errors on the order of tens of milliseconds, while remaining sensitive to structurally and spectrally complex passages such as cadential figures and ornamented sections.
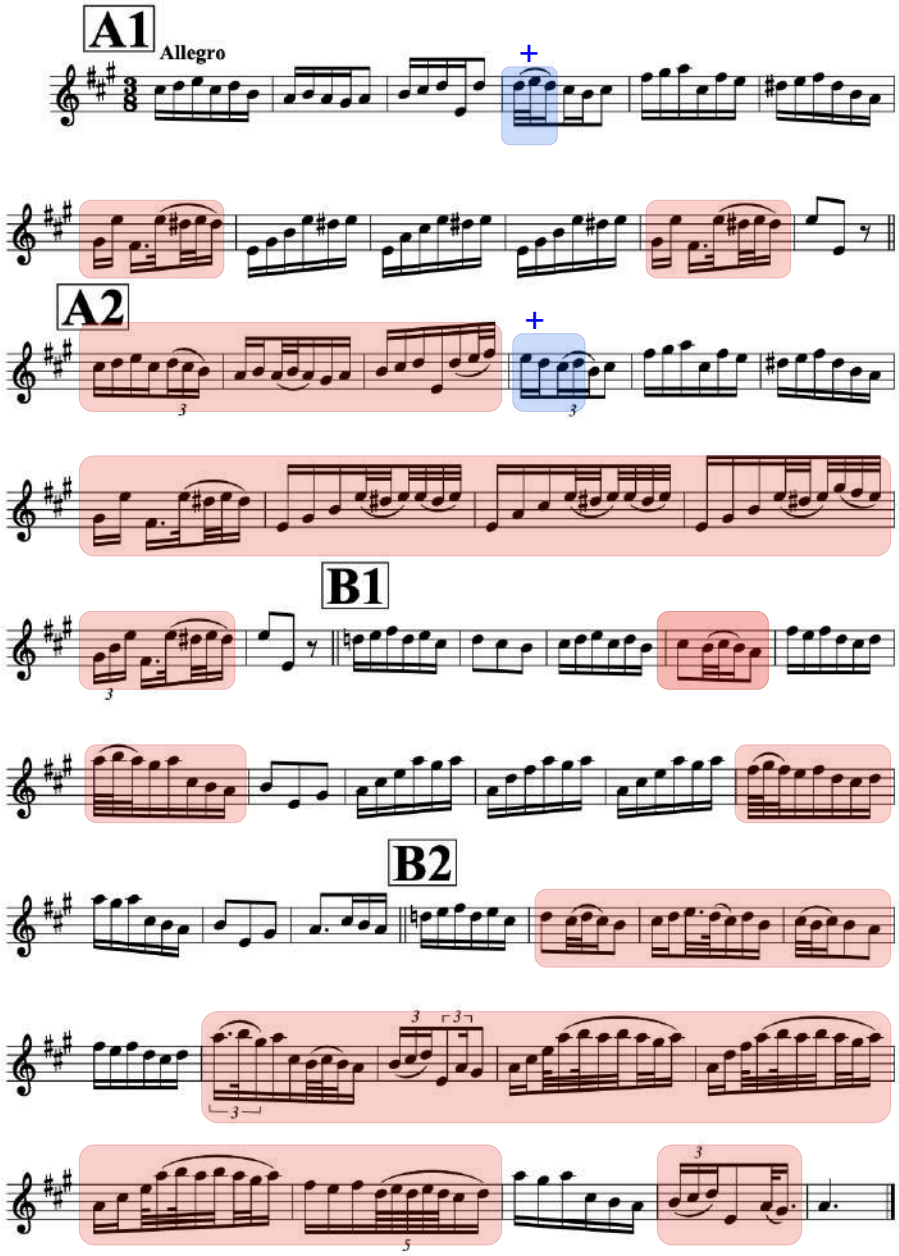


Figure 5.5: Transcription made by the author from Lazarevitch's performance of the Fantasia No. 1 *Passepied* 🔊. The two performed mordents present in the original score are annotated +. All the measures with added ornamentations are highlighted.

5.2 Comparison of Audio Features in Score-to-Audio Alignment

The previous section presented a score-to-audio alignment framework based on flute-specific harmonic spectral templates and evaluated its behaviour in detail on Fantasia No. 1. Although the obtained results demonstrated generally accurate note-level alignment, the analysis also revealed recurrent alignment deviations in passages characterised by ornamentation, arpeggiated figures, and less stable spectral configurations.

These observations raise a broader question: to what extent are such alignment errors specific to the harmonic-template representation, and to what extent do they reflect more general challenges inherent to monophonic woodwind recordings? Determining whether these deviations originate from the chosen audio representation requires a comparative evaluation across multiple feature types.

Woodwind instruments such as the flute introduce specific challenges for alignment due to soft or delayed onsets, less distinct attack transients often preceded by noisy components before pitch stabilisation [Castellengo (1999); Vergez (2010)]. These characteristics complicate the extraction of temporally precise representations and may reduce the robustness of standard alignment pipelines. To investigate this question, we now extend the study by using a subset of eight monophonic woodwind performances (🎧 in Tab. 4.3) of Telemann's *Twelve Fantasias for solo flute* with human measure-level annotations. We introduce a unified DTW-based evaluation framework comparing three feature representations: a chroma-based, a flute-specific spectral template-based, and a neural MPE representation. By keeping the alignment procedure unchanged and varying only the audio representation, we investigate how feature design influences alignment accuracy and temporal localisation in monophonic woodwind performances.

As our main contributions, we first empirically determine feature-dependent temporal shifts in score-to-audio alignment for monophonic woodwind instruments. We then introduce a straightforward global correction strategy to compensate for these systematic shifts. Finally, we demonstrate that both the observed shifts and the proposed correction generalise across instrument types, performances, recording conditions, annotators, and datasets.

5.2.1 Dataset of Monophonic Music Performances

Recordings. We consider eight recordings performed on different woodwind instruments (Tab. 5.2): not only 'flute-family' (modern flute, traverso, recorder) but also the oboe. The recordings range from 1973 to 2020.

For Verbruggen's recording, we only know that she used a set of recorders, which is a common practice to change between soprano, alto, or tenor recorder depending on the pitch range and

performer	year	instrument	hyperlinks
Jean-Pierre Rampal	1973	Flute	YT / MBID
Barthold Kuijken	1978	Traverso	YT / MBID
Marion Verbruggen	1997	Recorders	YT
François Leleux	1997	Oboe	YT
Marek Nahajowski	2014	Alto recorder	YT
Thomas Indermühle	2015	Oboe	YT
Emmanuel Pahud	2018	Flute	YT / MBID
Katja Pitelina	2020	Traverso	YT

Table 5.2: Overview of the subset of Telemann's *Twelve fantasias* audio recordings used in this Chapter (🎧 in Tab. 4.3).

key of each fantasia. For Nahajowski, the use of a single instrument (alto recorder in F) is explicitly mentioned in the audio booklet.

Fantasia No.	1	2	3	4	5	6	7	8	9	10	11	12
Original Keys	A	a	b	Bb	C	d	D	e	E	f#	G	g
Nahajowski	m3	m3	m3	P4	P4	m3	m3	m3	m3	m3	m3	P4
Verbruggen	m3	–	–	P4	P4	m3	–	m3	m3	m3	m3	m3

Table 5.3: Transpositions for Recorders performances. All transpositions are ascending intervals: m3 minor third, P4 perfect fourth.

Although our dataset is situated in the context of Baroque music, it exhibits a wide diversity of articulation and tempi across woodwind performances, making it suitable for evaluating generalisation.

Measure Annotations. For all fantasias in each performance, a single musically trained annotator–a flautist (A1)–creates measure annotations by marking measure onset positions using *Sonic Visualiser*. In total, 14 602 measures are annotated (ranging from 1 592 for Kuijken to 1887 for Pahud, Rampal, and Leleux). Only fantasias No. 1, 8, and 11 share the same structural organisation across all performances. As repeats vary across recordings, annotators must carefully check the performed structure before annotating, resulting in nearly 7.5 hours of annotated music (Fig. 5.6).

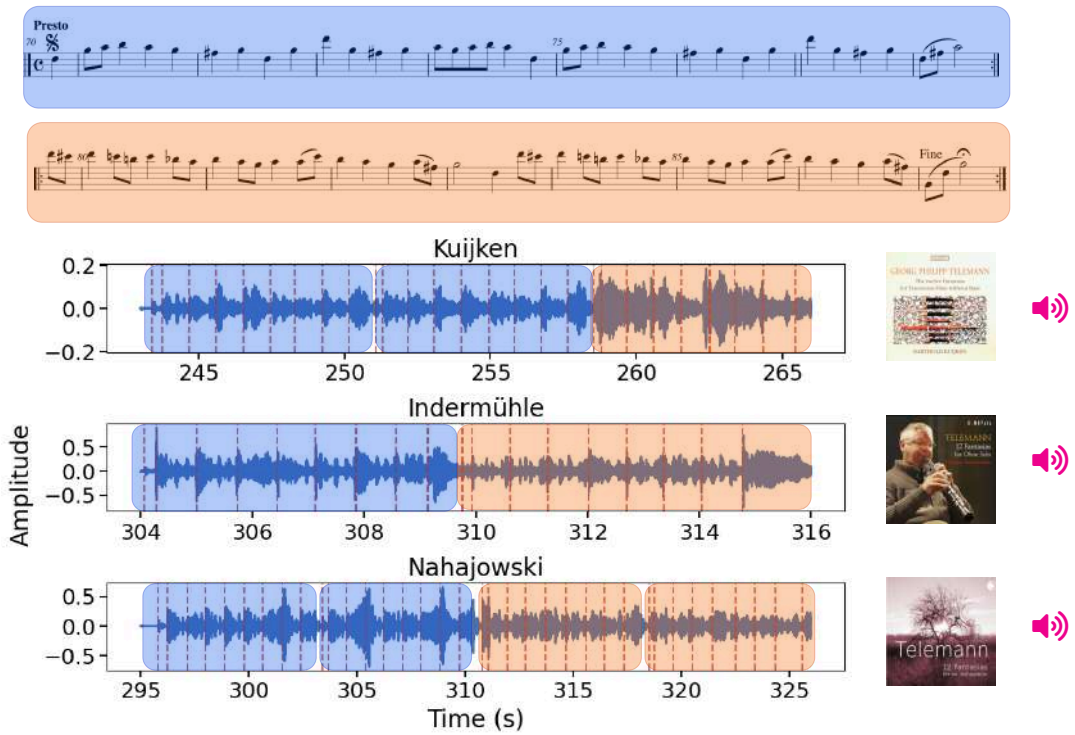


Figure 5.6: Three different realisations of the final repeats in the closing section (following the *Dal Segno al Fine* instruction) of the last movement (*Rigaudon*) from Fantasia No. 12 (TWV 40:13). Highlighted segments illustrate structural variants (`AAB`, `AB`, and `AABB`), with measure annotations overlaid on the waveforms.

The initial annotations were carried out at 80 % playback speed and subsequently refined in a second pass at 30 % speed with increased temporal resolution, enabling precise alignment of measure boundaries with acoustic landmarks observed in the waveform. This iterative process ensured accurate synchronisation between perceptual judgment and visual analysis of the signal.

While relying on a single annotator ensures internal consistency across all recordings, human annotation inevitably involves a degree of subjectivity, which limits the generalisability of the results. To study cross-annotator generalisation (Section 5.2.5), we included a second set of annotations for Fantasia No. 1 across all performances, done by a second annotator (A2) from a different country and with a different woodwind instrument background (clarinet and oboe). To ensure independence between the two annotators, A2 did not receive detailed methodological instructions regarding measure annotation. As a result, a total of $88 \times 8 = 704$ annotations

were produced independently of A1. After receiving the assigned subset, A2 was asked to describe the annotation workflow. The process is reported as a primarily perceptual, two-pass procedure, very similar to A1: an initial pass at near-real-time speed (not below 90 %) to locate approximate measure boundaries, followed by a second pass in which each position is refined individually at reduced speed (70–50 %). A2 adopts a more perceptually driven strategy with less dependence on visual features, mainly because the recordings are stereo and visual boundaries may differ slightly across channels. Interestingly, both annotators make similar subjective decisions regarding ornamentations, generally placing them on the beat unless their duration makes a pre-beat interpretation perceptually more plausible. Finally, A2 follows a structured order of annotation across instruments (starting with oboe, then flute, recorder, and ending with traverso), guided by an increasing difficulty in onset perception. All the annotations are available at https://github.com/patthmus/phd-manuscript-links/tree/main/audiofeatures-sec5.2/annotations.

5.2.2 Alignment Score-to-Audio

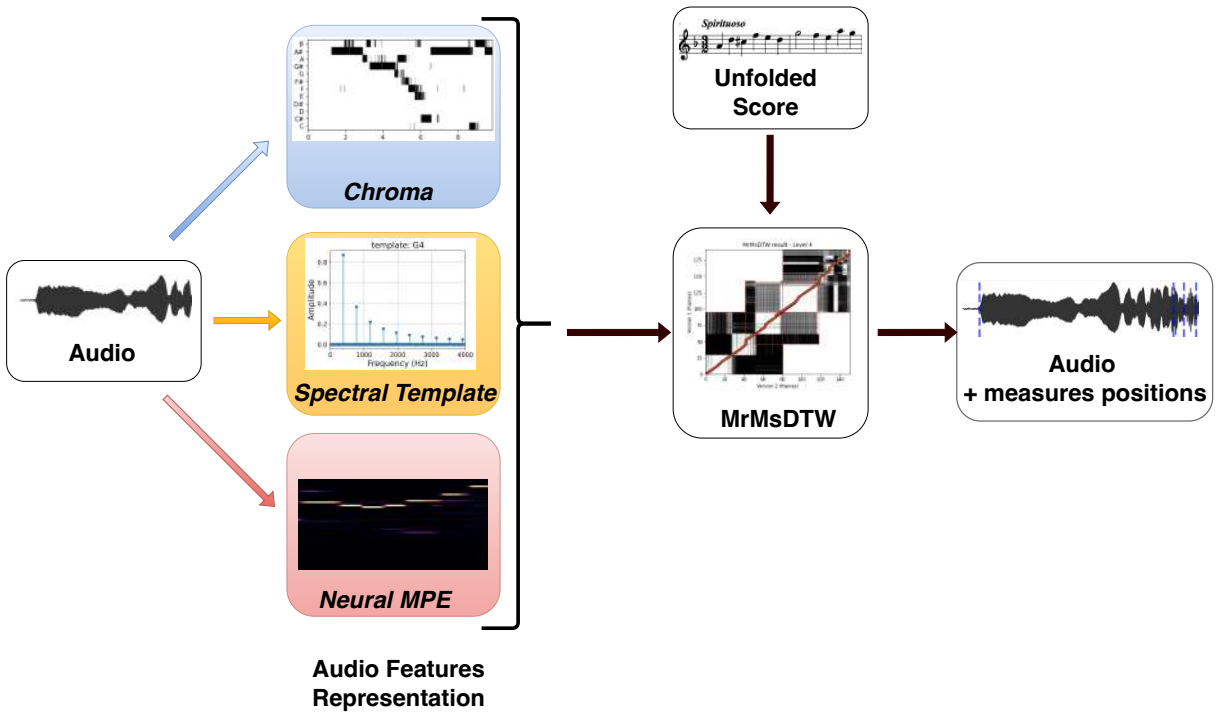


Figure 5.7: Score-to-audio pipeline used for comparing *Chroma*, *Spectral Template*, and *Neural MPE* audio features representations.

We propose a score-to-audio alignment framework between audio recordings and symbolic

unfolded scores using three audio feature representations (Fig. 5.7). Both modalities are projected into a shared feature space derived from descriptors computed with *SyncToolbox*. Spectrograms are computed at 22.05 kHz using a 1024-sample window and 512-sample hop. Prior to alignment, an estimated global tuning correction ensures pitch consistency between audio and score representations. A frame-wise cost matrix measuring the similarity between audio and score features is processed by a DTW-based algorithm (MrMsDTW) to estimate an optimal alignment path under temporal continuity constraints. No onset-based (DLNCO) features are used in the alignment stage to ensure a controlled comparison of audio feature representations. The three audio feature representations considered in this study are described in the following section.

Audio Feature Representations

Chroma refers to chroma-based features implemented in *SyncToolbox* (see Section 3.4.2), computed using an Infinite Impulse Response Filter Bank (IIR) and projected onto the 12 pitch classes. The resulting chroma vectors are L1-normalised and quantised.

Spectral Template is built on spectral templates adapted to the modern flute. We first compute a linear-frequency spectrogram via STFT, and for each candidate fundamental frequency, we generate discrete spectral flute templates as Gaussians centred at the harmonic frequencies nf_0 ($n = 1, \ldots, 10$). The amplitude of each partial follows a power-law decay of the form $f^{-\alpha}$, where the decay exponent is fixed to $\alpha = 1.25$ (see Section 5.1.1). The similarity between templates and audio frames produces a note likelihood matrix, projected onto 12 pitch classes, normalised and quantised.

Neural MPE is derived from a neural network trained for multi-pitch estimation on three multi-version datasets comprising different instrumentations (strings, piano, singing) as outlined in [Venohr et al. (2025); Weiß et al. (2021a); Zeitler et al. (2024b)]. The architecture is based on a convolutional neural network with additional residual connections (ResNet) as described in [Weiß and Peeters (2022)]. It produces a 72-semitone vector representing probabilistic pitch activations per frame (frame rate: 43.07 Hz). We project the normalised and quantised output onto 12 pitch classes.

Symbolic Features and Alignment

To obtain corresponding features for the symbolic score, we extract note events from *MusicXML* files using *music21* [Cuthbert and Ariza (2010)], based on unfolded scores. Symbolic note durations are converted into a continuous time axis using cumulative timing information. The resulting time representation is discretised to match the temporal resolution of the audio features, ensuring computational efficiency during alignment. MIDI note numbers are then mapped to 12-dimensional chroma vectors at each time step. The resulting symbolic feature matrices are directly comparable to the audio-derived *Chroma*, *Spectral Template*, and *Neural*

	Flute		Oboe		Traverso		Recorder	
Model	Pahud	Rampal	Leleux	Indermühle	Kuijken	Pitelina	Verbruggen	Nahajowski
Chroma	-6 ± 22	-6 ± 20	-4 ± 19	-13 ± 16	-6 ± 28	-9 ± 26	-3 ± 23	-5 ± 27
Spectral Template	5 ± 22	4 ± 19	5 ± 16	$-2 \pm \mathbf{14}$	7 ± 25	3 ± 20	7 ± 19	7 ± 20
Neural MPE	$-11 \pm \mathbf{17}$	$-12 \pm \mathbf{16}$	$-6 \pm \mathbf{14}$	$-11 \pm \mathbf{14}$	$-13 \pm \mathbf{20}$	$-13 \pm \mathbf{19}$	$-9 \pm \mathbf{16}$	$-12 \pm \mathbf{16}$

Table 5.4: Median alignment **signed** errors (in ms) and interquartile ranges expressed as median ± IQR/2 for the three audio feature representations across performers and their instruments. For each performer, bold indicates lowest **IQR/2 values** and gray-shading indicates the median closest to zero.

MPE representations. We compute this comparison using chroma-based features, where cosine distance is used to construct the cost matrix, enabling score-to-audio alignment via MrMsDTW. Among the parameters of MrMsDTW, the step weights are by default $(1.5, 1.5, 2.0)$ and the memory threshold is set to $\tau = 10^6$.

The following sections present a systematic experimental evaluation of the proposed alignment framework, structured in three stages. First, we compare the alignment accuracy of different audio feature representations (Section 5.2.3). Second, we investigate the presence of systematic temporal shifts in alignment errors and evaluate different correction strategies (Section 5.2.4). Finally, we examine the robustness and generalisation of these findings across annotators and datasets (Section 5.2.5).

5.2.3 Audio Features Evaluation

For each recording and each audio feature representation, we evaluate alignment accuracy by computing the deviation between the onset times of the first beats of each measure m estimated by the alignment procedure $t^{(m)}_{\text{align}}$ and those obtained from annotations $t^{(m)}_{\text{manual}}$. Thus, we obtain both signed errors ϵ and absolute errors $|\epsilon|$ where $\epsilon(m) = t^{(m)}_{\text{align}} - t^{(m)}_{\text{manual}}$.

We present comparative statistics of signed alignment errors in Table 5.4, complemented by boxplots in Fig. 5.8. Overall, all representations lead to a high alignment precision, given that frame rates are in the order of 43.07 Hz, i. e., frame-to-frame distance is already around 23 ms, but the higher IQR/2 values point to a substantial amount of outliers. On closer inspection, the results reveal a clear trade-off between median accuracy and robustness across feature representations.

Spectral Template and *Chroma* achieve the lowest median deviations depending on the recordings (from −6 to 5 ms). In particular, *Chroma* tends to yield median errors closer to zero for recorder (−3 and −5 ms), while *Spectral Template* performs best for flute (4 and 5 ms). This indicates that classical signal-processing-based representations remain competitive for woodwind alignment and can be well adapted across instrument families.

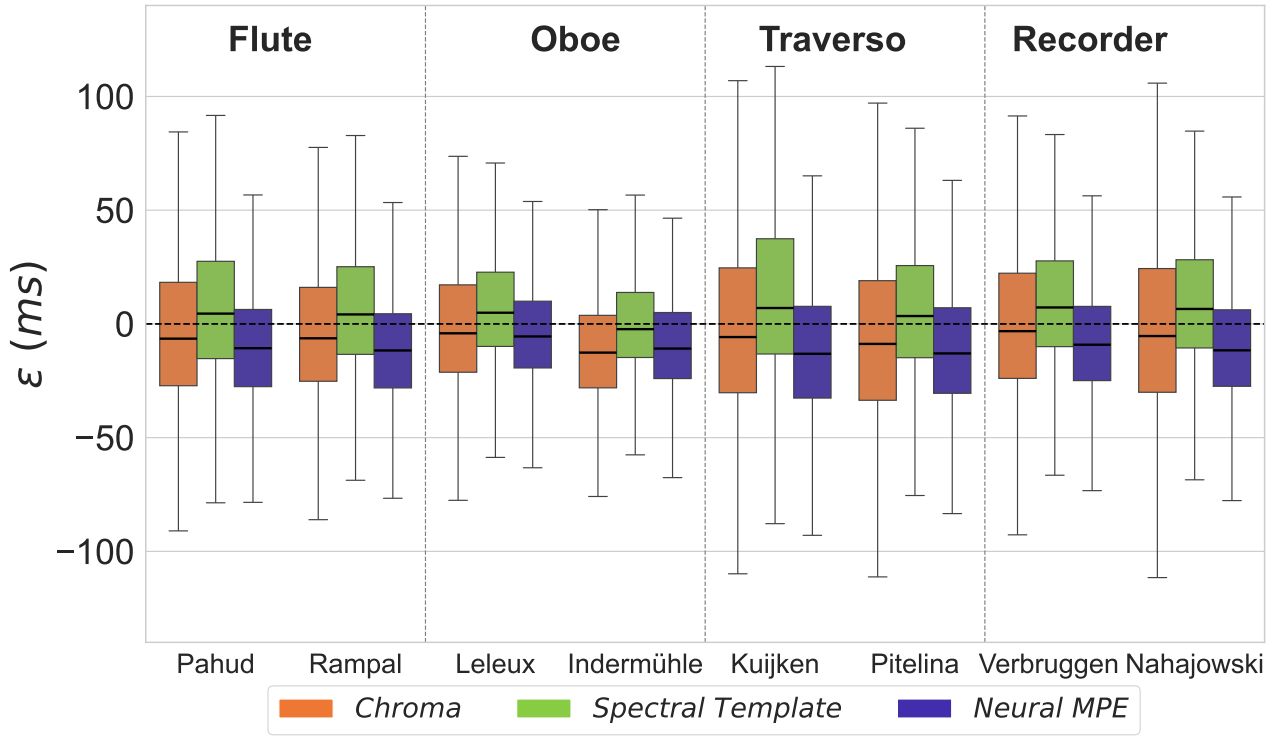


Figure 5.8: Box plots of the signed alignment errors grouped by performer and by audio feature representation for Telemann's fantasias recordings. The boxes span from the first to the third quartile, while the whiskers extend to the most extreme data points within 1.5 times IQR. A horizontal line indicates the median.

In contrast, *Neural MPE* representation systematically achieves the lowest inter-quartile ranges across all instruments (from 14 to 20 ms), indicating more stable and consistent alignment behaviour despite slightly larger median deviations in most cases. This suggests that learned representations provide improved temporal robustness, which may come at the expense of a systematic shift. Importantly, this representation exhibits a consistent shift pattern across all instruments, with predominantly negative median deviations (from −6 to −13 ms). This indicates a systematic tendency toward earlier alignment compared to the reference annotations, which provides promising opportunities for improvement through shift correction.

When comparing the direction of median shifts across representations, a structured ordering emerges: *Chroma* tends to produce small negative or near-zero shifts (average signed median across all recordings: −6.8 ms), *Spectral Template* shows slightly positive shifts in several cases (4.1 ms), while *Neural MPE* consistently introduces negative shifts (−10.5 ms). This ordering suggests that feature design influences not only error magnitude but also the direction of

Shift Types	unshifted	track	performer	instrument	representation
Pahud	-11 ± 17	0 ± 17	0 ± 17	1 ± 17	0 ± 17
Rampal	-12 ± 16	0 ± 16	0 ± 16	0 ± 16	-1 ± 16
Leleux	-6 ± 14	0 ± 14	0 ± 15	3 ± 15	5 ± 15
Indermühle	-11 ± 14	0 ± 15	0 ± 15	-2 ± 15	0 ± 15
Kuijken	-13 ± 20	0 ± 19	0 ± 20	0 ± 20	-3 ± 20
Pitelina	-13 ± 19	0 ± 19	0 ± 19	0 ± 19	-2 ± 19
Verbruggen	-9 ± 16	0 ± 16	0 ± 16	1 ± 16	1 ± 16
Nahajowski	-12 ± 16	0 ± 16	0 ± 17	-1 ± 17	-1 ± 17

Table 5.5: Median $\pm$ IQR/2 (in ms) of **signed** alignment errors for *Neural MPE* representation, by performer and shift type.

systematic temporal deviations. While such median deviations may appear small in absolute terms, they are significant in microtiming analysis, where even offsets of a few milliseconds can distort the interpretation of expressive timing and performer-specific articulation strategies. This motivates explicit modelling and correction of these deviations, which we investigate in the following section.

5.2.4 Optimal Shift Investigation

To assess the impact of systematic temporal shifts and estimate an upper-bound alignment accuracy, we conduct an oracle-style experiment where we simulate four shift correction strategies based on median estimates computed at different levels: track, performer, instrument, and representation. The **track-based** approach assigns a different shift to each fantasia recording (96 in total), whereas the **performer-based** one uses a single shift consistently across all tracks of a given performer (8 shifts per representation). The **instrument-based** strategy defines a specific shift per instrument (4). Finally, the **representation-based** one relies on a single median over all tracks and performers within the fantasias corpus for one audio feature (1 shift per representation). We first analyse the distribution of signed errors across performers for *Neural MPE*, followed by a discussion of absolute errors across all feature representations and performers.

The results, reported in Table 5.5, first show that the **track-based shifts** remove all systematic deviations. After this shift, the median error for each performer is exactly 0 by construction, but the $IQR/2$ remains unchanged. Knowing the exact track-level shifts provides an upper bound on the achievable alignment accuracy.

The **performer-based shifts** also subtract the median deviations for each performer. While performer-specific shifts are removed, IQR does not change compared to the initial alignment.

	Chroma		*Spectral Template*		*Neural MPE*	
Shift Type	none	rep.	none	rep.	none	rep.
Pahud	24±20	23±19	22±18	22±17	21±15	17 ±**15**
Rampal	22±18	21±18	18±14	19±14	19±14	17 ±**11**
Leleux	20±18	18±18	16±13	16±**12**	16±**12**	15 ±13
Indermühle	21±16	17±16	14 ±13	16±13	18±12	14 ±**11**
Kuijken	28±26	27±25	22±24	22±23	24±17	20 ±**16**
Pitelina	29±24	26±24	20±17	20±17	24±17	19 ±**16**
Verbruggen	23±23	23±23	19±17	19±17	19±14	16 ±**12**
Nahajowski	27±25	26±26	19±16	19±16	21±14	17 ±**13**

Table 5.6: Median ± IQR/2 (in ms) of **absolute** alignment errors by performer. For each representation, we show unshifted (none) and representation-shifted (rep.) results.

This strategy yields similar values while being simpler to apply than a track-based shift, but it still needs to estimate a specific shift for each performer.

With an **instrument-specific shift**, the same shift is used for all tracks of a given instrument. The medians remain close to 0, and the IQR values do not change. However, the approach appears less meaningful, as no instrument-specific trend is observed in our initial results.

Finally, a single **representation-based shift** substantially reduces systematic deviations across recordings, leading to consistently improved alignment results. After applying this correction, the *Neural MPE* representation achieves the best overall alignment accuracy in terms of both median error and robustness. This global correction (10.5 ms), independent of the corpus or individual tracks, is simple to apply and proves effective (e.g., from −11 to 0 ms for Pahud), although some residual deviations remain in certain cases (e.g., from −6 before to 5 ms after for Leleux).

Taking the behaviour of the individual audio features into account allows the removal of most systematic alignment errors while maintaining manageable precision, without requiring track-level shift correction. Such a correction is particularly important for microtiming-sensitive applications, as removing a global temporal shift is a prerequisite to ensure that observed timing deviations reflect musical expression rather than artifacts of the alignment process. This is a promising finding, from which we conclude that high-quality alignment can be obtained by combining our pipeline based on *Neural MPE* features with a shift correction that can be derived from a small, manually annotated fine-tuning dataset.

The study of the absolute errors (Tab. 5.6) across all representations confirms these conclusions. For each performer, the absolute alignment errors and their corresponding spread decrease when applying a representation-specific shift on the *Neural MPE* output (a gain of

at least 4 ms for 5 performers). The advantage of applying this method to the other audio features has a much weaker effect (no gain for *Spectral Template*, a few for *Chroma*).

5.2.5 Annotator Agreement and Robustness of Alignment Shift

To test the robustness of the observed alignment shift, we include an additional annotator A2 (see Section 5.2.5) and then extend the analysis to an independent dataset.

Cross-Annotator Evaluation

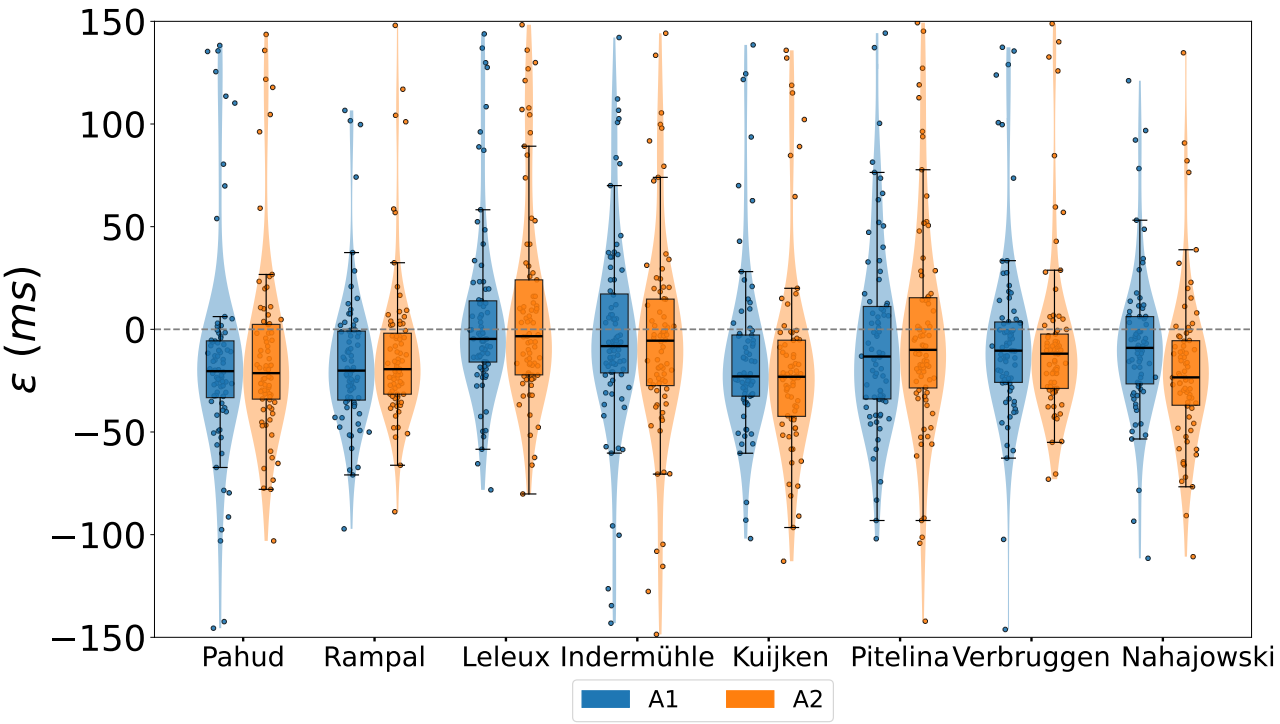


Figure 5.9: *Neural MPE* alignment signed errors for Fantasia No. 1 across all performers and annotators (A1, A2). Alongside the boxplots, violin plots depict error distributions, and jittered points reveal individual alignment deviations. Values are clipped to ±150 *ms* before visualisation.

To ensure a consistent comparison between different annotators (A1 and A2), we select Fantasia No. 1 (TWV 40:02), as its performed repeats exhibit a consistent musical measure structure across all recordings. For each annotator, 88 measure-level alignment errors are computed.

Figure 5.9 presents *Neural MPE* results. Certain outliers in the error distributions can be explained by musical characteristics such as ornamentations, rests at the beginning of measures, or octave leaps, which can locally challenge both automatic alignment and manual annotation consistency. To improve readability, we restrict the visualisation to the range ± 150 ms, excluding extreme deviations (approximately 10 events per distribution). Results are consistent across performers, with a median signed error of -11 ± 23 ms when aggregating A1 and A2 annotations. This average shift closely matches the shift computed over the entire corpus using annotator A1. When separating annotators, A1 yields -10 ± 23 ms and A2 yields -12 ± 23 ms for Fantasia No. 1, indicating a high level of agreement between annotators despite local deviations induced by expressive or structurally complex passages.

Building on these results, we show that the previously observed shift in alignment errors is not annotator-dependent. Indeed, the very close agreement between A1 and A2, both in terms of median error and IQR, indicates that the systematic deviations observed in earlier experiments are not driven by annotation variability. Instead, it reflects a consistent behaviour of the alignment method—based on a specific feature type—itself. The near-identical statistics across annotators therefore support the robustness of the reported shift, suggesting that it originates from the audio feature representation or alignment procedure rather than from subjective annotation differences.

Cross-Dataset Evaluation

To evaluate the generalisation of the proposed alignment correction beyond the Telemann Fantasias corpus, we conduct an additional cross-dataset experiment on the *ChoraleBricks* dataset [Balke et al. (2025)]. This dataset consists of modular multitrack recordings of 10 chorales from the German sacred repertoire (including Bach, Telemann, and Vulpius), covering four-part SATB writing. It includes isolated instrumental recordings from a diverse set of wind instruments (woodwinds and brass), which can be freely combined. In this work, we focus exclusively on the flute parts.

The annotations in *ChoraleBricks* are derived from a semi-automatic pipeline based on *F0* tracking using *pYIN* [Mauch and Dixon (2014)], followed by a manual correction of octave errors and ambiguous note transitions. According to the dataset description, the resulting onset-level annotations achieve an estimated temporal precision below 50 milliseconds [Balke et al. (2025)]. This contrasts with our Fantasias' annotations, which are based on perceptual, measure-level alignment performed manually by an expert listener.

piece ID	events	none	rep.
AN1	59	-13 ± 16	-2 ± 16
BA1	55	-19 ± 10	-8 ± 10
CR1	60	-17 ± 12	-6 ± 12
DR1	36	-15 ± 8	-5 ± 8
GE1	52	-12 ± 14	-2 ± 14
GE2	43	-12 ± 8	-2 ± 8
JA1	52	-6 ± 70	5 ± 70
TE1	58	-23 ± 12	-12 ± 12
VU1	26	-17 ± 7	-7 ± 7
VU2	32	-17 ± 13	-7 ± 13
corpus	473	-16 ± 13	-5 ± 13

Table 5.7: Median ± IQR/2 (in ms) of signed alignment errors for individual flute soprano parts across the *ChoraleBricks* corpus, comparing the unshifted *Neural MPE* model with the representation-shifted version. Piece IDs are formed from the composer's abbreviation (e.g., BA for Bach) and the corresponding work number in the dataset. Although each piece includes several recordings for the Soprano, Alto, Tenor, and Bass parts, only the flute parts are reported here.
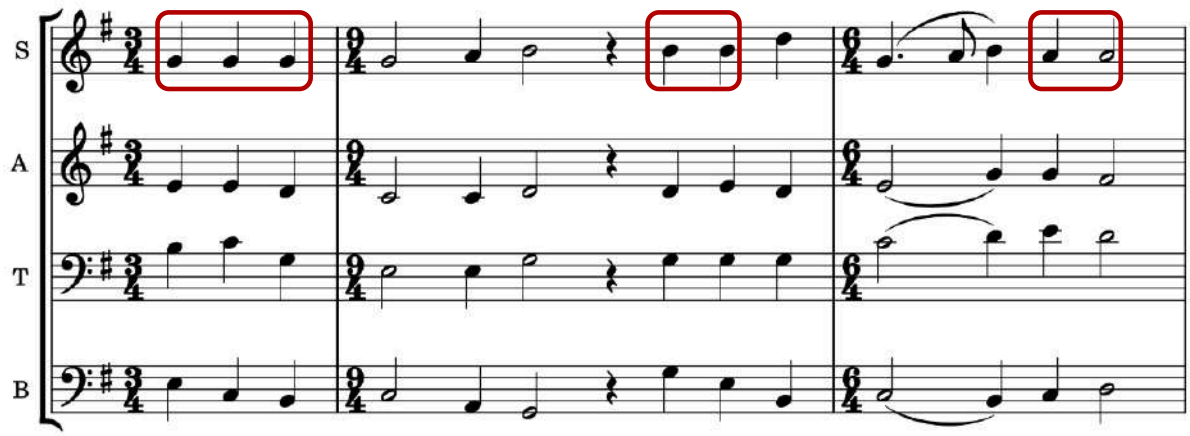


Figure 5.10: Mm.1–3 from "*Du großer Schmerzensmann*" by Martin Jan (1663). Six different recordings of the soprano part are available in the dataset (alto saxophone, clarinet, flugelhorn, flute, oboe, and trumpet). The presence of repeated notes introduces alignment ambiguities.

Therefore, *ChoraleBricks* provides an independent testbed for assessing whether the systematic temporal shift observed in our previous experiments transfers to unseen musical material and recording conditions. We apply the same representation-based shift estimated from the Fantasias corpus (e.g. 10.5 ms from A1) to all flute parts and report signed alignment errors before and after correction in Table 5.7. In contrast to our main experiments, which operate at the measure level, here we evaluate alignment at the onset level due to the limited number of events per excerpt (26–60), which nonetheless provides sufficient resolution to assess local alignment accuracy at a finer temporal scale. The results show a consistent reduction of the median error across all pieces (from −16 to −5 ms), while the IQR remains unchanged. Importantly, this improvement is observed uniformly across all excerpts, indicating that the estimated temporal shift generalises beyond the original dataset and evaluation scale. However, certain pieces, notably JA1, exhibit higher variability in the alignment errors, which persist after correction (Fig. 5.10). A closer inspection shows that outliers mostly arise from closely spaced repeated notes (19% of the onsets in JA1 are groups of 3 or 4 repeated pitches). This reflects an inherent ambiguity in the alignment problem rather than a failure of the correction itself.

Overall, these results confirm that a single representation-level temporal correction can generalise effectively across datasets and evaluation granularities (measure- and note-level), providing a simple yet robust post-processing strategy that improves alignment accuracy without requiring dataset-specific recalibration.

Section summary – Comparison of Audio Features in Score-to-Audio Alignment

In this section, we study score-to-audio alignment for monophonic woodwind instruments, where the absence of sharp onsets and polyphonic information makes temporal alignment particularly challenging. Within a unified alignment framework, we systematically evaluate three feature representations – chroma-based, flute-specific spectral template, and neural multi-pitch estimator – using Telemann's twelve Fantasias for solo flute as a case study. Among these, the neural representation achieves the lowest interquartile error, indicating the most stable performance across recordings. We show that all representations exhibit consistent, feature-dependent temporal shifts rather than random errors. Our main finding is that these systematic shifts are stable across recordings and can be accurately modelled with a simple post-processing correction, which significantly improves alignment accuracy without modifying the underlying representations. We further demonstrate that these shifts generalise across annotators and datasets, highlighting their feature representation-specific nature.

5.3 Soft-Late Fusion Strategy for Audio-to-Audio Alignments

In the case of pieces such as *Syrinx* or *Density 21.5*, direct score-to-audio alignment remains a particularly challenging task. These works frequently contain rapid melodic patterns composed of closely spaced pitches, making local distinctions between successive notes more ambiguous from a spectral perspective. In addition, tempo variations are often directly linked to these musical motifs and expressive gestures, further increasing temporal instability across performances. This discrepancy between the symbolic and acoustic domains limits the effectiveness of traditional score-following approaches based on spectral representations or harmonic templates. By contrast, audio-to-audio alignment benefits from a much stronger consistency between the compared signals: both performances share similar timbral and expressive characteristics despite interpretative differences. Breathing patterns and dynamic variations, and phrasing contours therefore, become highly informative temporal anchors for alignment algorithms such as DTW. In this context, annotation transfer through audio-to-audio alignment often proves more robust than direct score-to-audio alignment for highly expressive solo flute works.

5.3.1 Audio-to-Audio Alignment

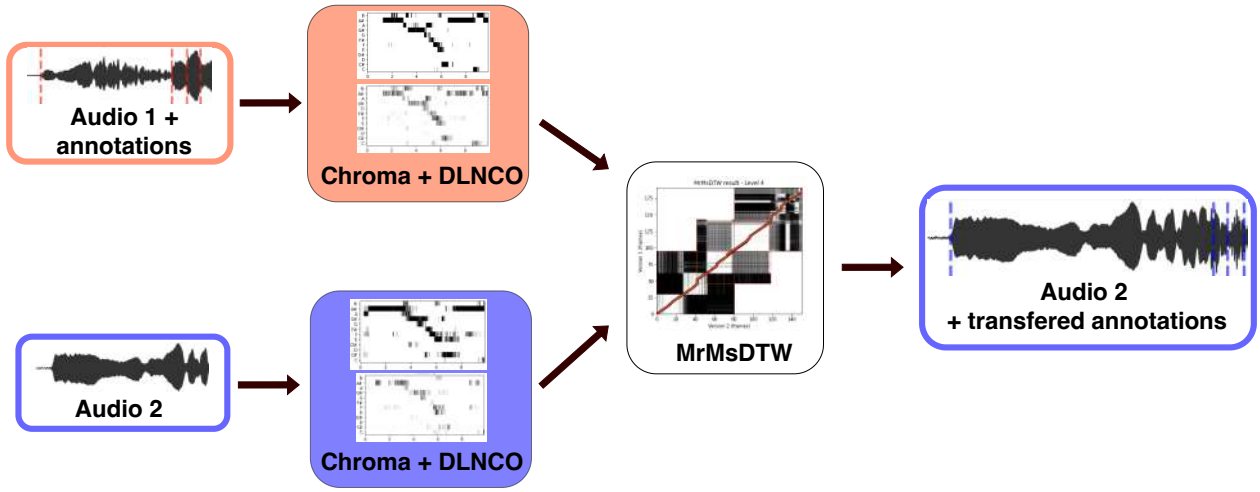


Figure 5.11: Audio-to-audio alignment pipeline.

We adopt an audio–audio alignment approach motivated by the challenges of accurately measuring expressive microtiming variations that often occur in performed music. To obtain

high-quality alignments, we choose to combine several reference annotations following the late-fusion strategy introduced by Zalkow et al. [Zalkow et al. (2017)], in which multiple bars' annotation transfers are combined based on a triple-error criterion to estimate local alignment reliability. In the present work, we investigate extensions of this framework aimed at improving the robustness of the fusion step while preserving its fully unsupervised nature not only for downbeat but for all onsets.

All alignments are performed directly on the audio signals using *SyncToolbox*. Time–frequency representations are computed at a sampling rate of 22.05 kHz with a 1 024-sample analysis window and a 512-sample hop size, followed by a tuning correction step to mitigate slight pitch deviations. From these representations, chroma features are extracted using an IIR and projected onto 12 pitch classes. In parallel, pitch-onset features are estimated and converted into DLNCO features (see Section 2.2.3) to emphasise note attacks and transitional events. The resulting features are *L1*-normalised and quantised to produce compact, scale-invariant representations. Audio–to-audio alignment is then performed using MrMsDTW (Fig. 5.11).

Triple Transfer

Let $\mathcal{T}_{R_k} = \{t_n^{R_k} \mid n = 1, \ldots, N\}$ denote the set of onset times of a reference performance R_k with $k \in \{1, 2\}$, N the total number of annotated onsets and $t_n^{R_k}$ the n-th onset time. These onset annotations are transferred by an alignment procedure to each target performance V_i where $i \in \{3, \ldots, 23\}$. This alignment procedure $R_k \longrightarrow V_i$ produces $2 \times 21 = 42$ single transfers per musical piece.

While direct transfers already provide plausible onset estimates in the target performances, individual alignments may suffer from local mismatches or performance-specific artifacts. To assess the reliability of these transfers, we adapt the triple transfer protocol proposed by Zalkow et al. (Section 3.5.2) and apply it to all available onsets.

Starting from a reference performance R_k, onset times are first transferred to a target performance V_i, then to a third and distinct unannotated version V_j, before being transferred back to the original reference:

$$R_k \longrightarrow V_i \longrightarrow V_j \longrightarrow R_k,$$

where $(i, j) \in \{3, \ldots, 23\}^2$ with $i \neq j$, yielding a total of $42 \times 20 \times 2 = 1680$ triple transfer paths per piece (Fig. 5.12).

These transfer paths serve to assess the self-consistency of the direct transfer $R_k \longrightarrow V_i$ by exploiting redundancy across multiple performances.

For a given onset o_n occurring at time t_n in the reference performance, the onset time obtained after a triple transfer via V_j is denoted $t_n^{(R_k, V_i, V_j)}$. The deviation induced by this triple

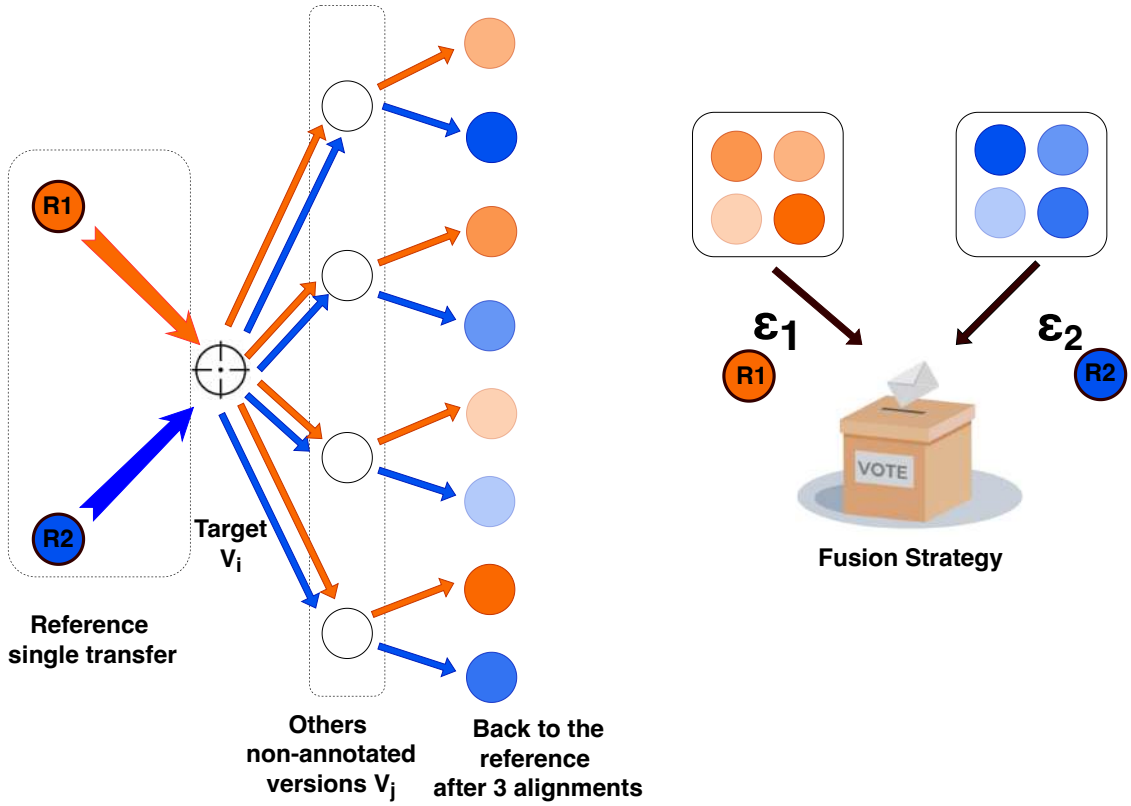


Figure 5.12: Triple transfer pipeline.

transfer is quantified by the *triple error*:

$$e_{triple}^{(R_k,V_i,V_j)}(o_n) = \left| t_n^{(R_k,V_i,V_j)} - t_n \right| \tag{5.3}$$

For each onset o_n, target performance V_i, and reference performance R_k, we summarise the reliability of the direct transfer across all intermediate versions V_j using the median triple error:

$$\epsilon_n^{(V_i,R_k)} = \text{median}_j \left(e_{triple}^{(R_k,V_i,V_j)}(o_n) \right) \tag{5.4}$$

This quantity serves as a confidence measure for the onset estimate obtained from the direct transfer $R_k \longrightarrow V_i$.

Late-Fusion Strategies

In parallel, for each target performance V_i and reference performance R_k, we obtain direct onset estimates denoted $t_n^{(V_i,R_k)}$, computed from the alignment $R_k \longrightarrow V_i$. The reliability of the

corresponding alignment is assessed using triple-transfer cycle consistency across intermediate versions V_j. These triple-transfer paths do not directly modify onset estimates but provide a robustness measure based on alignment consistency.

Hard Late-Fusion. Since two independent reference performances are available, the late-fusion strategy of [Zalkow et al. (2017)] selects, for each onset, the reference performance that yields the smallest median triple error, i.e., the most self-consistent transfer.

The final fused onset is obtained by selecting the direct alignment from the best reference:

$$R_k^* = \underset{R_k \in \{R_1, R_2\}}{\arg\min} \; \epsilon_n^{(V_i, R_k)}, \qquad \widehat{t}_n{}^{(V_i)} = t_n^{(V_i, R_k^*)} \tag{5.5}$$

Soft Late-Fusion. The hard selection of a single reference performance proposed by Zalkow et al. can be sensitive to local fluctuations of the triple-error criterion, potentially leading to suboptimal alignment decisions in specific musical passages. In practice, when transferring temporal annotations from multiple reference performances to a single target performance, no single reference consistently provides the most reliable alignment throughout the entire piece.

To alleviate this effect, we propose a soft fusion strategy that preserves the same reliability measure $\epsilon_n^{(V_i, R_k)}$ while allowing multiple reference performances to jointly contribute to the final estimate. For each onset n, we convert the median triple errors into normalised confidence weights using a *softmax* function:

$$w_n^{(V_i, R_k)} = \frac{\exp\Big(-\alpha\, \epsilon_n^{(V_i, R_k)}\Big)}{\sum_{R_k \in \{R_1, R_2\}} \exp\Big(-\alpha\, \epsilon_n^{(V_i, R_k)}\Big)} \tag{5.6}$$

where $\alpha > 0$ controls the sharpness of the weighting. In the following, we empirically chose $\alpha = 10$ and discuss this choice later in the chapter (see Section 5.3.2).

The final fused onset time is then obtained as a weighted average of the reference-specific onset estimates:

$$\widehat{t}_n{}^{(V_i)} = \sum_{R_k \in \{R_1, R_2\}} w_n^{(V_i, R_k)}\, t_n^{(V_i, R_k)} \tag{5.7}$$

This generalises the hard late-fusion strategy, which is recovered when α tends to infinity.

5.3.2 Evaluation of Alignment Quality

To evaluate annotation transfer quality and quantify the alignment accuracy gain from the proposed late-fusion strategy, we measure alignment errors against additional performance annotations E across several transfer strategies.

Strategies Comparison

Although the triple transfer error alone cannot guarantee absolute alignment accuracy – since error compensation effects may occur – it remains a useful criterion for comparing competing annotation transfers. For each onset of the evaluative performance E, we compute the absolute alignment error between a given strategy S and the reference annotation as $e_n = \left|t_n^S - t_n^E\right|$.

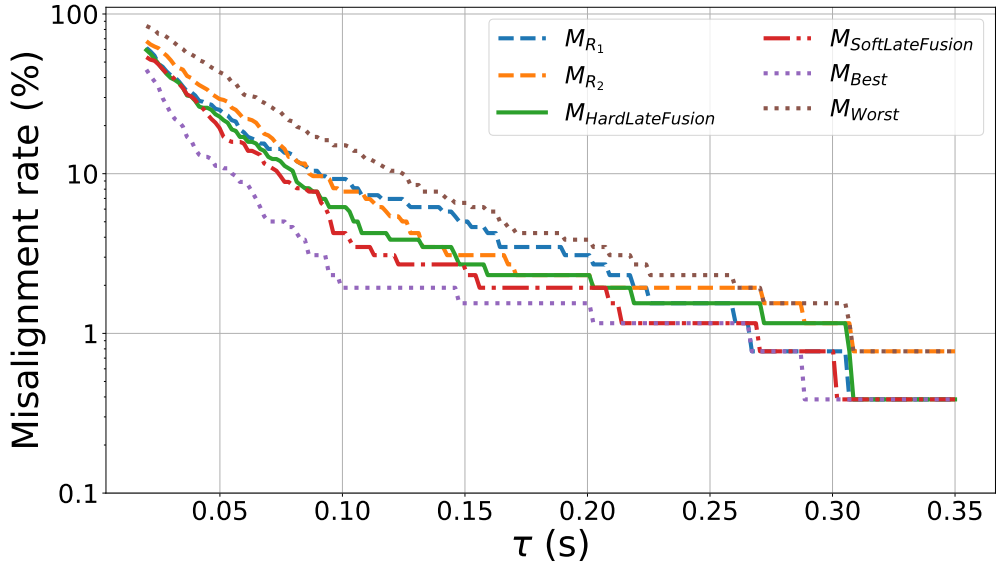
Six annotation transfer strategies are evaluated based on their *misalignment rate* M, defined as the proportion of onsets for which the error e_n exceeds a threshold τ. As illustrated in Figure 5.13, the proposed soft late-fusion strategy yields substantially fewer large errors than all non-oracle baselines. For instance, in *Density 21.5*, only 5.6% of onsets exceed a 150 ms error threshold, confirming a strong reduction of extreme misalignments.

Syrinx												
Strategy	P_{25}	P_{50}	P_{75}	P_{85}	P_{95}	P_{98}	P_{99}	$P_{99.25}$	$P_{99.75}$	$P_{99.9}$	$\bar{e}$	σ_e
R_1	13	24	50	68	147	216	263	269	350	398	42	54
R_2	15	30	56	75	124	197	296	316	737	1068	49	92
Hard-LF	11	24	46	65	104	195	286	307	351	398	38	50
Soft-LF	10	24	44	60	95	154	236	271	**326**	**353**	35	44
Worst	25	44	72	98	164	255	307	316	737	1068	64	94
Best	**7**	**17**	**29**	**39**	**70**	**98**	**229**	**268**	338	393	**26**	**41**
Density 21.5												
Strategy	P_{25}	P_{50}	P_{75}	P_{85}	P_{95}	P_{98}	P_{99}	$P_{99.25}$	$P_{99.75}$	$P_{99.9}$	$\bar{e}$	σ_e
R_1	9	27	52	84	191	238	263	274	458	614	48	70
R_2	14	31	78	117	247	372	565	576	610	630	67	101
Hard-LF	11	23	47	77	197	238	296	338	362	363	44	60
Soft-LF	11	24	55	78	165	246	273	280	314	347	44	56
Worst	26	48	98	148	281	552	582	598	671	699	87	110
Best	**6**	**15**	**36**	**52**	**99**	**201**	**211**	**215**	**234**	**235**	**29**	**40**

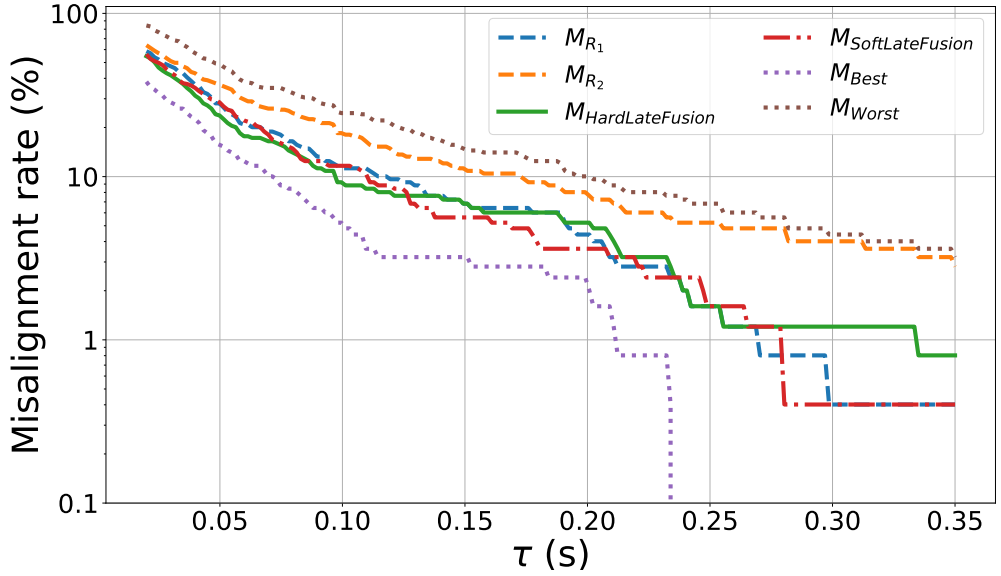
Table 5.8: Absolute alignment errors (in ms) for six strategies using Auer's performance as evaluation playing *Syrinx* and Zukerman's performance for *Density 21.5*. Columns indicate percentiles (25–99.9th), mean $\bar{e}$, and standard deviation σ_e. Cells shaded in gray show the value closest to the Best local oracle strategy.

Beyond threshold-based analysis, we further investigate the full error distributions using percentile statistics (Tab. 5.8). This analysis is conducted on both *Syrinx* and *Density 21.5*, which exhibit different levels of intrinsic alignment difficulty.

In both cases, late-fusion strategies consistently outperform single-reference transfers across most percentiles, particularly in the low and mid-range of the distribution (25th to 85th percentiles), where the majority of musical events lie. At higher percentiles (95th to 99.9th),



(a) *Syrinx*



(b) *Density 21.5*

Figure 5.13: Misalignment rates as a function of the tolerance threshold τ for six alignment strategies on (a) *Syrinx* by using Auer performance for evaluation and on (b) *Density 21.5* against Zukerman performance. M_{R_1} and M_{R_2} are single-reference baselines, $M_{HardLateFusion}$ the original late-fusion, and $M_{SoftLateFusion}$ our softmax-weighted variant. M_{Best} and M_{Worst} are local oracle strategies selecting, for each onset, the reference with the smallest and largest alignment error, respectively.

where errors are dominated by ambiguous or difficult alignment regions, soft late fusion remains systematically closer to the oracle than hard fusion and single-reference strategies. For example, in *Syrinx*, at the 99.9th percentile a gain of 0.72 s compared to direct transfer from R_2 is observed (45 ms improvment compared to Hard-Late fusion)

However, the magnitude of these gains depends on the intrinsic difficulty of the target performance. *Syrinx* exhibits relatively stable baseline alignment, with moderate dispersion even in the upper percentiles, which limits the observable improvement margin. In contrast, *Density 21.5* shows a significantly heavier error tail, reflecting a more challenging alignment scenario. In this case, the proposed fusion strategy yields stronger improvements in the upper quantiles, substantially reducing extreme errors compared to single-reference approaches.

Overall, these results indicate that the primary benefit of the proposed approach lies in its ability to control error distribution tails, particularly in difficult alignment scenarios. Rather than uniformly shifting the entire distribution, late fusion acts as a mechanism that improves robustness by reducing sensitivity to ambiguous or unstable alignment configurations.

Cross-Evaluation

Piece	**Configuration**	**References (R_1, R_2)**	**Target (E)**
Syrinx	S0	Pahud, Ferguson	Auer
	S1	Auer, Ferguson	Pahud
	S2	Auer, Pahud	Ferguson
Density 21.5	D0	Barth, Hurel	Zukerman
	D1	Barth, Zukerman	Hurel
	D2	Hurel, Zukerman	Barth

Table 5.9: Cross-evaluation configurations based on reference–target permutations for *Syrinx* and *Density 21.5*.

To complement this evaluation protocol, we further introduce a cross-evaluation scheme based on reference–target rotations. In contrast to the previous setting where a single evaluative performance E was fixed, we systematically rotate the roles of the available performances such that each recording alternately serves as reference and evaluation target. This allows us to assess the robustness of the proposed alignment strategies across different interpretative realisations and to reduce potential selection bias induced by a single target choice (Tab. 5.9).

Under this extended protocol, alignment errors are computed in the same manner as previously defined and aggregated across all reference–target configurations. The global statistics (Tab. 5.10) confirm the trends observed in the single-target evaluation, but also reveal an important additional effect on the variability of the alignment errors. While the mean error

Strategies	R_1	R_2	Hard-LF	Soft-LF	Best	Worst
Full dataset	55 ± 120	54 ± 108	44 ± 92	43 ± 86	**26 ± 79**	80 ± 136
Syrinx	50 ± 122	50 ± 125	43 ± **93**	43 ± 104	**25** ± 99	72 ± 140
Density 21.5	60 ± 117	59 ± 87	46 ± 89	45 ± 63	**27 ± 50**	89 ± 130

Piece	**Target**	R_1	R_2	Hard-LF	Soft-LF	Best	Worst
Density 21.5	Barth	67 ± 140	47 ± 62	39 ± 49	37 ± 48	**23 ± 37**	89 ± 142
	Hurel	65 ± 128	64 ± 92	54 ± 81	68 ± 80	**34 ± 68**	92 ± 137
	Zukerman	48 ± 70	67 ± 101	44 ± 57	56 ± 56	**25 ± 39**	87 ± 110
Syrinx	Auer	42 ± 54	49 ± 92	38 ± 44	44 ± 44	**23 ± 39**	64 ± 94
	Ferguson	62 ± 177	59 ± 174	57 ± 168	34 ± 165	**23 ± 165**	84 ± 182
	Pahud	46 ± 103	41 ± 89	34 ± 44	45 ± 44	**18 ± 23**	67 ± 130

Table 5.10: Alignment error statistics across strategies (top) and cross-evaluation per target performance (bottom). Values are mean ± standard deviation in ms.

decreases moderately with the proposed fusion strategies, a more consistent improvement is observed in terms of standard deviation.

In particular, the soft late-fusion strategy not only achieves the lowest mean alignment error (43 ms compared to 44 ms for hard fusion and 55 ms for single-reference transfers), but also yields a reduced variability (86 ms compared to 92 ms for hard fusion and 108–120 ms for single-reference strategies). This reduction in dispersion indicates that the fusion mechanism improves the stability of the alignment process across heterogeneous musical events, rather than only optimising average performance.

The cross-evaluation analysis further supports this observation (Tab. 5.10, lower panel). Across all reference–target configurations, standard deviations remain consistently lower for both fusion strategies compared to single-reference transfers, even in cases where mean differences are small. This suggests that the main benefit of the proposed approach lies in reducing alignment uncertainty across interpretations rather than solely improving average accuracy. Importantly, this effect is not uniform across all evaluation targets. A more detailed inspection reveals that improvements are particularly pronounced for targets exhibiting higher intrinsic alignment difficulty, as reflected by larger baseline variances and higher single-reference errors (e.g., Ferguson in *Syrinx* and Hurel in *Density 21.5*). In these cases, both hard and soft fusion strategies provide a noticeable reduction in error dispersion compared to single-reference transfers, suggesting that the fusion mechanism is especially effective when individual references are unreliable or highly variable. In the case of Ferguson recording, we know that she played live on a historical flute made by Louis Lot in Paris in 1879. This unique flute uses a "958" silver (high-purity) that was worked from a flat sheet instead of drawn in a tube [Botha (2016)].

Conversely, for targets that already exhibit strong alignment consistency (e.g., Auer or Barth), the relative gain of fusion is more limited. This suggests that the method mainly improves stability by reducing variability in the alignment and that its advantage becomes larger when the alignment task is more difficult.

Importantly, this rotational evaluation confirms that the observed improvements are not dependent on a specific evaluation target, but instead reflect a systematic property of the fusion process. The increased stability observed across heterogeneous conditions further supports the robustness of the proposed method, particularly in challenging alignment scenarios where single-reference approaches are less reliable.

Sensitivity of Parameter α

To assess the sensitivity of the proposed soft-late fusion strategy to the parameter α, we perform a systematic evaluation over a range of values $\alpha \in \{5, 10, \ldots, 60\}$.

For each configuration, we recompute alignment outputs using the same cross-evaluation protocol and measure absolute alignment errors against our annotations. Absolute mean error and standard deviation are yielded by aggregation across all pieces, rotations, and evaluation targets. Additionally, we analyse the relative *fusion gain*, defined as the difference between the best single-reference transfer and the soft-late fusion output. This quantity measures how closely the proposed strategy approaches an oracle-like local reference selection. Negative values therefore, indicate that the oracle remains superior, which is expected since it represents an ideal lower bound constructed with access to ground-truth information.

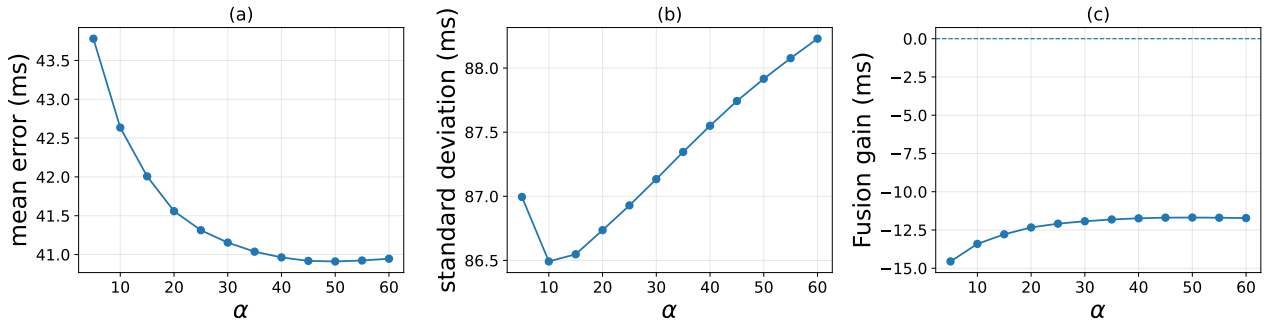


Figure 5.14: For different values of α parameter, soft-late fusion alignment : (a) absolute mean error ; (b) standard deviation error; and (c) gain against the best single-reference transfer.

The results (Fig. 5.14) show that increasing α progressively improves the mean alignment accuracy of the soft-late fusion strategy, with the average error decreasing from approximately 44 ms at $\alpha = 5$ to about 41 ms for larger α values. However, these improvements rapidly become marginal beyond $\alpha \approx 20$. At the same time, the standard deviation of the alignment

error slightly increases with larger α values. A similar tendency is observed for the fusion gain statistics: larger α values bring the fusion output marginally closer to the oracle reference selection.

These observations reflect the trade-off induced by the soft-late fusion parameter. Small α values produce smoother weighting between references, leading to more stable but slightly less selective fusion behaviour. Conversely, large α values increasingly approximate hard local selection, which can improve local accuracy while also amplifying local estimation noise and variability. Within this context, the choice $\alpha = 10$ appears to provide a suitable compromise between alignment accuracy, robustness, and numerical stability. Although slightly larger values achieve marginally lower mean errors, the observed gains remain extremely small relative to the overall error scale, while the associated variability gradually increases.

By combining onset estimates through a weighted average rather than selecting a single reference, the proposed softmax fusion can implicitly compensate for small systematic deviations, resulting in more stable onset estimates, particularly in ambiguous or expressive passages. The accuracy of our alignments allows us to produce consistent results that will support the following discussion on performance comparison.

Section summary – Soft-Late fusion Strategy for Audio-to-Audio Alignments

Building on a triple-transfer protocol by Zalkow et al., we estimated the reliability of each onset through cycle-consistency across multiple performance paths, providing a principled measure of alignment confidence. We then proposed a soft late-fusion strategy that generalises hard reference selection by combining multiple transfers in a probabilistic manner. This approach reduces sensitivity to local alignment errors and improves robustness.

Chapter summary – Improving Alignment Methods for Solo Flute

This chapter introduces and evaluates several alignment strategies for solo flute and related woodwind performances. We first developed a score-to-audio framework based on harmonic templates. We then show that, in this scenario, alignment accuracy depends strongly on the choice of audio representation, with neural features providing the most stable performance but introducing a systematic temporal shift. This shift can be effectively compensated using a simple correction.
In parallel, we addressed audio-to-audio alignment for a highly expressive repertoire using a multi-version framework based on triple-transfer consistency. We proposed a soft late-fusion strategy that improves robustness by combining multiple references, reducing extreme alignment errors.

To the question *"How can onset-level correspondences be robustly inferred across diverse performances?"*, the answer is by adapting alignment strategies to the nature of the musical representations being compared and by combining appropriate audio features, alignment models, and complementary reference information. For score-to-audio scenarios, reliable correspondences can be obtained by exploiting score-derived musical constraints. For audio-to-audio scenarios, where expressive timing and structural ambiguity make direct comparison more challenging, robustness can be improved through multi-version strategies that integrate information from several performances. Together, these approaches demonstrate that fine-grained alignment requires methods capable of accounting for both the temporal variability of performance and the heterogeneity of musical representations.

Chapter 6

Performance Analysis of Expressive Timing

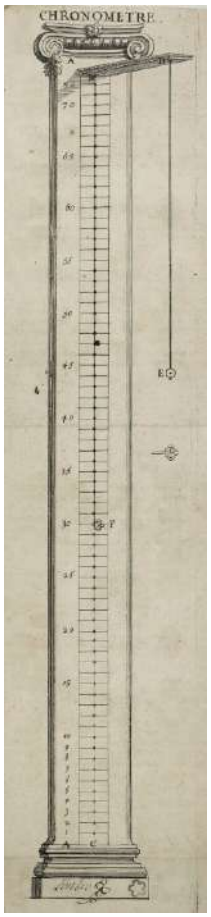


Figure 6.1: Chronomètre from [Loulié (1696)].

Contents of this chapter

6.1 Levels of Temporal Analysis 139
6.2 From Notated Rhythm to Performed Timing in Fantasias . 140
6.2.1 Rhythmic Deviation Metrics 140
6.2.2 Fantasias Across Performances 142
6.3 Timing Across Patterns in 20th Solo Flute Music 149
6.3.1 Musical Patterns Selection 149
6.3.2 Timing Descriptors 152
6.3.3 Inter-Performer Analysis of Patterns in Syrinx . . . 153
6.3.4 Timing Analysis of Patterns in Density 21.5 156

Expressive timing refers to deviations from notated temporal structures in musical performance. These deviations operate at multiple temporal scales, ranging from global tempo trajectories (rubato, accelerando, etc.) to local micro-timing variations between adjacent notes.

At the global level, tempo fluctuations have been extensively studied in performance analysis and are often used to describe expressive shaping at the phrase or section level. In contrast, much less attention has been devoted to fine-grained temporal deviations occurring within short rhythmic groupings.

How are local timing deviations structured within musical performance, and how do they interact with metric structure, voice-leading, and motivic organisation?

This interest in the precise measurement of tempo is not new. As early as the late 17th century, Loulié introduced the *chronomètre* (Fig. 6.1) as a device intended to standardise musical tempo and enable reproducible performance practices in the absence of the composer.

Despite such early attempts at formalising tempo, expressive deviations remain a fundamental and constitutive aspect of musical performance practice.

In this chapter, we focus on local timing deviations measured at the level of IOI, and analyse how they relate to different musical dimensions. Beyond the description of individual timing variations, we propose a methodological framework for analysing expressive micro-timing based on metrics that relate performed timing to musically meaningful temporal references. This approach aims to characterise how local deviations are organised rather than simply quantify their presence. The chapter is structured as follows: we first introduce a comparison between metronomic and performed timing, then we investigate timing patterns associated with pattern structures.

Research outputs associated with this chapter

When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In 26th International Society for Music Information Retrieval Conference (ISMIR), 2025

Soft Late-Fusion Alignment and Pattern-Based Performance Analysis of Early 20th-Century Solo Flute Music

Patrice Thibaud, Louis Bigo, Mathieu Giraud, Yann Teytaut

Accepted to ISMIR 2026

6.1 Levels of Temporal Analysis

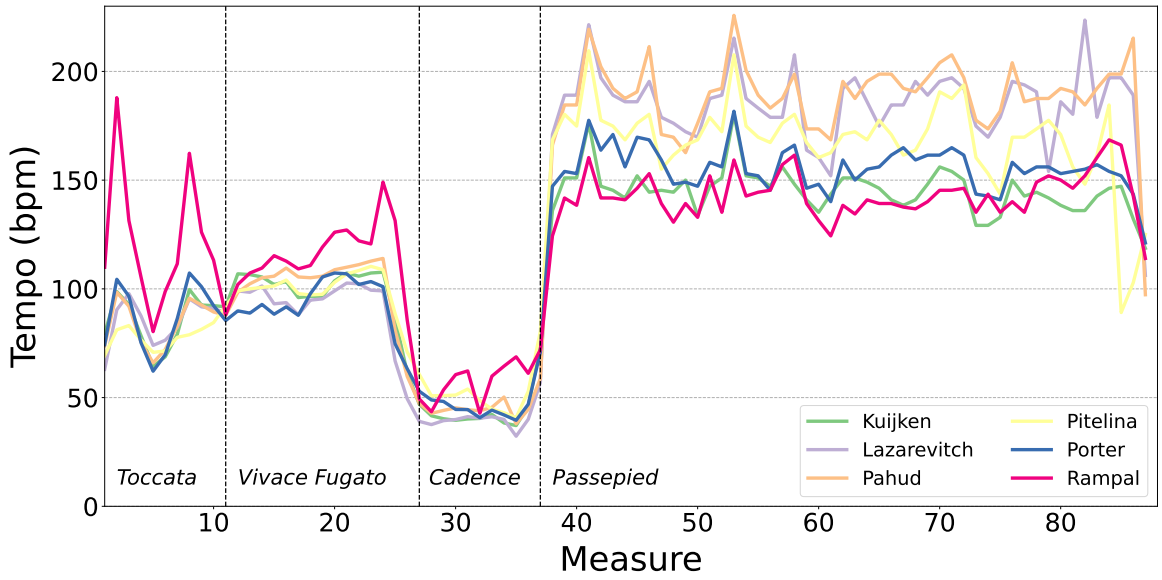


Figure 6.2: Tempo by measure in *bpm*, computed on three consecutive measures, for six recordings on traverso and modern flute of Fantasia No. 1 (TWV 40:2) (🕓 in Tab. 4.3). All performers underline the four movements of this Fantasia.

Performances of Fantasias exhibit global tempo variations (as illustrated in Fig. 6.2), including gradual slowdowns or speedups.

We can distinguish two levels of temporal organisation:

- a **global tempo level:** long-term tempo fluctuations across entire movements or sections. As here, Rampal plays faster in the *Toccata* and slower in the *Cadence*.
- a **local timing level:** fine-grained deviations between beats, within beats, or across short rhythmic groupings. At this level, the tempo curves becomes considerably more irregular, reflecting continuous micro-adjustments made by the performer. These local variations are generally associated with expressive timing strategies.

The present analysis focuses primarily on local timing deviations, allowing us to study them in relation to musical structures rather than as isolated deviations. This perspective provides a framework for investigating how expressive timing emerges from the interaction between temporal deviations and musical organisation. We first investigate the comparison between metronomic and performed versions of the Fantasias (Section 6.2) looking for strategies on beats, rhythmic structures and voices. Then, we study timing across patterns in 20th-century solo flute music (Section 6.3).

6.2 From Notated Rhythm to Performed Timing in Fantasias

Cover	Performer	Year	Instrument	Recording Time (hh:mm:ss)	Hyperlinks
	Jean-Pierre Rampal	1973	Flute	00:49:14	YT / MBID
	Barthold Kuijken	1978	Traverso	00:47:54	YT / MBID
	Amy Porter	2008	Flute	00:57:00	YT
	François Lazarevitch	2016	Traverso	00:58:50	YT / MBID
	Emmanuel Pahud	2018	Flute	01:00:44	YT / MBID
	Katja Pitelina	2020	Traverso	00:56:13	YT

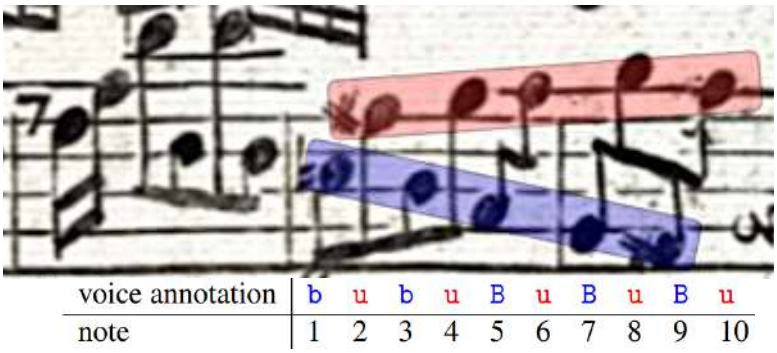
Table 6.1: Overview of the subset of the six recordings of Telemann's *Twelve Fantasias* used for the expressive timing study (see 🕓 and Tab. 4.3).

Using the subset 🕓 of Fantasias Dataset presented in Chapter 4 together with the score-to-audio alignment outputs for the corresponding six complete performances, we first introduce the metrics employed to measure timing deviations. These metrics form a general approach for analysing expressive timing by comparing performed timing with a locally defined metronomic reference. Rather than assuming deviations to be random fluctuations around the score, this framework enables the identification of systematic temporal strategies associated with rhythmic and musical organisation. We then investigate expressive timing across performances at several score levels, ranging from individual notes to larger musical structures.

6.2.1 Rhythmic Deviation Metrics

After aligning the flute and traverso performances (Tab. 6.1) to the corresponding unfolded scores (as described in Section 5.1), each note n has a *performance onset* $o_p(n)$, from which we derive a *performance IOI*, denoted as $IOI_p(n)$.

We rely on the metric structure as a reference framework for normalisation and cross-performance comparison. Thus, we choose to compare $o_p(n)$ and $IOI_p(n)$ to their theoretical counterparts, assuming no timing deviations.



voice annotation	b	u	b	u	B	u	B	u	B	u
note	1	2	3	4	5	6	7	8	9	10

Figure 6.3: The second episode of the *Vivace fugato* movement of Fantasia No. 2 (TWV 40:3) begins at mm. 38. It features two interleaved voices, a descending bass line and an upper voice 🔊. In the autograph score, Telemann occasionally highlights the bass voice in these measures by using stems pointing in opposite directions within the same beam, as here on notes 5, 7, and 9.

To account for this, we use *each measure* as a reference point for the tempo. For each note n, we thus estimate its *metronomic onset* $o_m(n)$, assuming that the entire measure $M(n)$ encompassing it is played at a constant tempo:

$$o_m(n) = o_m(n_1) + \sigma(n)\big(o_m(n_2) - o_m(n_1)\big) \tag{6.1}$$

where n_1 and n_2 are the first beats of the measures $M(n)$ and $M(n)+1$, and $\sigma(n) \in [0,1[$ represents the *symbolic position* ratio of n within this measure. We also compute the *metronomic IOI* as $IOI_m(n)$. The deviations in IOI and onset timing are then computed:

$$\Delta IOI(n) = \frac{IOI_p(n) - IOI_m(n)}{\mathrm{dur}(M(n))} \tag{6.2}$$

$$\Delta o(n) = \frac{o_p(n) - o_m(n)}{\mathrm{dur}(M(n))} \tag{6.3}$$

where $\mathrm{dur}(M(n))$ is the duration of the enclosing measure $M(n)$.

For example, in measures 38 and 39 of Fantasia No. 2 (Fig. 6.3), Porter plays the second eighth note 50 ms later than its metronomic position (Fig. 6.4a). This delay corresponds to $\Delta IOI(n_1) = \Delta o(n_2) = 3.3\,\%$ of the measure duration. Note that, by definition, $\Delta o(n_1) = 0$ on the first note (strong beat).

The inequality ratio is here 1.30 between the two firsts eighth (`C`: note 1 and `F#`: note 2), indicating that the first inter-onset interval is 30 % longer than it would be under a perfectly even subdivision. Because the ratio is computed from successive onset times, it is sensitive to alignment inaccuracies. Small errors in the estimated note boundaries may have a limited

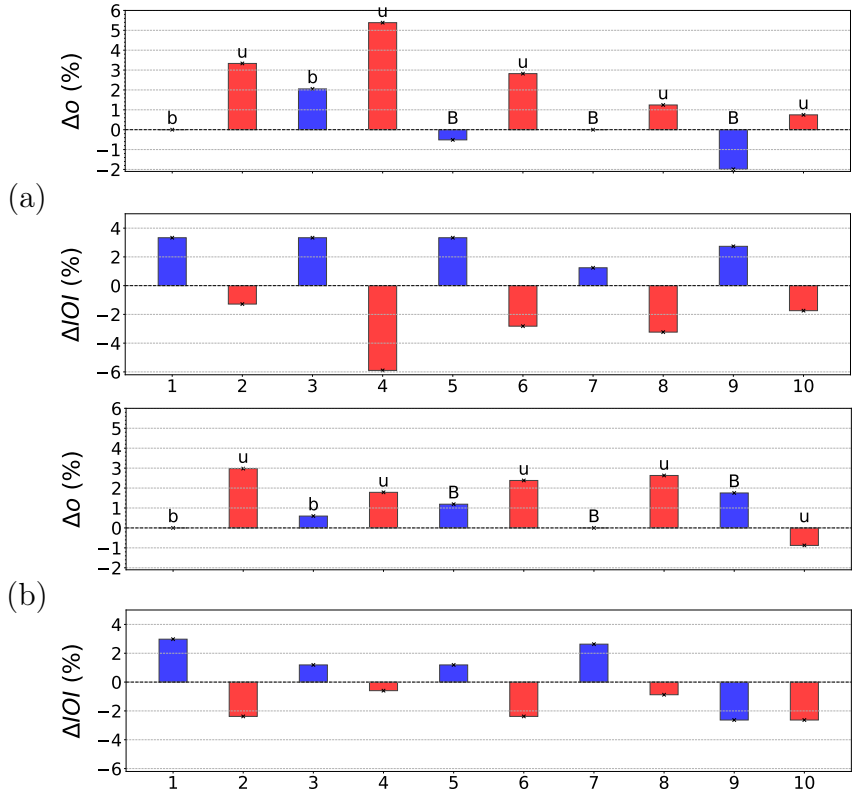


Figure 6.4: The ΔIOI and Δo values for the passage shown in the Figure 6.3 are compared across two performances on a modern flute. In Porter's performance (a), the average over absolute values of ΔIOI is 2.9 %. In contrast, Rampal (b), who plays this movement at a faster tempo, has less deviations (1.9 %).

effect on long note values but can noticeably alter the ratio for short durations. Consequently, isolated extreme ratio values should be interpreted with caution, especially when they occur near regions where onset detection or score-to-audio alignment is less reliable.

Nevertheless, when aggregated across many note occurrences and performances, the ratio provides a robust indicator of systematic expressive timing strategies.

6.2.2 Fantasias Across Performances

We now examine how the defined metric measures vary across performances of the Fantasias to identify systematic expressive timing behaviours shared by performers or associated with particular score contexts. We focus more specifically on eighth and sixteenth notes. Notes were grouped according to several score-derived criteria, including on-beat/off-beat position, interleaved voice annotation, metre, position within the beat or combinations of these.

		annotated voices					on beat		off beat	
performer	x	b	B	b+ B	u	m	all	♪	all	♪
Rampal	-0.27 ± 9.64	-0.47 ± 6.95	-0.14 ± 5.85	-0.29 ± 6.36	0.26 ± 7.15	-0.60 ± 7.30	-0.53 ± 10.63	-0.28 ± 9.47	0.05 ± 7.46	0.26 ± 9.26
Kuijken	0.07 ± 9.23	-0.17 ± 4.59	0.85 ± 5.46	0.40 ± 5.12	0.52 ± 7.06	-0.68 ± 7.57	0.69 ± 10.40	**1.13** ± 9.19	-0.29 ± 6.75	0.09 ± 8.49
Porter	-0.09 ± 9.78	0.50 ± 5.26	-0.09 ± 5.65	0.17 ± 5.49	0.20 ± 7.06	-0.57 ± 8.22	0.01 ± 10.90	0.88 ± 9.49	-0.07 ± 7.25	0.28 ± 9.00
Lazarevitch	-0.19 ± 10.76	-0.66 ± 5.43	0.26 ± 6.31	-0.14 ± 5.95	**1.28** ± 8.54	**-1.45** ± 8.57	0.79 ± 12.02	**1.55** ± 10.90	-0.73 ± 8.03	-0.90 ± 9.71
Pahud	-0.14 ± 9.15	0.05 ± 4.70	**1.20** ± 5.30	0.69 ± 5.07	0.35 ± 7.42	**-1.26** ± 7.66	0.73 ± 10.43	**1.29** ± 8.81	-0.62 ± 6.59	-0.06 ± 8.20
Pitelina	0.00 ± 9.80	**-1.23** ± 4.87	**-2.32** ± 6.55	**-1.84** ± 5.89	**1.96** ± 8.58	-0.28 ± 7.61	-0.26 ± 11.20	0.45 ± 10.09	0.23 ± 7.38	0.62 ± 9.31

	♫ in duple metre				♪♪♪♪ in duple metre			
performer	1st♪ interleaved	2nd♪ interleaved	1st♪ x	2nd♪ x	1st♪	2nd♪	3rd♪	4th♪
Rampal	-0.71 ± 6.28	0.26 ± 7.82	-0.38 ± 10.55	-0.10 ± 10.72	-0.07 ± 3.42	-0.02 ± 3.65	-0.54 ± 4.26	0.24 ± 4.10
Kuijken	0.56 ± 4.22	-0.29 ± 4.74	0.76 ± 10.95	-0.42 ± 9.56	**1.23** ± 3.84	-0.55 ± 2.99	-0.60 ± 3.49	-0.15 ± 3.68 7
Porter	**1.21** ± 5.35	-0.63 ± 5.35	0.13 ± 10.49	0.43 ± 10.49	0.52 ± 3.31	-0.53 ± 2.94	**-1.14** ± 2.79	0.51 ± 3.41 2
Lazarevitch	**1.72** ± 6.62	**-1.42** ± 6.08	**1.12** ± 11.79	**-1.52** ± 11.00	**1.45** ± 4.05	**-1.01** ± 3.68	-0.16 ± 4.55	-0.65 ± 4.97
Pahud	0.29 ± 4.29	-0.33 ± 4.07	0.01 ± 9.86	0.02 ± 10.24	**1.58** ± 4.19	**-1.06** ± 2.62	-0.90 ± 2.66	-0.34 ± 3.20
Pitelina	**1.21** ± 8.73	-0.75 ± 6.91	0.03 ± 9.91	**1.17** ± 11.12	0.56 ± 3.84	-0.43 ± 3.29	-0.36 ± 4.10	0.67 ± 4.70

	♪♪♪ in triple metre (Fig. 6.6)		
performer	1st♪	2nd♪	3rd♪
Rampal	-0.36 ± 9.22	-0.91 ± 7.69	0.61 ± 7.03
Kuijken	0.71 ± 6.10	-0.74 ± 7.40	0.78 ± 9.27
Porter	-0.03 ± 6.52	-0.43 ± 8.02	0.49 ± 7.12
Lazarevitch	**2.16** ± 10.09	**-2.99** ± 8.78	0.75 ± 9.31
Pahud	**2.28** ± 6.45	**-2.95** ± 4.99	0.55 ± 7.17
Pitelina	0.97 ± 9.12	**-2.50** ± 7.44	**1.77** ± 7.06

Table 6.2: Mean and standard deviation of the $\Delta IOI(n)$ of notes (excluding rests), in percentage on the length of each measure, for the six performers on all Telemann Fantasias. Values above 1% are in bold. (Top). The non-annotated notes (column x) are played with small deviations in contrast to the upper and lower voices (columns u, b+B). (Midle and Bottom). Focus on groups of notes within the same beat. Several performers deviate annotated notes written with interleaved voices more than when they are not. They also emphasise the first note of the group.

Means and standard deviations of ΔIOI are presented in Table 6.2. These distributions exhibit notable dispersion and deviate from normality due to several factors, including inherent performance variability, the systematic accentuation of metrically strong beats, and grouping effects induced by rhythmic structure. This deviation from normality is confirmed by a Shapiro–Wilk test [Shapiro and Wilk (1965)].

Given this non-Gaussian behaviour, parametric comparisons are not appropriate. We therefore adopt non-parametric methods to compare selected populations (Tab. 6.3), specifically the Mann–Whitney U test [Mann and Whitney (1947)], alongside Cohen's d as an effect size measure [Cohen (1988)]. In our analysis, we consider an effect size greater than 0.5 as large.

In the following, we first examine generic statistical tendencies across the dataset before focusing on targeted comparisons between populations defined by score-based criteria, in particular the influence of interleaved voices.

	Rampal	*Kuijken*	*Porter*	*Lazarevitch*	*Pahud*	*Pitelina*
all the corpus						
b+ B/ x	–	0.55	0.54	0.54	0.58	0.44
B/ x	–	0.57	–	0.56	0.60	0.41
u/ x	0.54	0.53•	–	0.57	0.52•	0.58
B/ b	–	0.56•	–	0.55•	0.55	0.44
u/ b+ B	0.53•	–	0.46	0.53•	0.43	0.65
all the corpus						
all: on beat / off beat	–	0.57	0.54	0.60	0.61	–
♪: on beat / off beat	–	0.55	0.54	0.61	0.58	–
duple metre						
♫ interl.: 1st / 2nd	–	0.58	0.64	0.67	0.57•	–
♫ x: 1st / 2nd	–	–	–	0.60	–	–
𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯: 1st / 2nd	–	0.69	0.64	0.72	0.75	0.61
𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯: 2nd / 3rd	0.59	–	0.59	0.44•	–	–
𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯: 3rd / 4th	0.40	0.45•	0.28	–	0.43	0.39
triple metre						
♫♪: 1st / 2nd	–	0.58•	–	0.69	0.79	0.63
♫♪: 2nd / 3rd	0.42•	–	0.43•	0.36	0.31	0.29

Table 6.3: Cohen's d effect size for Mann-Whitney U test measuring the magnitude of the difference between two groups on the metric $\Delta IOI(n)$, focusing on cases where the p-value is $< 10^{-2}$ (•) or $< 10^{-3}$ (all others).

Group of Notes in a Same Beat
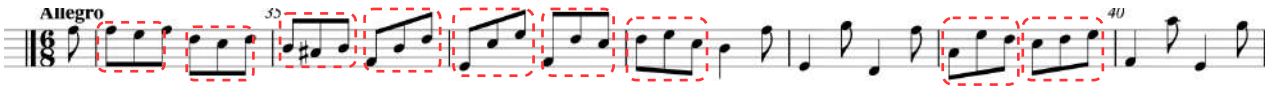


Figure 6.5: First measures of Fantasia No. 3 *Gigue* with groups of three eighths within a beat (♫)

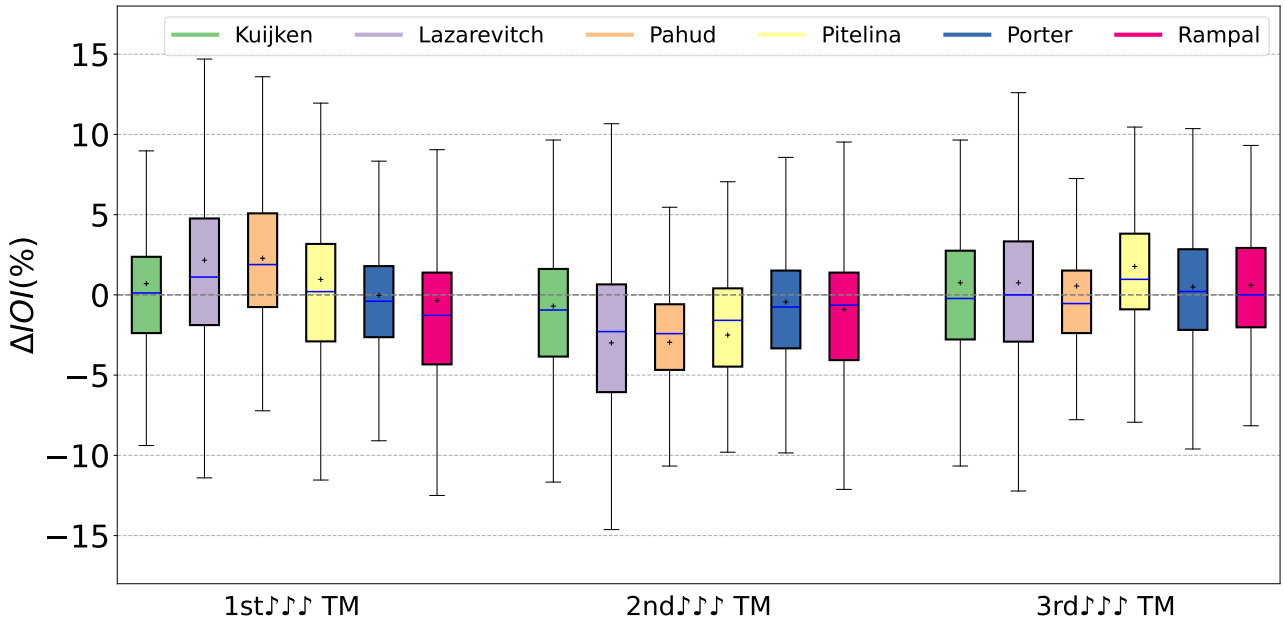


Figure 6.6: $\Delta IOI(\%)$ of the 1st, 2nd, and 3rd eighth notes within a beat in groups of three, across all movements in a triple metre and for all performers. The boxes span from the first to the third quartile, while the whiskers extend to the most extreme data points within 1.5 times the interquartile range (IQR). The median is indicated by a horizontal line, and the mean by a '+' sign.

Beyond beat-level timing deviations, performers also shape the internal organisation of note groups within a beat. By comparing the relative durations and onset positions of notes belonging to the same metrical unit, we can identify recurring patterns of rhythmic articulation in both triple and duple contexts.

In the final movement of Fantasia No. 3 (Fig. 6.5) Pahud systematically emphasises the first eighth note of each triple subdivision, particularly on the first beat 🔊. Remarkably, Pahud plays the first eighth note of these groups $\Delta IOI_1 = 5.23\,\%$ longer than the second ones (strong

effect size of 0.79 see Tab. 6.4) on average across all the fantasias. This lead to a $\Delta o_2 = 2.35\,\%$ deviation on the onset of the second notes. According to Garrison, Pahud's approach to articulation "is similar to that of traverso players, he slurs less often and varies his tonguing style" [Garrison (2020)].

Comparing all the performances, in in triple metre, when considering groups of three consecutive eighth notes within a beat (♫), the second eighth note is, on average, played shorter by all performers (Fig. 6.6). Moreover, the third notes are played slightly longer compared to the second ones (up to 4.27 % for Pitelina), although the effect is less pronounced (0.29–0.43). These deviations reflect established performance practices in Baroque triple-metre dances, where strong rhythmic articulation can lead to expressive displacement of notes toward the second or third subdivision of the beat [Little and Jenne (2001)].
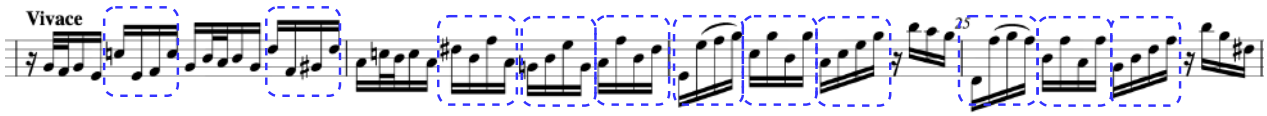


Figure 6.7: First measures of Fantasia No. 3 *Vivace Fugato* with groups of four sixteenth notes within a beat (♬).

In the *Vivace fugato* of Fantasia No. 3 (Fig. 6.7), Kuijken plays in average the first note of each group of sixteenth longer 🔊. In duple metre, when considering these groups (♬), the differences are significant for all performers except for Rampal.

Interleaved and Opposite-Stem Notes

Interleaved passages most commonly occur in fugato movements, particularly in duple metre within this corpus, such as the two *Vivace fugato* sections of Fantasia No. 3, separated by four measures of a *Largo*.

We compute here the "♫ interleaved" metrics for all beats in duple metre where a pair of eighth notes appears, each with a different voice annotation (excluding cases with a single voice annotation or an eighth rest). For four performers, there is a significant difference in timing between the first and second eighth notes, with strong effect sizes (0.57–0.67). Remarkably, for three performers (Kuijken, Pahud and Porter), no significant difference is observed when comparing the same situation in non-interleaved passages.

Lazarevitch's interpretation is particularly interesting, as he plays the first eighth note 3.14 % longer in interleaved passages and 2.64 % longer in non-interleaved ones. This could be explained by his consistent approach to inequality in timing within this Baroque corpus, and sometimes a slightly low tempo like in the first movement of Fantasia No. 6 (*Adagio*), which he plays in 3'45" 🔊, while the fastest performers played it in less than 2'30". For four performers (Kuijken, Lazarevitch, Pahud, Pitelina), the effects are moderate to strong (0.41–0.60) when one focuses

on the *opposite stemmed notes* notated in Telemann's autograph (Fig. 4.3), as shown on the `B/x` line in Table 6.3. The specific effect of these notes against the bass notes not annotated by Telemann (`B/b`) is remarkably observed (0.44–0.56) in the same recordings.

The specific effect of interleaved can also be seen on annotated sixteenth-note groups (𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯𝅘𝅥𝅯), 9.86 % of them featuring interleaved voices (Tab. 4.2). As in the non-interleaved case, the first sixteenth ΔIOI is quite different from the others, but, notably, there is for Rampal and Pitelina a significant difference (p-value $< 10^{-3}$), between the second and the third sixteenths with strong effect sizes of 0.83 and 0.74, whereas there is no statistical difference in groups without interleaved voices.

Studies could further investigate these cases, in particular by separating them according to their interleaving patterns, beginning with the upper voice or with the bass voice.

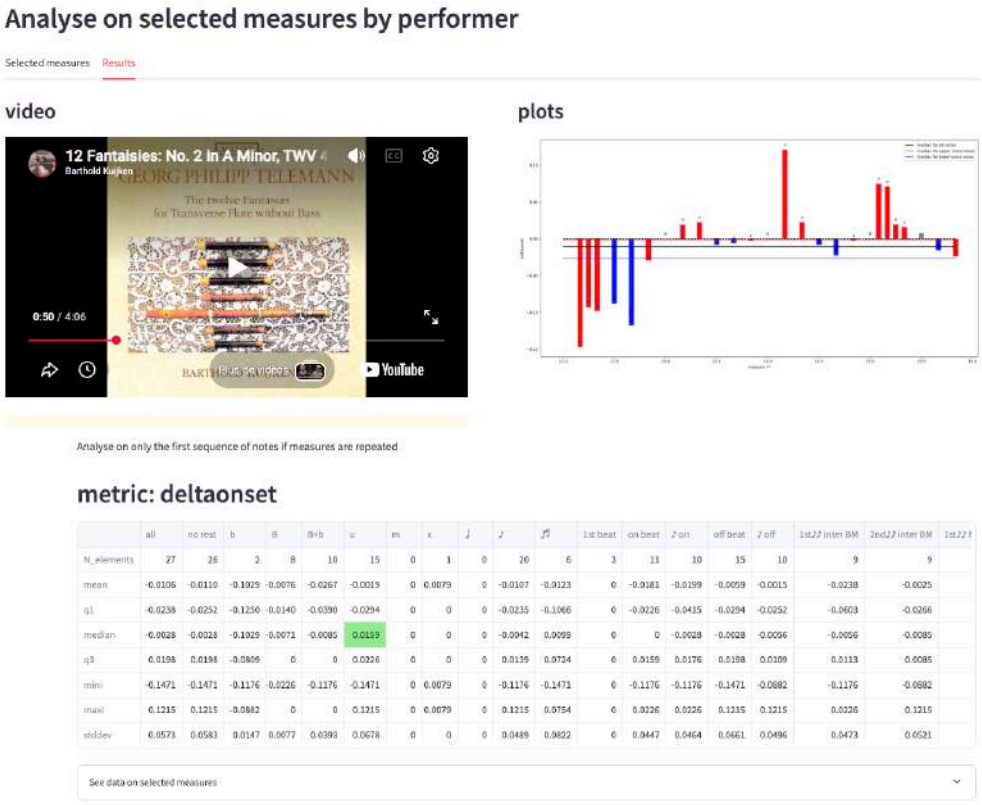


Figure 6.8: Overview of the Streamlit web application (*Measures by performer* module) accompanying the fantasias performances comparison (https://telemannvoicesinterleave.streamlit.app/).

Interpretation and Limitations of the Metric

The proposed deviation measures assume that the metronomic reference is locally constant within each measure. This choice provides a musically meaningful compromise between strict beat-level synchronisation and global tempo modelling.

However, this assumption may oversimplify performances exhibiting continuous tempo drifts within a measure. In such cases, deviations capture a mixture of expressive timing and local tempo curvature.

Despite this limitation, measure-based normalisation allows consistent comparison across heterogeneous rhythmic contexts and performers and provides a stable reference frame for statistical analysis.

Code Availability and Web Application

The availability of the dataset is described in Section 4.3.

We also distribute code enabling the reproduction of results and figures via the git repository at https://gitlab.com/algomus.fr/telemann-voices-interleave.

All the metrics values can be browsed and queried. A web application built with *Streamlit* at https://telemannvoicesinterleave.streamlit.app (Fig. 6.8) is published online to browse and query the metrics values across performances.

Section summary –From Notated Rhythm to Performed Timing in Fantasias

We introduced normalised onset and IOI deviation metrics that compare performed timing with a locally metronomic reference defined at the measure level. Applied to six complete performances of Telemann's Fantasias, these metrics reveal consistent expressive timing strategies across performers. Statistical analyses show systematic lengthening of the first note within duple and triple subdivisions. Moreover, passages annotated with interleaved voices exhibit distinctive timing patterns.

6.3 Timing Across Patterns in 20th Solo Flute Music

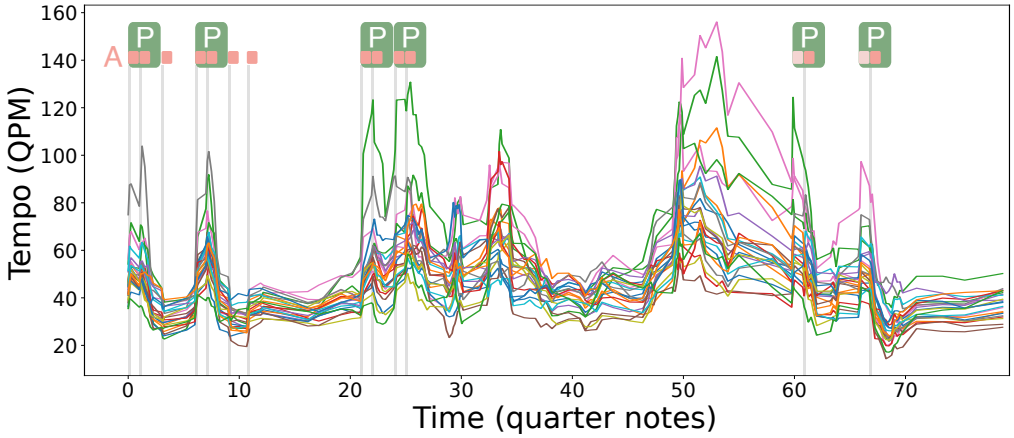


Figure 6.9: Tempo curves for the 23 *Syrinx* performances with patterns `A` and `P` (see Fig. 6.10). The tempo is expressed in Quarter notes per minutes (QPM).

Precise timings for every note allow the detailed study of musical expressivity. The tempo curves of *Syrinx* reveal significant flexibility in tempo (Fig. 6.9) and expressive variations across specific musical structures. Going beyond tempo curves, we propose a systematic analysis of pattern-level timing. This extension adapts the proposed microtiming framework to repertoires where expressive flexibility is less constrained by regular metre, allowing us to study how performers shape timing within recurring musical structures.

We first present the methodology used to select and identify recurring patterns in our Early 20th Century dataset (see Chapter 4). We then introduce the timing descriptors used to characterise expressive timing patterns. Thus, we can analyse results on *Syrinx* and *Density 21.5* using onset estimates derived from the late-fusion audio-to-audio alignment strategy described in Chapter 5.

6.3.1 Musical Patterns Selection

Rather than addressing automatic pattern discovery, we focus on the analysis of patterns previously identified in the literature. The descriptions of the patterns are done in Section 4.2.3. In this expressive timing study, we consider for *Syrinx* (Fig. 6.10):

- four short *cell-level* patterns: `A` the lead pattern, `B`, `E`, and `C`
- two *phrase-level* patterns: `P` the main *Syrinx* theme and `Q`

In the case of *Density 21.5*, we also examine a subset of occurrences of the motif α in the first measures (Fig. 6.11).

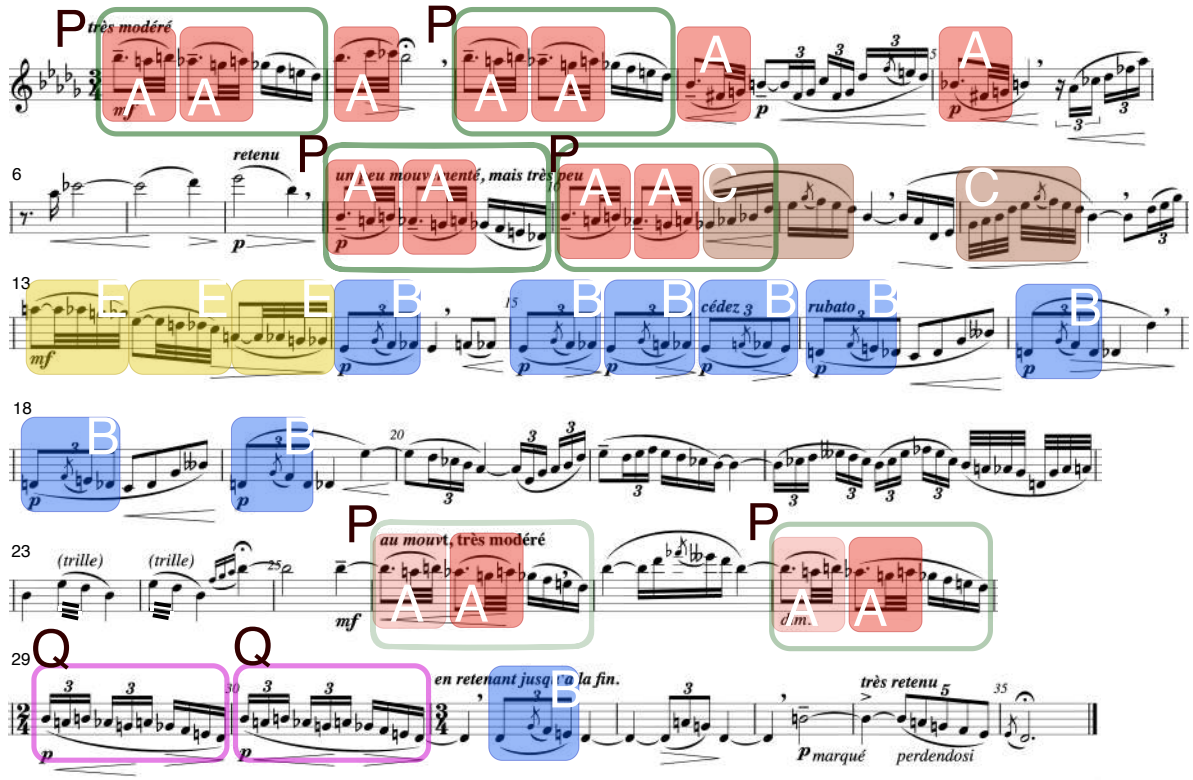


Figure 6.10: All occurrences of cell-level (`A,B,C,E`) and phrase-level (`P,Q`) patterns in Debussy's *Syrinx*.
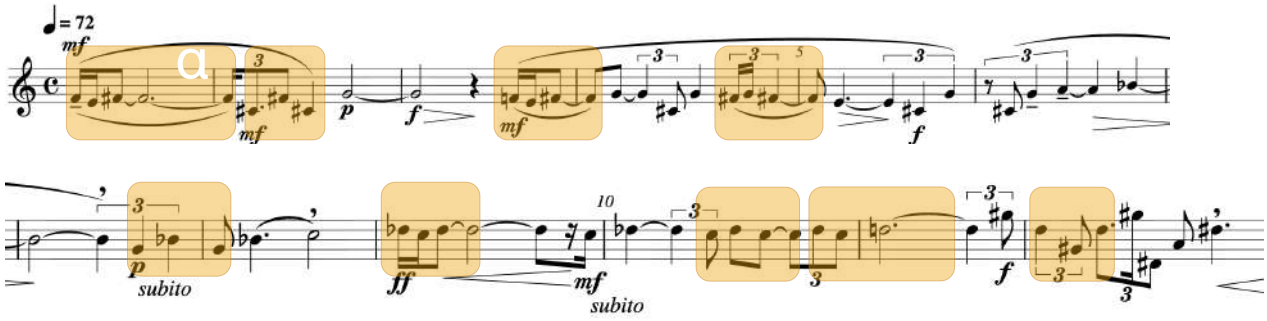


Figure 6.11: Beginning (mm 1–13) of *Density 21.5* with 9 pattern α occurrences as defined by Ohriner [Ohriner (2019)] (see Fig 4.11) 🔊.

Constraint-Based Pattern Identification

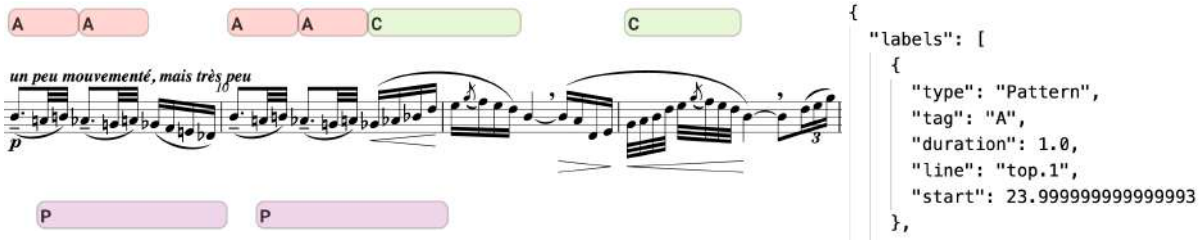


Figure 6.12: 8 patterns are identified in mm. 9–12 of *Syrinx* in Dezrann (https://www.dezrann.net/~/early20thsoloflute/syrinx?m=9). The first notes of P pattern are omitted (see footnote 1). An excerpt of exported *dezfile* code is given for the first cell-level pattern A.

To identify the previous patterns, we formulate the evaluation of constraint functions over sliding windows of a symbolic score. Let $\mathcal{E}$ denote the space of musical events, and let $S = (e_1, \ldots, e_T)$ with $e_t \in \mathcal{E}$ be a score representation. For a window size k, we define

$$W_t^{(k)} = (e_t, \ldots, e_{t+k-1}) \in \mathcal{E}^k.$$

A musical pattern is modelled as a Boolean-valued function

$$M : \mathcal{E}^k \to \{0, 1\}, \quad M(W) = \bigwedge_{i=1}^{n} c_i(W),$$

where each constraint $c_i : \mathcal{E}^k \to \{0, 1\}$ encodes a musical property.

A window $W_t^{(k)}$ is retained as a valid occurrence if and only if $M(W_t^{(k)}) = 1$, yielding the set

$$\mathcal{T}_M = \{t \mid M(W_t^{(k)}) = 1\}.$$

Pattern identification in the pieces is built on musical constraints that may capture (i) structural properties (e.g., absence of rests), (ii) intervallic relations between pitches, (iii) duration ratios, or (iv) exact rhythmic patterns specified at given positions. This unified formulation supports both strictly notated rhythmic figures and more abstract motivic configurations.

Optional parameters further control overlap between occurrences and restrict admissible starting positions.

This framework can be viewed as a declarative, constraint-based formalism for musical motif specification, where patterns are defined through compositions of symbolic constraints rather than procedural rules, enabling transparent and controllable modelling of musical structure and facilitating the explicit formulation and testing of musicological hypotheses.

All detected occurrences were exported in JavaScript Object Notation (JSON)-like format (*dezfile*) for visualisation in Dezrann [Ballester et al. (2025)] and then manually verified (Fig. 6.12).

6.3.2 Timing Descriptors

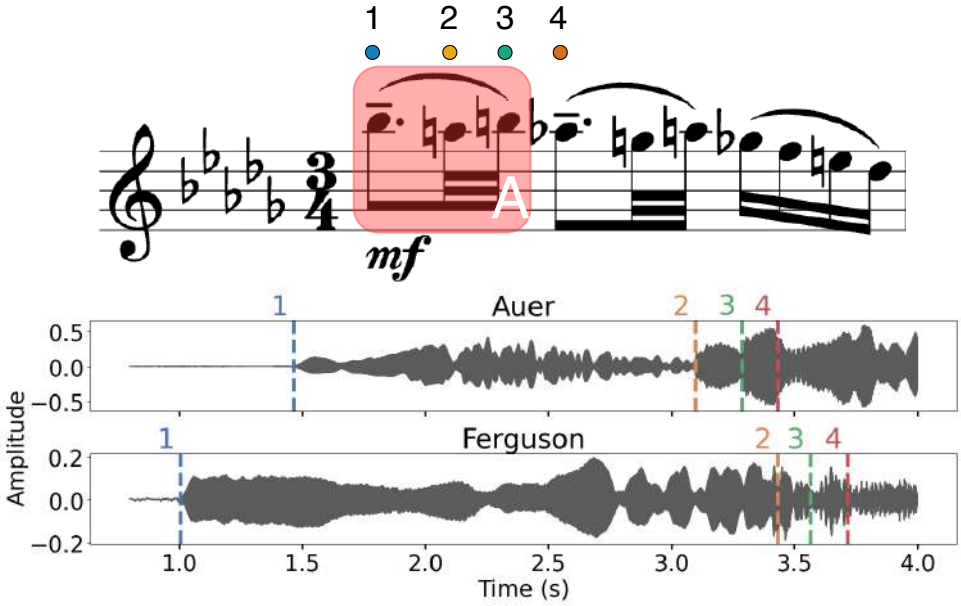


Figure 6.13: First occurrence of *Syrinx* pattern `A` and corresponding waveform extracted from the start of Auer's and Ferguson's recordings. The pattern considers three notes (1-3) and the onset of the next note (4).

This section revisits the notion of expressive timing at the pattern level and relates it to existing metric representations, including IOI-based descriptors and ratio-based measures. While ΔIOI captures local deviations between successive events (as used in Section 6.2.1), it remains insufficient on its own to characterise the internal structure of timing within multi-note patterns, where relative temporal organisation plays a central role.

We consider a pattern with N notes at performed onsets $o_1 \dots o_N$, together with o_{N+1}, the next onset after the pattern. To characterise expressive timing, we compute temporal descriptors derived from the inter-onset intervals $IOI_i = o_{i+1} - o_i$

The **performed pattern duration** is defined as:

$$\mathrm{D} = \sum_{i=1}^{N} IOI_i = o_{N+1} - o_1 \tag{6.4}$$

The **local timing** is defined for each onset as $t_i = s_i / IOI_i$, where s_i denotes the corresponding score duration. To characterise fine-grained temporal variability within patterns, we define the **timing fluctuation** as a measure of changes in local timing, independent of their direction, normalised by the local score duration $w_i = s_i + s_{i+1}$:

$$\mathrm{TF} = \frac{\sum_{i=1}^{N-1} w_i \left(\log \frac{t_{i+1}}{t_i}\right)^2}{\sum_{i=1}^{N-1} w_i} \tag{6.5}$$

TF values close to zero indicate stable local timing.

As an example in Figure 6.13, Auer plays the first occurrence of pattern A in $D_{A1} = 1.95$ s, faster than Ferguson (2.71 s). Within this occurrence, Auer has a more stable local timing ($\mathrm{TF}_{A1} = 0.3$) than Ferguson (0.98).

6.3.3 Inter-Performer Analysis of Patterns in Syrinx

Pattern		$D(s)$				TF			
	N	η^2	d	min	max	η^2	d	min	max
A	13	0.65	3.72	*11*: 0.92	*3*: 2.28	0.19	2.28	*10*: 0.08	*1*: 0.54
B	9	0.34	1.88	*8*: 1.18	*4*: 2.13	–	1.14	*1*: 0.05	*4*: 0.23
C	2	–	2.31	*2*: 1.37	*1*: 2.25	–	–	*2*: 0.13	*1*: 0.14
E	3	0.18	1.29	*2*: 0.94	*1*: 1.23	–	–	*2*: 0.09	*1*: 0.11
P	6	0.44	2.3	*4*: 2.2	*6*: 4.46	0.17	1.16	*4*: 0.22	*5*: 0.43
Q	2	0.04	–	*1*: 2.55	*2*: 2.75	–	–	*1*: 0.18	*2*: 0.21
α	9	–	5.3	*9*: 1.22	*1*: 4.09	0.07	–	*6*: 0.05	*3*: 0.16

Table 6.4: Pattern occurrences analysed across 23 performances of *Syrinx* (A–E) and *Density 21.5* (α). For each metric (D, TF), eta-squared (η^2) is computed per pattern. Min and max occurrences denote the lowest and highest mean values across performers, respectively (index and value are reported). Cohen's d quantifies the contrast between them. Effect sizes are omitted (–) when p-value $> 10^{-2}$.

Given the 46 aligned performances of *Syrinx* and *Density 21.5*, we choose to compare D and TF across patterns and performances.

We compute eta-squared (η^2) [Cohen (1973)], estimating for each metric, the proportion of variance explained by *occurrence organisation* across performers. For duration, high η^2 (above

0.14) indicates systematic differences across occurrences : the performers lengthen/shorten/vary the same occurrences of the pattern. Low η^2 (below 0.06) reflects weak differentiation.

For each pattern, we also contrast the fastest vs. slowest occurrences (D) and the most vs. least regular ones (TF), using Cohen's d [Cohen (1988)], computed across performances. Statistical significance is assessed via a permutation test (Tab. 6.4).

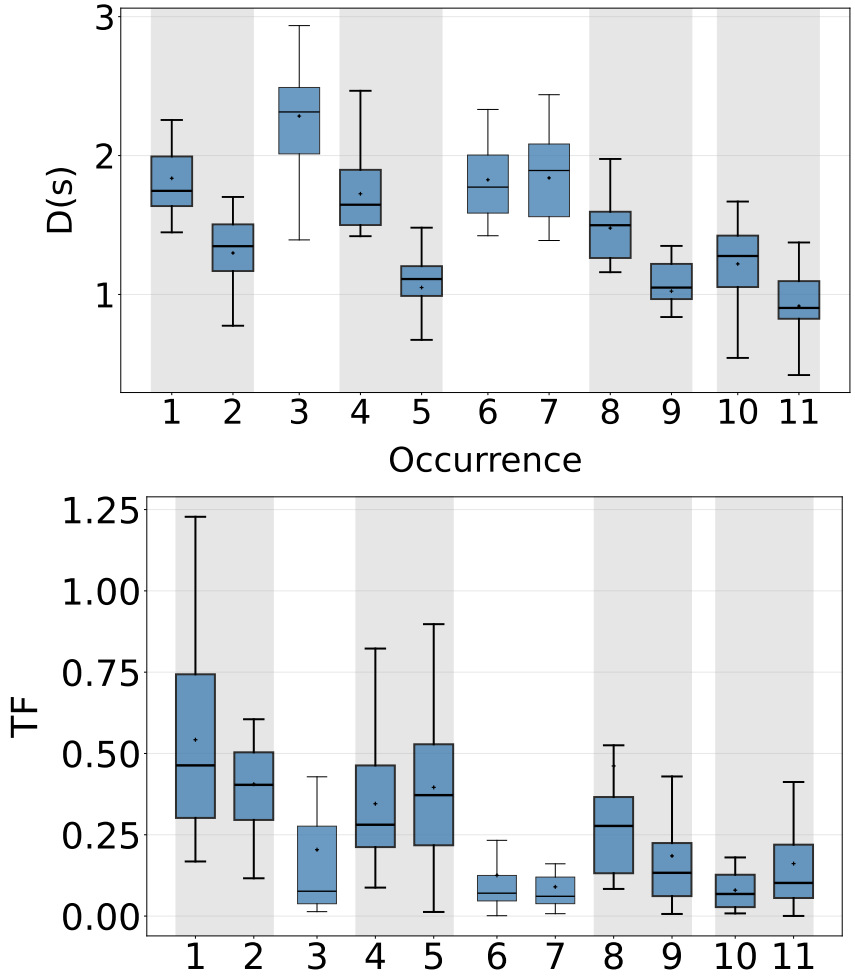


Figure 6.14: Duration and timing fluctuation for the 11 first occurrences of pattern `A` in *Syrinx* (see Fig. 6.10). The boxes show the Interquartile Range (IQR), whiskers extend to $1.5 \times IQR$, and horizontal line marks the median. The occurrences 1–2, 4–5, 8–9, and 10–11 are colocated in a large `P` phrase-level pattern.

Patterns `A` and `P` exhibit the most consistent organisation across both duration and timing fluctuations [1]

Pattern `A` shows strong organisation in duration ($\eta^2 = 0.65$), with a large contrast ($d = 3.72$) between occurrences `A11` (0.92 s) and `A3` (2.28 s). `A3` occurs at the end of the opening phrase before a fermata introducing the Arabesque theme [Trezise (2003)], whereas `A11` appears in a more unstable, *"un peu mouvementé"* section preceding pattern `C`. This confirms that expressive timing is not purely pattern-dependent but strongly conditioned by local context.

Indeed, considering occurrences of `A` appearing in pairs (1–2, 4–5, 8–9, 10–11), the first element in each pair is systematically longer than the second (Fig. 6.14).

In addition, TF shows a slightly lower, yet still substantial, organisation ($\eta^2 = 0.19$), accompanied by a strong contrast between occurrences ($d = 2.28$). Notably, the occurrences `A6` and `A7` show the lowest IQR in TF (0.08), along with low and closely aligned mean values (0.06 and 0.07). This suggests a marked convergence across performers and a consistently stable local timing in these segments. These occurrences belong to a more relaxed phrase (mm. 4–5), where the pattern A is developed through an upward chromatic expansion [Larson (1990)], potentially offering less flexibility than in the preceding thematic material.

The larger theme **pattern** `P` also exhibits substantial organisation in duration ($\eta^2 = 0.44$, $d = 2.30$ between `P4` and `P6`). Here again, expressive timing appears linked to structural transitions, with `P4` preceding a passage involving an ascending pentatonic figure [Larson (1990)]. TF also shows a significant effect ($\eta^2 = 0.17$).

Pattern `B` exhibits a moderate degree of temporal organisation ($\eta^2 = 0.34$, $d = 1.88$), along with a clear contextual influence. This is particularly evident at position `B4` (2.13 s), which coincides with the *"cédez"* marking, suggesting a slight relaxation of tempo by the performer. A comparison of interpretations highlights this effect (Fig. 6.15) : in Bourdin's 1959 performance, the second and third eighth notes of the triplet are lengthened. Conversely, some performers – such as Galway (1986) – override this indication, delaying the local deceleration until after the triplet groups. In contrast, TF shows no consistent organisational pattern overall, suggesting heterogeneous timing fluctuations across performers.

[1] 15 occurrences of `A` appear in the score, but only 13 are considered: nos. 12 and 14 are excluded due to a tied initial note. Consequently, for `P`, the first note of the pattern is omitted, yielding 6 occurrences.

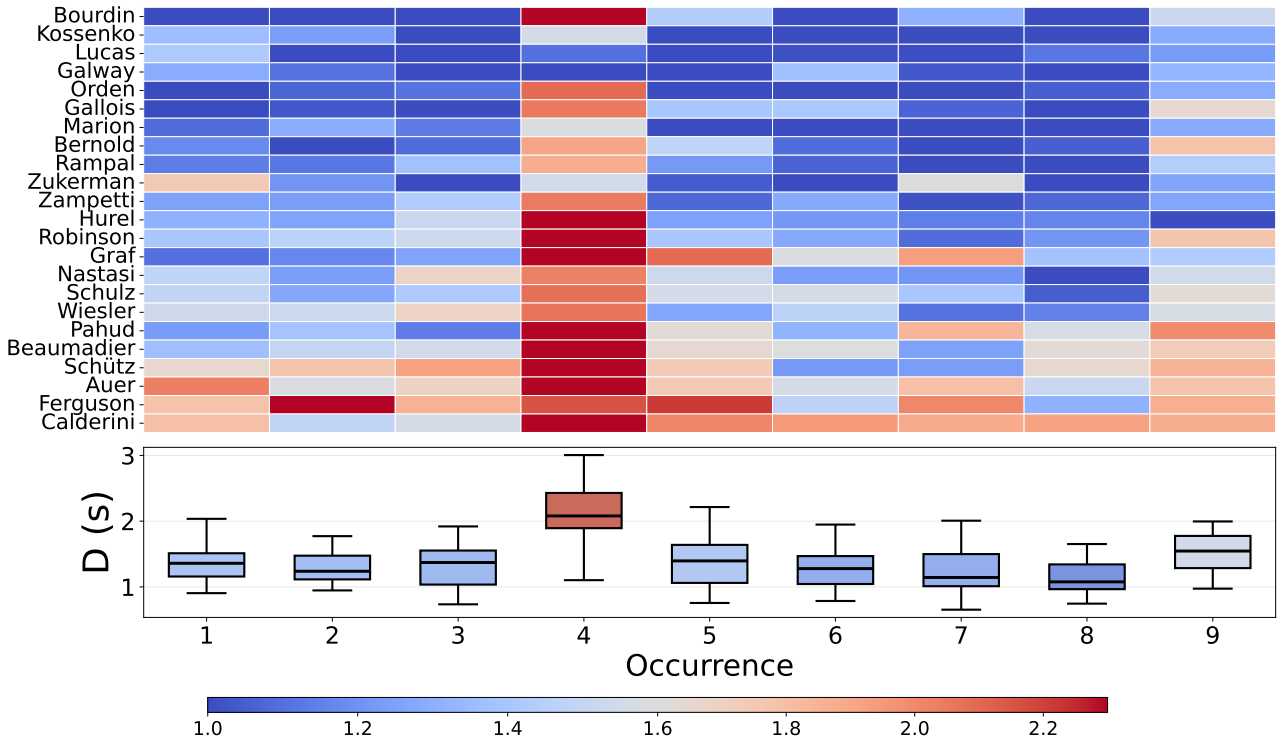


Figure 6.15: Duration of pattern `B` in *Syrinx* (see Fig. 6.10), across performers and occurrences (top, heatmap), and summarised by occurrence (bottom, boxplots as in Fig. 6.14).

Pattern `E` has a strong organisation ($\eta^2 = 0.18$), the first occurrence being significantly slower, and **patterns `C` and `Q`** show weaker organisation. While `E` shows moderate contrast in duration ($d = 1.29$), `C` and `Q` are too sparsely represented to support strong conclusions. None of these patterns exhibits significant TF organisation.

In *Syrinx*, musical time is strongly shaped by performer agency, where tempo fluctuations, rubato, and breath contribute to a fluid, quasi-narrative unfolding that resists strict metric regularity. This aligns with Jankélévitch's conception of Debussian musical time as "not a calligraphy projected into space" but an "experience lived within life itself" [Jankélévitch (1976)] grounded in an expressive temporality.

6.3.4 Timing Analysis of Patterns in Density 21.5

Compared to *Syrinx*, the main pattern in *Density 21.5* is quite heterogeneous regarding the score durations (Fig. 6.11), limiting the interpretability of the performed duration-based organisation. We thus focus here on timing fluctuations, which offer a more relevant perspective on local variability in such a context.

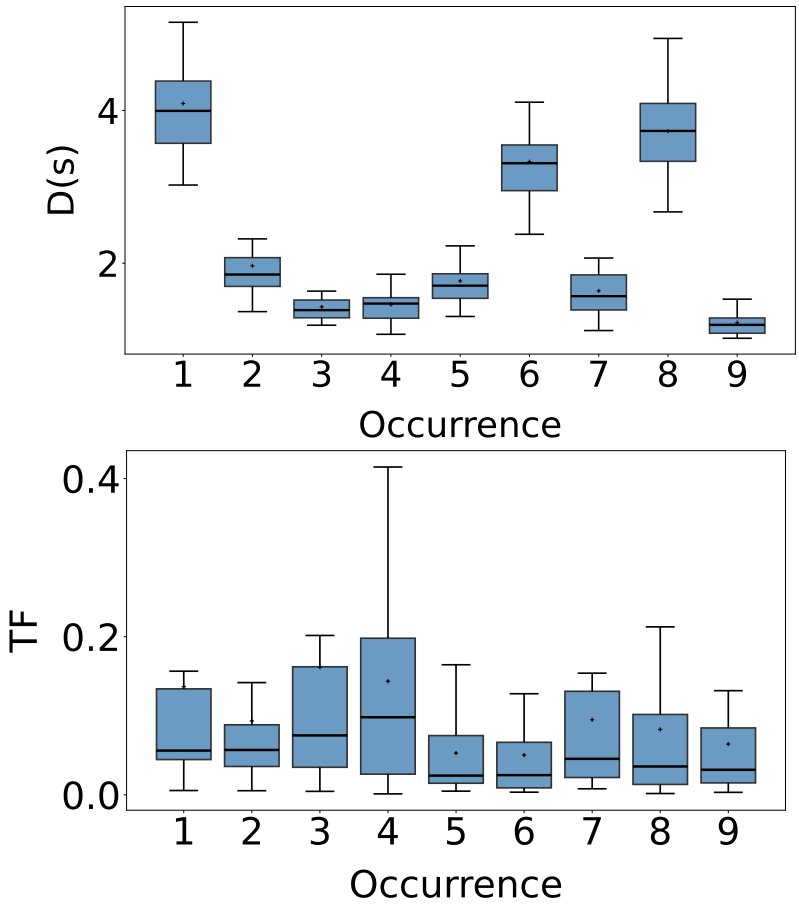


Figure 6.16: Duration and timing fluctuation for the 9 first occurrences of pattern α in *Density 21.5* (see Fig. 6.11).

Certain occurrences ($\alpha 5$ and $\alpha 6$) exhibit particularly low mean and median values (Fig. 6.16), indicating strong inter-performer convergence, whereas others show higher IQR, reflecting locally more unstable realisations. In particular, $\alpha 6$ introduces a new pitch (`Db5`) and thus constitutes a structurally salient moment in the performance 🔊. This same conclusion has been previously proposed by Ohriner [Ohriner (2019)] based on duration ratios computed over local event pairs. In contrast, the proposed TF measure does not rely on pairwise or interval-specific comparisons, but integrates temporal variability over the full occurrence, making it less dependent on the local size or segmentation of the pattern. Nevertheless, TF shows weak organisation ($\eta^2 = 0.07$) and no substantial contrast.

This result is consistent with Varèse's conception of "organised sound" [Varese and Wen-chung (1966)] : expressivity is quite fixed in the score–through registral contrasts, intervallic structures, and sonic blocks–leaving the performer with limited freedom.

On the Limitations of the Proposed Metrics and on the Importance of Musical Meaning

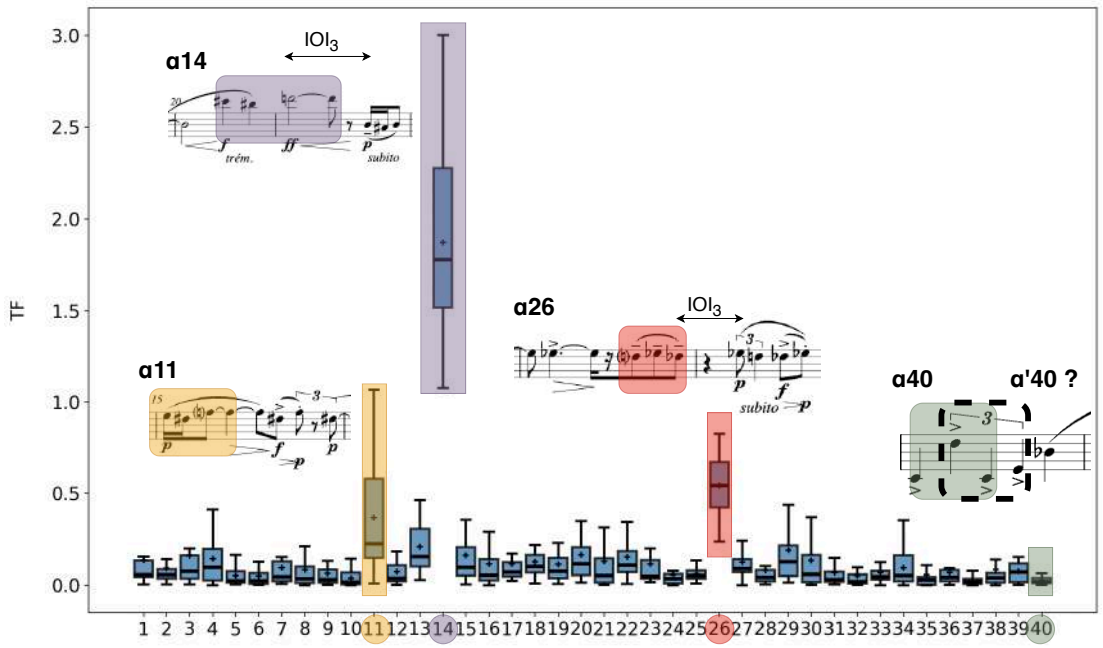


Figure 6.17: 40 occurrences of α pattern in *Density 21.5*.

If we attempt to apply Ohriner's definition of the α-pattern beyond the opening section of the piece, we identify 40 occurrences (Fig. 6.17). The order of magnitude of the *timing fluctuation* values is generally consistent with our previous findings, with two notable exceptions: α14 and α26, which yield TF values of 1.78 and 0.54, respectively. A closer examination reveals that both patterns are followed by a rest. Since our current methodology does not account for note offsets, the final IOI used to compute the metric is biased by the onset of the note following the rest. Consequently, these TF values should be interpreted with caution. This situation did not occur in the selected patterns from *Syrinx*, but it affects 5 out of the 40 identified α patterns in *Density 21.5*.

Another point worth discussing concerns the strict pattern definition proposed by Ohriner (Fig. 6.11). In [Ohriner (2019)], he compares the durations of the first nine motives across a corpus of 19 recordings. However, the recordings are not specified, and the annotations were almost certainly produced manually rather than derived from alignment outputs. When extending the analysis to the entire piece, the question arises as to whether this definition remains generalisable.

Among the identified occurrences, $\alpha 11$ appears particularly interesting. Its TF value of 0.23 is slightly stronger than that of the preceding occurrence and may be musically meaningful. According to the score, this pattern introduces a new pitch at its beginning, and performers seem to emphasise it. Although this interpretation remains speculative, it could be related to the segmentation proposed by Nattiez [Nattiez and Barry (1982)].

At the same time, the full-piece analysis suggests the possibility of eliminating certain false positives. For example, $\alpha 40$ satisfies Ohriner's formal definition. However, from a musical perspective, the pattern that listeners are more likely to perceive begins immediately after, with the `D-C-E` triplet. The current definition may therefore be overly restrictive, as it does not account for octave-equivalent relationships or for the rhythmic coherence created by the quarter-note triplet figure.

More generally, the selection of musically meaningful patterns is essential if performance studies are to produce results that are useful for musicological research, analytical interpretation, or pedagogical applications. One advantage of our pipeline is that such pattern definitions can be refined and curated within Dezrann or simply manually chosen, allowing analysts to incorporate musicological expertise directly into the annotation process.

Section summary – Timing Across Patterns in 20th Solo Flute Music

In this section, we study expressive timing in solo flute performances using from pattern-level analysis point of view which enables a finer-grained characterisation in flexible metronomic repertoires. Results show that pattern duration supports a stable global structure, while timing fluctuations reflect more local, context-dependent control. In *Syrinx*, expressive timing remains closely tied to performed patterns, whereas in *Density 21.5*, it aligns more closely with score-based sonic organisation. This contrast is made visible through the pattern-level analysis,

Chapter summary – Performance Analysis of Expressive Timing

The chapter introduces a methodological framework for analysing expressive microtiming in musical performance. Instead of considering timing deviations as isolated differences from notation, it proposes context-aware measurements that relate local temporal variations to rhythmic, metric, and motivic structures. This framework is applied to performances of the *Twelve Fantasias for Solo Flute* and to 20th-century solo flute works. It distinguishes global tempo variations from local micro-timing deviations, focusing mainly on the latter. A measure-based normalisation of IOI and onset deviations is introduced to enable systematic comparison across performances and rhythmic contexts. Results reveal consistent expressive strategies linked to metrical structure and rhythmic grouping, including systematic timing asymmetries within beats and effects related to interleaved voices. The framework is further extended to recurring patterns in *Syrinx* and *Density 21.5*, showing that expressive timing is strongly context-dependent: more flexible in *Syrinx*, and more structurally constrained in *Density 21.5*.

To the question *"How are local timing deviations structured within musical performance, and how do they interact with metric structure, voice-leading, and motivic organisation?"*, the answer is that expressive micro-timing variations are not random deviations from the score but systematic performance phenomena shaped by multiple musical dimensions. They reflect interactions between metric organisation, rhythmic grouping, implied or explicit voice-leading, and musical patterns. Across different repertoires and performances, the proposed framework reveals how expressive timing emerges from the interaction between local temporal deviations and musical organisation.

Chapter 7

Conclusion and Perspectives

Contents of this chapter

7.1 Contributions and Discussion . . . 162
7.2 Perspectives . . . 163
7.2.1 Tonguings . . . 163
7.2.2 Extended Techniques . . . 164
7.2.3 Performance Capture . . . 165
7.2.4 Respiratory Control and Sound Production . . . 166
7.3 Toward Richer Representations of Musical Performance . . . 168



Figure 7.1: Ian Anderson performance (https://patthmus.github.io/phd-manuscript-links/chap7.html) of *The Locomotive Breath* (Jethro Tull, live at the Rockpop In Concert 10.7.1982, Dortmund, Germany).

Locomotive Breath (Fig. 7.1) is a progressive rock composition by Jethro Tull, characterised by expressive timing, and unconventional sound events played by the flute. This type of dense audio material highlights some of the challenges faced by automatic alignment methods when dealing with complex audio material, including less clearly defined onsets, breath noises, and extended techniques.

7.1 Contributions and Discussion

This thesis investigated alignment and expressive timing through the case study of solo flute performance. While the proposed methodologies were developed and evaluated on flute recordings, they are based on general principles that are not specific to this instrument. The flute was chosen because its soft onsets, breath noise, and expressive micro-timing make it a particularly challenging testbed for alignment methods. Consequently, the solutions proposed here may also benefit the analysis of other monophonic woodwind instruments.

The contributions of this thesis revolve around three complementary axes: (1) the construction of annotated resources; (2) the development of alignment methodologies; and (3) the application of these methods to the study of expressive timing.

First, we introduce two complementary annotated datasets covering a broad historical distinct repertoire, from G.P. Telemann's *Twelve Fantasias for Solo Flute* to early twentieth-century works including Debussy's *Syrinx* and Varèse's *Density 21.5*. These corpora provide multi-level annotations ranging from structural descriptions to note-level alignments, establishing a reproducible benchmark for evaluating score-alignment methods on solo flute repertoire. Their main limitation lies in their relatively limited size, reflecting the considerable effort required to produce high-quality expert annotations. Extending these resources to larger repertoires, additional performers, and more diverse playing techniques therefore remains an important direction for future work.

Second, we developped several alignment strategies. These include a score-to-audio approach based on harmonic templates, a post-processing correction that compensates for systematic onset shifts in feature-dependent alignments, and an audio-to-audio framework exploiting multi-reference consistency through soft late-fusion. Although these methods substantially improve alignment accuracy, they can remain dependent on human annotation and existing synchronising tools.

Finally, rather than merely analysing expressive microtiming in flute performance, we proposed a methodological framework enabling its systematic investigation. The resulting analyses reveal consistent timing strategies related to metrical structure, rhythmic grouping, implied polyphony, and recurring musical patterns.

While the validation presented in this thesis is limited to solo flute recordings, the underlying principles are not specific to this instrument: the use of musically informed temporal references and context-dependent descriptors provides a basis for studying expressive timing in other

monophonic instruments. However, instrument-specific acoustic properties may require adapted strategies for onset estimation and alignment.

7.2 Perspectives

While this thesis focuses primarily on timing as a dimension of musical expression, flute performance involves a broader interaction of expressive parameters, including airflow control, embouchure-related timbral variations, microtonal pitch adjustments, vibrato, articulation, dynamics, and phrasing. In contemporary repertoire, expression is further enriched by extended techniques such as multiphonics, aeolian sounds, and key clicks. Capturing these dimensions jointly remains a significant challenge for MIR.

7.2.1 Tonguings

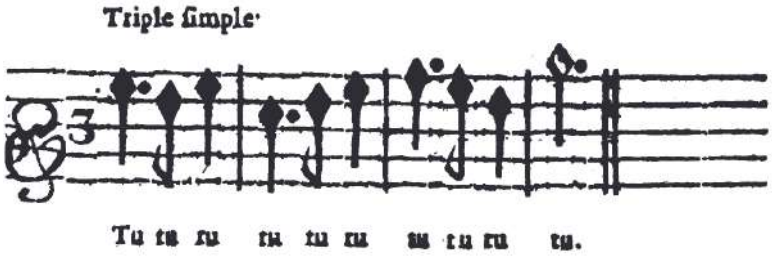


Figure 7.2: In [Chapter 8, p. 23] [Hotteterre (1719)], Hotteterre describes two tonguing techniques: *articulez* (**Tu**) and *coulez* (**Ru**), together with their usage. The original text reads: *"On doit prononcer Ru sur la Note qui fuit la Croche quand elle monte ou descend par degrez conjoints."* (author's translation: "One should pronounce *Ru* on the note that follows the eighth note when it ascends or descends by conjunct degrees").

The alignment frameworks proposed in Chapter 5 primarily focuses on estimating the temporal correspondence between score events and audio recordings. While this provides accurate information about note timing, it does not explicitly model other aspects of flute performance, such as articulation, tonguing, or phrasing.

Historical treatises provide useful insight into how these gestures were understood and transmitted across performance traditions (Fig. 7.2). However, from a contemporary MIR standpoint, tonguing remains a largely underrepresented dimension. In this context, integrating tonguing into performance capture frameworks would open new avenues for modelling expressive timing and articulation at a symbolic and physiological level. As reference annotations, we can use historical treatises but also more recent pedagogical suggestions (Fig. 7.3). Future work

could explicitly encode tonguing as a structured layer aligned with score events, enabling more interpretable models of phrasing, pedagogy, and stylistic variation in flute performance.
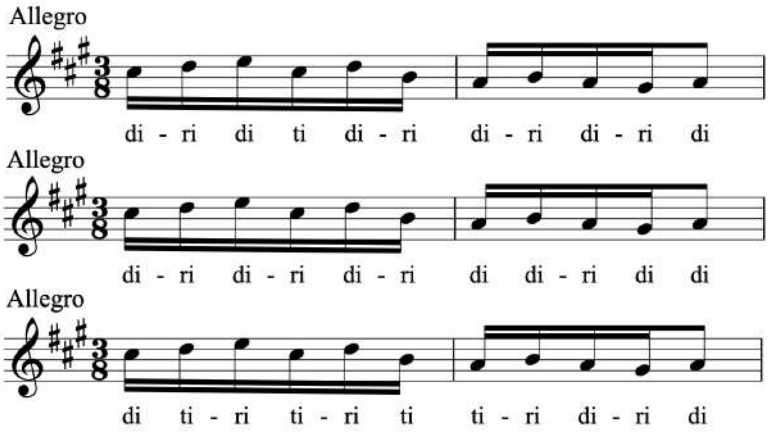


Figure 7.3: *Passepied* of Fantasia No. 1 in *A major* (TWV 40:02), mm. 1–2. Different tonguing patterns proposed by Brown rely on Quantz and Hotteterre treatises. Choosing adapted articulations "plays a vital part in bringing out one voice to the foreground and placing another in the background" [Brown (2015)].

However, capturing tonguing directly from audio signals remains challenging. Standard acoustic features only provide indirect and often ambiguous correlates of articulation gestures. As a result, audio-only representations are often insufficient and further motivate the integration of complementary data sources (breath pressure) within performance capture frameworks.

7.2.2 Extended Techniques

The datasets introduced in Chapter 4 mainly cover conventional flute playing. Applying the same annotation principles to performances involving extended techniques represents a natural continuation of this work.

Extended techniques on the flute encompass a range of non-conventional playing methods that significantly expand the instrument's sonic possibilities beyond traditional tone production (Fig. 7.4). Extending the exploration of flute possibilities initiated by Debussy (*Syrinx*) and Varèse (*Density 21.5*) [Radio France and Vilarem (2025)], many composers like Berio in *Sequenza I* (1958), explore flatterzunge, trills, or key clicks [Halfyard (2017)].

Multiphonics, a technique used on wind instruments in which a performer produces more than one pitch simultaneously, is not yet extensively studied [Riera et al. (2014)]. In recent years, however, several initiatives have tried to document this extended technique. In particular, the emergence of online databases (https://www.multiphonics.app/) or sample libraries such as *Hyperflute* (https://hyperflute.com/) aim to provide musicians with direct access

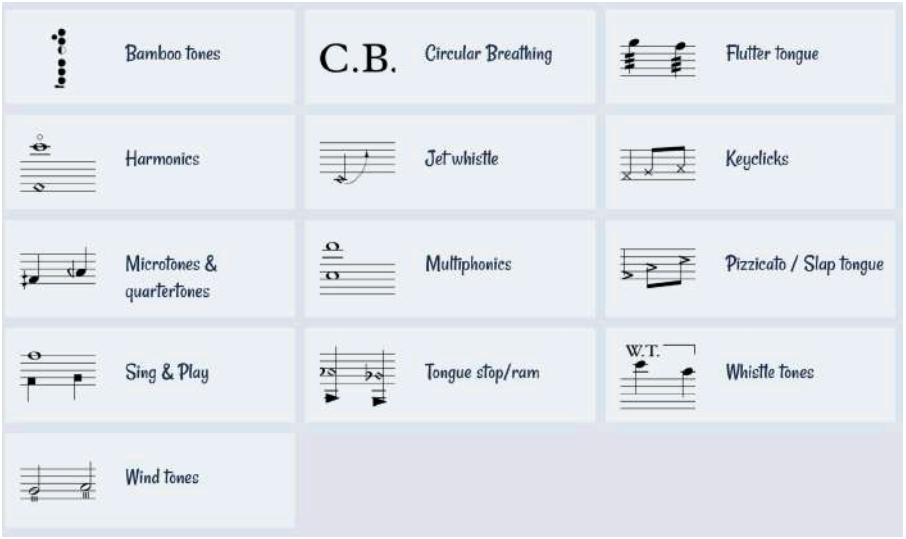


Figure 7.4: Extended Techniques for flute panel from https://www.flutecolors.com/techniques/.

to recorded examples of extended techniques. A further relevant effort in this direction is the dataset: *Chinese Bamboo Flute Performances DataBase* (CBF-periDB), which contains monophonic recordings of four types of isolated playing techniques–vibrato, tremolo, trill, and flutter-tongue– and twenty expert full-length performances of classic Chinese bamboo flute pieces [Wang et al. (2019)].

Despite these developments, a lot of resources remain relatively isolated from the principles of open science. Access is often restricted by subscription models, limited data export options, or lack of standardised metadata formats, which hinders reproducibility and integration into computational MIR workflows. This situation suggests that multiphonics could represent a particularly valuable direction for the development of structured datasets that will help in identifying such techniques in audio performances.

7.2.3 Performance Capture



Figure 7.5: Commercial picture of *Aerophone Brisa* from Roland (https://www.roland.com/fr/products/aerophone_brisa/).

A natural extension of this thesis is the development of a dedicated *performance capture*

framework for the flute, similar to existing systems for piano instruments (e.g., Disklavier-based technologies). Such systems enable the collection of high-resolution performance data and have been widely used to build large-scale datasets for performance analysis and machine learning evaluation.

A key motivation for such an approach is the limitation of current alignment-based methods, which attempt to synchronise audio recordings or audio-score representations but remain constrained by timing uncertainty. Direct acquisition of performance data would provide more accurate temporal information and support the development of models specifically designed for flute performance. In this context, digital wind instruments such as the *Aerophone Brisa* offer a promising platform. Equipped with dual breath sensors and motion sensors[1], it can capture expressive performance parameters and transmit them via MIDI, for which a complete implementation is provided by the manufacturer[2].

Although performance-capture devices provide more accurate temporal information, score alignment would not become obsolete. Instead, its role would evolve from estimating note-level correspondences between score and audio towards synchronising heterogeneous modalities, such as audio, MIDI, respiration, motion capture, and score representations. Alignment would therefore remain essential as a common temporal reference enabling the integration and analysis of multimodal performance data.

Despite these perspectives, several obstacles remain before such systems can be deployed at scale. The fidelity of the captured expressive parameters may not fully reflect the nuances of acoustic flute performance, particularly regarding embouchure-based timbral control and subtle pitch inflections. This raises questions about the validity of the collected data.

7.2.4 Respiratory Control and Sound Production

The timing analyses presented in Chapter 6 only consider observable acoustic events. Yet, many of these temporal variations are likely to originate from underlying physiological processes, particularly respiratory control.

Beyond timing and control parameters, respiration plays a central role when playing wind instruments. Previous studies highlight the importance of respiratory and physiological processes in flute performance. Important limitations remain when using an aerophone, as its embouchure differs significantly from that of an acoustic flute. Nevertheless, the study of breathing behaviour and its temporal organisation remains highly relevant. Expert flautists demonstrate coordination between breath pressure, embouchure control, and airflow shaping, enabling independent control of dynamics and pitch [Cossette et al. (2000)]. They also achieve more efficient and stable airflow by using a slower and more precisely directed air jet compared

[1] https://www.roland.com/fr/products/aerophone_brisa/specifications/

[2] https://www.roland.com/fr/support/by_product/aerophone_brisa/owners_manuals/c19aa546-a4d4-4488-b51c-0b6b7ea29819/

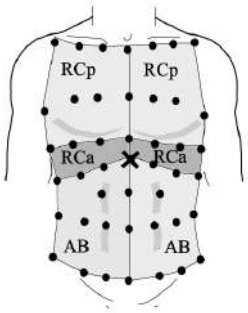


Figure 7.6: Sketch from [Vauthrin et al. (2015)] showing reflective markers placed on the flautist's thorax to track chest wall movements recorded using an optoelectronic system with infrared cameras.

to novices [De et al. (2008)].

More recent work has further shown that respiration is tightly coupled with musical cognition. Visual anticipation during reading is linked to breathing regulation [Cara and Mitrovic (2024)], while respiratory and aerodynamic parameters dynamically adapt to musical structure, including phrase length and register [Vauthrin et al. (2015)].

In flute pedagogy, teachers explicitly guide students on where to breathe in order to preserve musical phrasing and structure, highlighting breathing as an essential component of musical interpretation rather than a purely physiological constraint. Such pedagogical practices suggest that breathing strategies are learned, structured, and closely tied to musical meaning.

This raises the possibility of systematically comparing breathing strategies across skill levels, as well as investigating how pedagogical instruction shapes respiratory behaviour over time. In particular, the timing of inhalation in anticipation of note onsets appears to be a key aspect of learning, as it is closely related to expressive timing and performance preparation. Rather than studying these strategies acoustically, it could be really interesting to link them to the score. This could provide useful insights for musicologists and performance analysts.

Related work by Choi *et al.* [Choi et al. (2025)] further highlights the role of bodily signals in musical coordination. In a duet setting with piano, they investigate how flautists use respiratory and gestural cues to synchronise with the accompanying musician at the beginning of a piece. Their results suggest a dual mechanism of synchronisation, combining global preparation with local physical cues, and show that repetition improves the efficiency of this non-verbal coordination between performers.

In terms of limitations, sensor integration is complex and potentially intrusive for performers. Building synchronised multimodal datasets introduces non-trivial calibration and alignment issues between heterogeneous data streams (audio, MIDI, respiration, and motion), especially under real performance conditions.

7.3 Toward Richer Representations of Musical Performance

More broadly, these perspectives illustrate a shift in MIR from representing musical performances primarily through audio or symbolic data toward richer multimodal descriptions of performance. For wind instruments in particular, musical expression emerges from the continuous interaction between sound production, gesture, and musical intention.

Building open, multimodal, and interoperable datasets could foster collaborations between music information retrieval, musicology, acoustics, cognitive science, and performance studies. Rather than focusing solely on improving prediction accuracy, future research may increasingly aim to develop computational models that are both expressive and interpretable, offering new ways to study the complexity of human musical performance.

Capturing these complementary dimensions may not only improve computational models but also provide new tools for understanding performance practice and better understand why no two performances of the same musical work are ever exactly alike.

Appendix A

Interleaved Voices Detection

As we show in chapter 4 (Section 4.1), the Fantasias exhibit a frequent use of interleaved voices, where a single melodic line suggests the presence of multiple implicit parts. This raises the question of whether such voice changes can be automatically identified from the symbolic score representation.

In this appendix, we investigate a rule-based approach for detecting voice transitions and evaluate its performance against reference annotations. The goal is first to assess to what extent perceptually motivated rules can capture the annotated interleaved voice structure in Telemann's writing. Then, we apply this approach to another neo-baroque monodic piece written for oboe.

Research outputs associated with this appendix

Towards detecting interleaved voices in Telemann Flute Fantasias

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In DMRN+ 19: Digital Music Research Network One-day Workshop, 2024

When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute

Patrice Thibaud, Mathieu Giraud, Yann Teytaut

In 26th International Society for Music Information Retrieval Conference (ISMIR), 2025

A.1 Algorithm for Voice Change Detection

Whereas voice separation in polyphonic music is well studied [Cambouropoulos (2008); Chew and Wu (2004)], it is not the case for implied polyphony. Davis [Davis (2006)] proposed a perceptually motivated model to identify potential voice transitions in monodic melodic lines (see Section 2.4.2).

A.1.1 Feature Design and Scoring Function

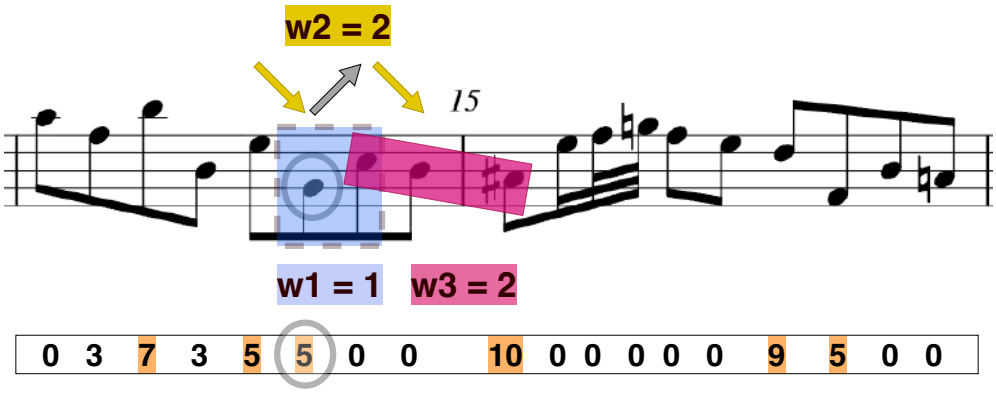


Figure A.1: Vivace from Fantasia No. 1 (TWV 40:2), mm 14–15. Each component of $\mathbf{w}_i$ (where i corresponds to the note index in the score) is computed according to Davis's rules. In this excerpt, using a detection threshold of 4, the model detects six voice changes, notably between G and C (mm. 14).

For each interval $I_i = (p_i, p_{i+1})$ of a monophonic line, we compute a feature vector $\mathbf{w}_i$ that summarises local and contextual melodic information, including pitch-related, contour-based, and motion-based descriptors derived from the surrounding melodic context.

In this work, $\mathbf{w}_i$ is composed of four perceptually motivated components:

- w_1 (Pitch distance): number of diatonic step increments between p_i and p_{i+1}, thresholded to zero for intervals smaller than a third;
- w_2 (Contour change): number of directional inconsistencies in the surrounding intervals (0, 1, or 2);
- w_3 (Conjunct motion): length of stepwise melodic runs before and after the interval;
- w_4 (Extended contour alternation): number of interval direction changes before and after I_i until the alternation stabilises.

Depending on the chosen model, the feature vector $\mathbf{w}_i$ is instantiated using different subsets of these components, which are then combined through a scoring function f.

Davis originally proposed the first three descriptors to predict voice transitions (Fig. A.1).

We define a scoring function applied to each inter-note interval, which is then thresholded to obtain a binary classification of voice transitions. The corresponding scoring function is defined as:

$$S_i = w_1 + w_2 + w_3, \quad \text{with the constraint } w_1 > 0.$$

A voice transition is then detected if $S_i \geq 4$

A.1.2 Algorithm

Algorithm 1 Voice change detection in monodic melodic lines

Require: pitch sequence $(p_1, \ldots, p_N)$
Ensure: binary sequence $(b_1, \ldots, b_{N-1})$
  Compute interval sequence $I_i = (p_i, p_{i+1})$ for $i = 1 \ldots N-1$
  for $i = 1$ to $N-1$ **do**
    Extract local and contextual feature vector $\mathbf{w}_i$
    Compute score $S_i = f(\mathbf{w}_i)$
    if $S_i > \tau$ **then**
      $b_i \leftarrow 1$
    else
      $b_i \leftarrow 0$
    end if
  end for

  return $(b_1, \ldots, b_{N-1})$

Algorithm 1 summarises the generic voice-change detection procedure. The feature extraction stage and scoring function f depend on the selected model.

We implement this model using *music21* and compare it to extended formulations that incorporate the additional descriptor w_4 (Section A.2.3). This formulation defines a binary classification over all inter-note intervals.
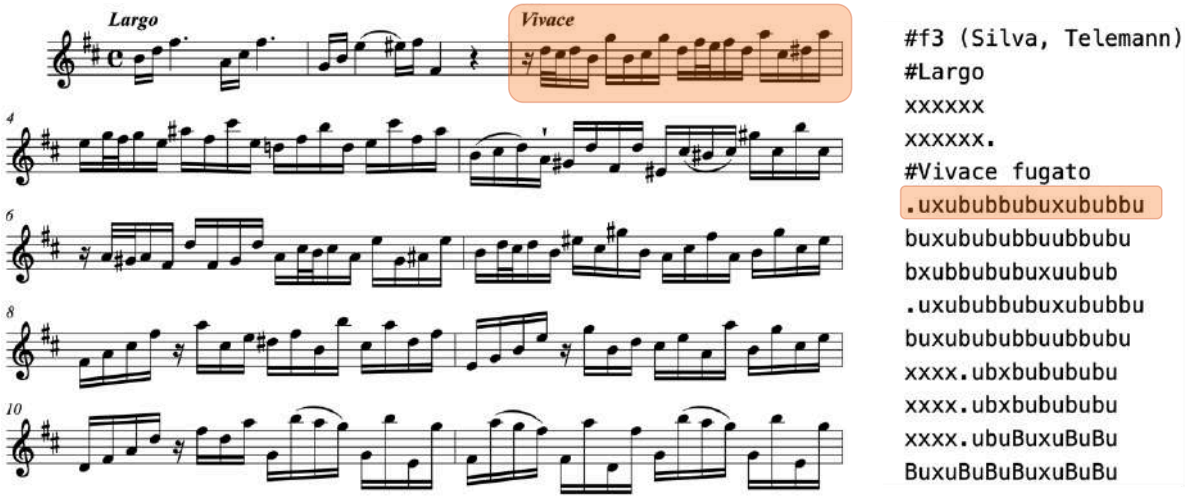


Figure A.2: Annotation raw text file of the first 11 measures of Fantasia No. 3 (TWV 40:4) with movement separations and annotations sources. The first measure of *Vivace fugato* is highlighted.

A.2 Voice Change Detection in the Fantasias

A.2.1 Reference Annotations

References annotations are extracted from structured text files in which each measure is encoded on a single line (Fig. A.2). Score repeats are not unfolded (see Section 4.1.2) to avoid artificial duplication of events. Voice transitions are derived from changes between consecutive annotated voice labels.

Rest-like symbols `.` indicate silence, while `x` denotes unannotated or unspecified voice content. These symbols are therefore excluded from transition detection. This procedure yields a binary ground-truth sequence aligned with the interval representation of the score.

A.2.2 Evaluation Protocol

The evaluation compares the binary prediction sequence with the ground-truth sequence at the interval level. A prediction is considered correct if it matches a ground-truth transition at the same position, under a one-to-one matching constraint that prevents multiple detections of the same event.

From this alignment, we compute the numbers of true positives, false positives, and false negatives, from which *precision* (the proportion of correct detections among all predictions)

and *recall* (the proportion of true changes successfully detected) are derived.

A.2.3 Results

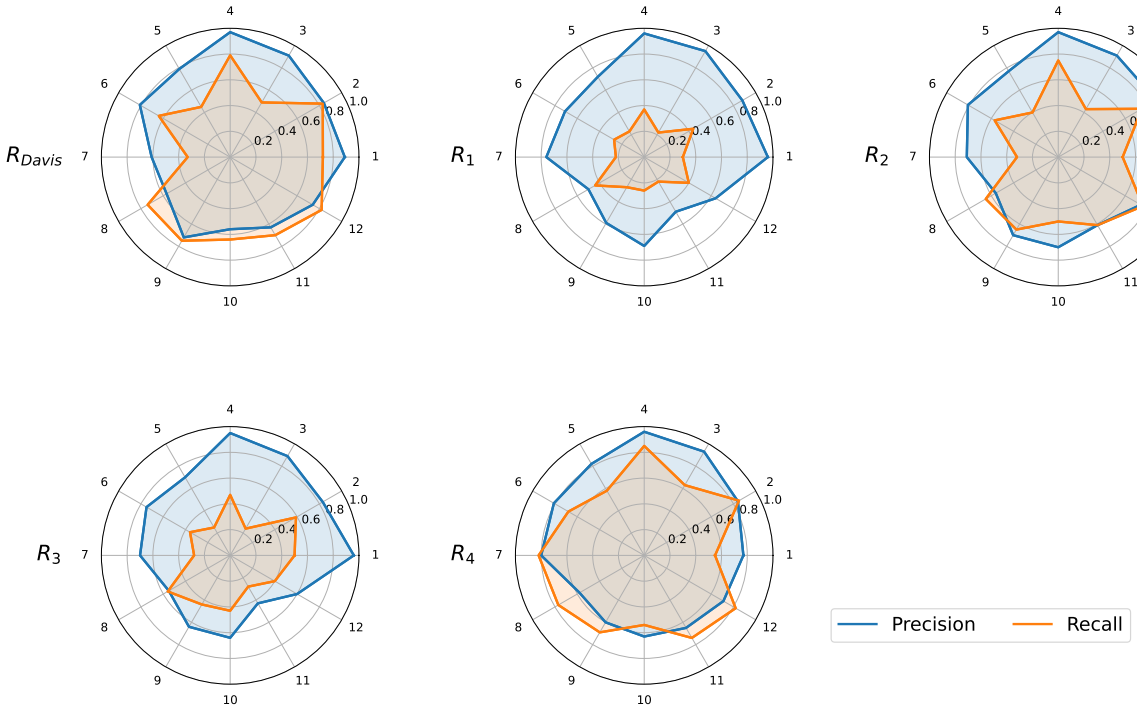
The original rule achieved a precision of 78% (Tab. A.1), though with a slightly lower recall (64%). This model is highly selective: the rule w_1 plays a decisive role. Movements with large intervals between the interleaved voices, such as the Fantasia No. 4 with sixths and tenths, perform better (more than 90% precision, see Fig. A.3 with R_{Davis}).

rules		**precision**	**recall**	F_1**-score**
	All movements			
R_{Davis}	$w_1 > 0$ and $w_1 + w_2 + w_3 \geq 4$	0.78	0.64	0.70
R_1	$w_1 \geq 4$	0.77	0.30	0.43
R_2	$w_1 > 0$ and $w_1 + w_2 \geq 4$	**0.81**	0.57	0.67
R_3	$w_1 > 0$ and $w_1 + w_3 \geq 4$	0.75	0.39	0.51
R_4	$w_1 + w_4 \geq 4$	0.80	**0.72**	**0.76**
	Fugatos only			
R_{Davis}	$w_1 > 0$ and $w_1 + w_2 + w_3 \geq 4$	0.76	0.60	0.67
R_4	$w_1 + w_4 \geq 4$	**0.78**	0.69	**0.73**

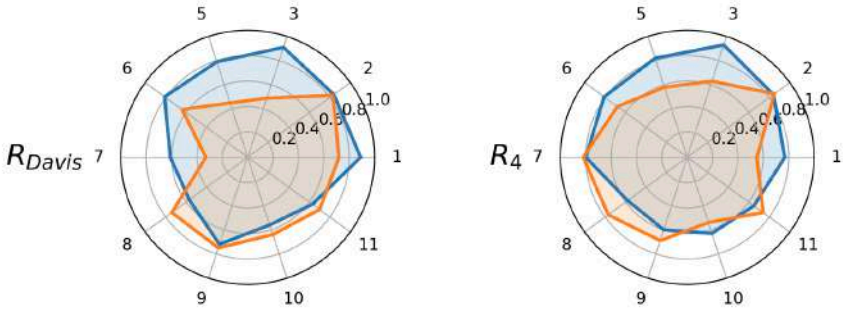
Table A.1: Classification results on the detection of interleaved voices using several sets of rules.

Conversely, the most difficult cases arise when voices are closely spaced, such as in some pedal passages like the opening measures of Fantasia No. 5, where a `C` pedal is played on off beats (Fig. A.4). To address the limitations Davis identified, we introduce a new weight w_4 and study more combinations of these weights. In this kind of excerpts, taking into account of the quick alternation of interval direction changes on a larger scale is benefit. We achieved better results while considering this new weight, the highest recall reaching 72% across the corpus (and 69% when considering only fugato movements).

The Davis model was initially designed and tested on relatively small musical monophonic examples. She evaluated the accuracy of her model on the *Allemande* from Bach's Violin Partita in *D minor* (BWV1004) by asking a group of fifteen second-year university music students to analyse the implied polyphony in this piece. In this work, we extend its application to a larger corpus: Telemann's fantasias, characterised by greater structural and contrapuntal diversity. This extension reveals several limitations of the original formulation, motivating the introduction of refined decision rules and alternative combinations of perceptual weights to improve robustness in more complex musical contexts. This naturally opens perspectives for



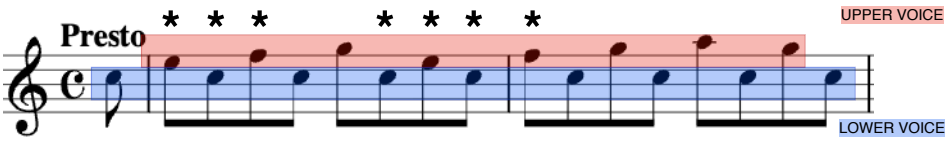
(a) Evaluation results for all fantasias.



(b) Evaluation results restricted to fugato sections.

Figure A.3: *Precision* and *recall* by fantasia for each combination of rules used in Table A.1.

future work, especially regarding the adaptation of perceptual voice change rules to larger and



w_1	0	1	1	2	2	0	0	1	1	2	2	3	3	2	2
w_2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
w_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
w_4	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
R_{Davis}	0	3	3	4	4	0	0	3	3	4	4	5	5	4	4
R_4	15	16	16	17	17	15	15	16	16	17	17	18	18	17	17

Figure A.4: *Presto* overture of the Fantasia No. 5 (TWV 40:6). The original rules R_{Davis} failed in detecting the quick alternation between voices when the interval is lower than a perfect fourth (3 times between `C` and `E`) or if even if no conjunct motion are present (4 times between `C` and `F`).

more structurally diverse corpora, as well as the integration of more robust decision mechanisms.

All necessary code for reproducing the results presented is available on https://gitlab.com/algomus.fr/telemann-voices-interleave

A.3 Study Case: Oboe 3 Parts Fugue

rules		**precision**	**recall**	F_1**-score**
R_{Davis}	$w_1 > 0$ and $w_1 + w_2 + w_3 \geq 4$	0.56	**0.39**	**0.46**
R_1	$w_1 \geq 4$	**0.70**	0.13	0.22
R_2	$w_1 > 0$ and $w_1 + w_2 \geq 4$	0.60	0.27	0.37
R_3	$w_1 > 0$ and $w_1 + w_3 \geq 4$	0.54	0.25	0.34
R_4	$w_1 + w_4 \geq 4$	0.59	0.31	0.41

Table A.2: Oboe results

Antal Doráti composed the monodic piece *Cinq pièces pour le hautbois* in 1980, dedicated to the oboist Heinz Holliger. In the third movement, *Fugue à trois voix*, Doráti adopts a neo-baroque approach. The writing is based on the systematic imitation of a subject across three voices.

Unlike Telemann's Fantasias, where voice structure is only implicitly encoded and must be inferred from surface features such as pitch proximity, register continuity, and local voice-

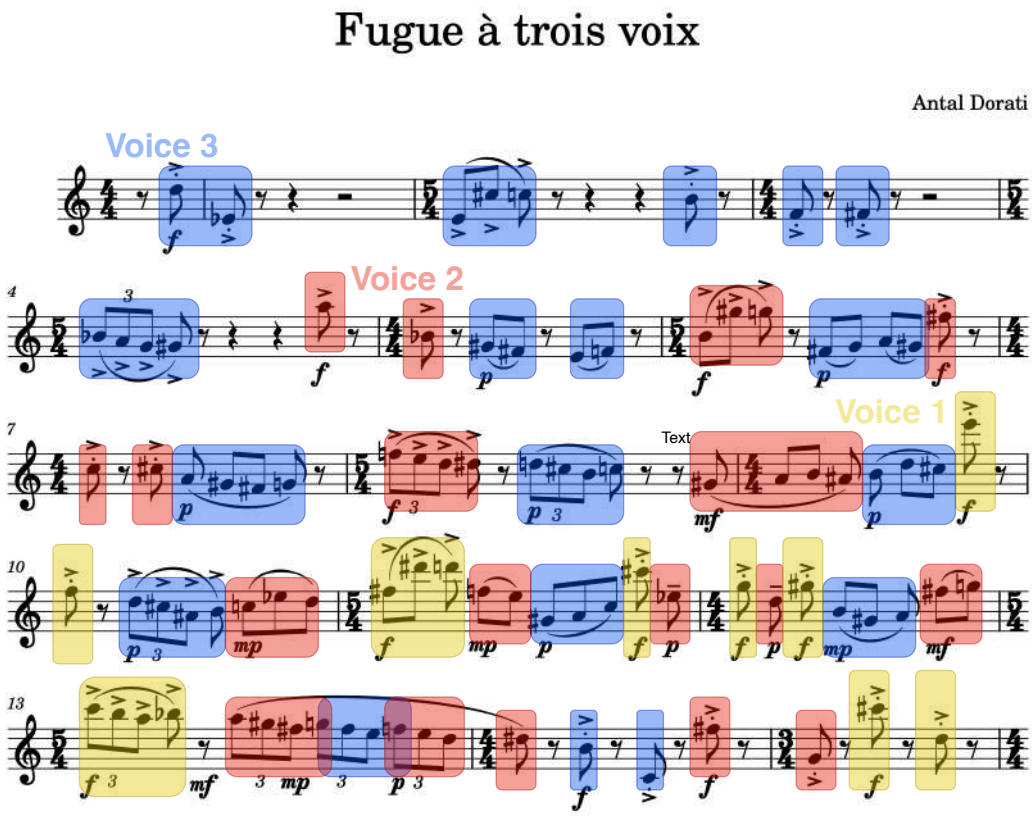


Figure A.5: First measures of *Fugue à troix voix* by Antal Doráti. The three voices are identified in appendix by the composer. Sometimes like in mm. 13 a note can be assigned up to two voices.

leading consistency, the Doráti fugue exhibits explicitly notated voices in an open-score representation [Doráti (1981)] where the three contrapuntal lines are explicitly separated. We convert this excerpt into MusicXML and annotate each voice (Fig. A.5).

A.3.1 Evaluation of the Voice-Fission Algorithm

We propose to evaluate the voice-fission algorithm introduced in Section A.1 on this 541-note excerpt, which features frequent time-signature changes. Results are reported in Table A.2.

Rule R_1 achieves a precision of 0.70 but a low recall of 0.13, while Rules R_2, R_3, and R_4 remain similarly recall-limited (0.27, 0.25, and 0.31, respectively), indicating overly conservative thresholding in dense contrapuntal contexts.

In contrast, on Telemann's Fantasias, the same rules achieve higher recall (up to 0.72 for R_4) while maintaining competitive precision (up to 0.80), yielding a best F_1-score of 0.76. The

drop in best F_1-score to 0.46 on the fugue highlights a regime shift: from sparse, implicitly structured polyphony to strict contrapuntal writing where voice identity is less aligned with pitch-based separability.

The explicit separation of voices proposed by Doráti does not correspond to an easier computational structure: the voices frequently cross and exchange registral positions, and the contrapuntal organisation relies more on motivic imitation than on stable pitch stratification. As a consequence, voice changes are not consistently aligned with clear separations in pitch space. This weakens the assumptions underlying pitch-driven voice-fission methods, which typically rely on the correlation between voice identity and pitch proximity. The voice-fission detection becomes more challenging.

List of Figures

1.1 First measures of "ECHOS Pour la Flûte Traversiere seule" by Jacques Hotteterre (1708) [[Hotteterre (1708)]], the earliest known score for solo flute. You can listen a performance by Guillermo Peñalver at traverso (https://patthmus.github.io/phd-manuscript-links/chap1.html). 12

1.2 Drawing and description of the Böhm flute in an english translated edition of the french *"Famous Complete Flute Method"* ALTES (https://imslp.org/wiki/Special:ReverseLookup/280140). 14

1.3 Quantz's double keywork from the Supplement to the Diderot's Encyclopédie (http://ark.bnf.fr/ark:/12148/cb30340683w.public). 15

1.4 Personal photograph of Dr. Heinz Prager's flute collection exposed at the Deutsches Museum (Munich, Germany). This collection was compiled by Heinz Prager together with Ludwig Böhm's (great great-grandsom of Theobal Böhm). We can see the evolution of the models, from a conical Flute (T. Böhm 1838–1839) at the top to a cylindrical flute, Böhm system No. 67 (T. Böhm 1852) at the bottom. 18

1.5 Thesis Outline. 19

2.1 Score and interpretation examples by Rampal and Leleux (https://patthmus.github.io/phd-manuscript-links/chap2.html) of Fantasia No. 7 by G.P. Telemann (TWV 40:8) in *D Major*, mm. 89–95. 25

2.2 Fantasia for solo flute No. 4 (TWV 40:5) by G. P. Telemann, mm. 1–16, autograph score. 26

2.3 Fantasia No. 2 (TWV 40:3), mm 1–2. 27

2.4 Tree-based score representations of Figure 2.3 exerpts, in MusicXML and MEI. . 27

2.5 Piano-roll representation corresponding to the Figure 2.3, displayed in GarageBand (DAW). 29

2.6 Various audio representations of the first 6 seconds of Fantasia No. 2 (corresponding score Fig. 2.3). Recorded by the author on a Yamaha YFL-587 modern Flute (Würzburg, Germany, April 2025). From top to bottom: waveform, STFT, STFT (logarithmic frequency axis). The following audio representations (Fig. 2.7, 2.8, 2.9, 2.10, 2.11) are based on the same audio signal and the same magnitude scale and correspond to a performance derived from the score excerpt presented in Fig. 2.3. 32
2.7 Pitch features computed via an IIR filterbank aggregated as short-time mean-square power. 33
2.8 Constant-Q transform (CQT) of the audio signal, showing the time–pitch representation with the vertical axis expressed in MIDI pitch values. 35
2.9 Various chroma-based representations. From top to bottom: chroma, CENS, and DLNCO. 38
2.10 Mel-frequency cepstral coefficient (MFCC) representation of the audio signal, with the vertical axis corresponding to cepstral coefficient indices. 39
2.11 Multi-pitch estimation representation of the audio signal, with estimated pitches shown over time [[Weiß et al. (2021b)]]. 40
2.12 Spectral flux of the introduction recording of Fantasia No. 2 (Fig. 2.6), computed using *libfmp* [[Müller and Zalkow (2021)]]. The vertical lines indicate the positions of the manually annotated reference onsets. Peaks in the spectral flux novelty function highlight potential onset locations. 41
2.13 Interactions in performance (Images of the flautist and the audience generated using OpenAI DALL·E, 2026). 44
2.14 Excerpt of Chapter "De entrelaceur, & de la maniere de chanter les demiminimes" from "Le Droict Chemin de musique" by Loys Bourgeois [[Bourgeois (1550)]]. 47
2.15 Analysis by the author of the first three measures of the last movement, *Rondo*, from Fantasia No. 6 (TWV 40:7) for solo flute by G. P. Telemann, according to Narmour's *Implication–Realization* model. The labels indicate process (`P`), intervallic process (`IP`), and vectorial process (`VP`). 49
2.16 From [[Davis (2006)]]: excerpt of J. S. Bach's *Allemande* from Violin Partita No. 2 in *D minor*, mm. 4–7. Numbers represent point totals generated by Davis's rule system; circles indicate intervals that mark the transition from one implied voice to another. 51

3.1 Waveforms of the first 30 s of *Syrinx* (L. 129 - CD. 137) by Claude Debussy, as performed by Emi Ferguson (top) and Emmanuel Pahud (bottom). Corresponding score onsets are connected by red lines (https://patthmus.github.io/phd-manuscript-links/chap3.html). 55

3.2 From left to right: *Dezrann* an open-source web platform to interact with corpora while sharing music and music analysis [[Ballester et al. (2025)]] https://www.dezrann.net/~/bach-fugues/bwv851?audio=1; *Metronaut* (screenshot of a commercial video): a real-time score-following application using *Antescofo* https://metronautapp.com/; An overview of WSD: a large-scale annotated and audio-to-audio synchronised operas [[Weiß et al. (2023)]]. 59

3.3 Cost matrix, chroma vectors and DTW warping path of the Intro of Flute 2, Grave movement (see Fig. 2.3), performed by J.-P. Rampal and E. Pahud on modern flutes. 64

3.4 Multilevel refinement of the warping path corresponding to Fig. 3.3 using the MrMsDTW algorithm. From coarse (top-left) to fine resolution (bottom-right), the alignment corridor progressively narrows, enabling accurate local synchronisation under strict memory constraints. In our experiments, we used four resolution levels with smoothing windows [201, 101, 21, 1] and downsampling factors [50, 25, 5, 1]. Step weights were set to [1.5, 1.5, 2.0] to favour diagonal transitions, and the memory threshold was set to 10^6. Visualisations are produced by *SynctoolBox* (Section 3.4.2): white regions are absent at intermediate levels due to full cost-matrix coverage, while finer levels use locally constrained regions defined by neighbouring anchors, leading to partial coverage of the alignment cost matrix. 69

3.5 Triple transfer principle as proposed by Zalkow et al. (schema from [[Zalkow et al. (2017)]]). 74

3.6 Word cloud of MIR dataset names from https://www.audiocontentanalysis.org/datasets.html. 76

4.1 Density 21.5 is a solo flute piece by Edgard Varèse, composed in 1936 and revised in 1946. It was written for Georges Barrère's platinum flute made of a 90/10 platinum–iridium alloy with a similar density (∼21.6 g/cm³). Listen to Claire Chase's performance (2019, Cambridge, MA) at https://patthmus.github.io/phd-manuscript-links/chap4.html. 80

4.2 Autograph score of Fantasia No. 2 (TWV 40:3) by G.P Telemann. Four Sections are outlined corresponding to different movements. The boxed excerpt is shown in Fig 4.3. 82

4.3 In the *Vivace* of Fantasia No. 2, some of the eighth notes in the same beamed group have stems down while the others have stems up. Telemann uses these typographic changes in direction to emphasise the interleaved lines. 83

4.4 Treemaps of the repartition of the movements in the fantasias 85

4.5 The *Vivace fugato* from Fantasia No. 1 (TWV 40:2), mm. 11-16 with harmonic labels and identification of the fugato form available on Dezrann. After exposition of a first subject (`S`), the answer (`A`) is combined with the counter-subject (`CS`). Later an episode (`E`) develops the preceding material through a sequence of harmonic progressions. 87
4.6 The first measure of Fantasia No. 3 (TWV 40:4) implies harmonies, tonic (`I`) then dominant (`v6`). The local key being there the main key *B minor*. 89
4.7 Performed repeats structure of Fantasia No. 12 (TWV 40:13) last movement (*Rigaudon*). 92
4.8 Scrapbook of programmes and press articles on Gabriel Mourey's Psyché (Act III, Scene 1), Bibliothèque nationale de France, Performing Arts Department, 8-RF-68061. 94
4.9 *Sonic Visualiser* screenshot during the annotation process. 97
4.10 Cell-level (`A,B,C,E`) and phrase-level (`P,Q`) patterns in Debussy's *Syrinx*. Full identification of all patterns in the score is provided in Fig. 6.10. 99
4.11 *Density 21.5* with 40 pattern α occurrences identified based on Ohriner's definition. 100

5.1 Vermont Counterpoint by Steve Reich linked to a performance by flautist Joe LaRocca. 104
5.2 Score-to-audio alignment pipeline. 106
5.3 Structure of Fantasia No. 1 (TWV 40:2) with movements names. Four *mordents* are indicated by a "+" sign in the score. 108
5.4 Signed alignments errors by onset in Fantasia No. 1 performed by Lazarevitch. . 109
5.5 Transcription made by the author from Lazarevitch's performance of the Fantasia No. 1 *Passepied*. The two performed mordents present in the original score are annotated +. All the measures with added ornamentations are highlighted. . 112
5.6 Three different realisations of the final repeats in the closing section (following the *Dal Segno al Fine* instruction) of the last movement (*Rigaudon*) from Fantasia No. 12 (TWV 40:13). Highlighted segments illustrate structural variants (`AAB`, `AB`, and `AABB`), with measure annotations overlaid on the waveforms. 115
5.7 Score-to-audio pipeline used for comparing *Chroma*, *Spectral Template*, and *Neural MPE* audio features representations. 116
5.8 Box plots of the signed alignment errors grouped by performer and by audio feature representation for Telemann's fantasias recordings. The boxes span from the first to the third quartile, while the whiskers extend to the most extreme data points within 1.5 times IQR. A horizontal line indicates the median. . . . 119
5.9 *Neural MPE* alignment signed errors for Fantasia No. 1 across all performers and annotators (A1, A2). Alongside the boxplots, violin plots depict error distributions, and jittered points reveal individual alignment deviations. Values are clipped to ± 150 ms before visualisation. 122

5.10 Mm.1–3 from "*Du großer Schmerzensmann*" by Martin Jan (1663) 🔊. Six different recordings of the soprano part are available in the dataset (alto saxophone, clarinet, flugelhorn, flute, oboe, and trumpet). The presence of repeated notes introduces alignment ambiguities. 124
5.11 Audio-to-audio alignment pipeline. 126
5.12 Triple transfer pipeline. 128
5.13 Misalignment rates as a function of the tolerance threshold τ for six alignment strategies on (a) *Syrinx* by using Auer performance for evaluation and on (b) *Density 21.5* against Zukerman performance. M_{R_1} and M_{R_2} are single-reference baselines, $M_{HardLateFusion}$ the original late-fusion, and $M_{SoftLateFusion}$ our softmax-weighted variant. M_{Best} and M_{Worst} are local oracle strategies selecting, for each onset, the reference with the smallest and largest alignment error, respectively. 131
5.14 For different values of α parameter, soft-late fusion alignment : (a) absolute mean error ; (b) standard deviation error; and (c) gain against the best single-reference transfer. 134

6.1 Chronomètre from [[Loulié (1696)]]. 137
6.2 Tempo by measure in *bpm*, computed on three consecutive measures, for six recordings on traverso and modern flute of Fantasia No. 1 (TWV 40:2) (🕓 in Tab. 4.3). All performers underline the four movements of this Fantasia. 139
6.3 The second episode of the *Vivace fugato* movement of Fantasia No. 2 (TWV 40:3) begins at mm. 38. It features two interleaved voices, a descending bass line and an upper voice 🔊. In the autograph score, Telemann occasionally highlights the bass voice in these measures by using stems pointing in opposite directions within the same beam, as here on notes 5, 7, and 9. 141
6.4 The ΔIOI and Δo values for the passage shown in the Figure 6.3 are compared across two performances on a modern flute. In Porter's performance (a) 🔊, the average over absolute values of ΔIOI is 2.9 %. In contrast, Rampal (b) 🔊, who plays this movement at a faster tempo, has less deviations (1.9 %). 142
6.5 First measures of Fantasia No. 3 *Gigue* with groups of three eighths within a beat (♩♫) 145
6.6 $\Delta IOI(\%)$ of the 1st, 2nd, and 3rd eighth notes within a beat in groups of three, across all movements in a triple metre and for all performers. The boxes span from the first to the third quartile, while the whiskers extend to the most extreme data points within 1.5 times the interquartile range (IQR). The median is indicated by a horizontal line, and the mean by a '+' sign. 145
6.7 First measures of Fantasia No. 3 *Vivace Fugato* with groups of four sixteenth notes within a beat (♬♬). 146

6.8 Overview of the Streamlit web application (*Measures by performer* module) accompanying the fantasias performances comparison (https://telemannvoicesinterleave.streamlit.app/). 147
6.9 Tempo curves for the 23 *Syrinx* performances with patterns `A` and `P` (see Fig. 6.10). The tempo is expressed in QPM. 149
6.10 All occurrences of cell-level (`A,B,C,E`) and phrase-level (`P,Q`) patterns in Debussy's *Syrinx*. 150
6.11 Beginning (mm 1–13) of *Density 21.5* with 9 pattern α occurrences as defined by Ohriner [[Ohriner (2019)]] (see Fig 4.11) 🔊. 150
6.12 8 patterns are identified in mm. 9–12 of *Syrinx* in Dezrann (https://www.dezrann.net/~/early20thsoloflute/syrinx?m=9). The first notes of `P` pattern are omitted (see footnote 1). An excerpt of exported *dezfile* code is given for the first cell-level pattern `A`. 151
6.13 First occurrence of *Syrinx* pattern `A` and corresponding waveform extracted from the start of Auer's 🔊 and Ferguson's 🔊 recordings. The pattern considers three notes (1-3) and the onset of the next note (4). 152
6.14 Duration and timing fluctuation for the 11 first occurrences of pattern `A` in *Syrinx* (see Fig. 6.10). The boxes show the IQR, whiskers extend to $1.5 \times IQR$, and horizontal line marks the median. The occurrences 1–2, 4–5, 8–9, and 10–11 are colocated in a large `P` phrase-level pattern. 154
6.15 Duration of pattern `B` in *Syrinx* (see Fig. 6.10), across performers and occurrences (top, heatmap), and summarised by occurrence (bottom, boxplots as in Fig. 6.14).156
6.16 Duration and timing fluctuation for the 9 first occurrences of pattern α in *Density 21.5* (see Fig. 6.11). 157
6.17 40 occurrences of α pattern in *Density 21.5*. 158

7.1 Ian Anderson performance (https://patthmus.github.io/phd-manuscript-links/chap7.html) of *The Locomotive Breath* (Jethro Tull, live at the Rockpop In Concert 10.7.1982, Dortmund, Germany). 161
7.2 In [Chapter 8, p. 23] [[Hotteterre (1719)]], Hotteterre describes two tonguing techniques: *articulez* (**Tu**) and *coulez* (**Ru**), together with their usage. The original text reads: *"On doit prononcer Ru sur la Note qui fuit la Croche quand elle monte ou descend par degrez conjoints."* (author's translation: "One should pronounce *Ru* on the note that follows the eighth note when it ascends or descends by conjunct degrees"). 163
7.3 *Passepied* of Fantasia No. 1 in *A major* (TWV 40:02), mm. 1–2. Different tonguing patterns proposed by Brown rely on Quantz and Hotteterre treatises. Choosing adapted articulations "plays a vital part in bringing out one voice to the foreground and placing another in the background" [[Brown (2015)]]. 164

7.4 Extended Techniques for flute panel from https://www.flutecolors.com/techniques/. 165
7.5 Commercial picture of *Aerophone Brisa* from Roland (https://www.roland.com/fr/products/aerophone_brisa/). 165
7.6 Sketch from [[Vauthrin et al. (2015)]] showing reflective markers placed on the flautist's thorax to track chest wall movements recorded using an optoelectronic system with infrared cameras. 167

A.1 Vivace from Fantasia No. 1 (TWV 40:2), mm 14–15. Each component of $\mathbf{w}_i$ (where i corresponds to the note index in the score) is computed according to Davis's rules. In this excerpt, using a detection threshold of 4, the model detects six voice changes, notably between G and C (mm. 14). 170
A.2 Annotation raw text file of the first 11 measures of Fantasia No. 3 (TWV 40:4) with movement separations and annotations sources. The first measure of *Vivace fugato* is highlighted. 172
A.3 *Precision* and *recall* by fantasia for each combination of rules used in Table A.1. 174
A.4 *Presto* overture of the Fantasia No. 5 (TWV 40:6). The original rules R_{Davis} failed in detecting the quick alternation between voices when the interval is lower than a perfect fourth (3 times between `C` and `E`) or if even if no conjunct motion are present (4 times between `C` and `F`). 175
A.5 First measures of *Fugue à troix voix* by Antal Doráti. The three voices are identified in appendix by the composer. Sometimes like in mm. 13 a note can be assigned up to two voices. 176

List of Tables

1.1 Significant Solo Flute Works: Baroque to the Mid-20th Century 16

2.1 Typical usages of common audio representations organised by abstraction level, ranging from low-level signal descriptors to higher-level harmonic and timbral features. 43

4.1 The musicological sources collected in this dataset analyse some or all of the Telemann's *twelve fantasias* referred by their number. For the movements names some sources propose either full identification by fantasia or just sparse information (partial). For the fugatos movements, in the same way, the fugato can be just identified (partial) or fully described. Concerning voices annotations only three fantasias are exhaustively analysed , in the other cases either voices are fully identified on specific movements or are given in a sparse way . Harmony annotations include local keys *(LK)* and chord progressions *(CP)*. . . 84

4.2 Distribution of main voice annotations in groups of four sixteenths in duple metre movements of the corpus. Bold values show alternating upper/bass voices. . . . 90

4.3 Complete audio recordings of the 12 Telemann fantasias used in this thesis. Recorded during the COVID-19 lockdown, the Pitelina recording is the only non-commercial one and also contains videos. The hyperlink column reports, when available, YouTube playlists (YT) and MusicBrainz identifiers (MBID). The last columns indicate inclusion in two different subsets (,) and their corresponding main applications. 91

4.4 Min/max measures played per fantasia across performances. 92

4.5 Among the 23 performances used in this study for each piece, three are manually annotated: R_1 and R_2 are the reference performances and E (also designed as V_3) the evaluative one. 95

4.6 Number of onsets; minimum, maximum and total duration among the 46 recorded performances used. 96

4.7 Comparison of IOI-based timing statistics across the six manually annotated performances of *Syrinx* and *Density 21.5*. Mean μ, median $\tilde{x}$, and standard deviation σ of IOI are reported in seconds. The coefficient of variation (CV), defined as $CV = \sigma/\mu$, provides a normalised measure of temporal variability across performances. 98
4.8 Overview of the datasets: symbolic labels, aligned performances, and manual onset annotations. 101

5.1 Absolute errors mean ($\bar{e}$) , standard deviation (σ), and absolute errors median (μ_e) by movement with and without taking account of ornaments on score-to-audio alignments of Fantasia No. 1 performed by Lazarevitch. We also report the number of onsets and annotated voices. 110
5.2 Overview of the subset of Telemann's *Twelve fantasias* audio recordings used in this Chapter (🎧 in Tab. 4.3). 114
5.3 Transpositions for Recorders performances. All transpositions are ascending intervals: `m3` minor third, `P4` perfect fourth. 114
5.4 Median alignment **signed** errors (in ms) and interquartile ranges expressed as median ± IQR/2 for the three audio feature representations across performers and their instruments. For each performer, bold indicates lowest **IQR/2 values** and gray-shading indicates the median closest to zero. 118
5.5 Median ± IQR/2 (in ms) of **signed** alignment errors for *Neural MPE* representation, by performer and shift type. 120
5.6 Median ± IQR/2 (in ms) of **absolute** alignment errors by performer. For each representation, we show unshifted (none) and representation-shifted (rep.) results. 121
5.7 Median ± IQR/2 (in ms) of signed alignment errors for individual flute soprano parts across the *ChoraleBricks* corpus, comparing the unshifted *Neural MPE* model with the representation-shifted version. Piece IDs are formed from the composer's abbreviation (e.g., BA for Bach) and the corresponding work number in the dataset. Although each piece includes several recordings for the Soprano, Alto, Tenor, and Bass parts, only the flute parts are reported here. 124
5.8 Absolute alignment errors (in ms) for six strategies using Auer's performance as evaluation playing *Syrinx* and Zukerman's performance for *Density 21.5*. Columns indicate percentiles (25–99.9th), mean $\bar{e}$, and standard deviation σ_e. Cells shaded in gray show the value closest to the Best local oracle strategy. . . 130
5.9 Cross-evaluation configurations based on reference–target permutations for *Syrinx* and *Density 21.5*. 132
5.10 Alignment error statistics across strategies (top) and cross-evaluation per target performance (bottom). Values are mean ± standard deviation in ms. 133

6.1 Overview of the subset of the six recordings of Telemann's *Twelve Fantasias* used for the expressive timing study (see 🕓 and Tab. 4.3). 140
6.2 Mean and standard deviation of the $\Delta IOI(n)$ of notes (excluding rests), in percentage on the length of each measure, for the six performers on all Telemann Fantasias. Values above 1% are in bold. (Top). The non-annotated notes (column `x`) are played with small deviations in contrast to the upper and lower voices (columns `u`, `b`+`B`). (Midle and Bottom). Focus on groups of notes within the same beat. Several performers deviate annotated notes written with interleaved voices more than when they are not. They also emphasise the first note of the group. 143
6.3 Cohen's d effect size for Mann-Whitney U test measuring the magnitude of the difference between two groups on the metric $\Delta IOI(n)$, focusing on cases where the p-value is $< 10^{-2}$ (•) or $< 10^{-3}$ (all others). 144
6.4 Pattern occurrences analysed across 23 performances of *Syrinx* (`A`–`E`) and *Density 21.5* (α). For each metric (D, TF), eta-squared (η^2) is computed per pattern. Min and max occurrences denote the lowest and highest mean values across performers, respectively (index and value are reported). Cohen's d quantifies the contrast between them. Effect sizes are omitted (–) when p-value $> 10^{-2}$. . . 153

A.1 Classification results on the detection of interleaved voices using several sets of rules. 173
A.2 Oboe results 175

Bibliography

John Aach and George M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, June 2001. (p. 61)

Claude Abromont and Eugène de Montalembert, editors. *Guide de la théorie de la musique*. Les indispensables de la musique. Fayard ; H. Lemoine, Paris, 2001. (pp. 88 and 89)

Ruchit Agrawal, Daniel Wolff, and Simon Dixon. A Convolutional-Attentional Neural Framework for Structure-Aware Performance-Score Synchronization. *IEEE Signal Processing Letters*, 29:344–348, 2022. (p. 62)

Andreas Arzt. *Flexible and robust music tracking / submitted by Andreas Arzt*. PhD thesis, Universität Linz, Austria, November 2016. (p. 60)

Andreas Arzt, Sebastian Bock, Sebastian Flossmann, Harald Frostel, Martin Gasser, Cynthia C S Liem, and Gerhard Widmer. The Piano Music Companion. In *Proccedings of the 21st European Conference on Artificial Intelligence (ECAI'14)*, Prague, Czech Republic, 2014. (p. 60)

Stefan Balke, Axel Berndt, and Meinard Müller. ChoraleBricks: A modular multitrack dataset for wind music research. *Transaction of the International Society for Music Information Retrieval (TISMIR)*, 8(1):39–54, 2025. (pp. 77 and 123)

Stefan Balke, Johannes Zeitler, Vlora Arifi-Müller, Brian McFee, Tomoyasu Nakano, Masataka Goto, and Meinard Müller. RWC Revisited: Towards a Community-Driven MIR Corpus. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 9(1), 2026. (p. 76)

Charles Ballester, Baptiste Bacot, Louis Bigo, Vanessa Nina Borsan, Louis Couturier, Ken Déguernel, Quentin Dinel, Laurent Feisthauer, Klaus Frieler, Mark Gotham, Richard Groult, Johannes Hentschel, Alexandre d'Hooge, Dinh-Viet-Toan Le, Florence Levé, Francesco Maccarini, Ivana Maričić, Gianluca Micchi, Meinard Müller, Alexandros Stamatiadis, Tom

Taffin, Patrice Thibaud, Christof Weiß, Rui Yang, Emmanuel Leguy, and Mathieu Giraud. Interacting with Annotated and Synchronized Music Corpora on the Dezrann Web Platform. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 8(1), May 2025. (pp. 21, 59, 71, 73, 76, 152, and 180)

Barthol Kuijken. 12 Fantasias TWV 40:2-13 (with Facsimile), 1987. MR 2167. (p. 84)

M.A. Bartsch and G.H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, February 2005. (pp. 36, 37, and 58)

Jonathan W. Bernard. *The music of Edgard Varèse*. Composers of the twentieth century. Yale University Press, New Haven, 1987. (p. 94)

Axel Berndt. The Music Performance Markup Format and Ecosystem. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 50–57, Online, 2021. (p. 46)

DJ Berndt and J Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994. (p. 62)

Janice Dockendorff Boland. *Method for the one-keyed flute, baroque and classical*. University of California Press, Berkeley, 1998. (p. 48)

Stanley Boorman, Eleanor Selfrodge-Field, and Donald W. Krummel. Printing and publishing of music, 2001. URL https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040101. (p. 26)

Christopher Bork, Kenneth Ng, Yinhan Liu, Alex Yee, and Michael Pohlscheidt. Chromatographic peak alignment using derivative dynamic time warping. *Biotechnology Progress*, 29(2):394–402, 2013. (p. 61)

Michael C Botha. fluteworx.co.za/articles/LouisLotArticle.pdf, October 2016. URL https://www.fluteworx.co.za/articles/LouisLotArticle.pdf. (p. 133)

Loys Bourgeois. *Le droict chemin de musique, composé par Loys Bourgeois, avec la manière de chanter les Pseaumes par usage ou ruse...* s.n., Geneva, 1550. (pp. 47 and 179)

Judith C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, January 1991. (p. 35)

Rachel Brown. Telemann Fantasias: a feat of ingenuity and inspiration, 2015. (pp. 84, 164, and 183)

George W. Bullen. The Galin-Paris-Cheve Method of Teaching Considered as a Basis of Musical Education. *Proceedings of the Musical Association*, 4:68–93, 1877. (p. 25)

Donald Byrd and Tim Crawford. Problems of music information retrieval in the real world. *Information Processing & Management*, 38(2):249–272, March 2002. (p. 84)

Sebastian Böck and Gerhard Widmer. Maximum Filter Vibrato Suppression for Onset Detection. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, September 2013. (p. 41)

Emilios Cambouropoulos. Voice And Stream: Perceptual And Computational Modeling Of Voice Separation. *Music Perception*, 26:75–94, September 2008. (p. 170)

Chris Cannam, Christian Landone, and Mark Sandler. Sonic visualiser: an open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1467–1468, Firenze Italy, October 2010. ACM. (p. 73)

Michel A. Cara and Divna Mitrovic. Coupling of anticipation and breathing in expert flute performance: the influence of musical structure and practice. *Frontiers in Cognition*, 3, September 2024. (p. 167)

M Castellengo. Acoustical Analysis of Initial Transients in Flute Like Instruments. *acta acustica*, 85, 1999. (p. 113)

Antoine Chaigne and Jean Kergomard. *Acoustique des instruments de musique*. Échelles. Belin, Paris, 2008. (p. 15)

Elaine Chew and Xiaodan Wu. Separating Voices in Polyphonic Music: A Contig Mapping Approach. In *Computer Music Modeling and Retrieval: Second International Symposium, CMMR 2004*, pages 1–20, Esblerg, Denmark, May 2004. (p. 170)

Jaeran Choi, Taegyun Kwon, and Juhan Nam. Predicting Flutist Onset Timing in Duet Performance: A Multimodal Analysis of Gesture and Breath Cues. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, Daejeon, South Korea, September 2025. (p. 167)

Eric F. Clarke. The perception of expressive timing in music. *Psychological Research*, 51(1): 2–9, June 1989. (p. 47)

Jacob Cohen. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33(1):107–112, 1973. (p. 153)

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic Press, 1988. (pp. 144 and 154)

Darrell Conklin and Mathieu Bergeron. Discovery of Contrapuntal Patterns. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010. (p. 98)

Arshia Cont. ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music. In *International Computer Music Conference (ICMC)*, pages 33–40, 2008. (p. 59)

Arshia Cont. A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, June 2010. (pp. 34 and 60)

Nicholas Cook. *Beyond the score: music as performance*. OxfordUniversity Press, New York, 2013. (p. 13)

Isabelle Cossette, Pawel Sliwinski, and Peter T. Macklem. Respiratory parameters during professional flute playing. *Respiration Physiology*, 121(1):33–44, June 2000. (p. 166)

Michael Scott Cuthbert and Christopher Ariza. Music21: a toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 637–642, 2010. (pp. 96 and 117)

Marco Cuturi and Mathieu Blondel. Soft-DTW: a Differentiable Loss Function for Time-Series. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 894–903. PMLR, July 2017. (p. 62)

Mary Cyr. *Performing baroque music*. Routledge, London, 2016. (p. 48)

Roger B. Dannenberg. An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, volume 84, pages 193–198, 1984. (pp. 59 and 60)

Roger B. Dannenberg and Ning Hu. Polyphonic Audio Matching for Score Following and Intelligent Audio Editors. *Carnegie Mellon University Research Showcase*, February 1987. (pp. 57 and 60)

Stacey Davis. Implied Polyphony in the Solo String Works of J. S. Bach: A Case for the Perceptual Relevance of Structural Expression. *Music Perception*, 23(5):423–446, July 2006. (pp. 51, 170, and 179)

Patricio De, Patricio de la Cuadra, Benoit Fabre, Nicolas Montgermont, and Laurent de Ryck. Analysis of Flute Control Parameters: A Comparison Between a Novice and an Experienced Flautist. *Acta Acustica*, 94, September 2008. (p. 167)

Robert de Bree. Telemann's Fantasias: a Fantastic Source of Inspiration. How can I improve my improvisatory skills , using Telemann's Fantasias ?, 2020. (pp. 84 and 90)

Peter Desain and Henkjan Honing. Tempo curves considered harmful. *Contemporary Music Review*, 7(2):123–138, January 1993. (p. 45)

Johanna Devaney and Hubert Léveillé Gauvin. Encoding music performance data in Humdrum and MEI. *International Journal on Digital Libraries*, 20, March 2019. doi: 10.1007/s00799-017-0229-3. (p. 46)

Johanna Devaney, Michael Mandel, and Ichiro Fujinaga. A Study of Intonation in Three-Part Singing Using the Automatic Music Performance and Comparison Toolkit (AMPACT). In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, 2012. (p. 71)

Johanna Devaney, Daniel McKemie, and Alex Morgan. pyAMPACT: A Score-Audio Alignment Toolkit for Performance Data Estimation and Multi-modal Processing, January 2026. (pp. 46, 59, and 71)

Sofia B. Dias and Leontios J. Hadjileontiadis. Telemann's Fantasia No 2 à Travers Sans Basse: Compositional Intentions, Performance Characteristics and Audio Footprint Signal Processing Analysis. In Luísa Correia Castilho, Rui Dias, and José Francisco Pinho, editors, *Perspectives on Music, Sound and Musicology: Research, Education and Practice*, Current Research in Systematic Musicology, pages 31–47. Springer International Publishing, Cham, 2021. (pp. 39 and 45)

Simon Dixon. Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research (JNMR)*, 30(1):39–58, March 2001. (p. 46)

Simon Dixon. Live tracking of musical performances using on-line time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx)*, 2005. (p. 71)

Simon Dixon and Gerhard Widmer. MATCH: A Music Alignment Tool Chest. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, London, United Kingdom, 2005. (p. 71)

Anton Doráti. *Cinq pieces pour le hautbois*. Boosey & Hawkes, 1981. (p. 176)

W. J Dowling. The perception of interleaved melodies. *Cognitive Psychology*, 5(3):322–337, November 1973. (pp. 50, 51, and 52)

J Stephen Downie. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003. (p. 13)

William Drabkin. Motif, 2001. URL https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000019221. (p. 98)

Zhiyao Duan and Bryan Pardo. A state space model for online polyphonic audio-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 197–200, Prague, Czech Republic, May 2011. IEEE. (p. 62)

Katelyn Emerson and Peter M. C. Harrison. Multimodal Datasets for Studying Expert Performances of Musical Scores. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 8(1), December 2025. (p. 13)

Eppinger, S. Georg Philipp Telemann : 12 Fantasien für Flöte solo, Teil I. *TIBIA*, 2:86–99, 1984. (pp. 84 and 90)

Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, 2011. (p. 37)

Sebastian Ewert, Meinard Muller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1869–1872, April 2009. (pp. 38 and 58)

Neville H. Fletcher and Thomas D. Rossing. *The physics of musical instruments*. Springer, New York Heidelberg, 2. ed., corr. 7. print edition, 2008. (p. 14)

Claire A. Fontijn. Quantz's unegal: implications for the performance of 18th-century music. *Early Music*, XXIII(1):54–64, February 1995. (p. 48)

Francesco Foscarin, Andrew McLeod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai. ASAP: A Dataset of Aligned Scores and Performances for Piano Transcription. In *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada, 2020. (p. 77)

Christian Fremerey, Meinard Müller, and Michael Clausen. Handling Repeats and Jumps in Score-performance Synchronization. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 243–248, Utrecht, Netherlands, 2010. (p. 93)

Anders Friberg and Giovanni Battel. Structural Communication. In *The Science & Psychology of Music Performance: Creative Strategies for Teaching and Learning*. Oxford University Press, April 2002. (p. 46)

Joachim Fritsch and Mark D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 888–891, Vancouver, BC, Canada, May 2013. IEEE. (p. 77)

David Fuller. *Notes inégales*, volume 1 of *The New Grove Dictionary of Music and Musicians*. Oxford University Press, 2001. (p. 47)

Johann Joseph Fux, Alfred Mann, and John Edmunds. *The study of counterpoint from Johann Joseph Fux's Gradus ad parnassum*. Number N277 in The Norton library. W. W. Norton, New York, rev. ed., 32. print edition, 1971. (pp. 86 and 88)

François-Joseph (1784-1871) Auteur du texte Fétis. *Traité du contrepoint et de la fugue contenant l'exposé analytique des règles de la composition musicale depuis deux jusqu'à huit parties réelles,... / par F.-J. Fétis,...* Janet et Colette, Paris, France, 1825. URL https://gallica.bnf.fr/ark:/12148/btv1b52501394b. (p. 87)

D. Gabor. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93 (26):429–441, November 1946. (p. 31)

Gabrielsson, Alf. Studying emotional expression in music performance. *Bulletin of the Council for Research in Music Education*, pages 47–53, 1999. (p. 46)

Pierre Galin. *Exposition d'une nouvelle méthode pour l'enseignement de la musique*. Baron, 1835. (p. 26)

Leonard Garrison. Telemann's Twelve Fantasias: A Survey of Editions and Recordings. *Flutist Quarterly*, 45(5):24–31, 2020. (pp. 91 and 146)

André (1856-1926) Auteur du texte Gedalge. *Traité de la fugue. 1ere partie : De la fugue d' école. Partie 1*. Enoch et Co, Paris, France, 1901. (p. 87)

Mathieu Giraud, Richard Groult, Emmanuel Leguy, and Florence Levé. Computational Fugue Analysis. *Computer Music Journal*, 39(2):77–96, June 2015. (pp. 86 and 87)

M Good. MusicXML: An internet-friendly format for sheet music. In *Xml conference and expo*, pages 03–04, 2001. URL https://wpmedia.musicxml.com/wp-content/uploads/2012/11/xml2001-talk.pdf. (p. 27)

Masataka Goto. RWC Music Database: Popular, Classical, and Jazz Music Databases. In *Proceedings of the 3rd International Society for Music Information Retrieval (ISMIR)*, pages 287–288, Paris, France, 2002. (p. 76)

Masataka Goto. RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In *Proceedings of the 4th International Society for Music Information Retrieval (ISMIR)*, pages 229–230, Baltimore, USA, 2003. (p. 76)

G.P. Telemann. *12 Fantasias for Flute without Bass, TWV 40:2-13*. Number T 5823 W in Royal Conservatoire of Brussels, Library (B-Bc). Telemann, Hamburg, 1732. (pp. 84 and 85)

Maarten Grachten, Martin Gasser, Andreas Arzt, and Gerhard Widmer. Automatic Alignment of Music Performances with Structural Differences. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013. (p. 59)

P Grosche, M Müller, and S Ewert. Combination of Onset-Features with Applications to High-Resolution Music Synchronization. In *Proceedings of the International Conference on Acoustics (NAG/DAGA)*, pages 357–360, 2009. (pp. 38 and 58)

Janet K. Halfyard, editor. *Berio's Sequenzas: Essays on Performance, Composition and Analysis*. Routledge, 1 edition, July 2017. (p. 164)

Yoonchang Han and Kyogu Lee. Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players. In *Proccedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014. (p. 45)

Andrew Hankinson, Perry Roland, and Ichiro Fujinaga. The Music Encoding Initiative as a Document-Encoding Framework. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, USA, 2011. (p. 28)

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019. (p. 77)

George A. Heise and George A. Miller. An Experimental Study of Auditory Patterns. *The American Journal of Psychology*, 64(1):68–77, 1951. (pp. 51 and 52)

Henkjan Honing. Evidence for tempo-specific timing in music using a web-based experimental setup. *Journal of Experimental Psychology: Human Perception and Performance*, 32(3): 780–786, 2006. (p. 48)

Jacques Hotteterre. *Pièces pour la flûte traversière, et autres instruments, avec la basse continue par Mr Hotteterre le Romain... Livre 1er. Oeuvre second.* Christophe Ballard, ch. ballard edition, 1708. (pp. 12 and 178)

Jacques Hotteterre. L'Art de préluder sur la flûte traversière, sur la flûte à bec, sur le haubois et autres instrumens de dessus avec des préludes tous faits sur tous les tons dans différens mouvements et différens caractères., 1719. URL https://gallica.bnf.fr/ark:/12148/btv1b8538436m. (pp. 163 and 183)

Jacques Hotteterre. Principes de la flute traversière ou flute d'Allemagne, de la flute à bec ou flute douce, et du haut-bois, 1722. (p. 16)

George Houle. V- Articulation of Quantative Meter. In *Meter in music, 1600-1800: performance, perception, and notation*, Music–scholarship and performance. Indiana University Press, Bloomington, 1987. (p. 47)

Andrew Hugues and Edith Gerson-Kiwi. Solmization, 2001. URL https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000026154. (p. 25)

Lauren Hunt. *A Horn Transcription of the twelve fantasias for flute without bass, TWV 40:5-8, by Georg Philipp Telemann.* PhD thesis, University of Georgia, 2016. (pp. 84 and 90)

David Huron. Music Information Processing Using the Humdrum Toolkit: Concepts, Examples, and Lessons. *Computer Music Journal*, 26(2):11–26, 2002. (p. 71)

David Huron and Matthew Royal. What Is Melodic Accent? Converging Evidence from Musical Practice. *Music Perception: An Interdisciplinary Journal*, 13(4):489–516, 1996. (p. 52)

Vladimir Jankélévitch. *De la musique au silence. 2: Debussy et le mystère de l'instant.* Plon, Paris, 1976. (p. 156)

Cyril Joder, Slim Essid, and Gaël Richard. A Conditional Random Field Framework for Robust and Scalable Audio-to-Score Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2385–2397, November 2011. (p. 62)

Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, March 2005. (p. 61)

Anssi Klapuri. Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. In *Proccedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, pages 216–221, Victoria, Canada, 2006. (p. 34)

K. Koffka. *Principles of Gestalt psychology*. Principles of Gestalt psychology. Harcourt, Brace, Oxford, England, 1935. Pages: 720. (p. 49)

Verena Konz, Meinard Muller, and Andi Scharfstein. Extracting Expressive Tempo Curves from Music Recordings. In *Proceedings of the 35th International Conference on Acoustics (NAG/DAGA)*, pages 1–3, Rotterdam, 2009. (p. 45)

Verena Konz, Meinard Müller, and Rainer Kleinertz. A Cross-Version Chord Labelling Approach for Exploring Harmonic Structures—A Case Study on Beethoven's Appassionata. *Journal of New Music Research (JNMR)*, 42(1):61–77, March 2013. (p. 73)

Hendrik Vincent Koops, W. Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research (JNMR)*, 48(3):232–252, May 2019. (p. 73)

Carol L. Krumhansl. *Cognitive foundations of musical pitch*. Number 17 in Oxford psychology series. Oxford Univ. Press, New York, NY, 2001. (p. 88)

Frank Kurth and Meinard Muller. Efficient Index-Based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, February 2008. (p. 37)

Kyoung Joo Min. *Analyses of the twelve fantasies for solo flute(1732-1733) and the twelve fantasies for solo violin(1735) of G. Ph. Telemann*. PhD thesis, University of Washington, 1998. (p. 84)

Julia Larson. *Flute without accompaniment: Works from Debussy: "Syrinx" (1913) to Varese: "Density 21.5" (1936)*. PhD thesis, University of Maryland College Park, 1990. (p. 155)

Olivier Lartillot. Automated Extraction of Motivic Patterns and Application to the Analysis of Debussy's Syrinx. In Timour Klouche and Thomas Noll, editors, *Mathematics and Computation in Music*, volume 37, pages 230–239. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. (p. 98)

Alison Latham, editor. *The Oxford companion to music*. Oxford reference online premium. Oxford University Press, Oxford, online-only revised ed edition, 2011. (p. 45)

Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. Music Performance Analysis: A Survey. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, 2019. (p. 45)

Cynthia C S Liem and Alan Hanjalic. Expressive Timing from Cross-Performance and Audio-based Alignment Patterns: An Extended Case Study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 519–524, Miami, Florida, USA, October 2011. University of Miami. (pp. 59 and 73)

Meredith Little and Natalie Jenne. *Dance and the Music of J. S. Bach, Expanded Edition*. Music: Scholarship and Performance. Indiana University Press, Bloomington, IN, 2001. (p. 146)

Beth Logan and Cambridge Ma. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the 1st International Society for Music Information Retrieval Conference (ISMIR)*, Plymouth, Masachussets, USA, 2000. (p. 39)

Etienne Loulié. *Elements ou principes de musique, mis dans un nouvel ordre*. Christophe Ballard, Paris, France, 1696. (pp. 48, 137, and 182)

Ben Maman and Amit H Bermano. Unaligned Supervision for Automatic Music Transcription in-the-Wild. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 14918–14934, 2022. (p. 40)

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. (p. 144)

Marion Beyer, Rachel Brown. Georg Philipp Telemann Douze Fantaisies pour flûte solo TWV 40:2-13, 2013. URL https://www.henle.de/fr/Douze-Fantaisies-pour-flute-solo-TWV-40-2-13/HN-556. (p. 84)

Matthias Mauch and Simon Dixon. Approximate Note Transcription for the Improved Identification of Difficult Chords. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010. Pages: 140. (p. 34)

Matthias Mauch and Simon Dixon. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, May 2014. (pp. 33 and 123)

Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *SciPy, 2015*, pages 18–24, Austin, Texas, 2015. (p. 97)

David Meredith, Kjell Lemström, and Geraint Wiggins. Algorithms for Discovering Repeated Patterns in Multidimensional Representations of Polyphonic Music. *Journal of New Music Research (JNMR)*, 31, April 2003. (p. 98)

George A. Miller and George A. Heise. The Trill Threshold. *The Journal of the Acoustical Society of America (JASA)*, 22(5):637–638, September 1950. (p. 51)

Dirk Moelants. The Performance of Notes Inégales: The Influence of Tempo, Musical Structure, and Individual Performance Style on Expressive Timing. *Music Perception: An Interdisciplinary Journal*, 28(5):449–460, 2011. (p. 48)

Marcel Mongeau and David Sankoff. Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175, June 1990. (p. 65)

J. Montagu, H. Brown, J. Frank, and A. Powell. Flute. In Laurence Libin, editor, *The Grove dictionary of musical instruments*, volume 2, pages 323–339. Oxford University Press, New York, second edition edition, 2014. (p. 15)

Meinard Muller and Sebastian Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, March 2010. (p. 39)

Meinard Muller, Frank Kurth, and Tido Roder. Towards an Efficient Algorithm for Automatic Score-to-Audio Synchronization. In *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, Spain, 2004. (p. 57)

Meinard Muller, Henning Mattes, and Frank Kurth. An efficient multiscale approach to audio synchronization. In *Proccedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, pages 192–197, 2006. (p. 67)

Meinard Müller. *Fundamentals of music processing: using Python and Jupyter notebooks*. Springer, Cham, Switzerland, 2nd edition, 2021. (pp. 13, 33, 37, 38, 41, 57, and 60)

Meinard Müller and Ching-Yu Chiu. A Basic Tutorial on Novelty and Activation Functions for Music Signal Processing. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 7(1), September 2024. (p. 40)

Meinard Müller and Frank Kurth. Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations. *EURASIP Journal on Advances in Signal Processing*, 2007 (1):089686, December 2006. (p. 33)

Meinard Müller and Frank Zalkow. libfmp: A Python Package for Fundamentals of Music Processing. *Journal of Open Source Software (JOSS)*, 6(63):3326, July 2021. (pp. 41 and 179)

Meinard Müller, Frank Kurth, and Michael Clausen. Audio Matching via Chroma-Based Statistical Features. In *Proceedings of the 6th International Society for Music Information Retrieval Conference, (ISMIR)*, pages 288–295, London, United Kingdom, January 2005. (pp. 37 and 58)

Meinard Müller, Yigitcan Özer, Michael Krause, Thomas Prätzlich, and Jonathan Driedger. Sync Toolbox: A Python Package for Efficient, Robust, and Accurate Music Synchronization. *Journal of Open Source Software (JOSS)*, 6(64):3434, August 2021. (p. 72)

Monica S. Nagawade and Varsha R. Ratnaparkhe. Musical instrument identification using MFCC. In *2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2198–2202, May 2017. (p. 39)

N. J. Nalini and S. Palanivel. Music emotion recognition: The combined evidence of MFCC and residual phase. *Egyptian Informatics Journal*, 17(1):1–10, March 2016. (p. 39)

Jyoti Narang, Marius Miron, Ajay Srinivasamurthy, and Xavier Serra. Analysis of Musical Dynamics in Vocal Performances using Loudness Measures. In *International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 301–311, 2022. (p. 45)

Eugene Narmour. *The analysis and cognition of basic melodic structures: the implication-realization model*. University of Chicago Press, Chicago, 1990. (pp. 48 and 51)

Eugene Narmour. *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*. University of Chicago Press, November 1992. (p. 48)

Jean-Jacques Nattiez. *Fondements d'une sémiologie de la musique*. Number 1017 in 10/18. Esthétique. Union générale d'éditions, Paris, 1975. (p. 98)

Jean-Jacques Nattiez and Anna Barry. Varese's 'Density 21.5': A Study in Semiological Analysis. *Music Analysis*, 1(3):243–340, 1982. (pp. 94 and 159)

Mitchell Ohriner. Chapter 14 Expressive Timing. In Alexander Rehding and Steven Rings, editors, *The Oxford handbook of critical concepts in music theory*, pages 369–394. Oxford University Press, New York, NY, 2019. (pp. 99, 150, 157, 158, and 183)

Nicola Orio. Music Retrieval: A Tutorial and Review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, November 2006. (p. 13)

Nicola Orio and Diemo Schwarz. Alignment of Monophonic and Polyphonic Music to a Score. In *International Computer Music Conference (ICMC)*, pages 1–1, Havana, Cuba, September 2001. (p. 57)

Nicola Orio, Serge Lemouton, and Diemo Schwarz. Score Following: State of the Art and New Developments. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, Montreal, Canada, 2003. (p. 60)

Caroline Palmer. On the Assignment of Structure in Music Performance. *Music Perception: An Interdisciplinary Journal*, 14(1):23–56, 1996a. (p. 47)

Caroline Palmer. Anatomy of a Performance: Sources of Musical Expression. *Music Perception: An Interdisciplinary Journal*, 13(3):433–453, 1996b. (p. 47)

Steven Par, Martin McKinney, and André Redert. Musical Key Extraction from Audio Using Profile Training. In *Proccedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, 2006. (p. 88)

Marcus T. Pearce and Geraint A. Wiggins. Expectation in Melody: The Influence of Context and Learning. *Music Perception*, 23(5):377–405, July 2006. (p. 50)

Walter Piston. *Counterpoint*. W. W. Norton & Company, New York London, 1947. (pp. 50 and 83)

Amy Porter. Telemann: 12 Fantasias for Flute without Bass – A Study Guide DVD, 2008. URL https://www.youtube.com/watch?v=T_PW8GcIdQI&list=PLpc7xR4RabCanZbs_fQz_JDtBeS863ck6. (p. 84)

Ardal Powell. *The flute*. The Yale musical instrument series. Yale University Press, New Haven ; London, 2002. (pp. 15 and 17)

Thomas Pratzlich and Meinard Muller. Freischütz Digital: a case study for reference-based audio segmentation of operas. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 589–594, Curitiba, Brazil, 2013. (p. 73)

Thomas Pratzlich and Meinard Muller. Triple-based analysis of music alignments without the need of ground-truth annotations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270, Shanghai, China, March 2016. IEEE. (pp. 59 and 74)

Thomas Pratzlich, Jonathan Driedger, and Meinard Muller. Memory-restricted multiscale dynamic time warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 569–573, Shanghai, March 2016. IEEE. (p. 68)

Laurent Pugin, Rodolfo Zitellini, and Perry Roland. VEROVIO: A Library for Engraving MEI Music Notation into SVG. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 107–112, Taipei, Taiwan, 2014. (p. 28)

Quantz, Jean Joachim. Essai D'une Méthode Pour Apprendre À Jouer De La Flûte Traversière Avec Plusieurs Remarques Pour Servir Au Bon Gout Dans La Musique Le Tout Eclairci Par Des Exemples Et Par Xxiv. Tailles Douces, 1752. (pp. 17 and 110)

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., USA, 1993. (pp. 39, 61, and 67)

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. (p. 65)

Radio France and Laurent Vilarem. *Programme notes for LUCIANO BERIO Sequenze pour instrument seul*. Radio France, Paris, France, 2025. Concert programme, Maison de la Radio et de la Musique. (p. 164)

Betsy Rajesh and D. G. Bhalke. Automatic genre classification of Indian Tamil and western music using fractional MFCC. *International Journal of Speech Technology*, 19(3):551–563, September 2016. (p. 39)

Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270, Beijing China, August 2012. ACM. (p. 62)

Jean Philippe Rameau. *Traité de l'harmonie reduite à ses principes naturels;: divisé en quatre livres. Livre I. Du rapport des raisons & proportions harmoniques. Livre II. De la nature & de la proprieté des accords et de tout ce qui peut servir à rendre une musique parfaite. Livre III. Principes de composition. Livre IV. Principes d'accompagnement*. de l'imprimerie de Jean-Baptiste-Christophe Ballard ... A Paris, 1722. (p. 88)

Iris Yuping Ren, Hendrik Vincent Koops, Anja Volk, and Wouter Swierstra. In Search of the Consensus Among Musical Pattern Discovery Algorithms. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017. (p. 98)

Bruno H. Repp. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America (JASA)*, 92(5):2546–2568, 1992. (pp. 45 and 46)

Bruno H. Repp. A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *The Journal of the Acoustical Society of America*, 104(2 Pt 1):1085–1100, August 1998. (p. 46)

Pablo E. Riera, Martin Proscia, and Manuel C. Eguia. A Comparative Study of Saxophone Multiphonics: Musical, Psychophysical and Spectral Analysis. *Journal of New Music Research (JNMR)*, 43(2):202–213, April 2014. (pp. 45 and 164)

Joseph Rothstein. *MIDI: a comprehensive introduction*. Number 7 in The computer music and digital audio series. A-R Ed, Madison, Wis, 2. ed edition, 1995. (p. 28)

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978. (p. 61)

Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. In *KDD Workshop on Mining Temporal and Sequential Data*, 2004. (p. 67)

E. Glenn Schellenberg. Simplifying the Implication-Realization Model of Melodic Expectancy. *Music Perception: An Interdisciplinary Journal*, 14(3):295–318, 1997. (p. 50)

Jan Schlüter and Sebastian Böck. Improved musical onset detection with Convolutional Neural Networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983, May 2014. (p. 42)

Susan Schreibman, Raymond George Siemens, and John Unsworth, editors. *A companion to digital humanities*. Number 26 in Blackwell companions to literature and culture. Blackwell Pub, Malden, MA, 2004. (p. 13)

Arnold Schönberg. *Fundamentals of musical composition*. Faber and Faber, London, 1967. (p. 98)

Carl E. Seashore. The Psychology of Music. *Music Educators Journal*, 23(4):30–33, February 1937. (p. 45)

Joan Serra, Emilia Gomez, Perfecto Herrera, and Xavier Serra. Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, August 2008. (p. 59)

C.E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, January 1949. (p. 31)

S. S. Shapiro and M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4):591, December 1965. (p. 144)

Portela da Silva and Antonio Carlos. *A performance guide to three of Telemann's 12 fantasias for flute*. PhD thesis, University of Alabama Libraries, 2012. (pp. 83 and 84)

Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. Audio-to-Score Alignment Using Deep Automatic Music Transcription. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, October 2021. (p. 62)

Ferreol Soulez, Xavier Rodet, and Diemo Schwarz. Improving Polyphonic and Poly-Instrumental Music to Score Alignment. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, USA, 2003. (p. 57)

Yann Teytaut, Antoine Petit, Céline Chabot-Canet, and Axel Roebel. A musicological pipeline for singing voice style analysis with neural voice processing and alignment. In *Journées d'Informatique Musicale (JIM)*, Saint-Denis, France, May 2023. (p. 57)

Patrice Thibaud, Mathieu Giraud, and Yann Teytaut. Towards Detecting Interleaved Voices in Telemann Flute Fantasias. In *DMRN+ 19: Digital Music Research Network One-day Workshop}*, London, December 2024. (p. 22)

Patrice Thibaud, Mathieu Giraud, and Yann Teytaut. When Voices Interleave: Timing Deviations in Six Performances of Telemann's Fantasias for Solo Flute. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, Daejeon, South Korea, 2025. (pp. 21 and 22)

John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning Features of Music From Scratch. In *International Conference on Learning Representations*, 2017. (p. 77)

Verena Thomas, Christian Fremerey, Meinard Müller, and Michael Clausen. Linking Sheet Music and Audio - Challenges and New Approaches. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Dagstuhl Follow-Ups*, volume 3, pages 1–22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2012. (p. 57)

Simon Trezise, editor. *The Cambridge Companion to Debussy.* Cambridge Companions to Music. Cambridge University Press, Cambridge, 2003. (pp. 94 and 155)

Robert J. Turetsky and Daniel P. W. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, United States, 2003. (p. 57)

L.P.A.S van Noorden. *Temporal Coherence in the Perception of Tone Sequences.* Ph.D. thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1975. (p. 51)

Perrine Vantalon, Mathieu Giraud, Richard Groult, and Thierry Lecroq. Towards modeling alternating patterns through inter-notes relations. In *23rd International Society for Music Information Retrieval (ISMIR), Late-Breaking Demo Session*, page LV10, Bagalore, India, December 2022. (p. 98)

Edgard Varese and Chou Wen-chung. The Liberation of Sound. *Perspectives of New Music*, 5 (1):11, 1966. (p. 157)

Camille Vauthrin, Benoît Fabre, and Isabelle Cossette. How Does a Flute Player Adapt His Breathing and Playing to Musical Tasks? *Acta Acustica united with Acustica*, 101(2):224–237, March 2015. (pp. 167 and 184)

Yannik Venohr, Yiwei Ding, and Christof Weiss. Towards Robust Music Transcription by Measuring Cross-Version Consistency in Western Classical Music. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, Daejeon, South Korea, September 2025. (p. 117)

Barry Vercoe. The Synthetic Performer in The Context of Live Performance. In *Proceedings of the 1984 International Computer Music Conference (ICMC)*, volume 1984, Paris, France, 1984. (p. 59)

Christophe Vergez. *Analyse du fonctionnement des instruments de musique à vent.* thesis, Université de la Méditerranée - Aix-Marseille II, June 2010. (p. 113)

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967. (p. 62)

Changhong Wang, Emmanouil Benetos, and Elaine Chew. CBF-PeriDB: A Dataset of Chinese Bamboo Flute Playing Techniques with Periodic Modulations. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, November 2019. (p. 165)

Siying Wang, Sebastian Ewert, and Simon Dixon. Robust Joint Alignment of Multiple Versions of a Piece of Music. In *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014. (pp. 58 and 73)

Christof Weiß and Geoffroy Peeters. Comparing Deep Models and Evaluation Strategies for Multi-Pitch Estimation in Music Recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2814–2827, 2022. (p. 117)

Christof Weiß and Thomas Pratzlich. Analyzing Measure Annotation for Western Classical Music Recordings. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, 2016. (p. 73)

Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald G. Grohganz. Schubert Winterreise Dataset: A Multimodal Scenario for Music Analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(2):1–18, June 2021a. (pp. 59, 77, and 117)

Christof Weiß, Johannes Zeitler, Tim Zunner, Florian Schuberth, and Meinard Müller. Learning Pitch-Class Representations from Score-Audio Pairs of Classical Music. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 746–753, Online, November 2021b. (pp. 40 and 179)

Christof Weiß, Vlora Arifi-Müller, Michael Krause, Frank Zalkow, Stephanie Klauk, Rainer Kleinertz, and Meinard Müller. Wagner Ring Dataset: A Complex Opera Scenario for Music Processing and Computational Musicology. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 6(1):135–149, October 2023. (pp. 59 and 180)

Max Wertheimer. Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung*, 4 (1):301–350, January 1923. (p. 49)

Matthew James Wright. *The shape of an instant: Measuring and modeling Perceptual Attack Time with probability density functions (if a tree falls in the forest, when did 57 people hear it make a sound?)*. PhD thesis, ProQuest Information & Learning, US, 2008. (p. 97)

Rui Yang, Mathieu Giraud, and Florence Levé. Music with Numbers: Jianpu Number-Based Notation in Cultural Heritage and Digital Humanities. In *Music Encoding Conference (MEC 2025)*, London, United Kingdom, 2025. (p. 26)

Frank Zalkow, Christof Weiß, Thomas Prätzlich, Vlora Arifi-Müller, and Meinard Müller. A Multi-Version Approach for Transferring Measure Annotations Between Music Recordings. In *Conference on Semantic Audio*, Erlangen, Germany, 2017. (pp. 58, 74, 127, 129, and 180)

Gioseffo Zarlino. *Le istitutioni harmoniche Di M. Gioseffo Zarlino da Chioggia*. s.n., 1558. (p. 88)

Johannes Zeitler and Meinard Müller. A Unified Perspective on CTC and Soft-DTW Using Differentiable DTW. *IEEE Transactions on Audio, Speech and Language Processing*, 34: 936–951, 2026. (p. 62)

Johannes Zeitler, Simon Deniffel, Michael Krause, and Meinard Müller. Stabilizing Training with Soft Dynamic Time Warping: a Case Study for Pitch Class Estimation with Weakly Aligned Targets. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, Milan Italy, 2023. (p. 62)

Johannes Zeitler, Ben Maman, and Meinard Müller. Robust and Accurate Audio Synchronization using Raw Features from Transcription Models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, USA, 2024a. (p. 40)

Johannes Zeitler, Christof Weiß, Vlora Arifi-Müller, and Meinard Müller. BPSD: A Coherent Multi-Version Dataset for Analyzing the First Movements of Beethoven's Piano Sonatas | Transactions of the International Society for Music Information Retrieval. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, January 2024b. (pp. 77 and 117)

Zhiyao Duan, B Pardo, and Changshui Zhang. Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, November 2010. (p. 77)

Steven David Zohn. *Music for a mixed taste: style, genre and meaning in Telemann's instrumental works*. Oxford University press, Oxford, 2008. (pp. 17, 83, 84, and 85)

Y. Özer, M. Krause, and M. Müller. Using the sync toolbox for an experiment on high-resolution music alignment. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021. (p. 41)