Capturing Musical Prosody Through Interactive Audio/Visual Annotations

Daniel Bedoya Ramos

Capturing Musical Prosody Through Interactive Audio/Visual Annotations

Dissertation to obtain a doctoral degree in

Information and Communication Science and Technology

submitted to

Sorbonne Université

École Doctorale Informatique Télécommunications et Électronique (ED130)
Sciences et technologies de la musique et du son (STMS–UMR9912)
Institut de Recherche et Coordination Acoustique/Musique (IRCAM)
Équipe Représentations Musicales (RepMus)

Author:

Daniel Bedoya Ramos

Defended publicly on October 18, 2023 in Paris, France
before a committee composed of:

Director:		
Carlos	Agón	Sorbonne Université, France
Supervisor:		
Elaine	Chew	King's College London, UK
Reviewers:		
Roberto	Bresin	KTH Royal Institute of Technology, Sweden
Pierre	Couprie	Université Paris-Saclay, France
Examiners:		
Jean-Julien	Aucouturier	FEMTO-ST, France (*President of the jury)
Louis	Bigo	Université de Lille, France
Muki	Haklay	University College London (UCL), UK
Anja	Volk	Utrecht University, Netherlands

In loving memory of

Anita Bustos

August 22, 1921 – May 24, 2023

Abstract

The proliferation of citizen science projects has advanced research and knowledge across disciplines in recent years. Citizen scientists contribute to research through volunteer thinking, often by engaging in cognitive tasks using mobile devices, web interfaces, or personal computers, with the added benefit of fostering learning, innovation, and inclusiveness. In music, crowdsourcing has been applied to gather various structural annotations. However, citizen science remains underutilized in musical expressiveness studies.

To bridge this gap, we introduce a novel annotation protocol to capture musical prosody, which refers to the acoustic variations performers introduce to make music expressive. Our top-down, human-centered method prioritizes the listener's role in producing annotations of prosodic functions in music. This protocol provides a citizen science framework and experimental approach to carrying out systematic and scalable studies on the functions of musical prosody. We focus on the segmentation and prominence functions, which convey structure and affect.

We implement this annotation protocol in CosmoNote, a web-based, interactive, and customizable software conceived to facilitate the annotation of expressive music structures. CosmoNote gives users access to visualization layers, including the audio waveform, the recorded notes, extracted audio attributes (loudness and tempo), and score features (harmonic tension and other markings). The annotation types comprise boundaries of varying strengths, regions, comments, and note groups.

We conducted two studies aimed at improving the protocol and the platform.

The first study examines the impact of co-occurring auditory and visual stimuli on segmentation boundaries. We compare differences in boundary distributions derived from cross-modal (auditory and visual) vs. unimodal (auditory or visual) information. Distances between unimodal-visual and cross-modal distributions are smaller than between unimodal-auditory and cross-modal distributions. On the one hand, we show that adding visuals accentuates crucial information and provides cognitive scaffolding for accurately marking boundaries at the starts and ends of prosodic cues. However, they sometimes divert the annotator's attention away from specific structures. On the other hand, removing the audio impedes the annotation task by hiding subtle, relied-upon cues. Although visual cues may sometimes overemphasize or mislead, they are essential in guiding boundary annotations of recorded performances, often improving the aggregate results.

The second study uses all CosmoNote's annotation types and analyzes how annotators, receiving either minimal or detailed protocol instructions, approach annotating musical prosody in a free-form exercise. We compare the quality of annotations between participants who are musically trained and those who are not. The citizen science component is evaluated in an ecological setting where participants are fully autonomous in a task where time, attention, and patience are valued. We present three methods based on common

annotation labels, categories, and properties to analyze and aggregate the data. Results show convergence in annotation types and descriptions used to mark recurring musical elements across experimental conditions and musical abilities. We propose strategies for improving the protocol, data aggregation, and analysis in large-scale applications.

This thesis contributes to representing and understanding performed musical structures by introducing an annotation protocol and platform, tailored experiments, and aggregation/analysis methods. The research shows the importance of balancing the collection of easier-to-analyze datasets and having richer content that captures complex musical thinking. Our protocol can be generalized to studies on performance decisions to improve the comprehension of expressive choices in musical performances.

Key words : musical prosody, web-based annotation, music performance, music structure analysis, music expressiveness, citizen science

Résumé

Des projets de science participative (SP) ont stimulé la recherche dans plusieurs disciplines au cours des dernières années. Des citoyens scientifiques contribuent à cette recherche en effectuant des tâches cognitives, favorisant l'apprentissage, l'innovation et l'inclusion. Bien que le crowdsourcing ait servi à recueillir des annotations structurelles en musique, la SP reste sous-utilisée pour étudier l'expressivité musicale.

On introduit un nouveau protocole d'annotation pour capturer la prosodie musicale, associée aux variations acoustiques introduites par les interprètes pour rendre la musique expressive. Notre méthode descendante, centrée sur l'humain, donne la priorité à l'auditeur dans la production d'annotations des fonctions prosodiques de la musique. On se concentre sur la segmentation et la proéminence, qui véhiculent la structure et l'affect. Ce protocole fournit un cadre de SP et une approche expérimentale pour réaliser des études systématiques et extensibles.

On met en œuvre ce protocole d'annotation dans CosmoNote, un logiciel web personnalisable, conçu pour faciliter l'annotation de structures musicales expressives. CosmoNote permet aux utilisateurs d'interagir avec des couches visuelles, y compris la forme d'onde, les notes enregistrées, les attributs audio extraits et les caractéristiques de la partition. On peut placer des frontières de niveaux différents, des régions, des commentaires et des groupes de notes.

On a mené deux études visant à améliorer le protocole et la plateforme. La première, examine l'impact des stimuli auditifs et visuels simultanés sur les frontières de segmentation. On compare les différences dans les distributions de frontières dérivées d'informations intermodales (auditives et visuelles) et unimodales (auditives ou visuelles). Les distances entre les distributions unimodales-visuelles et intermodales sont plus faibles qu'entre les distributions unimodales-auditives et intermodales. On montre que l'ajout de visuels accentue les informations clés et fournit un échafaudage cognitif aidant à marquer clairement les frontières prosodiques, bien qu'ils puissent détourner l'attention de structures spécifiques. À l'inverse, sans audio, la tâche d'annotation devient difficile, masquant des indices subtils. Malgré leur exagération ou inexactitude, les repères visuels sont essentiels pour guider les annotations de frontières en interprétation, ce qui améliore les résultats globaux.

La deuxième étude utilise tous les types d'annotations de CosmoNote et analyse comment les participants annotent la prosodie musicale, avec des instructions minimales ou détaillées, dans un cadre d'annotations libres. On compare la qualité des annotations entre musiciens et non-musiciens. On évalue la composante de SP dans un cadre écologique où les participants sont totalement autonomes dans une tâche où le temps, l'attention et la patience sont valorisés. On présente trois méthodes basées sur des étiquettes d'annotation, des catégories et des propriétés communes pour analyser et agréger les données. Les résultats montrent une convergence dans les types d'annotations et les descriptions utilisées pour marquer les éléments musicaux récurrents, pour toute condition expérimentale et aptitude musicale. On propose des stratégies pour améliorer le protocole, l'agrégation des données et l'analyse dans des applications à grande échelle.

Cette thèse enrichit la représentation et la compréhension des structures en musique interpré-

tée en introduisant un protocole et une plateforme d'annotation, des expériences adaptables et des méthodes d'agrégation et d'analyse. On montre l'importance du compromis entre l'obtention de données plus simples à analyser et celle d'un contenu plus riche, capturant une pensée musicale complexe. Notre protocole peut être généralisé aux études sur les décisions d'interprétation afin d'améliorer la compréhension des choix expressifs dans l'interprétation musicale.

Mots clés : prosodie musicale, annotation via web, interprétation musicale, analyse de structures musicales, expressivité musicale, sciences participatives

Acknowledgments

The following words are my attempt to acknowledge most of the people I owe my gratitude to.

First, I would like to thank Elaine Chew, my supervisor, for the opportunity to engage in this adventure and for her inspiring ideas and keen comments. Elaine's scientific and musical experience was essential for shaping the core of this manuscript. I most fondly recall sharing a sometimes crowded office with other teammates and her at the heart of Paris for almost three years. I must also thank Carlos Agón for generously stepping in as my thesis Director at the beginning of my fourth year, thus helping me to finish my PhD program in the best conditions.

My most profound appreciation goes to the reviewers and examiners in my dissertation committee for kindly agreeing to lend their vital expertise to this process. I am also grateful to my *Comité de Suivi de Thèse* for their constructive criticism and welcoming guidance. Similarly, I acknowledge all the people who kindly set aside their time to offer their precious feedback on this manuscript.

Special thanks to the members of the COSMOS team in Paris: Emily Graber, Emma Frid, Paul Lascabettes, Lawrence Fyfe, Corentin Guichaoua, Charles Picasso, and Gonzalo Romero. I thank Lawrence for his unwavering calmness, patience, responsiveness, and hard work, which has played a substantial role in this thesis. Thank you, Emily, Paul, and Gonzalo, for all our delightful, imaginative conversations and banter, musical and otherwise. It was a great pleasure working with Emma, Corentin, and Charles; I value our short time of shared meetings, meals, challenges, leisure, and laughter.

During my years at IRCAM and the STMS lab, I have encountered some gentle souls to whom I extend my gratitude. Firstly, I would like to thank Nicolas Misdariis, Gérard Assayag, Isabelle Bloch, Jean-Louis Giavitto, Jean-Julien Aucouturier, Patrick Susini, and Isabelle Viaud-Delmon for their valuable advice and kind words. Secondly, I wish to thank Vasso, Pablo, Victor, and Claire; all of you have a special place in my memories. I also want to thank the people of the *ludi-midi* sessions that helped fuel many afternoons with joy in difficult times and all the PhD students, researchers, and engineers with whom I shared unique experiences. I cannot leave IRCAM without mentioning the wonderful people making the *Ressources Humaines*, *Production*, *Médiathèque*, *Informatique*, and *Régie Bâtiment* departments.

To the dream team of the INSEAD-Sorbonne Université Behavioural Lab for their impeccable and amicable disposition: Huong, Germain, Jean-Yves, and Sébastien, whose professionalism and assistance in running demanding experimental setups cannot be overestimated.

For the last year of my PhD program, I accepted a research and teaching assistant position (ATER) with the LMSSC research laboratory at CNAM. I want to thank all the friendly people I had the opportunity to meet there, in particular, Éric, Jean-Baptiste, Alexandre, Sarah, Philippe, and Christophe. I greatly appreciated the help from Sarah, who thoughtfully answered my many questions about musical concepts and descriptions in French and their translation into English.

I would be remiss if I did not mention Rosalie, Christel, Bérengère, Hélène, Jean-Louis, and Laurent. You have all helped me get to this point in one way or another.

My parents, my sister, and the rest of my family and friends in Ecuador, particularly Fercho, Jaime, and David, deserve special credit for all the encouragement they have given me from afar. Finally, to my wife, unconditional partner, and best friend, Bernie, I am infinitely grateful for your unparalleled support. For all the highs and lows we have shared throughout these momentous years. Thank you for being my bedrock.

List of publications

Journal articles

1. Fyfe, L., **Bedoya, D.**, & Chew, E. (2022). Annotation and Analysis of Recorded Piano Performances on the Web. *Journal of the Audio Engineering Society*, 70(11), 962-978. https://doi.org/10.17743/jaes.2022.0057

2. **Bedoya, D.**, Fyfe, L., & Chew, E. (2022). A Perceiver-Centered Approach for Representing and Annotating Prosodic Functions in Performed Music. *Frontiers in psychology, 13*. https://doi.org/10.3389/fpsyg.2022.886570

Other publications and presentations

3. Fyfe, L., **Bedoya, D.**, & Chew, E. (2022). A Nutshell Guide to Annotating Recorded Piano Performances on the Web with CosmoNote. *7th International Web Audio Conference*, Cannes, France. https://hal.archives-ouvertes.fr/hal-03715778/document

4. **Bedoya, D.**, Fyfe, L., & Chew, E. (2022). Creating Experiments with Cosmonote: Advancing Web-Based Annotations for Performed Music. *Proceedings of the 19th Sound and Music Computing Conference*, Saint-Étienne, France. https://doi.org/10.5281/zenodo.6576284

5. **Bedoya, D.**, Fyfe, L., & Chew, E. (2022, May) CosmoNote-enabled active listening and transcribing of expressive music structures, 2nd edition [Conference workshop]. *Seventh International Conference on Technologies for Music Notation and Representation (TENOR) 2022*, Marseille, France.

6. **Bedoya, D.** Fyfe, L., & Chew, E. (2021, October). Towards a Set of Conventions for Representing and Annotating Musical Prosody [Conference presentation]. *International Symposium on Performance Science (ISPS) 2021*, Montreal, Canada. https://hal.science/hal-03419378

7. Fyfe, L., **Bedoya, D.**, Guichaoua, C., & Chew, E. (2021) CosmoNote: A Web-based Citizen Science Tool for Annotating Music Performances. *Proceedings of the International Web Audio Conference*, Barcelona, Spain. https://www.webaudioconf.com/posts/2021_25/

8. **Bedoya, D.**, Fyfe, L., Guichaoua, C., & Chew, E. (2021, May) CosmoNote-enabled active listening and transcribing of expressive music structures [Conference workshop]. *CitSciVirtual 2021*.

9. **Bedoya, D.**, Fyfe, L., Guichaoua, C., & Chew, E. (2021, May) CosmoNote: Visualizing expressive music information layers and public engagement in annotating performed music structures [Poster session]. *CitSciVirtual 2021*. https://hal.science/hal-03454713/

Personal contribution

Three publications from the list above have been integrated into the text of this manuscript, more specifically into Chapter 3 (articles 1 and 4) and Chapter 4 (article 2). A paragraph is written at the beginning of these chapters as a reminder. While several portions of the articles were added verbatim, most parts have been revised or re-written, adapted, extended, and contextualized to this thesis.

For all publications, I have recorded, edited, and annotated audio and video examples. I have created video tutorials, contributed to programming the extraction of music descriptors, and coded the web pages for CosmoNote's user agreement, audio calibration, and questionnaires, all of which will be described when relevant. Additionally, in articles I co-author, I contributed to the experimental design, data collection and analysis, and the writing and revision of the text, while the lead author carried out aspects related to web development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Music is a deeply personal experience. Most of the music we experience is either a live performance or a recording from one. These performances are built on studied, decomposed, modeled, and reconstructed structures, yet they are not fully understood. Music structure analysis, informed by fields such as music theory, music information retrieval (MIR), and musicology, can be used to study structures in music performance. Instead of focusing primarily on symbolic music representations, music performance research increasingly draws from computational music structure analysis techniques, particularly data-driven approaches and data collection methods from the citizen science paradigm, such as crowdsourcing. However, these techniques have yet to be fully explored for understanding music expressiveness. Consequently, advancing knowledge in music performance research requires applying these techniques to music performance through an extensive, adequately labeled annotation database for understanding the structures that make music expressive. Research able to characterize what in music moves people may be applicable in MIR tasks such as music recommendation, music categorization, or performer recognition, as well as personalization to the needs of individuals in music therapy.

1.1 Context

The first three years of this research were funded by the Computational Shaping and Modeling of Musical Structures (COSMOS[1]) European Research Council Advanced Grant (AdG) project 788960[2], while it was hosted by the French National Centre for Scientific Research (CNRS[3]), as part of the Science and Technology of Music and Sound (STMS[4]) joint research unit (UMR9912). The fourth year was founded by a temporary teaching and research associate (ATER) contract with the National Conservatory of Arts and Crafts (CNAM[5]), as part of the Structural Mechanics and Coupled Systems Laboratory (LMSSC[6]). The research was carried out within the Musical Representations team (RepMus) at the Institute for Research and Coordination in Acoustics/Music (IRCAM[7]), also part of the STMS Lab in Paris, France.

The experimental protocol in Chapter 4 received approval from the Ethics Research Com-

[1] https://cosmos.isd.kcl.ac.uk/
[2] https://cordis.europa.eu/project/id/788960
[3] https://www.cnrs.fr/en/
[4] https://www.stms-lab.fr/
[5] https://www.cnam.fr/
[6] https://lmssc.cnam.fr/en
[7] https://www.ircam.fr/

mittee (*Comité d'Éthique de la Recherche* CER[8]) at Sorbonne Université. (ID: CER-2020-87). The studies in Chapter 5 and Chapter 6 were conducted at the INSEAD–Sorbonne Université Behavioural Lab[9], where they were approved by the INSEAD Institutional Review Board (ID: 202063 and 2021-01), and funded by the French Excellence Initiative (Idex) at Sorbonne Université, in the context of the French Investment for the Future Program (PIA).

1.2 Aim of the Thesis

This research aims to capture music expressiveness in performed music, i.e., representing and marking the structures that make music expressive through digital **annotations**. To engage in a large-scale collection of annotations and explorations of performed music, we applied the principles of citizen science to create a workflow powered by CosmoNote, a web-based, highly customizable annotation platform. In conjunction with the CosmoNote platform, we introduce a novel, scalable annotation protocol, starting with controlled experiments and extending to the general public. We anchor our work on the study of expressiveness in performed music and, more specifically, in the concept of musical prosody, characterized by the acoustic variations that make music expressive, and its functions of segmentation and prominence (see Chapter 2 and Chapter 4).

The annotation platform rests on four main pillars: its representations and visualizations (how annotators first encounter the music), its annotation types (the core communication channel connecting annotators and researchers), its intuitiveness (taking inspiration from familiar controls and interface), and its interactivity (e.g., zooming, selecting, editing, filtering).

Our annotation protocol is centered on the listener's perception of prosodic functions in music performance. It is designed to gather segmentation annotations (how the music is divided into meaningful units) and prominence (characterizing how emphasis is signaled in music). We provide coherent instructions, musical examples, and basic annotation strategies to achieve this task.

1.3 Scope

This section defines the scope of this thesis. We describe frequently used terms, our approach for choosing performed music for our experiments and its characteristics, and the conditions in which annotations were prepared, collected, and analyzed.

Music performance can refer to a musical event where one or multiple performers and listeners are gathered to play or listen to someone else play music. However, we do not study performance in this sense or intend to recreate the live music performance experience in our methodology and experiments. We investigate the experience of listening to and subsequently annotating a recording of music played by a human. Consequently, music performance refers to recorded music performances in this document unless stated otherwise.

The work described in this thesis studies music with a written score. We do not study first-sight reading (i.e., the performer is already familiar with the music) nor improvisation (i.e., there are no intentional modifications to what is written in the score). Music performance data is obtained from a single performer playing solo piano. This instrument was chosen because its pitch and dynamic range, polyphonic nature, instrumental technique, and vast musical repertoire

[8] https://cer.sorbonne-universite.fr/le-cer
[9] https://www.insead.edu/insead-sorbonne-university-behavioural-lab

coalesce into a wide range of expressive variations that researchers can record, model, and reproduce through digitally controlled mechanical pianos.

Our annotation protocol (see Chapter 4) is adjustable. It can be applied to analyze performance by comparing numerous distinct performances played by the same person or multiple renditions of the same piece played by different performers. This procedure may be used to look for variations in music expressiveness. For instance, confronting information derived from music descriptors or annotations can be used to infer a performer's playing style. Furthermore, our methods are extensible to all music genres and periods. However, all the music in this document is part of the Western academic canon, from the Baroque to the contemporary periods.

We present annotation experiments where listeners mark structures in the music without necessarily having previous knowledge of its score. Our annotation protocol considers engaging listeners with music from databases from diverse sources (see Chapter 3). However, music data for the performances presented in the annotation studies (see Chapter 5 and Chapter 6) were all recorded on a Bösendorfer 280VC ENSPIRE PRO Disklavier in a recording studio at IRCAM. The music in these experiments was hand-picked from multiple takes of entire pieces played by Elaine Chew. We did not splice whole recorded performances from different takes, a common practice in music production that alters notes, rhythms, dynamics, or articulation deviating from a performance target to satisfy specific aesthetic or commercial music production goals. In contrast, we decided to leave the selected takes unaltered and incorporate them into our experiments as they were played.

To refine the annotation protocol for a large-scale application, we concentrate on the annotation process of individual pieces and their results rather than on the specifics of the music being annotated. We take subjectivity, ambiguities, and nuances in annotations as they are without aspiring to resolve these discrepancies. Additionally, our musical analyses focus on what is necessary to understand the annotations' context and illustrate the participants' possible reasoning when placing them. The analysis provided by the annotations may be partial. We do not attempt to carry out a comprehensive analysis of any particular performance.

1.4 Manuscript Overview

The rest of this document is structured as follows: Chapter 2 is an inventory of the core concepts and leading research in music structure analysis, music annotation, music performance, and citizen science. We communicate the perspective through which our research sees research in music performance. Chapter 3 is a deep dive into the conception and design of CosmoNote, the annotation platform used throughout the manuscript. We describe the various visual representations available to annotators (e.g., waveform, piano roll, tempo) and the four annotation types (boundaries, regions, comments, and note groups) built for capturing musical prosody. Chapter 4 details the protocol we created for annotating musical prosody, particularly segmentation and prominence in performed solo piano music, using the CosmoNote platform. The end-to-end process of the data collection is explained, from providing prosodic examples, structuring, and formatting the annotation data for analysis to applying techniques for preventing precision errors. We describe how this method can obtain reliable and coherent annotations that can be applied to theoretical and data-driven models of musical expressiveness and discuss the implications of this methodology in fostering community engagement. We later present two studies with 172 participants in total. Chapter 5 introduces a study investigating how the visual layers integrated into CosmoNote change the prosodic annotations. We compare boundary annotations produced by accessing a combination of auditory and visual stimuli to assess their importance in the annotation process. Chapter 6 presents a study using the full capabilities of

CosmoNote in a free-form setting and judging changes in annotations according to opposing levels of detail in experimental instructions. We analyze and aggregate different annotation types, closely examining their labels. Finally, Chapter 7 summarizes the overall results from the whole manuscript, delineates the future direction for this scientific research, draws global conclusions about the project, and presents closing remarks.

Chapter 2

Literature Review

The research presented in this thesis focuses on the annotation of expressiveness in performed music through musical prosody structures, with a protocol strongly linked to the principles of citizen science. Our work takes inspiration from different techniques and is based on extensive research across multiple disciplines. A complete, chronological review of all the contributions in each field is outside the scope of the manuscript. This chapter's primary focus is to present, contextualize, and link germane results and trends found in the scientific literature. At the same time, later chapters will specify and expand upon definitions and findings where they are relevant. In this chapter, we start by describing research on the annotation of music structures. Then, we explain how such annotations are conceptualized in the context of expressiveness in performance. Finally, we discuss historical and recent applications of citizen science to music. The relationship between composed and performed structures in music is explored throughout the chapter.

2.1 Musical Analysis

The analysis of music should aim to understand its meaning and, independently of the methods being used, provide pertinent conclusions about all the elements in a piece. Many approaches should be used in conjunction to serve the analysis (White, 1994). However, as we will see in this chapter, this is more difficult for computational analysis, where researchers must work with constraints of computer inputs and outputs and thus be limited to applying a single analysis framework (Marsden, 2016).

Traditional score-based methods of music analysis are divided into (1) the structure of music structure units (e.g., melody, rhythm, harmony) and (2) the study of the overall *Form* of a piece (Lorenzo de Reizábal & Lorenzo de Reizábal, 2009; Abromont & de Montalembert, 2010). Other known methods outside the scope of this work include Schenkerian analysis, twelve-tone analysis, set theory, and semiotic analysis (White, 1994; Cook, 1994).

2.1.1 Music Structure Units

This first approach in traditional analysis highlights the composed structures found in individual musical components. Once smaller-scale structures are identified, the analysis can continue to establish a relationship between them and move on to a larger scale. We present a simplified recount of elementary music components mentioned throughout the manuscript: melody, rhythm and meter, harmony, and texture.

- Melody is an ordered succession of notes built from rhythms and pitches. It is generally recognizable and is created with expressive intent. Involving an evolution in time, it is considered part of a horizontal musical dimension.

- Harmony studies the structure of superimposed sounds. Harmony is concerned with the configuration of notes in a frequency space and therefore is associated with a vertical dimension in music. However, part of harmony studies the concatenation and motion of chords (harmonic progression), which also involves the horizontal dimension of music.

- Rhythm results from the organization of durations, timbres, and successive accents having a pulse across time. The existence of distinctive accents and a pulse is crucial for the existence of rhythm. Meter is a surrogate measure of how pulses are divided and counted, typically using integer beats as in the bars/measures of a score.

- Texture refers to how musical elements are interwoven. Texture can be melodic (studying how melodies in different voices relate to one another), rhythmic, harmonic, or timbral.

These notions are used, linked, exemplified, and contextualized in specific cases throughout this work. We have adopted brief characterizations of these music structure units for simplicity while acknowledging a narrow focus and the existence of overlapping relationships between these and other musical concepts, such as counterpoint, temperament, and scales, among others. For a comprehensive examination of any of these terms' definitions, history, or classification, see Abromont, 2001; Lorenzo de Reizábal & Lorenzo de Reizábal, 2009; Kennedy & Kennedy, 2012.

2.1.2 Form in Western Academic Music

Form is determined by how structure units in music (Section 2.1.1) are organized, although different authors may disagree on the importance and number of its specific constituent elements. From another point of view, musical form is concerned with how music is presented, while musical content is concerned with what is presented. The concepts of form and content are complementary and may not exist independently (Lee, 2023).

Individual ideas or elements that constitute the form of a piece can be labeled using different conventions. Because of its recurrence for describing structures in annotations, as shown in the following chapters, the use of capital letters (e.g., A, B, C) for independent musical ideas is highlighted. With this convention, if an element is reminiscent of one that came before it but has been modified, the new element is usually denoted with a number or apostrophe (e.g., B′, B′′, B1). This organization of elements is used to construct rudimentary forms in Western Academic Music. For example, the form of a piece with two different ideas can be labeled AB, and if we include an altered version of the first idea at the end, the same piece can be labeled ABA′. Combinations of these characters will represent, at a glance, the large-scale organization of the musical ideas in a piece.

Since simple structures such as AB and ABBA and more complex structures such as ABACA and ABCBA were frequently found in tonal music, they were distinctly named to facilitate their analysis. Commonly studied forms in Western Academic Music include Binary, Three-part, Rondo, Minuet, Sonata, Variation, Song, and Fugue, among many others. We will not describe any of these forms in detail, except for cases when providing context for analyzing specific musical pieces becomes necessary (see Chapter 5 and Chapter 6). It is important to note that definitions of musical form help categorize and study the structure of musical pieces even though, in practice, a complete match between theoretical form and real music may sometimes be challenging to attain.

A classification of types of form in Western Academic Music was proposed by Lee to make the abundance of terms describing musical form easier to navigate. Additionally, we must distinguish between the terms form and genre, often used interchangeably. We will adapt Lee's conceptualization, where form evokes structure and organizational elements in music, while genre is a broader, richer notion involving context, history, medium, and form. In this sense, musical form can be divided into the following five types:

- Texture: Divides form according to how many voices or musical lines there are and how they interact. It mainly denotes homophonic (one voice and accompaniment) and polyphonic (multiple independent voices) forms. Monophonic (a single, unaccompanied voice) and heterophonic (simultaneous variations of a single voice) forms (that complete the texture space described by Huron, 2001) are neglected because they do not help discriminate between musical forms.
- Sections: Focuses on how music is divided into large blocks called sections (see Chapter 4). For example, this classification would include pieces with sections labeled ABA or those discussed above. This division is also explored in the concept of segmentation throughout this document.
- Size: Calls attention to the notion of nested structures, meaning that form can be studied at smaller and larger scales, and be named differently in each case.
- Indefiniteness: Distinguishes between defined and undefined forms. On the one hand, music with a defined form fits into existing categories, accommodating deviations from its definition more or less strictly. On the other hand, music with an undefined form does not fit any particular definition, thus requiring a unique definition of its form per piece.
- Instrumentation: Describes how music is played or sung, primarily dividing vocal and instrumental forms.

The research in this document is centered around instrumental music with defined forms. We will describe structures at small and large scales, using examples from different periods, from Baroque to contemporary, including homophonic and polyphonic forms.

2.2 Computational Music Structure Analysis

Computational music structure analysis is a field of research that studies how to break down music into its constituent elements and their structural functions, through computational methods that are based on traditional approaches or created specifically to execute a bespoke analysis. Research in this field, usually referring to computer-aided analysis, broadly overlaps between sub-disciplines such as mathematical music theory, systematic musicology, music information retrieval (MIR), and music cognition, among others. When referring to music structure analysis in this document, we will primarily refer to computational music structure analysis. Because automating music analysis is both far from trivial and an open question (Lartillot, 2021), we will illustrate how techniques in almost every domain need annotations from human participants, be it during data collection, identification, comparison, analysis, or validation.

2.2.1 Representations and Visualizations

Computational methods benefit from technological advances ranging from reliable, high-quality recording and playback devices to advanced digital signal processing techniques and versatile

MIR software packages in Python and Matlab. Despite the challenges of applying computational methods created in engineering to the study of music in the humanities, researchers now widely use them to visualize and analyze music structures, combining standard manual methods with modern digital tools and new music representations (Couprie, 2022). Music visualization techniques rely on their color, shape, and layout to convey visually identifiable information in service to the ear, indicative of large-scale structures and spatially/temporally logical (Lima, Santos, & Meiguins, 2021). However, each representation is created with a goal, selectively hiding and emphasizing features according to its design, thus possibly affecting how analysis is conducted. Some representations favor simplicity and conciseness, while others are brimming with information. Visualizations can also be multidimensional, interactive, or even animated in real-time with the music. For this reason, it is essential to be aware of what a particular analysis requires and orient the choice of representation accordingly.

Visualizations and representations are not only conceived for different purposes, but also created from different sources. One way to classify representations, visualizations, and techniques used for computer-aided musical analysis (inspired by Lalitte, 2011; Couprie, 2018), is to organize the origin of these representations into symbolic, sub-symbolic, signal-based, and mixed. We will describe these subdivisions below using seven concrete examples of two-dimensional visualizations used to represent the same 14-second musical excerpt of Chopin's Ballade No. 2[1], shown in Figure 2.1. However, we will not dive into the theoretical details defining individual representations.

Other examples of dedicated computer-aided music analysis software and projects that integrate music representations/visualizations for music structure analysis can be seen in the inventory[2] provided by the Musicology Research Institute (*Institut de recherche en Musicologie* – IReMus); see also Couprie, 2018; Lima et al., 2021; Couprie, 2022 for details on music representations not discussed in this chapter.

2.2.1.1 Symbolic

Symbolic techniques use information abstracted from musical properties. The prototypical example of this representation is the music score, also called sheet music. As shown in Figure 2.1a, scores represent notes, durations, dynamics, tempo, and technique, among other crucial aspects to perform a piece of music as a sequence of notes over time (Lima et al., 2021). Symbolic information can be encoded into markup languages with distinctive syntaxes such as musicXML[3], Common Music Notation (CMN[4]), Music Encoding Initiative (MEI[5]), and Humdrum **kern[6]. Another symbolic representation showing a sequence of notes over time is the widely used Musical Instrument Digital Interface (MIDI[7]) communication protocol. While still a symbolic representation, messages in MIDI are built for real-time applications, making it able to capture the nuances of performed music, as is the case of the example in Figure 2.1c. See Chapter 3 for more details on using MIDI to analyze music performance.

Symbolic representations can also illustrate harmonic progression. For instance, simplical complexes (based on the Tonnetz diagram) define spatial harmonic relationships among succes-

[1] Go to https://doi.org/10.6084/m9.figshare.23732649.v1 to hear the performance in Figure 2.1.

[2] https://www.iremus.cnrs.fr/fr/programme-de-recherche/analyse-musicale-assistee-par-ordinateur

[3] https://www.musicxml.com

[4] https://ccrma.stanford.edu/software/cmn/cmn/cmn.html

[5] https://music-encoding.org/

[6] https://www.humdrum.org/Humdrum/representations/kern.html

[7] https://midi.org/specifications

sive vertical slices of notes in a score, visualized along a grid as "infinite triangular tessellations" (see Bigo, Ghisi, Spicher, & Andreatta, 2015 for more details). Figure 2.1b shows the excerpt's harmonic trajectory[8] on the chord complex labeled $\mathcal{K}[3,4,5]$ (defined by its inclusion of major and minor triads), where played notes (circles) and chords (triangles) are connected with black lines and highlighted in dark yellow. In contrast, the last chord is highlighted in bright yellow.

2.2.1.2 Sub-Symbolic

Sub-symbolic techniques use descriptors derived from psychoacoustic models of human perception. They do not present musical information or physical dimensions (frequency, sound pressure, phase) per se but describe how humans perceive them as psychoacoustic concepts such as loudness, tempo, and brightness, among others. Figure 2.1f shows the loudness (Pampalk 2004) of the excerpt. This representation (see Chapter 3) is helpful to evaluate in one curve how loud a listener would perceive a given musical passage without having to relate amplitude values for individual notes or their frequency content.

2.2.1.3 Signal-Based

Signal-based techniques use audio signal processing to create representations later used for analysis. Figures 2.1e and 2.1d show the waveform and spectrogram of the excerpt, respectively. The waveform represents only the variation of amplitude over time. The spectrogram is a tridimensional representation (showing time, frequency, and amplitude) computed from the short-time Fourier Transform (STFT). Spectrograms are useful to visualize acoustic properties such as harmonics, which makes them helpful in speech analysis but impedes a clear distinction of the frequency of individual notes of instruments such as the piano, as seen in Figure 2.1d.

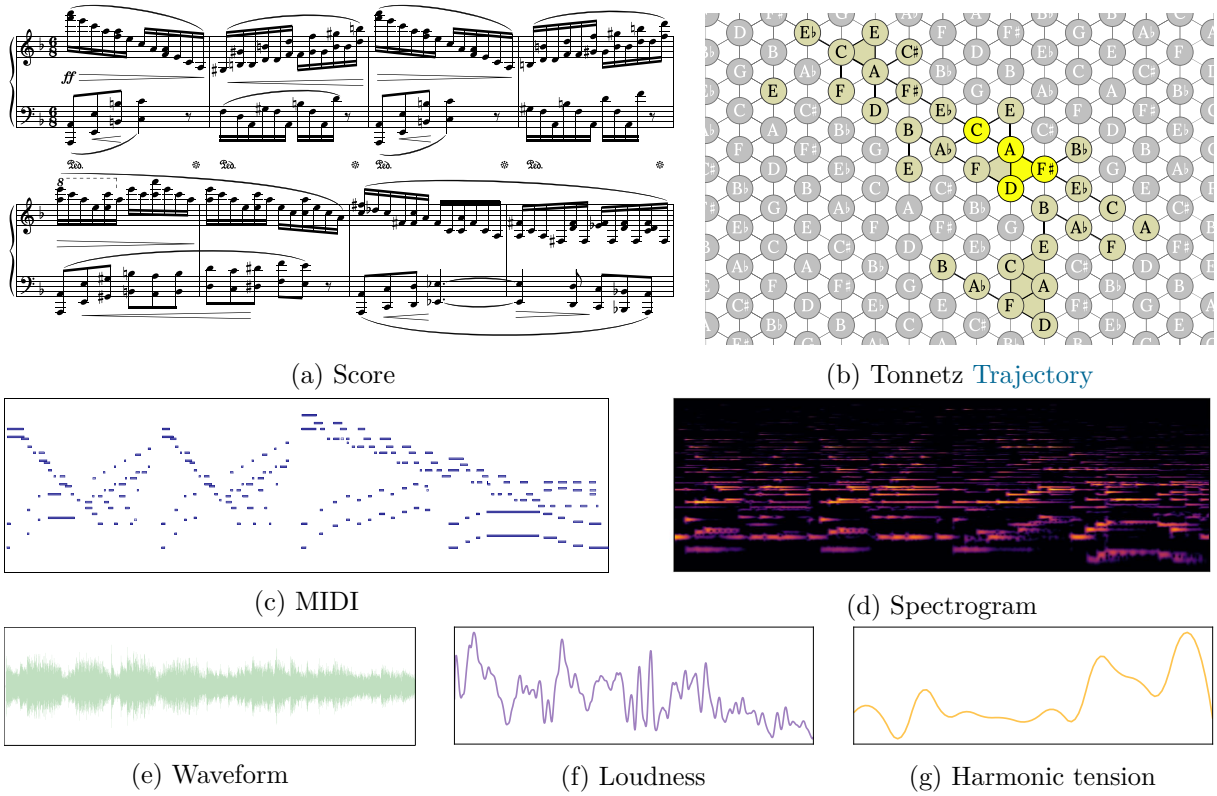
2.2.1.4 Mixed

A combination of symbolic, sub-symbolic, and signal-based techniques, which we call mixed techniques, is advantageous and common for computational music structure analysis applications. This approach combines the inputs and outputs of other techniques to derive new representations, using symbolic abstractions, signal processing, and psychoacoustic concepts to obtain a descriptor used in music structure analysis. As will be explored in further chapters, multiple representations may be presented at once to improve the analysis or to complement the information provided by one of them. For instance, an algorithm can be used on recorded MIDI (symbolic) to extract a representation of perceived musical tension aligned to the performance's tempo curve (sub-symbolic). Figure 2.1g displays one dimension of the harmonic tension model (Herremans & Chew, 2016) representing the distance from the global key, which in tonal music helps to evaluate dissonance at a glance.

2.2.2 Segmentation

The following section outlines the major influences, techniques, and trends in music structure segmentation, mainly of symbolic music (performance analysis is discussed in Section 2.3). For a comprehensive review of techniques used to analyze music structures, particularly structural segmentation, see Nieto et al., 2020; Lartillot, 2021.

[8] Recreated from the visualization option on the HexaChord software. To hear the trajectory in Figure 2.1b, go to `https://doi.org/10.6084/m9.figshare.23732646.v1`.



(a) Score (b) Tonnetz Trajectory

(c) MIDI (d) Spectrogram

(e) Waveform (f) Loudness (g) Harmonic tension

Figure 2.1: Different visualizations of the same excerpt from Chopin's Ballade No. 2, Op. 38 (*Presto con fuoco*). All axes are omitted for simplicity.

The publication in 1983 of the first edition of "Generative Theory of Tonal Music" (GTTM) marks a milestone in constructing an influential, listener-focused model for understanding music, later refined to include aspects of harmonic tension (Lerdahl & Jackendoff, 1996; Lerdahl, 2004). It proposed a rule-based decomposition of music structures based on: grouping structure (hierarchical segmentation into motives, phrases, and sections), metrical structure (hierarchy based on the alternation of strong and weak beats), time-span reduction (hierarchy concerning meter and grouping), and prolongational reduction (the added notion of tension/relaxation, continuity, and progression).

The problem of music segmentation relates to the concept and terminology (less used throughout this manuscript) of grouping structures in theoretical models. Many of the algorithms for analyzing music structures and automatic segmentation, both symbolically and in performance, have been inspired by the methodology of the GTTM. For example, the Melisma Music Analyzer program and its subsequent version (Temperley, 2004, 2009) are based on the GTTM model using preference rules to extract grouping, metrical, harmonic, and contrapuntal structures. Other approaches use statistical methods; this is the case of the Local Boundary Detection Model (LBDM), an algorithm to extract melodic patterns to determine probable segmentation points on music with clear segmentation boundaries (Cambouropoulos, 2006).

Segmentation music structure is mainly annotated using boundaries. These boundaries can be organized following a nested structure (i.e., hierarchical and dependent on one another) or not (i.e., flat and independent). The F-measure (Nieto, Farbood, Jehan, & Bello, 2014) is a common technique to compare boundary annotations in flat structures (see Section 5.2.5). Annotations with boundaries depending on nested hierarchical levels can be compared using the L-measure (McFee, Nieto, Farbood, & Bello, 2017). When segmentation is marked by humans, with no

difference in musical expertise, the relevance of their boundary annotations is correlated with both the number of boundaries close to the same time, and the weight attributed to individual boundaries, independent of their locations (Bruderer, McKinney, & Kohlrausch, 2009)

Besides the hierarchy, frequency, and salience of boundaries, it is necessary to mention the four segmentation principles most used in MIR for algorithmically marking segmentation boundaries (Nieto et al., 2020). These principles use self-similarity based on having one or more shared musical attributes. This technique compares every sample in a time series with itself and with all the other samples:

- Homogeneity: This principle is based on the assumption that within a section, musical attributes tend to be consistent locally, e.g., they share the same timbre.
- Novelty: Listeners are drawn to salient contrasts around sections of homogenous music and tend to identify boundary segments at salient peaks of the novelty curve.
- Repetition: Repeating sections is another common characteristic of music. Segment boundaries may be found at the start and end of repeated music regions.
- Regularity: It is based on the assumption that changes can be identified based on regular durations of one or more musical segments.

The fact that the perception of segmentation structure is ambiguous and annotators' interpretations are often divergent is well-known and mentioned throughout the literature. For example, different performance annotations can diverge because performers and listeners pay attention to different musical features like tempo or timbre (Chew, 2017). Another approach to the segmentation problem is to consider the listener's attention processes when annotations are produced (J. B. L. Smith, Schankler, & Chew, 2014) and try to reverse engineer the annotations based on combining multiple features. For instance, after annotating a recorded performance, automatic segmentation techniques can be used on self-similarity matrices from different musical features to generate boundaries, which are then compared to the actual annotations. Boundaries closer to those generated by a given feature may correlate with the feature the listener was likely paying attention to while annotating (J. B. Smith & Chew, 2013).

Current techniques for analyzing segmentation structures in music are based on models focusing on probabilistic and machine learning techniques, called data-driven methods. These annotations are used to collect the data necessary to train models to automate music analysis, creation, and recommendation applications, using large amounts of annotations to infer structural characteristics in music (Nieto et al., 2020). Even though our approach does plan to collect large-scale annotations of music structures using citizen science (Section 2.4), it is based on human annotators' interpretation of higher-level features in the music and uses flat boundary annotations primarily. Chapter 3 outlines selected projects with relevant contributions to music annotations that have influenced the development of our web-based annotation framework, CosmoNote. Chapter 5 will introduce our annotation protocol based on musical prosody and expressiveness (see Section 2.3), focusing on expressiveness in recorded performances, as discussed in the following section and further chapters.

2.3 Expressiveness and Performance

Music performance can be appreciated as part of a process where music is created, executed, and finally experienced. Thus, a three-part relationship between a composer, a performer, and a listener is created. The extent of these three roles, particularly that of performers, is sometimes

fuzzy. For example, performers may create improvised music in real-time without needing a written composition, composers may render music through software without a human performer, and experienced musicians may accomplish audiation, i.e., reading a score and experiencing music in their mind without hearing any sound.

As mentioned in Section 2.2, music structure analysis in symbolic representations, by concentrating solely on the abstract aspects of the composer's work, has relatively overlooked the performer's transformational contribution to the music. Thus, performers are often considered mediators who follow the instructions on the score merely to make them concrete (Cook, 2013).

Doğantan-Dack (2014) identifies Mathis Lussy's dynamics, timing, and phrasing annotations in musical scores during live performances and Carl Seashore's methodologies for capturing acoustic properties data on performances as the earliest instances of research studying expressiveness in music performance. These initial works established a paradigm where performance was seen as a deviation from the score and later as a deviation from a prototypical performance, a challenging concept to define. In contrast to that perspective, the approach we present in this section reviews research in performance from the point of view that considers deviation from the score as expected so that expressiveness is found in change (Leech-Wilkinson, 2009). This definition is aligned with Cook's outlook of music as performance over music as writing. Therefore, expressiveness in performance is tied to the thoughtful, timely executive decisions made by performers, impacting the quality of the music they play because of their engagement.

From the performer's perspective, according to (Palmer, 1997), playing a piece of written music involves cognitive processes categorized into three stages: (1) Interpretation, which entails the careful analysis of the meaning of a given piece of music, (2) Planning, retrieving from memory and deciding how to partition its structural components through specific actions, (3) Movement, the embodied delivery of these ideas by physically executing the actions mentioned above, generating the music that is heard. These steps constitute a mental model encapsulating premeditated structures, emotions, and slight human mistakes.

The model of the three cognitive stages in music performance helps separate areas of interest for performance research. For example, the mechanisms involved in performance, including cognitive or physical processes working together to generate sound, are included in the movement stage. Research in this area may examine instrumental technique and musical gesture, i.e., how fine motor movements are produced. Recent experiments extend work in biomechanical gesture models, using video recordings and motion capture to find which movements are responsible for the timbral characteristics of a pianist's sound (Valière, Lefèvre, Colloud, & Villard, 2019). Another example is the research on performance perception, which looks into the impact of performance on listeners while aiming to determine the origins of its structural ambiguities and their link to performer choices. Studies looking at the perception of changes in intensity and duration in music sequences have shown that there is not only a one-way relationship where perception constrains performance or vice versa. Instead, both perception and performance are constrained by inherent properties in music structures (Palmer, 1997).

Other aspects of scientific research in music performance, such as rehearsal, memory, performance anxiety, and performance practice, will not be covered in this work. We lean towards the interpretation and planning stages of music performance and are predominantly concerned with studying expressiveness in music performance. This broad concept is historically and culturally dependent and can be applied to human and computer performances. The term expressiveness, as used in this document, refers to auditory and musical elements, highlighting a delimited variation from standard performances without necessarily conveying any particular emotion, feeling, or mood (Fabian, Timmers, & Schubert, 2014).

Fabian et al. (2014) argued that expressiveness in music is found beyond a deviation from what is notated in a musical score. A performance can be expressive even without reference

to a score (as previously mentioned with improvisation). However, performers' and listeners' interpretations of certain structural relationships in music exist in a similar space of possible performances guided by the musical composition (Palmer, 1997). This relationship may be explored through multiple approaches, including the concept of expression layers in a musical work. It conceptualizes the difference between a compositional layer (based on the elements determined by the score) and a performance layer (belonging to what is shaped by performers) that contribute to the overall expressiveness of the music. In this model, additional dimensions map the expressiveness amount (from lacking to excessive) and the expressiveness appropriateness (whether it conforms to practices of a specific style) to both compositional layers (Schubert & Fabian, 2014). Although this way of looking at expressiveness may seem simplistic, it does capture the multidimensional nature of music performance in an accessible and adaptable framework where contributions from other fields may be understood and evaluated.

One concrete example of the interaction bewteen scientific disciplines studying music performance, is the influece of research in cognitive psychology, and more specifically on the expressive similarities between speech and music. The work by Juslin and Laukka (2003), among others, was significant in recognizing the commonalities in both speech and singing voice, for instance, shaping the creation of computational expressive performance models. This relationship is also encapsulated by the concept of musical prosody (Palmer & Hutchins, 2006), which describes how acoustic features can be manipulated for expressiveness without changing categorical information. Our protocol for annotating music structures in performance draws inspiration from this conceptualization. Chapter 4 expands on this rule-based form of auditory stimulation used to convey segmentation, prominence, coordination, and emotional response and, thus, to identify structures in performed music.

2.3.1 Types of Expressive Variations

As mentioned before, a limited palette of variations within performance expressiveness (Leech-Wilkinson, 2009) must be considered in research and viewed from a specific historical context and setting. Friberg and Battel (2011) classify ambiguities in expressiveness into two categories: (1) non-expressive variations arising from technical limitations of the performer (instrumental technique, instrument quality) or randomly (produced, for example, by imperfections in perceptual timing). (2) more interestingly, expressive variations indicate the intention to communicate expressiveness. These variations are deliberate but may be produced unconsciously, shaping the music using the expressive parameters discussed in the previous section.

Expressive variations can subsequently be divided into separate but not necessarily independent subcategories communicating music's emotional character (e.g., happy or sad), motional character (e.g., whether it is calm or urgent), or underlying structure (based on differentiating pitch and duration, and melodic, metric, and harmonic segmentation).

2.3.2 Communicating Expressiveness

Numerous external and internal factors may contribute to how expressiveness is communicated in a performance. Constraints can arise from a musical instrument's mechanics or condition and even from the room acoustics (Lerch, Arthur, Pati, & Gururani, 2019). For example, regardless of the pressure applied on the keys of a harpsichord, its strings are mechanically plucked with uniform force, limiting the player's ability to manipulate the volume of individual notes. Regarding an instrument's condition, piano players rarely play on instruments they own, so they must adapt to unknown instruments. Moreover, pianists who perform inside large, reverberant rooms may, for instance, have to lengthen pauses in a melody for the same

expressive effect. They might have to use the instrument's sustain pedal more sparingly than inside a small, sound-absorbent room. Even if ideal physical conditions and mastery of the instrument are assumed, the effective use of musical elements for expressiveness also requires some understanding of the music's historical context, cultural setting, and the audience who will hear the performance. Additionally, variations in performance may occur because of the performer's conscious or unconscious choices.

In this section, we will restrict the discussion of communicating expressiveness to musical elements independent of the previously mentioned factors. These elements, also called expressive parameters, capture different aspects of the performance, modeling standard practices that are then tested computationally. One example is musical phrases, which usually become faster and louder from the start while slower and quieter toward the end. Authors tend to describe different parameters, which leads to slight variations in the aspects of performance being represented (Cancino-Chacón, Grachten, Goebl, & Widmer, 2018). For the context of our research, we will examine pitch, dynamics, timing and tempo, timbre, and articulation. After a short characterization of these expressive parameters, we will see how these aspects are approximated, studied, evaluated as theoretical models of expressiveness, and even used in MIR tasks (transcription or recommendation) (Lerch et al., 2019). Because research presented in this manuscript focuses on attributes applicable to classical piano music that can be easily represented in MIDI data, some concepts contributing to expressiveness, such as ornamentation (ornaments would have to be hardcoded) and vibrato (not applicable to the piano), are not addressed in this description.

2.3.2.1 Pitch

Pitch is the perceptual attribute corresponding to the fundamental frequency of a sound. There exist acoustic musical instruments like the piano, where the pitch of each playable note is fixed to a given temperament and is altered only by re-tuning. In contrast, in others, like the violin or the human voice, performers can subtly control continuous pitch variations for artistic reasons. Expressive changes in pitch are usually notated in music scores with symbols/written instructions to execute techniques such as *vibrato*, *glissando*, and *portamento*.

A concept associated with pitch, often used to consider variations in pitch, is intonation. However, since we only use piano music in this document and neglect possible variations in pitch due to the tunning process of the instrument, we will not discuss this concept any further.

2.3.2.2 Dynamics

Variations in sound intensity represent one of the most effective ways to communicate musical changes, emotions, and intentions throughout a performance. In musical parlance, dynamics refers to the scope and movements in loudness. Italian subjective terms are standards in music by convention used to describe dynamics. For example, *piano* (***p***) for quiet and *forte* (***f***) for loud. Auxiliary words such as *poco* (a little), *mezzo* (moderately), and *più* (more) can be applied as modifiers to these words to change their meaning. Table 2.1 shows terminology used in dynamics ranging from quietest to loudest. Other denominations, such as *sotto voce* (hushed voice) or *mezza voce* (moderate voice) may also refer to musical dynamics.

Additional symbolic notations are often used to represent variations in loudness. Small-to-large hairpins indicate increases in dynamics (*crescendos*), while large-to-small hairpins illustrate decreases (*decrescendos*). However, the precise way to navigate such transitions is, in essence, up to the performer. Accent marks of varying kinds denote short spikes in loudness. For instance, sudden emphases in the sound may be marked by the term *sforzando* ***sf*** or ***sfz*** or preceded by the qualifier *subito*, as in *piano subito*. The gradations of such articulations differ in notation and practice, highlighting the performer's choice in communicating expressiveness.

Italian term	Abbreviation	English equivalent
pianissisimo	*ppp*	very very quiet
pianissimo	*pp*	very quiet
piano	*p*	quiet
mezzo piano	*mp*	moderately quiet
mezzo forte	*mf*	moderately loud
forte	*f*	loud
fortissimo	*ff*	very loud
fortississimo	*fff*	very very loud

Table 2.1: Common dynamic markings in Western Academic Music in Italian and English. Adapted from Abromont, 2001, p. 229.

2.3.2.3 Time and Timing

The three main aspects of time that performers have at their disposition are rhythm, meter, and tempo, as explained below. Related to how these three concepts interact when controlled by performers, the expressive parameter named timing often refers to the note-to-note or individual note duration variations (Cancino-Chacón et al., 2018).

Rhythm (as mentioned in Section 2.1.1) is linked to the duration of successive sounds (and silences), where they occur, and whether they are accented relative to each other. The sensation of pulse or beat within music is created by a steady pattern of sounds that are more or less important (strong or weak beats). In the Western canon, duration is notated as even subdivisions of the standard unit (whole note) by powers of two (e.g., 1/2, 1/4, and 1/8); writing a dot next to a note adds half its current duration. Researchers use the interonset interval (IOI), the time between two successive note starts (onsets), to estimate whether notes have been lengthened or shortened in performance (Friberg & Battel, 2011).

Metric structure arbitrarily groups sounds into musical measures based on their duration. A measure's meter or time signature is notated by counting how many sounds of the same duration fit into it according to a predetermined number of beats. Meter is used to give music a distinctive character and flow. To identify the meter of a piece, the listener must count the beats in a measure; more prevalent, regular time signatures are more straightforward to tap or dance to (e.g., a 3/4 waltz) than irregular (e.g., a 5/4 jazz piece) ones.

The term tempo is applied to describe the pace, rate, or speed at which the pulse of a piece unfolds in time, analogous to pace/rate in speech. Composers outline tempo globally in notation using a variety of standard Italian terms ranging from extremely slow *Grave* to extremely fast *Prestissimo*, as shown in Table 2.2. Additional modifiers can make tempo go faster (*accelerando*, *piu*), slower (*decelerando*, *ritardando*, *rallentando*), or steady. However, these markings are still subjective despite a broad agreement on each term's meaning. Metronomes grant a more precise delineation of tempo in beats per minute (BPM) for a particular note value, restraining the possibilities of what any given marker means. For example, a value indicating the tempo of a piece may be represented as a 'tempo equation' such as ♩=80, which is in the scope of *Andante* 76-108 BPM.

Tempo is controlled by the performer locally, i.e., they will consider the indications given by the composer while making their own decisions on where to change the pace of the music. It is worth noting that performers can also control rhythm by changing the duration ratio of successive notes (e.g., swing time - changing two eight notes into a dotted eight and a sixteenth) or may even use a different meter (to facilitate the execution of specific phrases). For performance analysis purposes, a curve tracing variations in tempo can be computed from a list (T) of timestamps (in seconds), containing N beats. The following formula calculates the inverse of the time difference

Italian term	Approximate BPM range	English equivalent
Grave	<40	extremely slow
Largo	40-60	very slow
Larguetto	60-66	less slow than *Largo*
Lento	52-60	slow
Adagio	66-76	slow and stately
Adagietto	70-80	less slow than *Adagio*
Andante	76-108	at a walking pace
Moderato	108-120	moderately
Allegretto	112-124	moderately fast
Allegro	120-168	fast and bright
Vivace	168-176	lively and fast
Presto	168-200	very fast
Prestissimo	>200	extremely fast

Table 2.2: Common tempo markings in Western Academic Music in Italian and English. Adapted from Abromont, 2001, p. 144.

between successive beats in one minute (60 seconds), producing a list where each value $b_{[i]}$ is expressed in BPM:

$$b_{[i]} = \frac{60}{T_{[i+1]} - T_{[i]}}, \quad \forall i \in \{1, 2, \ldots, N-1\}$$

2.3.2.4 Timbre

Timbre is the property of a sound that allows listeners to distinguish different sources (musical instrument, speaker, or other) reproducing the same pitch. Timbre perception is based on the temporal envelope and spectral components at the onset of a sound and on their evolution over time. Composers diversify timbre in music by using arrangement and orchestration besides directly specifying a performance technique. Controlling the timbre of an individual instrument is achieved differently in different musical instruments. For instance, in acoustic instruments, variations on the note's envelope or spectral content are controlled by the performer's sound producing gesture (e.g., light vs. heavy touch) or mechanically on the instrument (e.g., using the soft pedal on a piano or controlling the embouchure on a wind instrument). In MIDI, an entirely new timbre is accessible simply by explicitly indicating which musical instrument number to use (e.g., changing the number from 9–celesta to 10–glockenspiel). In score notation, modifying the timbre can be accomplished via articulation marks.

2.3.2.5 Articulation

Articulation in music specifies how one or more sounds are executed. It is notated on scores by drawing symbols above or below the note(s) that it alters. This musical element shapes how a sound is played, affecting other expressive parameters. Standard articulation marks notably influencing timbre and dynamics are *staccato* (sounds have a shorter attack and duration), *tenuto* (sounds are held for their entire length), *legato* (notes are linked together), *accents* and *marcato* (used to contrast sounds relative to their context).

2.3.3 Modeling and Evaluating Expressiveness

The following paragraphs describe the constituent ideas at the center of computational expressive performance models without dissecting their individual architectures and inner workings. We

will show how models utilize music representations (symbolic or not) and expressive parameters to generate performances and how they can be evaluated.

Models

Computational models of performance expressiveness can be constructed using different techniques and complexities. A framework to study expressive performance models, proposed by Kirke and Miranda (2012), deconstructs a performance model into a system with separate modules such as the analysis of musical features, the context defining the music's mood or style, the physicality of the musical instrument, the performance patterns in previous examples, and the ability to provide feeedback and adjust to inputs from other modules. The most important module in this framework is Performance knowledge, a core method controlling the whole model's behavior, guided by insights about the music, and capable of executing specific performance actions.

Among all Performance knowledge methods, Kirke and Miranda reported in their meticulous overview that it is worth noting a contrast between learning models and non-learning models. Learning models have Performance knowledge methods able to incorporate parameters from previous performances to affect the outcome of simulated ones. For example, machine learning models can train on the expressive parameters of recorded MIDI performances and apply them to new performances. Non-learning models are rule-based; they use a set of predetermined configurations of expressive parameters. These models are tractable, i.e., one can easily understand every step in their operation but may produce less generalizable expressive performances than machine learning techniques.

In a simplified appreciation, Cancino-Chacón et al. (2018) state that modeling expressive performance computationally involves a function mapping musical features at the input to predict expressive parameters that would produce an adequate acoustic realization of a piece at the output, reminiscent of the black box denomination given by Bresin and Friberg (2012). Their overview of the different techniques overlapped with that presented by Kirke and Miranda. As a result of their expressive performance model analysis, Cancino-Chacón et al. observed three noticeable trends in their development. (1) increasingly focusing on data-driven approaches, which infer the function's input from data (e.g., musicXML scores, performance MIDI data, or tempo curves), recognizing a growing use of machine learning approaches for their obtention and processing. (2) using human-computer interaction systems, where humans can influence the function's output. For example, performers can be directly responsible for shaping the expressive parameters. (3) including cognitively-based techniques, for example, using findings about musical perception to predict expressive parameters.

The KTH rule system for musical performance (Director Musices–DM) (Juslin, Friberg, & Bresin, 2001; Friberg, Bresin, & Sundberg, 2006) is a concrete example of a model that integrates cognitively-based techniques and real-time capabilities (in the pDM version, Friberg, 2006), within a rule-based Performance knowledge model. The KTH rule system was built on research combining the GTTM principles, emotional expression, musical motion, and random variations into performance rules to control expressive parameters. For instance, the phrasing rule alters dynamics and tempo following a similar arch-like motion, where the music is louder and faster after the start of a phrase while getting quieter and slowing down at the end. These performance actions also map to the emotional expression in music (Kirke & Miranda, 2012).

Evaluation

Once the expressive performance model has been built, it must be evaluated. However, since models are multidimensional processes constructed relying on Performance knowledge modules with varying assumptions, depending on different expressive parameters, and yielding ill-

matched musical outputs (e.g., only monophonic vs. non-monophonic), this question remains open (Kirke & Miranda, 2012). However, we will recount the approach Bresin and Friberg (2012) took for tackling this issue.

First, considering the simplified black box paradigm of expressive performance models, gaining relevant insights to define what is worth evaluating can be achieved by applying three criteria generalizability (how well it deals with different performances or styles), flexibility (how well it can adapt to novel performances or expressive intentions), and parameterization (how simple, explicit, or close-to-human it is). Then, as evaluation methods compare these criteria against distinctive standards, evaluation methods can be categorized according to what they compare. Some methods evaluate performances by comparing them against data, either ground truth annotations from experts analyzing expressiveness or measuring its closeness-of-fit to a particular metric or representation of expressive performances. Other methods evaluate expressiveness by conducting listening experiments and measuring specific parameters, such as applying a rule in the model or the emotional communication of a generated performance. Another method to evaluate expressiveness, interactive listening, is using the principle of human-computer interaction systems in models to control parameters in their evaluation.

The Performance Rendering Contest (RENCON), where different models generated competing expressive performances of the same score, was described as a 'musical Turing test' of performance (Bresin & Friberg, 2012; Kirke & Miranda, 2012). This competition's "ultimate goal" forecasted a performance-rendering machine winning the Chopin Concours[9] by 2050. Although RENCON was last held in 2013[10], their prediction seems reasonable, as recent advances in machine learning performance models have claimed generated music to be indistinguishable from human performances (Schubert, Canazza, De Poli, & Rodà, 2017).

While the extraction of performance features and evaluation expressive performance models has steadily pivoted to machine learning techniques, most of these evaluation methods still rely on manually annotated data, mainly because of the inadequate accuracy of automatic annotations. Bresin and Friberg and Lerch et al. mention that the importance of high-quality performance data (as opposed to limited, weakly labeled, and unlabeled data) is tied to the lack of generalizability of machine learning expressive models, despite the progress in recent years. Another crucial stage in their development is the interpretability of results, i.e., understanding what the model has learned about the performance. As discussed in further chapters, our work aims to provide a framework containing performance data, music features, and capturing structural annotations of musical prosody. Our protocol can generate scalable, high-quality rich annotations that may be used with data-driven methods to leverage their advantages while allowing improved interpretability.

2.4 Citizen Science

Citizen science is a research practice in which large amounts of people that do not necessarily have formal scientific training can participate in the creation, collection, classification, and analysis of scientific research (Strasser, Baudry, Mahr, Sanchez, & Tancoigne, 2019; M. M. Haklay et al., 2021). Different modalities of participants' involvement and motivations need to be considered on the side of the contributors, as do methodological issues like training and evaluation on the side of researchers.

Although the origin of the citizen science appellation is traced to 1989 (M. M. Haklay et al., 2021), civil society's contributions to academic research have been documented for far

[9] https://chopin2020.pl/en/

[10] https://web.archive.org/web/20150221235324/http://smac2013.renconmusic.org/

longer. Indeed, research was typically conducted in close collaboration between amateurs and professionals before the professionalization of science in the 19th century, when the role of non-scientists was diminished (Miller-Rushing, Primack, & Bonney, 2012). However, citizen scientists continued to impact research, notably when their well-being was at stake, as was the case during the AIDS epidemic, where activism impacted research development and its agenda (Senabre Hidalgo et al., 2021). In recent years, citizen science projects have proliferated to advance knowledge in astronomy, ecology, geography, biology, and medicine, among other fields.

Numerous scientific fields are involved in citizen science and different perspectives on cataloging its projects. In the context of the multiple approaches to categorizing citizen science, the concept of levels of participation (M. Haklay, 2013) is a valuable indicator developed for understanding the relationship between citizen scientists–involved in different stages of a research project–and stakeholders (citizens, researchers, policymakers) affected by its outcomes. M. Haklay denotes four levels of participation in citizen science projects:

1. Crowdsourcing: Researchers are responsible for the project from design to analysis. Participants contribute mainly with resources, for example, by using sensors to collect data and send it back for analysis or sharing unused computing power through a network. In this level, also called volunteered computing, the cognitive effort required by participants is minimal.

2. Distributed intelligence: Participants receive training on the project's protocol, still developed by the researchers, and contribute by collecting and interpreting data. The introduction of volunteered thinking means that researchers must consider participants' interests, questions, and expectations, going beyond the initial training.

3. Participatory science: This level extends the involvement of participants to define the research problem. Experts and participants collaborate on devising the protocol, collecting, and interpreting the data.

4. Extreme citizen science: It is characterized by blurring the line between experts and participants, combining all the aspects of previous levels into a fully collaborative effort where citizen scientists choose their implication level, being able to take part in the data analysis, publication, or deciding its ultimate usage.

Projects in citizen science vary significantly in scope and activities, ranging from observational tasks such as transcribing text to reporting the weather and folding proteins (Khatib et al., 2011). This diversity of activities and outcomes is why the levels of participation classification can be complemented by incorporating the notion of epistemic practices of citizen science projects (Strasser et al., 2019). This non-hierarchical list considers the actions of citizen scientists for knowledge production into five practices: sensing, computing, analyzing, self-reporting, and making. In this sense, the same project can integrate many epistemic practices throughout its lifespan, e.g., marking changes in loudness (sensing) at the start of a project and manipulating aggregated data (analyzing) to shape music performances once the project has matured.

Independently of epistemic practices and levels of participation, citizen scientists employ physical objects, such as mobile devices (equipped with sensors) or personal computers, coupled with virtual resources like specific software with social network features. To this end, web-based platforms grouping several projects in one place have emerged to maximize data collection and notoriety. For instance, the popular site called the Zooniverse grew from the *GalaxyZoo*

[10] https://www.zooniverse.org/

initiative in astronomy and now hosts projects in many disciplines. The site proposes its proprietary project builder tool, allowing to upload images, video, and audio files, giving creators and users basic annotation functionalities. Other websites, like *ParticipArc*[11] in France, have chosen instead to feature distinct citizen science projects without hosting any of their content or homogenizing their data collection process.

Citizen science projects focused on auditory tasks, such as research in acoustics and music, have often taken the form of crowdsourcing and distributed intelligence, where participants either volunteer their computing power or act as sensors, i.e., identifying animal calls (Shamir et al., 2014), classifying speech (Semenzin, Hamrick, Seidl, Kelleher, & Cristia, 2021), and participating in numerous music annotation tasks (Bruderer et al., 2009; Wang, Mysore, & Dubnov, 2017; Hartmann, 2017). However, the use of the term crowdsourcing is frequent even if paid sites like Amazon Mechanical Turk[12] (MTurk) are used, falling outside the definition of the volunteer contributions of citizen science.

Contribution from citizen scientists is especially relevant in music structure analysis, where automating annotation tasks is challenging, and machine-generated annotations sometimes fail to provide satisfactory data. Large databases containing analyses or annotations of musical pieces used to be impractical, given the tremendous effort and time needed for experts to mark each work manually. However, new technologies and the use of sensing, computing, and analyzing practices allowed the creation of new databases with MIDI, audio data, and many metadata properties for research. For example, projects such as the now renowned *RWC* (Goto, Hashiguchi, Nishimura, & Oka, 2003; Goto, 2006), *SALAMI* (De Roure, Downie, & Fujinaga, 2010; J. B. L. Smith, Burgoyne, Fujinaga, Roure, & Downie, 2011), and *Songle* (Goto, Yoshii, Fujihara, Mauch, & Nakano, 2011) databases contain structural annotations and new databases such as Music4all (Santana et al., 2020) or *GiantMIDI* (Kong, Li, Chen, & Wang, 2022) provide score alignment, lyrics, popularity and more. Recent dedicated efforts to gather annotations on musical pieces' expressive and emotional content (Gutiérrez Páez et al., 2021) prove the potential of using citizen science for studying expressiveness in music performance.

We will revisit topics related to web-based annotations of music expressiveness in Chapter 3. The citizen science project outlined in this manuscript (see Chapter 4) is situated in the distributed intelligence participation level. Scientists carry out the protocol design and data analysis, while participants contribute their volunteered thinking. As we will describe in later chapters, we contemplate the possibility of increasing the engagement of citizen scientists and the epistemic practices they implement to reach the participatory science level as the project evolves.

2.5 Connecting Threads – Summary

In this chapter, we have examined research and connections between disciplines dedicated to the traditional and computational analysis of music as writing and as performance, aligning our research methodology with the latter. In this sense, we are particularly interested in studying expressiveness in music as expressed through its prosodic functions. We have also seen how citizen science has been applied to research in music information retrieval, music perception, and music performance, among others.

The process of musical analysis is ancillary to the research described in this document. However, as explained in Chapter 1, analysis is not the goal of the research. We will present descriptive accounts of structures in performed music, supported by the conceptual framework

[11] https://www.participarc.net/
[12] https://www.mturk.com/

in the literature, providing specific insights about how musical elements (e.g., melody, harmony, rhythm) interact and, when necessary, how musical form influences these structures. Additionally, we have begun to explore how representations and visualizations of musical elements may influence how the music is studied and even inform the computational techniques that can be applied.

To introduce the study of annotations in performed music, we started by looking at the literature tackling the segmentation problem in music structure analysis via boundary demarcation. We showed, for instance, how boundaries can be identified mathematically (e.g., using the principles of homogeneity, novelty, repetition, and regularity) or perceptually using theories that have been validated experimentally (e.g., based on the GTTM rules). However, human perception of musical segmentation encompasses ambiguity and subjectivity that are not readily resolved. We will return to the question of ambiguity in subsequent chapters.

This chapter adopts the literature's definition of music as performance as an indispensable component in the music experience, where performers shape structure through interpretation, planning, and movement. We concentrate on the first two stages to study music expressiveness, defined as the change in performance, not necessarily tied to a score. We saw how expressive variations communicate structural elements through the performer manipulating expressive parameters (e.g., timing, dynamics, articulation) and how computational models can study and evaluate these variations.

The paradigm of citizen science was presented as an alternative to conventional data collection methods for studying structures in music. These practices involve volunteer contributions of nonprofessionals in scientific research projects. We noticed how participation in such projects could be classified into levels or epistemic practices and how citizen science initiatives have already been applied in music-related projects.

Current trends, including data-driven methods, bridge the gap in musical analysis, symbolic or otherwise, were also highlighted in this chapter. These methods use rule-based, cognitively-based, and machine learning techniques and are nourished from annotated musical data, demanding increasingly automated approaches for its collection, analysis, and evaluation. However, few studies have examined expressiveness in music performance. We have articulated the need for large databases with high-quality annotations and adequately labeled musical annotations, which can be achieved through citizen science projects at a distributed intelligence level or higher and by applying many epistemic practices. The following chapters will explore our proposal for tackling this problem.

Chapter 3

The Conception and Design of CosmoNote

As discussed in Chapter 2, there is a gap in online collaborative markup tools, having both standard music representations and other descriptors important for expressiveness. To enable large scale collection of annotations of performed music, we have created the web-based annotation platform CosmoNote, in tandem with a complete musical prosody annotation workflow. This chapter focuses on the technical aspects of the platform, which was used extensively in this thesis for reflecting about structures in performed music, designing musical annotation workflows, generating annotated music examples, and collecting annotations. Chapter 4 will dive deep into methodological aspects of using CosmoNote for collaborative research.

Lawrence Fyfe, the full-stack web developer of the COSMOS project, was responsible for coding the CosmoNote app. Throughout the whole development process, my involvement and contributions were primarily in the areas of: interface design, feature implementation, audio descriptor data computation, user interaction, testing/debugging, experiment design, experiment deployment, documentation, communication, and outreach. Work presented in this chapter was published in "Annotation and Analysis of Recorded Piano Performances on the Web" (Fyfe, Bedoya, & Chew, 2022) and "Creating Experiments with CosmoNote" (Bedoya, Fyfe, & Chew, 2022a).

3.1 Introduction

In the course of performing notated compositions, performers add their own expressive manipulations that may not be scripted in the score and which, if transcribed back into music notation, could be shown to be far from the written score (Chew, 2018). These expressive manipulations, including variations in timing, loudness, articulation, and timbre (Palmer & Hutchins, 2006), convey groupings and prominence of notes, forms of performed structures, to listeners (Chew, 2017). These structures may be conceived first in the mind of the performer, developed on the fly as they make sense of the music, or when performing the piece of music. In any of these cases, the structures are then eventually transmitted to the minds of listeners. While these structures may be perceptible by listeners, consciously or unconsciously, they may, however, be difficult to discern with automated analysis. For example, whether an accented note marks the beginning or the end of a note grouping may be ambiguous with automated analysis but not to a human listener. In the absence of reliable automated analysis, we use citizen science to study performed music structures as they are created in recorded music performances and in the minds of listeners.

The COSMOS project (see Section 1.1) was created to study such performed musical structures in performances of classical piano music. The video (in French), "Le piano virtuose," (CNRS, 2020), explains this research in general terms. To enable our research, we needed a workflow that allowed citizen scientists to annotate those perceived structures and a software tool that enabled that workflow. The workflow needed to start with presenting the recorded piano performances. Citizen scientists would then listen to the recorded performances and see the notes played as well as expressive features extracted from the recorded audio. A variety of annotation types would be provided including the marking of boundaries, regions, comments, and groups of notes. Finally, annotations collected from citizen scientists, ranging from musical novices to professional musicians, would then form the basis for studying how performance shapes or re-shapes perceived musical structures.

The research question that we address in this chapter is: how can we create a workflow for presenting recorded piano performances, creating annotations of those performances, and then analyzing those annotations? In answering this question, we developed CosmoNote[1], a web-based citizen science tool for visualizing and annotating expressive piano performances, to embody the workflow that we created. The basic design of the software was discussed in the article "CosmoNote: A Web-based Citizen Science Tool for Annotating Music Performances" (Fyfe, Bedoya, Guichaoua, & Chew, 2021). This chapter goes beyond a description of the software to describe the CosmoNote workflow in detail while including more detailed descriptions of the software functionality as well as including new features and significant changes since the introduction of the platform in 2021.

The next section (Section 3.2) describes related work and approaches to music annotation. The next four sections are organized according to the CosmoNote workflow. First, performance data is obtained for inclusion in CosmoNote (Section 3.3). Second, the performance data is presented to annotators as both audio and visuals (Section 3.4). Third, based on the performance data, annotators create their annotations (Section 3.5). Fourth, the annotations are collected and analyzed as shown in two pilot studies (Section 3.6). To conclude, the contributions in CosmoNote, the results of the two pilot studies, and directions for future work are summarized (Section 3.7).

3.2 Related Work

Multiple software tools with audio annotation capabilities have been developed for speech, music, and video applications (see Chapter 2 and Fyfe et al., 2021). This section provides an overview of related work in audio annotation software. Rather than providing a comprehensive list of all the audio annotation projects, we will describe only the music annotation tools most relevant to our work. Annotation applications tend to involve either human or automated annotations, and sometimes a combination of both. Since citizen science is a core part of our workflow, we only look at human or (a combination of) human-computer applications. Because CosmoNote was envisioned as a web-based project from the beginning, the projects reviewed are divided into non-web-based projects and web-based projects (most relevant to the CosmoNote workflow).

3.2.1 Non-Web-Based

Tzanetakis and Cook (2000) wanted to determine whether a computer-assisted human temporal segmentation annotation task benefited from automated segmentation. As part of a pilot user study, they presented users with an annotation application, based on their MARSYAS software

[1] https://cosmonote.isd.kcl.ac.uk//

framework (Tzanetakis & Cook, 1999), and asked them to mark temporal boundaries based on what they called sound "texture" or changes in instrument or speaker, etc. Similarly, Amatriain, Arumí, and Ramírez (2002) developed, using their CLAM audio framework, the CLAM Annotator (Amatriain, Massaguer, Garcia, & Mosquera, 2005), a combination human-computer annotation application where users could edit descriptors that were created automatically.

Notess and Swan (2004) developed Timeliner, a human-based annotation application, to allow users of a digital music library to create annotations for audio files in the library. Annotations included marking time regions and specific time points of interest. Text labels could be created for each of these annotations and the playback of the audio files could be tied to the annotations.

Herrera et al. (2005) developed the MUCOSA project to enable a variety of annotation workflows. MUCOSA was built on top of WaveSurfer (Sjölander & Beskow, 2000), a speech annotation tool that allowed for plugins. Annotations like structure markings were shown as squares in a panel vertically stacked below separate spectrogram and waveform visualizations. B. Li, Burgoyne, and Fujinaga (2006) wanted to establish a set of ground truth data for the segmentation of songs, i.e. by annotating regions for chorus and non-chorus parts of the songs. To do that, they built an annotation system on top of the Audacity audio editor[2] by adding separate tracks below Audacity's normal waveform visualization track. The annotation tracks contained region markers and labels for regions and, like the MUCOSA annotation system, the waveform visualization and the annotations were stacked vertically.

Cannam, Landone, and Sandler (2010) developed Sonic Visualiser to be "the first program you reach for when want to study a musical recording rather than simply listen to it". In addition to featuring different visualization layers like waveforms, spectrograms, and notes, Sonic Visualiser allowed annotations to be placed directly over the visualization layers, saving screen space. Of all the non-web tools that we examined, Sonic Visualiser had the set of features that matched most closely with our overall workflow.

3.2.2 Web-Based

The projects described in the previous section had interesting features but were not web tools, making them difficult to use in our citizen science-based workflow. So we looked explicitly at web-based annotation tools and their capabilities for music annotation. The following paragraphs describe some noteworthy examples.

Goto et al. (2011) introduced a public web service for active music listening, Songle. The platform automatically computes structural segmentation, beat structure analysis, melody, and chord extraction from the music. Users can visualize a video along with these music features and anonymously correct errors in them.

Cartwright et al. (2017), as part of their CrowdCurio project, wanted users to annotate soundscapes of varying complexity based on waveform visualizations, spectrograms, or no visualization at all. To do that, they created Audio-annotator (Mikloska, 2017), built on top of the WaveSurfer.js[3] waveform visualization library. Audio-annotator allowed users to create annotations by selecting a sound region. Annotations could be edited and users could listen to the sounds from the selected regions of their annotations. Meléndez Catalán, Molina, and Gómez Gutiérrez (2017), to allow users to annotate radio recordings with the goal of detecting music in radio broadcasts, created BAT (Catalán, 2022), another WaveSurfer.js-based web annotation tool. Annotators were asked to distinguish between music and speech in the recordings by selecting time regions and then identifying them as music or speech.

[2] https://www.audacityteam.org/
[3] https://wavesurfer-js.org/

Wang et al. (2017) used the CAQE toolkit (Cartwright, Pardo, Mysore, & Hoffman, 2016) to create a web interface for crowd-sourcing a music segmentation task in which they asked annotators on Amazon's Mechanical Turk to listen for changes between one part of a song and another and to mark a boundary there. Annotators would listen to 20 second clips from the songs and to mark boundaries by adjusting a slider that could be positioned anywhere in the time frame of the clip provided. Other than the slider and an audio progress bar, both of which only appear after the first listen, there is no other visual indicator since the authors wanted annotators to focus on what they heard in the clip.

Giraud, Groult, and Leguy (2018) used the Polymer 2.0 framework[4] to create the Dezrann annotation platform. This platform is mainly used in projects that annotate music structures using score notation (e.g., annotating music texture in piano music–Couturier, Bigo, & Levé, 2022), although it can also annotate audio/video files of performed music display a waveform visualization (e.g., annotating music for pedagogical purposes–Sauda, Giraud, & Leguy, 2022).

Gutiérrez Páez et al. (2021) proposed the Music Enthusiasts (ME) web annotation platform aimed for collecting emotional "ground truth" data to tackle the music emotion recognition (MER) problem in MIR. It presents excerpts from a database of mostly non-academic music from around the world and collects emotional ratings in two dimensions (valence/arousal), a binary question about music familiarity and preference, an emotional word tag, and a text box for explaining their reasoning. Their work applies citizen science aiming to create personalized music recommendations based on active learning techniques (Gómez-Cañón et al., 2023).

3.2.3 Relevance

While each of the projects described above offered some useful features, they did not offer enough features to be relevant for our workflow. In particular, all the web-based projects displayed their sound or music selections as waveforms and/or spectrograms, failing to include the note-based information that is crucial for a more fine-grained analysis of performed structures. The other projects were also limited in their annotation types with all the projects offering region selections and only some offering boundaries and/or comments. For our workflow, we needed all of these annotation options, including boundaries, regions, comments, and, along with our note visuals, the ability to select groups of notes. Our platform provides interactions for annotations (create, select, move, filter, delete) and visuals (zoom, pan, filter, toggle), increasing its potential for analysis and research applications (Lima et al., 2021). The need for a more customized web tool for annotations with a particular workflow led us to develop our own citizen science annotation tool, CosmoNote.

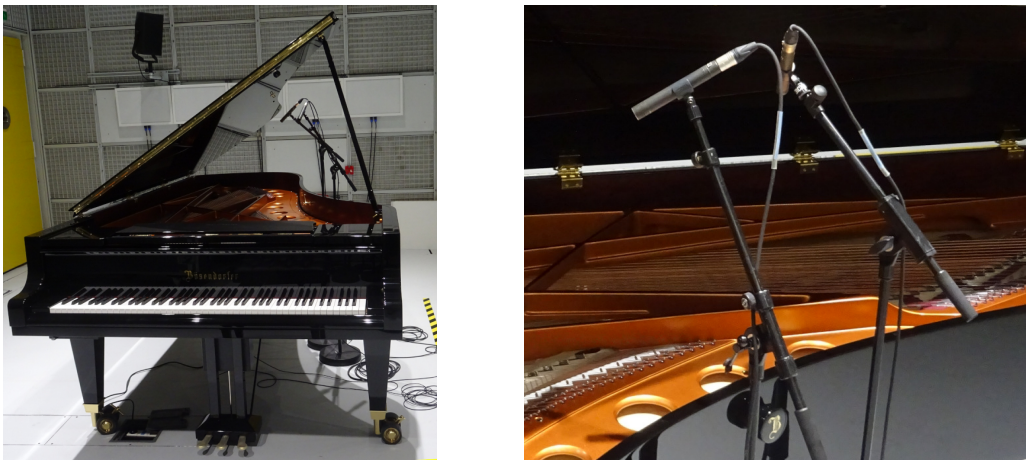
3.3 Obtaining Performance Data

The CosmoNote workflow starts with obtaining performance data for our citizen scientists to annotate. We obtain piano performance data in the form of both audio and/or MIDI recordings. These recordings are obtained via two paths: from existing recordings of audio and/or MIDI or from recordings made by the CosmoNote team. The recordings made by the CosmoNote team have the advantage of having synchronized audio and MIDI data, something that is not necessarily true for other recordings.

[4] https://polymer-project.org/

3.3.1 Recording the Performances

For recordings made by the CosmoNote team, we use a reproducing piano, the Bösendorfer Disklavier ENSPIRE[5] PRO Concert Grand 280VC, that is capable of recording performances as MIDI data in addition to producing high quality acoustic sounds and enabling fine control of musical expression. During the recording process, the MIDI data is streamed from the piano to the computer simultaneously with the audio, recorded via microphones, ensuring proper synchronization. Figure 3.1 shows an example of the microphone placement for recording audio, while Figure 3.2 shows, via CosmoNote, a performance of Christian Petzold's Minuet in G minor[6] (originally attributed to J.S. Bach because of its inclusion in Bach's Little Notebook for Anna Magdalena Bach). The performance, by Elaine Chew and recorded on the Bösendorfer piano, will be, unless otherwise noted, used throughout the remainder of the chapter to illustrate the different aspects of CosmoNote.



(a) Bösendorfer Disklavier piano. (b) Schoeps MK 4 condenser microphones.

Figure 3.1: Audio recording setup for the Bösendorfer piano at IRCAM's Studio 5. Two Schoeps MK 4 cardioid microphones are positioned using the ORTF stereo recording technique.

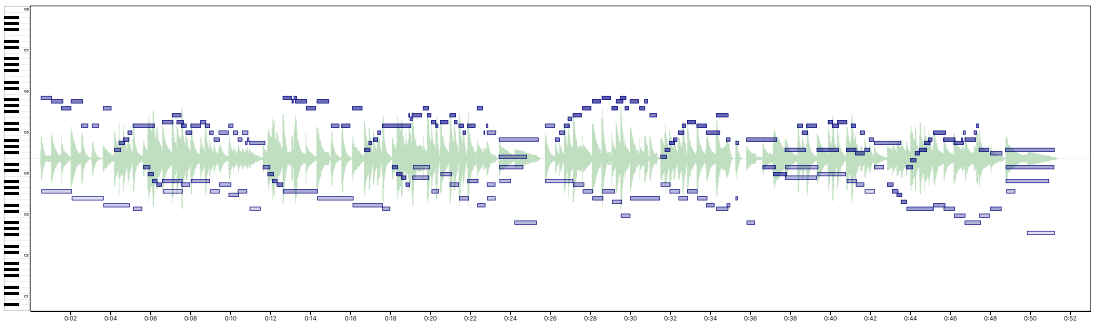


Figure 3.2: CosmoNote presenting a performance by Elaine Chew of the Minuet in G minor, BWV Anh. 115 from the Little Notebook for Anna Magdalena Bach as recorded on a Bösendorfer Enspire reproducing piano.

We also use pre-recorded MIDI from various performance collections, including the Bösendorfer Legendary Artists Library, or audio and MIDI from the Stanford Piano Roll Archive (SUPRA),

[5] https://www.boesendorfer.com/en/pianos/disklavier-edition

[6] Go to https://doi.org/10.6084/m9.figshare.23732634.v1 to hear the performance in Figure 3.2.

and Steinway's Glenn Gould Goldberg Variations MIDI files, meticulously created from Gould's 1955 recording for re-performances by Zenph Studios.

3.3.2 Preparing the Audio, Note, and Pedal Data

Audio files, whether recorded by the CosmoNote team or obtained from another source, start as uncompressed WAV or AIFF. To make the audio faster to download, we initially compressed the audio files using both FLAC[7] and OPUS (Valin, Maxwell, Terriberry, & Vos, 2013). FLAC files proved to be quite large for fast downloading and OPUS was not completely supported by all the browsers we tested. In the end, we decided to use MP3 for compressed audio since it provided a good trade-off between fast downloading and audio quality and because it was supported by all the browsers we tested.

Note data in CosmoNote comes from recorded MIDI files, either recorded by the CosmoNote team with our recording piano or from existing recordings. In MIDI, individual notes are split into pairs of note-on and note-off events. To visualize the notes in CosmoNote, we need each note to be a single event with a start time and an end time. To that end, we use a custom Python script that uses the Mido MIDI library (Bjørndalen, 2023), to convert the MIDI data to single-event note data.

Pedal data from the sustain, soft, and sostenuto pedals is also taken from MIDI files using the same script. The pedal data from MIDI is a series of control change events rather than pairs on on-off events as with the note events. For recordings from CosmoNote's recording piano, the extent of pedal depression is recorded with an 8-bit resolution of 0–127. For certain pre-existing MIDI files (not recorded with CosmoNote's recording piano) the control change messages for the pedals are only on-off events since the extent of pedal depression was not recorded.

3.3.3 Computing Feature Data

In CosmoNote, feature data is defined as the data computed from audio files, MIDI files, or the score for a given performance. To streamline the extraction and formatting of the feature data, we have created and released the first version of the Python package `cosmodoit`[8]. This package is composed of various custom Python scripts that bring together functionalities that come from different programming languages but were not available in one place. This tool is modular and extensible, and it is currently used to compute the following descriptors:

Loudness data (in sones) is computed from the audio file, estimated per frequency band using a psychoacoustic model using a module ported from the MATLAB MA Toolbox (Pampalk, 2004), as a global representation of the perceived intensity of the notes. It corresponds roughly to velocity data shown for the notes although the perceived loudness is influenced by the number of notes played and their pitches, as well as how quickly the key is depressed for individual notes as with velocity. Loudness data can be used to, for example, locate a group of notes highlighted by the performer to make a melody more salient than the contextual background or a note more prominent than its neighbors. The curve drawn in CosmoNote is normalized and smoothed so that it gives a global sense of the changes in loudness, without having abrupt local changes. However, the smoothing parameter can be adapted to different needs.

Tempo (in BPM) is computed using timestamps of the onset of each beat—where the beats are located by alignment to a MusicXML score using an automatic alignment module adapted from Nakamura, Yoshii, and Katayose (2017)—throughout a performance and is computed as

[7] https://xiph.org/flac/
[8] https://pypi.org/project/cosmodoit/

the inverse of the time between beats (see Section 2.3.2.3). The tempo can also be computed from manual beat annotations using the same approach. As an example of using tempo data, the parts where the curve shows a steep descent followed by an ascent could help in the identification of phrase boundaries (Stowell & Chew, 2013) or tipping points (Chew, 2016) (devices for heightening suspense in expressive performance).

Harmonic tension is computed from the score using the MIDI file of the performance, with a module that adapts the Python implementation of the XmlTensionVisualiser tool (Herremans, 2016; Guo, Herremans, & Magnusson, 2019) and has three dimensions (Herremans & Chew, 2016): cloud diameter representing dissonance, cloud momentum representing the rate of chord changes, and tensile strain representing the distance from the global key. Harmonic tension, for example cloud momentum, could be used by musically trained annotators to visualize large movements in tonality where non-diatonic chords are used.

3.3.4 Storing the Data

Once all the performance data is collected, it is loaded into a CouchDB[9] database server using custom Python scripts. CouchDB was chosen because it works well with web applications by supporting HTTP for transactions (with no database drivers needed) and by using JSON (Bray, 2017), a data format widely supported by a variety of tools and languages, for data storage.

3.4 Presenting The Performances

Next in the CosmoNote workflow, after the CosmoNote performance data is collected and stored, the data is presented to annotators as both audio and visuals via a client-side web application. CosmoNote audio uses the Web Audio API[10] and CosmoNote visuals use D3 (Bostock, Ogievetsky, & Heer, 2011), a highly-customizable, SVG-based[11] visualization and interaction library. CosmoNote clients get the performance data (as JSON) from the CouchDB server using PouchDB[12].

Performances are presented to annotators either as whole pieces of music or as fragments of pieces. In either case, all the performances are grouped into collections, where collections can be based around a performer, a composer, or some other theme. For example, the first sets of collections released in CosmoNote, starting in December 2021, were built around the 1955 recordings of Bach's Goldberg Variations by Glenn Gould and a collection of simple practice pieces for training new annotators. Figure 3.3 shows the collection selection page featuring four collections of performances by Glenn Gould along with the training collection which always appears at the top of the page. The performance of the Minuet in G minor by Elaine Chew, used as an example throughout this chapter, is included in the training collection.

Annotators select a collection before beginning to annotate its performances. Once they have selected a collection, annotators can access the performances in that collection with each recorded performance having its own page. A gray bar at the top of the page shows the name of the current collection and the current performance number (out of the total) along with forward and backward buttons for navigating the performances in the collection. Performances in a collection have a set order, but there is an option to present the performances in a given collection to each annotator in a randomized order with the order being stored per annotator for

[9] https://couchdb.apache.org/
[10] https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API
[11] https://www.w3.org/Graphics/SVG/
[12] https://pouchdb.com/

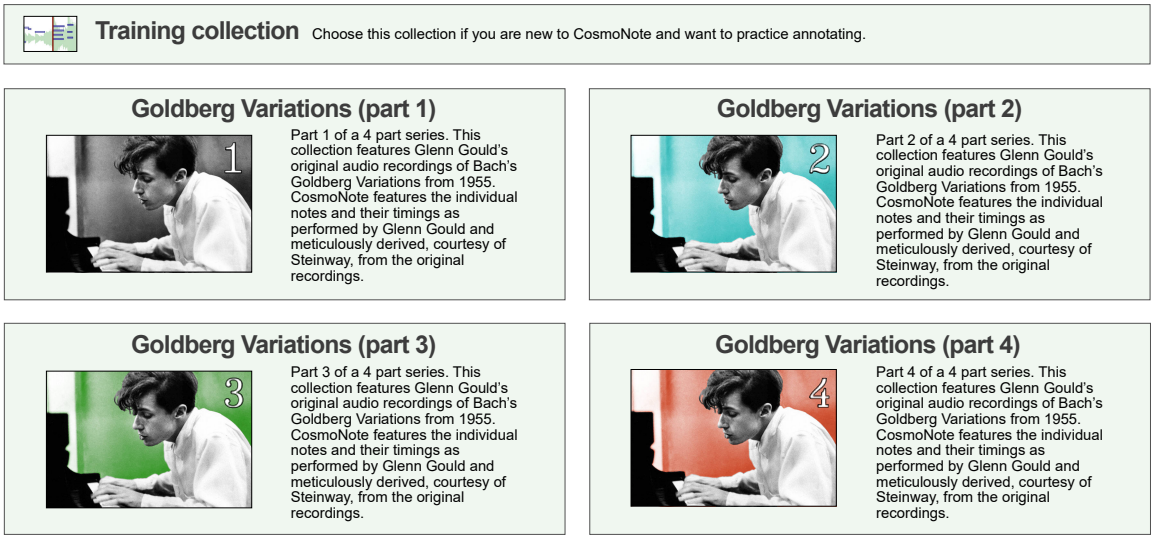


Figure 3.3: The CosmoNote collections page allows annotators to select a collection to annotate including an always-available training collection for new annotators.

later analysis. Below the collection navigation bar is information about the performance, usually title, performer, and composer. The display of performance information (or lack of display) is customizable, depending on the task. The controls, mostly buttons, are also highly customizable depending on the task (more on this functionality is detailed in Section 3.5.2). For example, there are buttons to turn various visuals like the waveform or notes on and off. These buttons can be removed by the researchers if, for a given task, annotators should always see those visuals. Figure 3.4 shows an example performance with the navigation bar, the performance information, and the controls highlighted. The remainder of the page layout for each recorded performance is taken up by the main visualization pane with a smaller zoom pane beneath.

CosmoNote presents the recorded performances to annotators via both audio and visual data, as described in the following subsections. In presenting the performances, CosmoNote has the option to combine these presentations in one of three forms: (1) with audio only, (2) with visuals only, or (3) with both audio and visuals, depending on the nature of the annotation task. Which of these three forms is presented to annotators is recorded per-annotator for later analysis.

3.4.1 Listening to the Performances

When an annotator wants to listen to the audio for a performance they can start with the controls shown in Figure 3.5a, and click play at any time. The audio playback can then be paused or stopped, again, at any time. The pause button replaces the play button while playback is ongoing as shown in Figure 3.5b, reverting to the play button when paused. The stop button is only active during playback and stopping resets the playback position to the beginning of the piece. The three buttons in between the play and stop buttons are for boundary annotations and are described in Section 3.5.1.1.

During playback, annotators can jump to any time point in the audio by clicking on that time position anywhere in the main visualization (vertical position makes no difference). The current audio playback will stop and then immediately start from the new time position. Furthermore, when playback is paused, an annotator can click on any time point and then, when restarted, playback will begin from that time point. CosmoNote has a zoom feature (see Figure 3.8) that

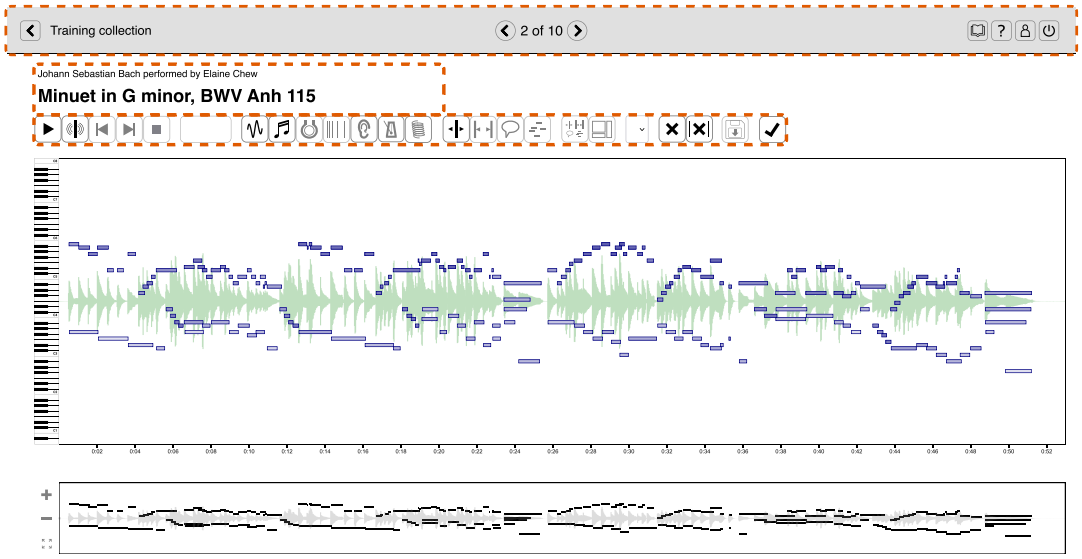
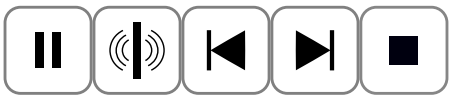


Figure 3.4: A complete view of CosmoNote's interface with three elements highlighted by orange dashed rectangles: (1) the navigation bar (top), (2) the information about the performance (middle), and (3) the controls (bottom).

(a) The CosmoNote audio controls *before* playback has started.



(b) The CosmoNote audio controls *after* playback has started.

Figure 3.5: A close-up view of the CosmoNote audio controls.

works for both visuals and audio. When a piece is zoomed in to a particular time range and an annotator hits the play button, only the corresponding time range in the audio will be played.

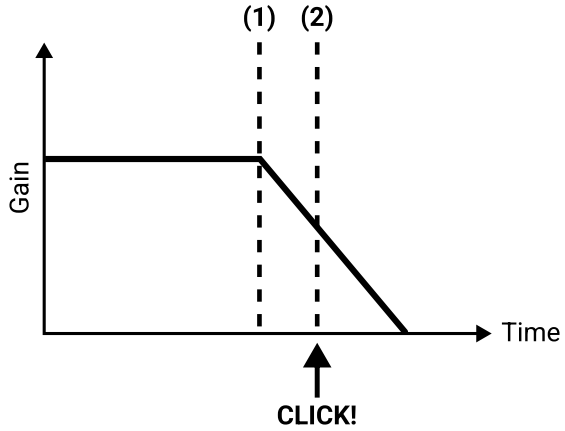
To show playback progress and pause time points, when playback begins, a green vertical line appears over the visuals as a play head. To keep the position of the play head synchronized with the audio playback, the play head is animated using `requestAnimationFrame()`[13], which enables animations to run at the frame rate currently used by the browser. On each animation frame request, a check is made on the amount of time that has passed in the audio file by subtracting the start time of the file from the current time (both obtained via the `AudioContext` object). The time passed is then converted into a position within the main note visualization. Using this technique, based on an idea by Wilson (Wilson, 2013), the play head is always in sync with the audio playback and the time increments for the movement of the play head are small enough to ensure smooth animation even for smaller audio files.

For audio file playback, CosmoNote uses the `AudioBufferSourceNode` interface, which will be abbreviated as `ABSN` from now on. One quirk of this interface is that playback on a particular

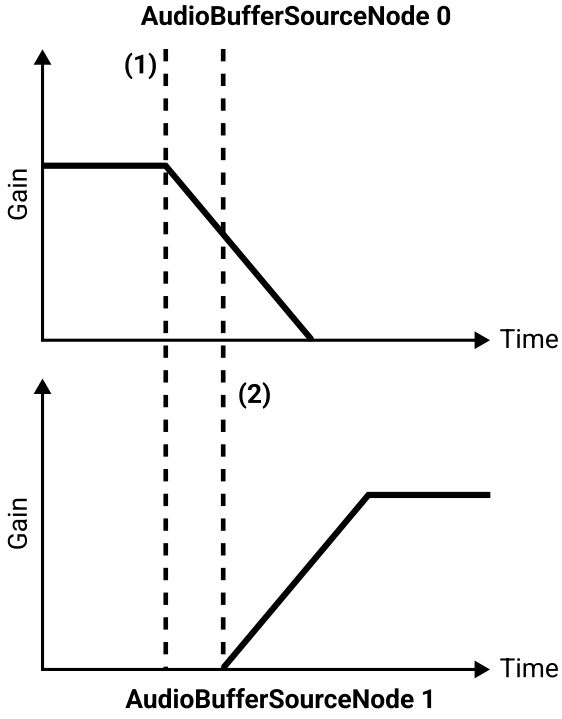
[13] https://developer.mozilla.org/en-US/docs/Web/API/window/requestAnimationFrame

file can only be started once[14]. This is not a problem if the audio file is played through and only once. However, it is sometimes necessary to stop playback on the file and to restart it again. For example, CosmoNote allows an annotator to click on the timeline during playback, causing the audio to stop and restart at that point in time. This feature requires that the `stop()` function be called on an `ABSN`, which in turn requires that a reference be kept to that `ABSN`. However, because of the quirk in that interface, to start the audio file playback again, a new `ABSN` needs to be created. When the existing source node reference (required for stopping) is pointed to a new `ABSN`, a click would normally be heard because the previous `ABSN` is now out of scope and and cannot continue to play during a gain ramp down created specifically to avoid clicks. This problem is shown graphically in Figure 3.6a.

To solve this clicking problem, CosmoNote uses an array of `ABSN` objects to separate a currently playing object `ABSN`$[n]$ that is being stopped from a new `ABSN`$[n+1]$ object that is being started. This allows for a currently playing `ABSN`$[n]$ that is ramping down after a stop to be kept in scope until the ramping down is completed. A new `ABSN`$[n+1]$ can be assigned, started, and ramped up before the ramp down of the stopping `ABSN`$[n]$ is complete. `ABSN`$[n]$ can then go out of scope while `ABSN`$[n+1]$ ramps up. The latest addition to the `ABSN` array is always the one that is stopped, immediately followed by the addition of a newly created `ABSN` that is started. Whenever playback needs to be stopped, then started again, the process is repeated. This solution is shown graphically in Figure 3.6b.



(a) A click happens when **(1)** the `stop()` function is called on an `ABSN` and then **(2)** the same reference is pointed to a new `ABSN` as required to restart playback.



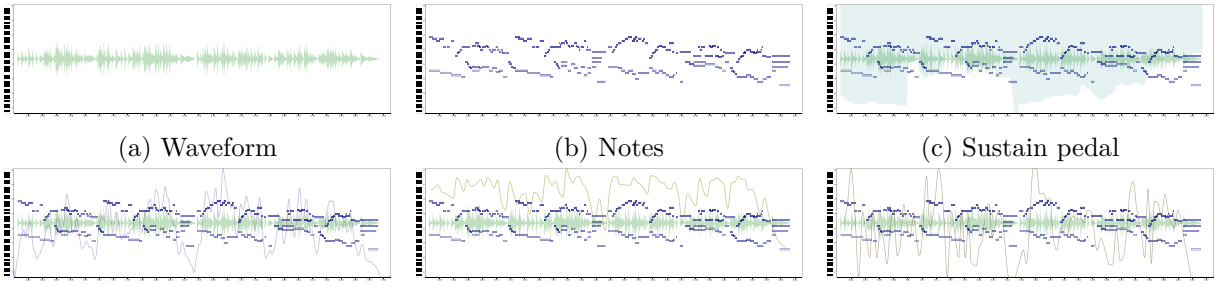
(b) An array holds two `ABSN` objects and then **(1)** the `stop()` function is called on `ABSN` 0 and **(2)** a new `ABSN`, 1, is created and then started. This avoids any clicks by creating a new `ABSN` instead of reassigning an existing reference to one.

Figure 3.6: `AudioBufferSourceNode` (`ABSN`) playback process.

[14] https://developer.mozilla.org/en-US/docs/Web/API/AudioBufferSourceNode

3.4.2 Visualizing the Performances

CosmoNote shows musical data in its main visual pane via waveforms, note and pedal data, and curves based on extracted feature data including loudness, tempo, and harmonic tension (all described in Section 3.3). Each type of visual data is a distinct layer in the overall visualization. With all the data layers displayed at the same time, the visualization can become dense and difficult to read, so transparency is used to varying degrees in all the layers. To further mitigate the problem of too much visual data, each information layer can be turned on and off individually. This also allows for a focus on particular data types as shown for different layers in Figure 3.7. For all the layers, the design of the visuals emphasizes general trends in each kind of data since annotators are tasked with annotating expressive structures in the performance rather than finding exact values for any given data type.



(a) Waveform (b) Notes (c) Sustain pedal

(d) Loudness (e) Tempo (f) Harmonic tension

Figure 3.7: CosmoNote data visualization layers: (a) waveform, (b) notes, (c) sustain pedal, (d) loudness (in sones), (e) tempo (in beats per minute), and (f) harmonic tension (the rate of chord changes). The first two layers, waveform and notes, are shown as backdrop for the rest are and maintained throughout the other examples.

Beneath the main visual pane is a zoom pane in which annotators can select a time range that zooms the main visualization panel to the selected time range, allowing for a more detailed look at the notes and other data layers for that part of the performance. When zooming, all the visible data layers in the main pane are zoomed. Zooming into specified time ranges not only zooms the visuals but the audio as well, allowing annotators to listen to the corresponding time range in the audio for a given performance. Figure 3.8 shows a time range selection (the gray area) in the zoom pane while showing the corresponding notes in that time range in the main visualization panel. The zoom pane shows only the waveform and the notes since it is meant to provide a basic context map of the full performance while the main visualization pane is zoomed.

3.4.2.1 Audio Data

The lowest layer for CosmoNote visuals is a basic waveform, plotted in pale green, taken from the audio file data from a given recorded performance; currently, only the information from the left channel in a stereo audio file is drawn. The waveform displays a classic indication of the intensity (vertical axis) and duration (horizontal axis) of sounds in the music. A function reduces the amount of audio data to match the pixel width of the main and zoom panes, allowing for efficient waveform drawing, especially with the zoom function. When zoomed, the waveform behaves as expected, showing just the waveform from the selected time range. The waveform layer is shown in Figure 3.7a.

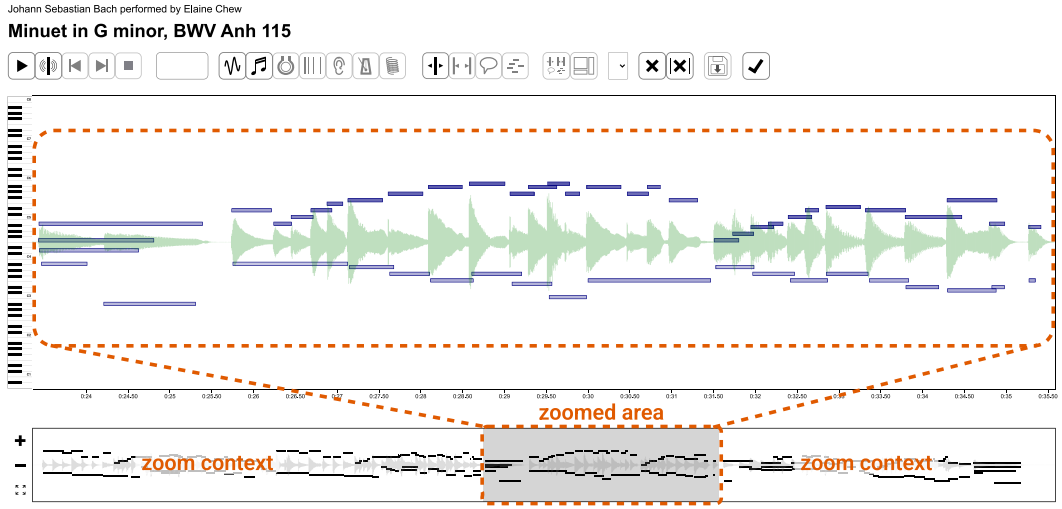


Figure 3.8: CosmoNote zoomed to show a short time range. The smaller visualization pane below shows the context as well as the zoomed notes (inside the gray square). The controls on the left, from top to bottom, increase the zoom, decrease it, and reset it.

3.4.2.2 Note and Pedal Data

The note visual layer includes note data derived from MIDI recordings with the note value on the vertical axis (from lowest to highest note on the piano) and the length of each note in seconds on the horizontal (time) axis. Since the actual values of the notes played are not needed for the annotation task, the MIDI note value is not shown. However, since all the performance data in CosmoNote comes from piano recordings, a piano graphic is shown along the left vertical axis that allows annotators to loosely infer the values of the notes. The MIDI velocity, associated with approximate loudness, of each note is depicted via the note's transparency with more opaque notes being louder and more transparent notes being quieter. Note data is shown in Figure 3.7b.

Sustain, soft, and sostenuto pedal data are taken from the MIDI recordings and shown as area curves, each with a different color. The displacement of each pedal is shown on the vertical axis, with the distance from the top of the graph representing the pedal displacement distance from the rest position. The closer the line is to the horizontal (time) axis, the more the pedal is depressed. With this orientation, it makes sense to show the data as an area graphic in which the area shows both that the pedal is being used and how much it is depressed at a glance. Figure 3.7c shows sustain pedal data as the teal-colored area curve.

3.4.2.3 Feature Data

Feature data, described in Section 3.3.3, is shown as curves layered over the waveform and note layers. Figure 3.7 shows the various information layers with the note layer as backdrop. The loudness curve is the purple curve in Figure 3.7d. Tempo (in beats per minute) is given as the olive curve in Figure 3.7e. The tension curve is the brown curve in Figure 3.7f. More specifically, Figure 3.7f shows cloud momentum (changing dissonance) though all three dimensions of tension can be shown with each dimension being assigned its own color.

Since feature data is already synchronized with the music, its associated time is shown on the horizontal axis of the main visualization pane. The vertical axis, as explained above, is exclusively dedicated to pitch. Amplitude information of feature data is drawn scaled, from

minimum to maximum value, into the visualization pane. While it would be convenient to have a depiction of values for feature data on the vertical axis, the interface would get visually cluttered and users could be confused about which curve that information belongs to. In practice, scaled visuals are enough to understand the evolution of the curves, so the current display is maintained while other options are being considered (see Section 3.7).

3.4.2.4 Instants Data

Instants are a type of annotation for pre-marking areas of interest, like score structures or other landmarks, for annotators. Since instants are a part of the recorded performance data, in contrast to the annotations created by annotators themselves (which are described next in Section 3.5), they cannot be edited or deleted. They mark specific times in a piece visually with a vertical line and a text label that appears when annotators mouse over the line. Figure 3.9 shows a series of instants (showing only one label) for a performance of the Minuet in G minor by Elaine Chew. In this case, the instants describe dynamic and tempo markings from the score labelled, in order, mezzo forte, diminuendo, forte, diminuendo, piano, mezzo forte, piano, crescendo, forte, and poco ritardando.

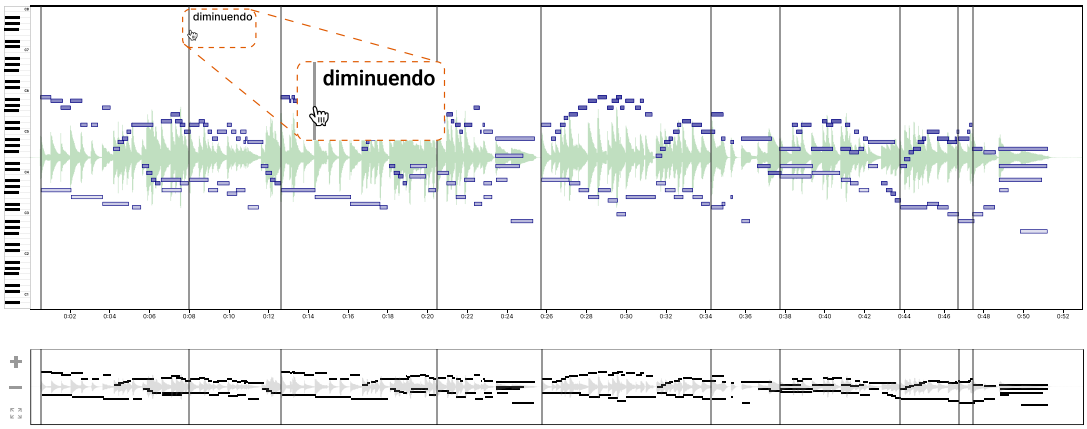


Figure 3.9: A set of instants, shown as gray vertical lines, for a performance. Labels are displayed when the mouse cursor hovers over the instants as with the 'diminuendo' label shown here.

The lines representing instants are shown in both the main visual pane and in the zoom pane below it (as seen in the zoom pane at the bottom of Figure 3.9). In the zoom pane, the instants delineate a series of clickable boxes that represent time ranges. By double-clicking in one of the boxes in the zoom pane, the main visual pane will zoom exactly to that time range, giving annotators a convenient way to zoom into the boxes delineated by instants.

3.4.2.5 Supplementary Data

In CosmoNote, supplementary data is defined as data that is not directly computed from a recording of a performance or even the score but may still be related to the performance. Essentially any time-based data can be included as supplementary data though that data is most useful if synchronized with the audio or note data. Supplementary data is presented as curves with customizable appearances and any number of curves can be added. For example, CosmoNote can incorporate related data such as physiological data from performers or listeners.

Figure 3.10 shows two electrocardiographic (ECG) signals recorded during a piano performance[15] with the red signal being the player and the blue signal the listener.

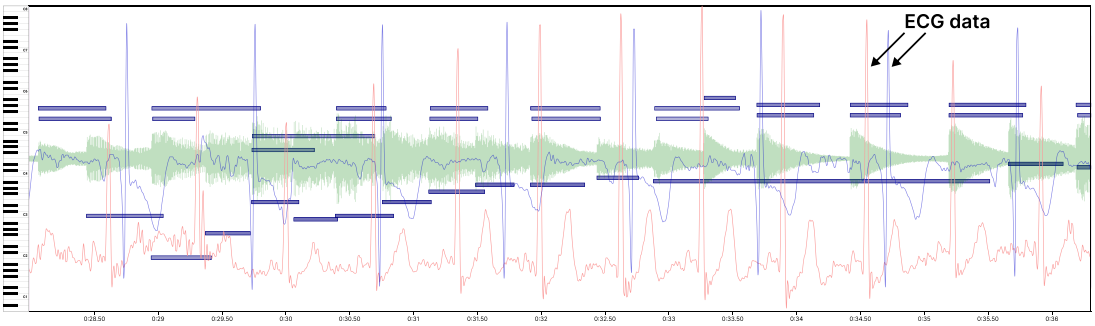


Figure 3.10: Supplementary data in the form of two electrocardiographic (ECG) signals recorded synchronously with the performance, by Elaine Chew, of Prokofiev's Juliet As A Young Girl excerpt from *Romeo And Juliet*, Op. 75.

3.5 Annotating The Performances

Once annotators have listened to and studied the visuals of the performance data, the next step in the workflow, at the heart of CosmoNote, is the annotation of the perceived structures in the performances.

3.5.1 Annotation Types

CosmoNote features four types of annotations: boundaries, regions, comments, and note groups. To create one of the annotation types, annotators select the corresponding button on the toolbar above the main visualization pane (shown in Figure 3.4) to set the annotation type mode. Once an annotation type mode is selected, annotators can place as many of that type of annotation as desired either by using mouse clicks or by using the keyboard. For each type of annotation, annotators can create a label with custom text. Labels are created for annotations by accessing an inspector pane (see Figure 3.11 for example labels for regions) and typing in the desired text for each annotation. For boundaries, regions, or comments, annotators can, when that annotation type mode is selected, hover the mouse over the annotation to see the label. Figure 3.12 shows examples of each the four annotation types.

3.5.1.1 Boundaries

Boundaries represent time points that separate the performed music into segments of coherent chunks of music, e.g. a complete musical idea or a musical thought. Boundaries communicated through performance not only separate a larger piece of music into smaller, meaningful units, they also help listeners make sense of the music. Annotators can place any number of boundaries for a given recorded performance, and they can be placed (or removed) at any time. Once placed, boundaries can be moved in time by clicking on the lines and dragging them or, when selected, by clicking and dragging the arrows that appear at the top of the boundary (as shown in Figure 3.12b). Boundary labels are created or edited via the annotation inspector like the

[15] Go to https://doi.org/10.6084/m9.figshare.23732637.v1 to hear the performances in Figure 3.10.

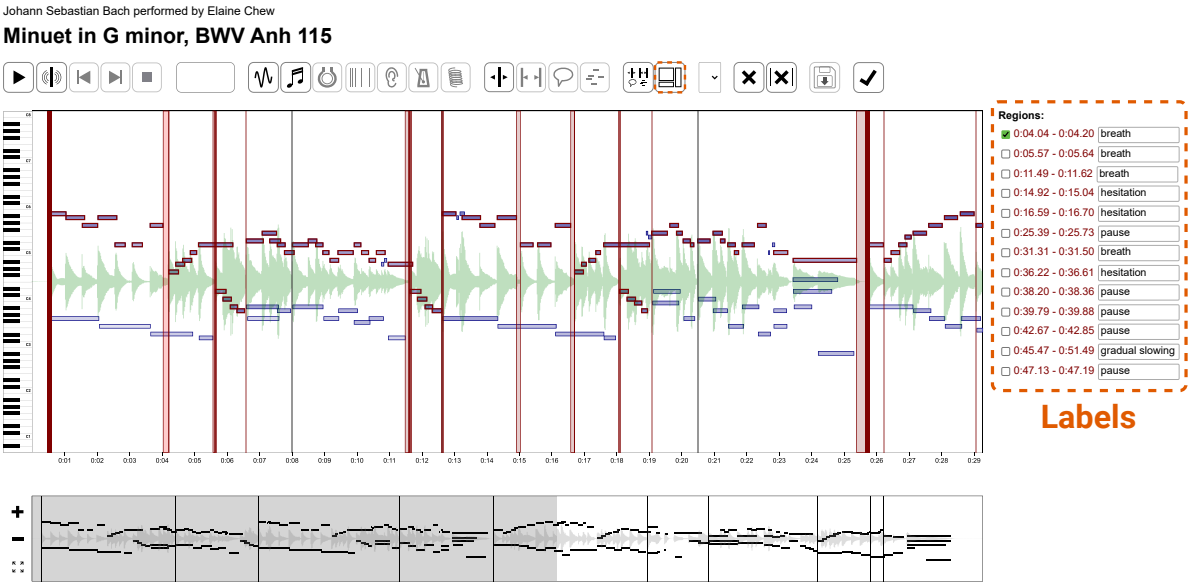


Figure 3.11: Various annotations with the annotation inspector showing labels for regions. All annotation labels appear to the right of the main pane when the inspector is activated. The annotation's starting time (and ending time for regions) is shown to the left of each label.

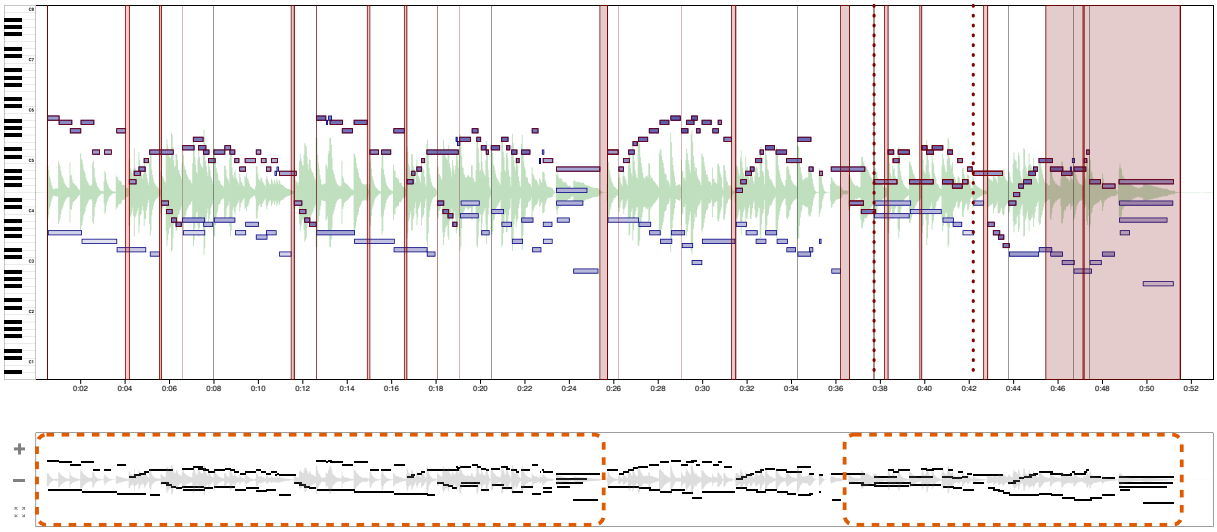
region inspector shown in Figure 3.11. The red vertical lines in Figure 3.12b are examples of boundaries.

Annotators can place boundaries of four different levels that represent segmentation of musical ideas at different time scales with the exact definition of each level depending on the specific annotation task. The boundary levels are indicated visually via the thickness of the boundary and by transparency with opacity increasing as the level increases as shown in Figure 3.12b. During audio playback, boundaries can be placed at the current location of the play head by selecting one of the numbers from 1 to 4 on the keyboard, with the numbers corresponding to the four levels.

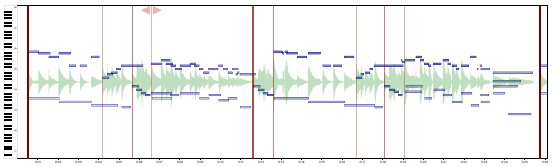
When an annotator listens to a performance after boundaries have been placed, they can hear (if enabled) a small woodblock sound effect when the play head reaches a boundary. The gain of the sound effect is louder for each of the four boundary levels with level one being fairly quiet and level four being the loudest. When the audio is playing, an annotator can use the skip forward or backward buttons to skip the playback to the next boundary or back the previous boundary, allowing annotators to hear the results of their boundary placement (see Figure 3.5).

3.5.1.2 Regions

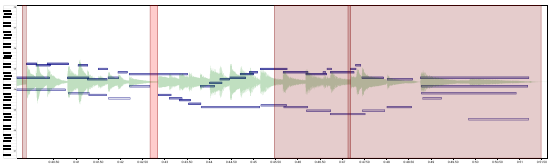
Regions delineate entire sections or areas of interest in a performance and, though they perform a function similar to boundaries, they instead encompass all the notes between two boundaries rather than simply denoting the boundaries themselves. They can be used, for example, to mark transitions or the lead up to a tipping point (see Section 4.3.2). Any number of regions can be placed, and they can be moved, resized, or deleted after placement. Regions can overlap each other, enabling the annotation of, for example, overlapping phrases where one phrase begins before the other ends. Since regions use transparency, when regions do overlap, the overlapping area will be darker, showing the overlap clearly at a glance. Regions are shown as the semi-transparent red squares in Figure 3.12c and region labels are shown in Figure 3.11.



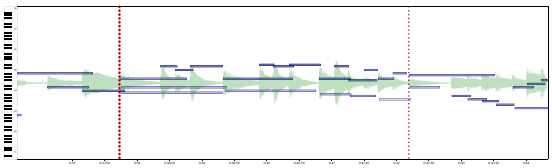
(a) The fully annotated Minuet in G minor



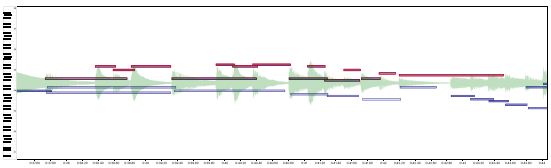
(b) Boundaries (fourth boundary selected)



(c) Regions (second region selected)



(d) Two comments (first comment selected)



(e) Two note groups (top group selected)

Figure 3.12: The four types of annotations with (a) showing a complete piece with every annotation type. The two dashed areas in the zoom pane are the selected time ranges for boundaries and regions shown in (b) and (c), respectively. Comments (d) and groups (e) are further zoomed to better show their placement.

3.5.1.3 Comments

Comments enable annotators to mark elements of interest and to write custom text about them. Though similar to the labels associated with the other annotation types, comments provide a means to point at something of interest that is not captured by the other types. Comments are presented as dotted lines to distinguish them from boundaries as shown in Figure 3.12d. During audio playback, comments can also be placed at the current location of the play head by pressing c or C on their keyboard. The text can be edited from the annotation inspector, in comment mode, similar to what is shown in Figure 3.11.

3.5.1.4 Note Groups

Note groups provide a way for annotators to select and highlight one or more notes of interest. Any number of notes can be selected to create a group and any number of groups can be created. To create a group, annotators create a selection square (or rectangle) with the mouse that approximately encompasses the notes that they want to select for their group. Once a

group selection is created, annotators can add or remove notes from the group as either (1) another selection square (by holding down the Shift key), or (2) individually (by clicking on each note). This is especially useful if the selection square, because of its shape, did not exactly create the desired group. As with the other annotations, annotators can create and edit labels for groups by enabling the inspector when in group mode. Figure 3.12e shows two note groups where the two groups are highlighting the upper and lower melodies played by the right hand of the performer.

3.5.2 Conducting Experiments

CosmoNote's tools and visuals are designed to be highly re-configurable for performed music studies. This design choice is an essential part of CosmoNote. The software facilitates music annotation tasks through an accessible interface that is both familiar for music enthusiasts and powerful for advanced users. The interface has a modular construction to promote its use by researchers running different experimental protocols and applying various experimental conditions. For example, the buttons described in Section 3.4, which allow users to show/hide a particular visualization, are themselves a part of a system of a logic structure of boolean variables and dictionaries. To access these advanced functionalities, a researcher working with CosmoNote will be granted access to three main types of documents stored in the database server: pieces, collections, and users. This is how they relate to each other to facilitate the design of an experiment:

Pieces: When a performance is uploaded into the database, a JSON document with a unique ID that contains its properties (name, composer, performer, etc.), feature data (note events, loudness, instants), and a link to the audio file is created.

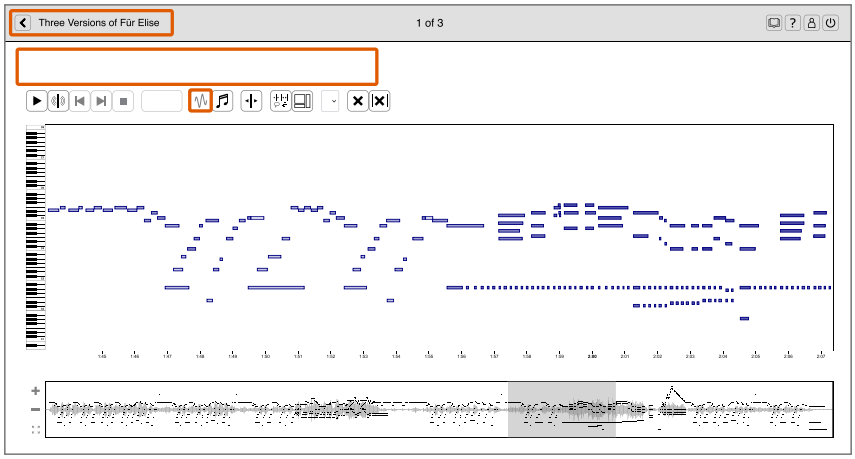
Collections: The set of pieces in given collection, chosen because they share individual properties (e.g., performer, composer) or an overall theme (e.g., music from the romantic era) are stored in a JSON file with a unique ID. Thus, a piece can belong to more than one collection at a time so that participants can navigate through collections and complete annotation tasks that are specific to that collection. The structure of each collection file includes customizable options for: presentation (show/hide piece information), navigation (change piece order and piece selection), controls (show/hide buttons), visualizations (show/hide visual representations), annotation types (allow one or more types), and writing to the database (grant/revoke saving privileges). Pieces in a collection can also be displayed in a fixed order, or can be shuffled for each participant, storing the presentation order for later analysis.

Users: To access CosmoNote, annotators create a user account or are provided with one. User accounts allow annotators to have a personalized experience in which they can, for example, save their annotations between listening sessions. Each user account creates a JSON document that encompasses many associated properties (identification, account setup, among others), and is linked to data files containing all annotations marked by respective users in CosmoNote. For experimental design purposes, a *role* property is associated with each user account. User roles give experimenters the possibility to control precisely what type of access users have. There are three main types of roles: (1) super users, who can see all the content inside the CosmoNote server, which is hidden for everybody else; (2) public users, who can only see active public campaigns (this is the default role assigned when creating an account); and (3) custom users, who will see only a subset of the data that is different from the public campaigns (e.g., a set of pieces for a custom experiment).

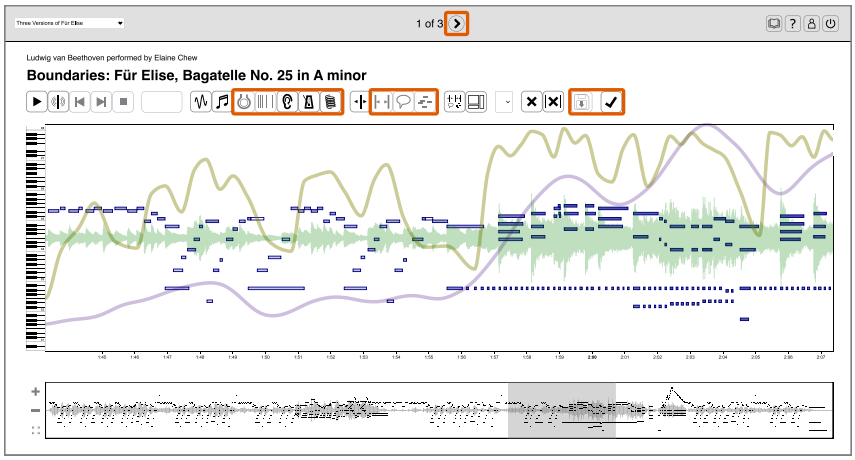
Understanding these documents[16] allows us to extend the scope of studies that are possi-

[16]See Figure A.1, Figure A.2, and Figure A.3 in Appendix A for examples of these JSON files.

ble with the platform. Experimenters can modify the properties in a collection document in creative ways, to require for example a collection that: (1) doesn't display piece titles or performers, (2) shows waveform and loudness visuals but doesn't allow users to hide the loudness data, (3) shuffles the order of the pieces, (4) forces navigation to the next piece when finishing annotations and (5) is visible only for users with the role `experiment-xyz`. A performance example[17], shown on Figure 3.13, emphasizes the differences between two configurations of the CosmoNote's interface. Figure 3.13a is a restricted version of the app: Only two visualizations are possible (waveform –toggled off– and piano-roll), only boundary annotations can be placed, and navigation is forced (users cannot advance to the next piece until they finish annotating the current piece). In contrast, Figure 3.13b shows a complete version of the app: All visualizations, annotation types and navigation options are available to the user. These examples are just a small selection of the configurations available in the CosmoNote interface. The experiments shown in Chapters 5 and 6 show how we can take full advantage of these modes.



(a) Restricted



(b) Not restricted

Figure 3.13: Excerpt from the C section of Ludwig van Beethoven's Bagatelle No. 25 in A minor - WoO 59. Variations of CosmoNote's interface in two experimental cases: (a) The user has restricted access, few options are enabled vs. (b) The user has full access, all options enabled. Orange boxes highlight the differences.

[17] Go to https://doi.org/10.6084/m9.figshare.23899527.v1 to hear the performance in Figure 3.13.

3.5.3 Annotation Tasks

Once annotators have understood the different annotation types, they can move on the annotation task itself. Annotators are provided with instructions for specific annotation tasks, accessible via a link at the top of the page, which can change depending on the nature of the collection or study. While the task instructions are variable, they are essential for helping annotators to focus on the significant elements of any given task. The difficulty though is to balance the task instructions such that they are neither too specific nor too general. A thorough description of the annotation task adopted in many studies is given in Section 4.3. The two pilot studies described below represent the first of a set of experiments that are reported in this document. Pilot studies 1 and 2 guide the reader through a progression steps that is common for all studies, starting from the presentation of a set of instructions to the analysis of the data and the user's evaluation of CosmoNote.

3.5.3.1 Pilot Study 1 Task

For the first pilot study task, eight music and audio researchers were presented a series of excerpts from Chopin's Ballade No. 2 as performed by Elaine Chew[18]. The ballade was split into eight musically coherent excerpts and presented to annotators in a shuffled order. For the annotation task, participants were asked to place boundaries as communicated in the performed music and indicate the strength of each boundary (levels 1 through 4) with the levels defined as:

1. **Motives:** the smallest indivisible succession of notes that may be delineated by accents
2. **Sub-phrases:** parts of a phrase, e.g. antecedent or consequent phrases
3. **Phrases:** complete self-contained musical statements
4. **Sections:** a major structural unit comprising a complete musical idea

To further help annotators with the task, the following suggestion about boundaries was included with the instructions:

> Performers may mark boundaries using pauses, stress, or contrast. For example, accents could mark the beginnings of groups of notes, pauses can separate musical ideas, phrases may be expressed by increasing then decreasing tempo and/or loudness, a change of timbre and loudness may mark the beginning of a new section.

The interface presented annotators with visual layers from note, pedal, loudness, tempo, and harmonic tension data. Participants were allowed to toggle any of these visualizations on and off at any time but were instructed to "let their ear be their main guide".

3.5.3.2 Pilot Study 2 Task

For the second pilot study task, we presented seven music and audio researchers with a single short music excerpt, Variation XXXII[19] in Beethoven's 32 Variations in C minor, WoO 80. In contrast to the free form nature of the task from the first pilot study, for this task, the annotators were asked, in the task instructions, to annotate **segmentation** and **prominence** as

[18] Go to https://doi.org/10.6084/m9.figshare.c.6755307.v1 to hear the excerpts of the performance in Figure 3.15.

[19] Go to https://doi.org/10.6084/m9.figshare.23732325.v1 to hear the performance in Figures 3.16, 3.17, and 3.18

paraphrased (for stylistic consistency) below. Both definitions are followed by lists of examples, also provided in the task instructions (details about these concepts are developed in Chapter 4).

Segmentation is the process of dividing something, in this case music, into meaningful units.

- Boundaries: time points that separate a music stream into segments representing meaningful chunks of music e.g., a musical idea or a musical thought. Boundaries not only separate a larger piece of music into smaller, coherent units, they also help listeners make sense of the music. There is a boundary annotation type in CosmoNote, with four levels of boundaries, defined from 1 (weakest) to 4 (strongest).
- Transitions: musical passages that set up a change that is coming in the music or that blur changes in the music by moving slowly through them, linking musical ideas. Transitions may be annotated using regions.
- Pauses: segments of time that add space between two adjacent structures. Executed by the performer by lingering on notes or using silence. Pauses may be annotated using regions.

Prominence characterizes an emphasis drawn towards a certain part of a whole in the music.

- Stress: an emphasis of a particular element to make it more prominent than those around it. Stress may be indicated by a combination of performer actions like an increase in sound intensity, duration, or a change in timbre. It may be marked using boundaries or note groups.
- Melodic salience: a special case of prominence dedicated to a sequence of notes that may be recognized by an increase in loudness and duration, or a variation in the timbre of the melody notes. Melodic salience may be marked using note groups.
- Tipping points: moments where musical time is suspended/stretched to a point beyond which a return to the pulse is inevitable. Tipping points may be marked using regions, boundaries or note groups.

3.6 Analyzing The Annotations

The last step in the CosmoNote workflow is the collection and analysis of the annotation data. The data is exported from the CosmoNote database with a custom Python script that extracts all annotation data as JSON[20]. Data for each annotation contains a pseudonymous annotator identifier, the collection, the performance, the annotation type (boundary, region, comment, group), a time (two times for regions), the boundary strength (for boundaries only), a label (if set), a creation timestamp, and an update timestamp (if subsequently edited after creation). In addition, note groups also have the MIDI note number and the start and end times for every note in the group.

To evaluate how citizen scientists use CosmoNote for annotations tasks, we conducted two pilot studies as described in Section 3.5.3. For the first pilot study, we asked annotators to focus

[20]See Figures A.4, A.5, A.6, and A.7 in Appendix A for examples of these JSON files.

strictly on boundaries. For the second, we asked annotators to use all four of the annotation types: boundaries, regions, comments, and note groups. In both pilot studies, we compared the annotations of participants with those of the performer.

3.6.1 Pilot Study 1 Results

For the first pilot study, whose task is described in Section 3.5.3.1, the results were compared to annotations made by the performer that served as a baseline for comparison. We first compared the number of boundaries of each level per excerpt. Figure 3.14 shows a distribution of the number of boundaries (by level) participants placed for each excerpt compared to the performer's annotations of the same excerpts. The x-axis shows the count (the number of boundaries placed) and the y-axis shows results for each of the eight excerpts. Individual box plots represent the boundary distribution of the eight listeners and the red dot represents the performer's. The panels contain the different distributions of the boundary levels from 1 to 4.

The results indicate a convergence for the higher boundary levels, 3 and 4; these levels were overall less used by both the listeners and the performer. Level 2 boundaries were also broadly comparable between them. In contrast, level 1 presented the highest variability in the number of boundaries and the lowest correspondence between how the performer and the listeners placed boundaries on each excerpt. In general, the performer placed fewer boundaries than the average listener for levels 1, 3, and 4.

We also analyzed the location of boundaries for every level over time. An aggregated representation of the placement of boundaries, by level, for all listeners is compared to that of the performer in Figure 3.15. The boundary profile curves by level were obtained using kernel density estimation (KDE) (Silverman, 2017) to indicate the concentration of boundaries marked by many listeners around a given time point (see Section C.2). Boundaries of level 4 occur more often close to score markings, which are often large scale sectional boundaries, as is also the case with some level 3 boundaries. Density profiles for levels 2 and 1, which may correspond to finer subdivisions into phrases or subphrases, are more spread over time and are distinct from score markings. Overall, all listeners marked boundaries close to the performer's annotations, although they sometimes used different boundary levels to demarcate the segmentation. For example, many listeners' level 1 markings can be seen after the score section marked *agitato*, whereas the performer used boundary level 2 annotations on that passage. Thus while listeners concur on the existence of a segmentation boundary, the precise boundary level varied.

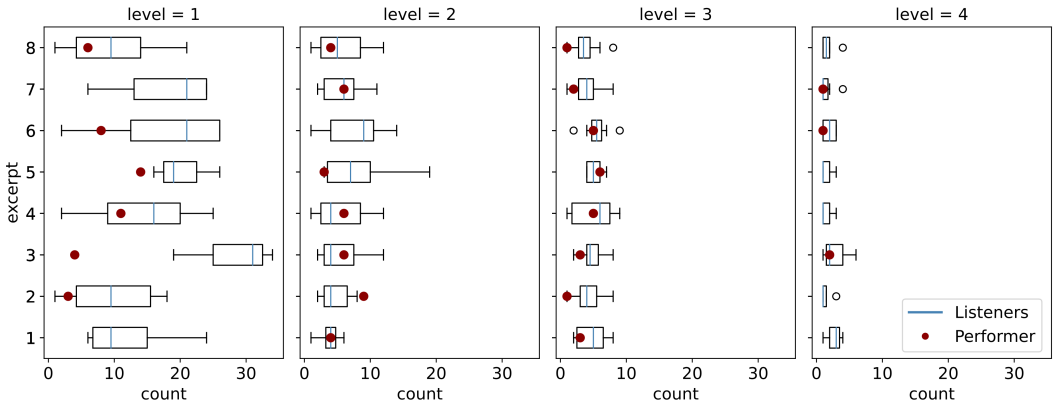


Figure 3.14: Pilot study results showing the number of boundaries of each level placed, per excerpt, by participants compared to the boundaries placed by the performer.

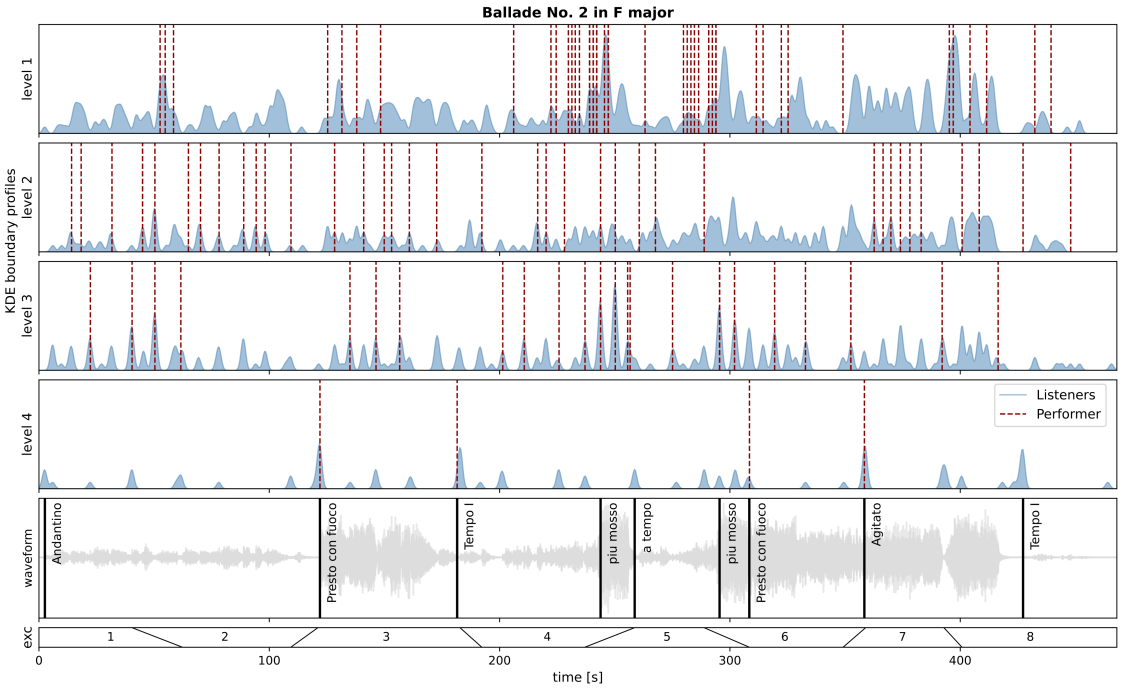


Figure 3.15: Boundary annotation placement profiles comparing the listeners (blue density curves over time) to the performer (red vertical dashed lines). Annotations are split into 4 levels; boundaries were aggregated per excerpt. The bottom panel shows how Chopin's Ballade No. 2 was split into 8 musically coherent excerpts; diagonal lines mark overlapping zones between excerpts, where the widest possible range of each trapezoid represents the time range of the excerpt. Score dynamics and tempo markings are overlaid on the waveform.

3.6.2 Pilot Study 2 Results

For the second pilot study (task described in Section 3.5.3.2) listener's annotations were also compared to those of the performer with, for this study, comparisons of boundaries (by level), regions, and groups for the same piece.

To analyze the boundary annotation results, we applied the same technique as demonstrated in Figure 3.15. Figure 3.16 shows the distribution of boundaries for the second study. Results for this piece were similar to those described in Section 3.6.1 with more boundaries of lower levels (1 and 2) than those of higher levels (3 and 4) for both the listeners and the performer. However, while we observe an overall agreement in the location of these boundaries, we do not observe agreement for their strength level. For example, almost all the performers markings are represented by listeners' annotations, yet the performer did not place any boundary level 4 annotations while listeners did. For the performer, this excerpt was one of 32 variations (plus the preceding theme for a total of 33) and, as such, did not require any level 4 boundary whereas the listener likely viewed the excerpt as a standalone piece, hence the level 4 boundaries.

The region annotations results are shown in Figure 3.17. The results showed that listeners marked regions according to quite different conceptualizations of their meaning. For example, listener 5 placed regions throughout the whole piece, essentially dividing it into chunks, while all other listeners (including the performer) used regions to mark only specific time selections of interest. Region starts and endings, which sometimes correspond to dynamic markings, were shared between most listeners and the performer.

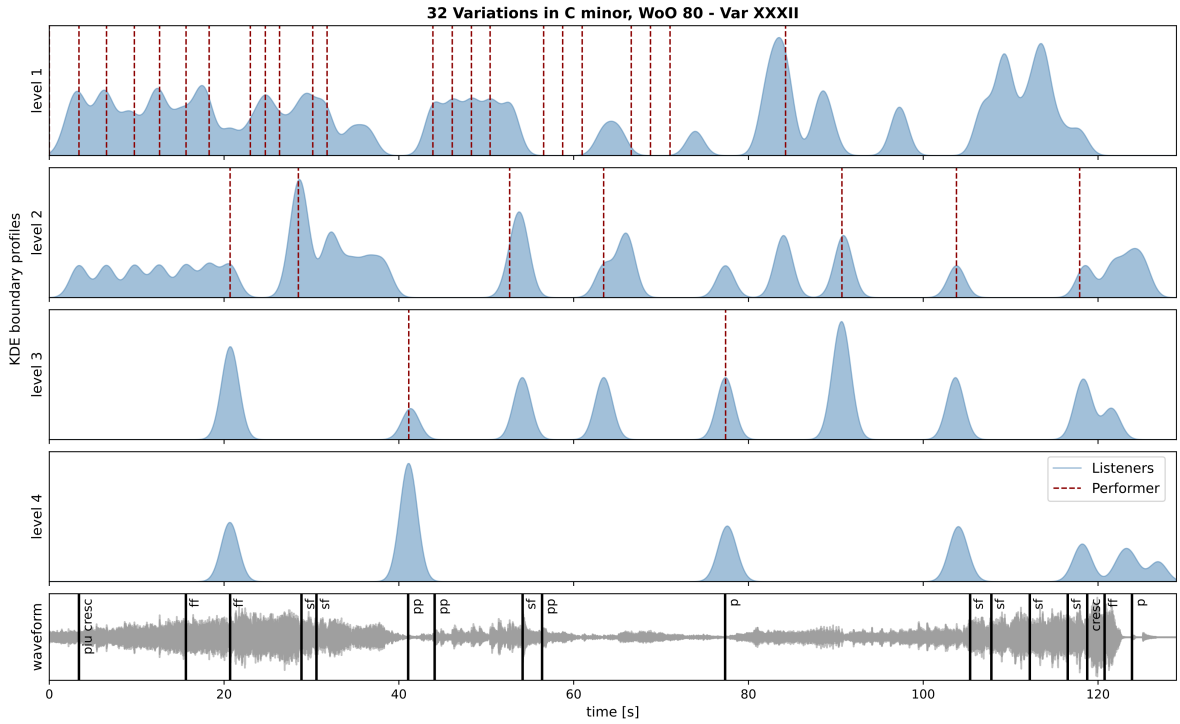


Figure 3.16: Boundary annotation placement profiles comparing the listeners (blue density curves over time) to the performer (red vertical dashed lines). Annotations are split into 4 levels. The bottom panel shows Beethoven's Variation XXXII waveform with score dynamic markings overlaid.

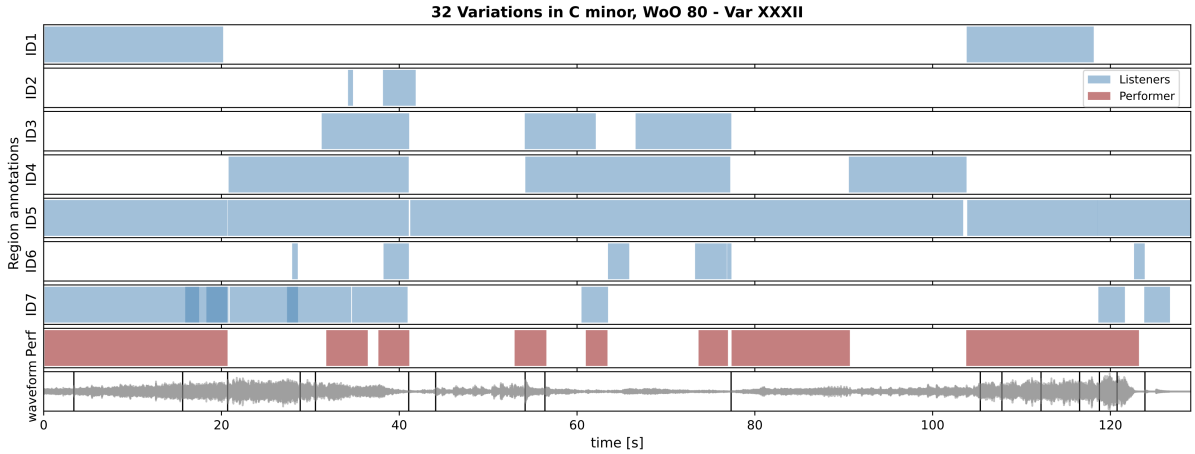


Figure 3.17: Region annotations comparing the listeners (blue patches over time) to the performer (red patches over time). Each row shows data for one listener ID. The bottom panel shows Beethoven's Variation XXXII waveform with score dynamic markings overlaid.

The group annotations are shown in Figure 3.18. Listeners were not required to use note group annotations and only five out of seven listeners did so for this task. As shown in Figure 3.18, all the listeners marked the series of ascending notes of the left-hand melody in the first 20 seconds of the piece while the performer only marked the last note. This and other notes the performer marked tended to be notes made prominent for structural significance, like an outline of the listeners' combined note groups. Common groups were marked by listeners 3

and 6 at around 70-80 seconds and again around 90-110 seconds. These corresponded with a few notes marked by the performer. The performer's note groups that coincided with listeners' note groups are outlined in red in Figure 3.18. The number of groups created (differentiated by the markers in Figure 3.18) is also of interest. While listener 4 marked 12 different groups, listener 3 and the performer created 8, and listener 6 created 5, and listeners 2 and 7 only created 1 group, with corresponding notes. This difference in group numbers and the discrepancy between listeners and the performer are larger than for the other annotation types which may be an indication of how listeners are understanding the concept of note groups and using them in their annotations.

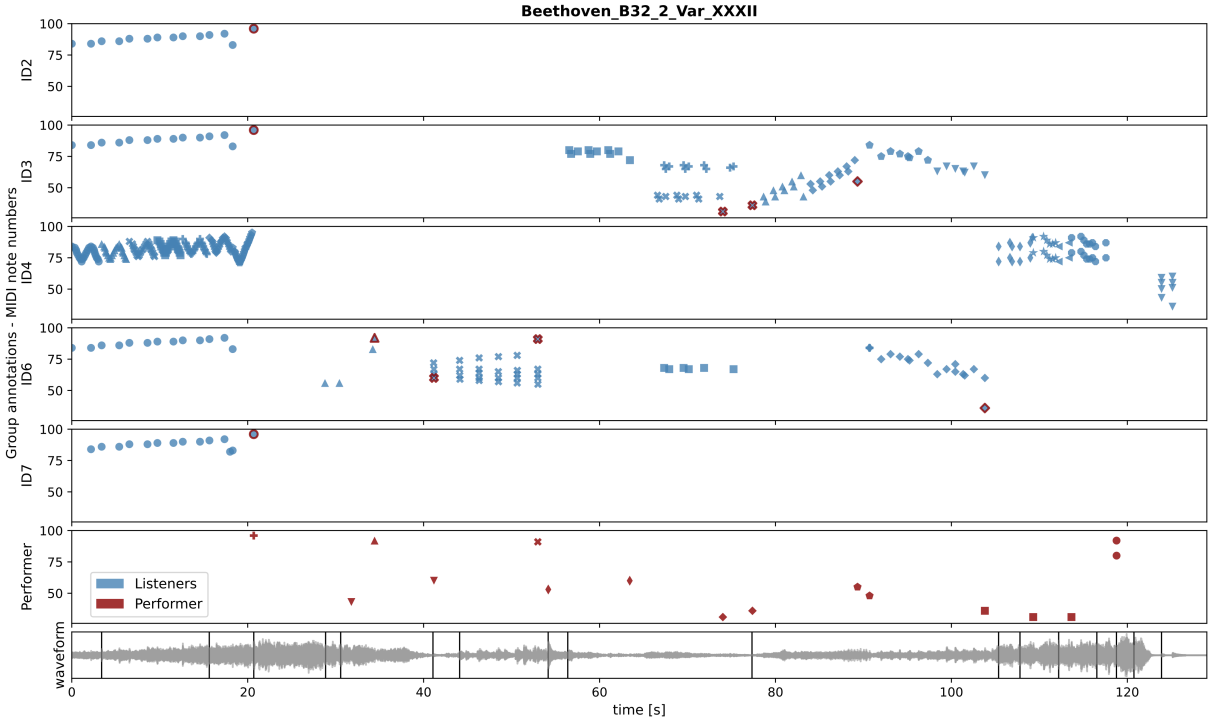


Figure 3.18: Group annotations comparing the listeners (blue markers) to the performer (red markers). Each row shows data for one listener ID; with different markers per group. The bottom panel shows Beethoven's Variation XXXII waveform with score dynamics markings overlaid.

3.7 Conclusions And Future Work

In this chapter, we described CosmoNote, a web-based citizen science annotation tool, and the workflow that it enables. In the course of developing and testing CosmoNote, the following novel features were created:

- Display of discrete note information for the performance visualization that is synchronized with the audio playback;
- Visual display of note velocity via transparency corresponding to loudness;
- Ability to start audio playback from any arbitrary time point with a simple click;
- Ability to zoom into a set of notes and to see and hear just the audio for those notes;

- Option to view information layers like piano pedals, loudness, tempo, and harmonic tension;
- Ability to zoom all data visuals including the waveform, notes, pedal data, loudness, tempo, and harmonic tension;
- Ability to easily zoom into sections of performances delineated by instants;
- Ability to flexibly incorporate other time series data as additional data visualization layers for analysis;
- Ability to create multiple types of annotations including boundaries, regions, groups, and comments; and,
- Ability to skip the audio playback from boundary to boundary.

The version of CosmoNote depicted in this chapter was used for designing the annotation protocol defined in Chapter 4 and running the studies described in Chapter 5 and Chapter 6. However, the following list of features enhances the annotation platform and complements those already mentioned. These ideas have been the object of internal design discussions and user feedback. We do not delve into the technical difficulties of their implementation. Instead, this list is a record for features that could be developed and included in future releases:

- An "undo" functionality. This feature is powerful and customary in a software workflow. However, it requires the ability to track every user action sequentially.
- Collaborative annotations. Allowing users to access the same document at the same time would take full advantage of the web-based nature of CosmoNote. With collaborative annotations, comment annotations would act more as a communicative tool and would thus need a redesign from their current form and function.
- Data upload and download. Currently, all files are stored on a server to which only the developer has access. Allowing users to both upload their own performance data, and download their annotations as a JSON file, would help draw attention from more users to CosmoNote's public campaigns.
- New musical representations. Some less common musical representations –computed from the music data using `cosmodoit`– could be added to the default options cited in this chapter. A non-exhaustive list of new, specialized visualization examples could include spectrograms with different scales (Mel, constant Q), self-similarity matrices, novelty curves, or interonset interval (IOI) curves.
- Advanced "play selection only". The current ability to zoom into a set of notes and to see and hear just the audio for those notes may be improved by the ability to heat only the notes of a selected group. Since the database already knows the properties of these notes (pitch, start time and length), the Web Audio API or other method could be used to synthesize audio from selected notes.
- New mouse and keyboard interactions. Playback, boundaries and comments already have keyboard shortcuts associated to them. Other actions (e.g., skipping to the previous/next boundary, saving, toggling visualizations) could also be mapped to keyboard shortcuts.

- Displaying the y-value of feature data. A small pane on the toolbar already displays the time position (x-value) of the cursor when it moves around the main visualization pane. Showing also the y value corresponding to a given feature could help to have more context without cluttering the visual field.

- Interactive tutorials and gamification. Current how-to guides and instructions are static. A more engaging and enjoyable experience may come from dynamic, animated tutorials that are displayed contextually and prompt specific actions only when they are relevant for the annotation task. In the same way, CosmoNote's capabilities could be progressively unlocked after correctly performing a given number of actions.

- Data aggregation visualizations. New annotators might benefit from the knowledge of previous ones by looking at a comprehensive summary of what other users have annotated before them. We discuss more about data aggregation in further chapters.

- Open collections. While annotators must have credentials (a username and a password) so we can collect and retrieve their annotations, sometimes we just want people to visualize and listen to a particular example. An open collection would function as direct access to a specific collection/piece that doesn't require logging in.

- Improved labels. Current labels allow unlimited text input from annotators. The next iteration of labels may be divided into title (restricted) and description (loose) text fields, plus an elective, optional tag field from a pre-existing list.

CosmoNote currently features the following collections of music performances, fully available to the public: four collections that gather Bach's Goldberg Variations 1955 performance by Glenn Gould, one collection that brings together composers from the Baroque to the Romantic eras, and two collections that collage atrial fibrillation rhythms and music. The list is periodically updated with other thematic campaigns featuring different performance collections and study the efficacy of this method of collecting annotations on musical expression.

In addition to a public release, CosmoNote is used to run private annotation studies accessible to specific users only. This chapter showcased two pilot studies we conducted to evaluate the use of CosmoNote for annotating performances.

1. In the first pilot study, we compared annotations of excerpts from a Chopin performance from study participants with those of the performer. Aggregating data from all participants allowed us to smooth out annotation behaviors of specific participants to concentrate on the global tendencies. Annotations from individual boundary levels provided consistent information showing listeners and the performer focusing on different time scales of the performance. The highest levels coincided more with score markings and matched more closely between participants and performer while performance subtleties were most evident at the lowest boundary levels, and we found that the most significant divergence between participants and performer were at these lowest levels, especially level 1. The placement of level 1 boundaries could be caused by a mismatch in listener's perception of subtle segmentation cues. Pilot study 1 provided the first feedback to help us understand how annotators used boundaries in CosmoNote and has given us clues about how to approach our annotation task instructions for future annotators, especially for marking the more subtle aspects of performances.

2. The second pilot study also compared listener's annotations to those of the performer. In this case however, listeners were free to use all of CosmoNote's annotation types, and

they were tasked with marking segmentation and prominence on a shorter (2 minute) piece. Boundary and region annotations were consistent with those of the performer on most occasions (with varying strength levels for boundaries and longer/shorter selections for regions). On the contrary, group annotations tended to be clustered and were most similar among participants but were different from those of the performer, whose individual markings were more sparse. While these results conveyed information about the structures listeners and performers were perceiving, more importantly for our analysis, they provided clues to help us fine-tune our annotation task instructions.

Chapter 4 details the research approach for collecting data from citizen scientists to establish relationships among the aggregated annotation times and high-level prosodic features, low-level musical features, and acoustic properties. Chapters 5 and 6 will elaborate on how of the platform that is used for experimental research in collecting music annotations. Other studies that make use of CosmoNote's supplementary data functionalities, notably annotating ECG signals synchronized to music, are being conducted at the time of writing.

Ultimately, we will use the citizen science annotation data to develop a vocabulary of expressive musical gestures to establish a common standard for transcribing expressive elements in performed music, and to facilitate the sharing of annotated databases for research and technological development to advance understanding of decision-making in expressive musical performances.

Chapter 4

Representing and Annotating Musical Prosody

In Chapter 3, we presented the CosmoNote web-based annotation platform, referencing its technical modalities, and its capabilities for music representation and annotation. This chapter dives deep into the concept of musical prosody, characterized by the acoustic variations that make music expressive, and explores a novel approach to capturing information about prosodic functions such as segmentation and prominence. This design is applied primarily to recordings of performed solo piano music. We discuss and appraise various annotation strategies, techniques for preventing precision errors, and the long-term objectives of our data collection process. The methodology in this chapter was published in "A Perceiver-Centered Approach for Representing and Annotating Prosodic Functions in Performed Music" (Bedoya, Fyfe, & Chew, 2022b).

4.1 Introduction

Acoustic variations in music can be studied through the concept of musical prosody. The term musical prosody is described by Palmer and Hutchins (2006) as the way acoustic properties are manipulated by performers to be expressive, without changing any existing categorical information (e.g., the pitch and duration categories of the notes in a score remain the same). Musical prosody can be applied to musical concepts such as melodic salience, expressive accents, musical phrases, and pauses. This view of music prosody borrows from speech, in particular, linguistics research that focuses on the phonetic features of speech sounds. In speech, prosody refers to phenomena involving the physical parameters of sound and the functions they serve in linguistic structure. The study of speech prosody is concerned, for example, with how frequency, duration, and intensity are used in stress, rhythm, and intonation (Speer & Blodgett, 2006), which has parallels in music. Music in and of itself can exhibit prosody, just like speech. Our focus here is on the prosody of the music content itself. However, the term musical prosody has also been used to refer to that of the lyrics embedded in a song, which is often related to, but distinct from, that of the music content. For example, Migliore and Obin (2018) defines and analyzes musical prosody as a function of the "syllables of the words and beats of the measure". In a contrasting example, more closely reflective of our approach in this chapter, Bauer (2014) examines speech-like prosodic characteristics embedded in melodic attributes of jazz solos, such as swing and phrasing, manipulated by performers to communicate expressiveness in music. In the remainder of the chapter, we use the definition of musical prosody as given by Palmer and Hutchins and applied to instrumental music.

In Palmer and Hutchins (2006), musical prosody is described as carrying out four main

functions: (1) segmentation (separating relevant elements in the music), (2) prominence (highlighting important events), (3) coordination (communicating with other performers while playing together), and (4) emotional response (reactions that the audience experiences when listening to the performance). We concentrate on the first two functions of musical prosody: segmentation and prominence. It should be pointed out that although segmentation and prominence are two distinct functions of musical prosody, they are not mutually exclusive. For instance, prominence is employed extensively by performers with the goal of highlighting segmentation, and by segmenting the music, changes in it are made more prominent. For the purpose of this thesis, we present applications of our method to solo instrumental piano music, so we will not include studies of coordination. Emotional response is outside the scope of this chapter, although our method could also be used to study and explain emotional reactions to music.

The definition of musical prosody as a topic is generally centered on studying how specific musical constructs, with or without considering their acoustic correlates, serve a given prosodic function. Palmer and Hutchins (2006) report how tempo, intensity and pitch modulations indicate the hierarchy of phrases, how articulations mark metrically important events, and how a tone duration can be mistaken when placed in an unexpected position in a phrase, to name just a few examples. This can be viewed as a bottom-up approach to musical prosody. Keeping the definition of musical prosody, we propose to explore its meanings differently, from a top-down perspective. That is, using the functions of segmentation and prominence as a starting point, given these constructs, we seek to understand how they are created. Ultimately, this will allow us to model the relationships between prosodic functions and the acoustic properties that form them, giving us a more complete understanding of these structures in performed music.

In developing our method, we asked two questions: (1) What kind of framework is necessary to represent and annotate segmentation and prominence? and (2) how can we harness human perception to find these structures in performed music?

To answer the first question, we took inspiration from annotation protocols in speech, more specifically, from the ToBI (Tones and Break Indices) annotation standard. ToBI is a system that enables the transcription and annotation of speech prosody based on two concepts: tonal events (pitch accents, boundary tones, and phrase accents) and break indices (cues about phrase segmentation into words). Annotations are created in tiers and have their own labels. For example, the prosodic grouping of words is marked by vertical lines that have five different levels from 0 (most conjoint) to 4 (most disjoint) Beckman and Ayers (1997). Large communities of annotators have created their own versions of ToBI for multiple languages, each one with its own rules based on a specific language's prosodic structure. Similarly, we have created the CosmoNote annotation platform (described in Section 4.2.1), drawing from the ToBI logic, for creating complex performed music structure visualizations and annotations.

To respond to the second question, we employ the citizen science paradigm to devise an experimental protocol aimed at gathering annotations on musical prosody. The umbrella term citizen science, is broadly defined across different fields encompassing many scientific goals. We defer to its characterization by M. M. Haklay et al. (2021) of participatory practices where people, called *citizen scientists*, get involved in research (e.g., collecting and analyzing data) without it being a part of their paid work. Some projects in citizen science encourage participants to adopt larger roles as they learn about the research subject, others foster co-creation of the research goals with members of the community to solve real-world problems Senabre Hidalgo et al. (2021). This type of research is of interest because it allows researchers to reach a diverse population of people (who in our case enjoy music), and who have the potential to be involved in a project beyond the role of data gatherers, for example, by volunteering their thinking and reasoning capacities.

The rest of this chapter is organized as follows: Section 4.2 refers back to the CosmoNote web-

based citizen science platform and other technical details of the protocol; Section 4.3 introduces the annotation method as applied to the prosodic functions of segmentation and prominence. While the specific musical structures that make up these functions are not fully known, we mention some plausible performed music structures within these categories, and recount some strategies for marking these structures systematically. Sections 4.4 and 4.5 discuss the anticipated results from collected data and the implications of answering our research questions for music performance science.

4.2 Materials

In order to study what are the structures that musicians create in performance, and how they could be represented, we need a set of tools to annotate listeners' and performers' perceptions of structures. Our approach is to record people's annotations of performed structures through a computer interface in intuitive and human-friendly ways as close as possible to how performers and listeners might think of these structures. The conveniences afforded by a digital platform that can be exploited include being able to see multiple layers of representations encoding different types of information, the ease of recognizing patterns over small and large time scales, and the ability to automate and scale certain annotation actions. Furthermore, annotations created in software have the advantage of being easily editable, accessible, and shareable.

As previously mentioned (see Section 3.2), in the realm of audio, several well-known computational tools for annotating, representing, and analyzing time-based media offer ways to markup structures in visual representations of audio descriptors, such as waveform, spectrogram, fundamental frequency, spectral centroid, or self-similarity matrix. For example, linguists use Praat (Boersma & van Heuven, 2001) and ELAN (Brugman & Russel, 2004) to analyze, transcribe and annotate speech and video; musicologists use Sonic Visualiser (Cannam et al., 2010) and its vamp plugins, Eanalyse (Couprie, 2012), and Telemeta (Fillon et al., 2014) to study, analyze, and mark up electroacoustic music with or without collaborative web access.

4.2.1 The CosmoNote Annotation Tool

To present our performance data to participants, we load the data into CosmoNote. The platform was described in detail in Chapter 3. Nonetheless, it is worth reexamining some of its capabilities in the context of this particular protocol.

4.2.1.1 Annotations

Participants mark performance structures using four annotation types (see Section 3.5.1 for more details): boundaries of varying strengths (numbered 1 through 4), regions, note groups, and comments. Each of these constructs can be assigned their own labels and may be combined to denote a given performed structure. Annotations are displayed in shades of red on top of other visualization layers and can be toggled on and off independently (see Section 3.4.2). Although allowing custom color coding for annotations could be useful, the visualization layers already use different colors. We thus decided to represent all annotations with red hues for a high visual contrast against other data. Figures 4.1A-E show the entire CosmoNote interface including its elements and an annotated performance example[1]. Visual layers (waveform, piano roll, loudness, and tempo) and annotations (boundaries, a region, and group notes) are superimposed. The four annotation types (see Figure 4.2) operate in conjunction, as follows:

[1] Go to https://doi.org/10.6084/m9.figshare.23732667.v1 to hear the performance in Figure 4.1

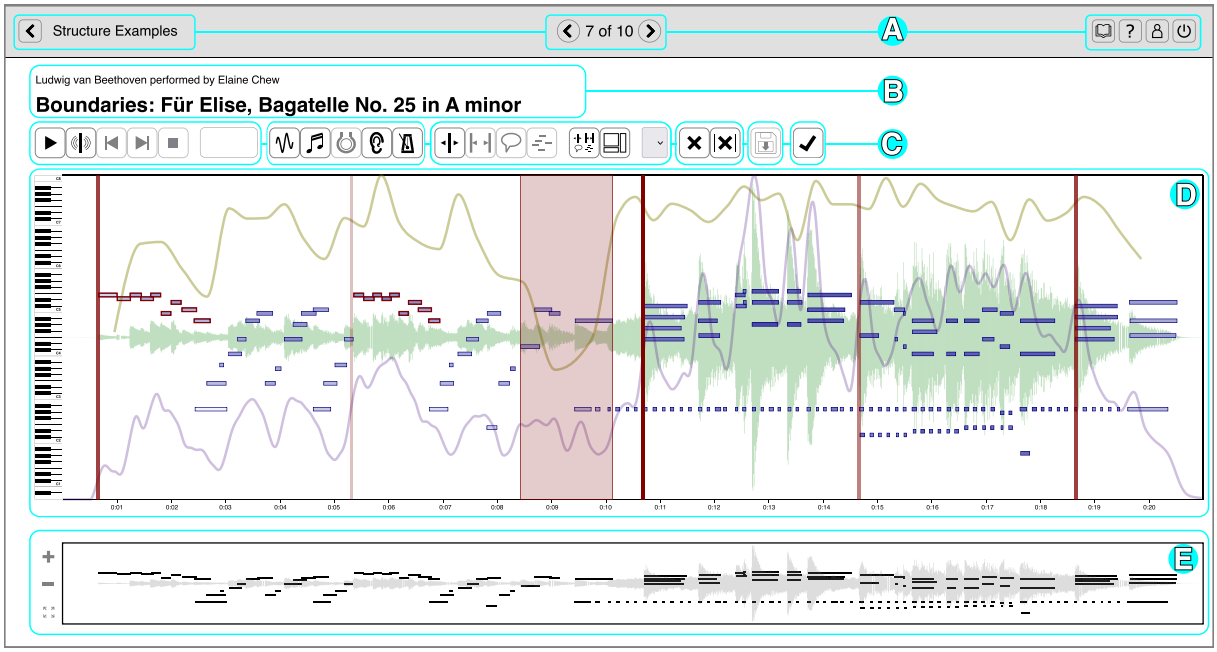


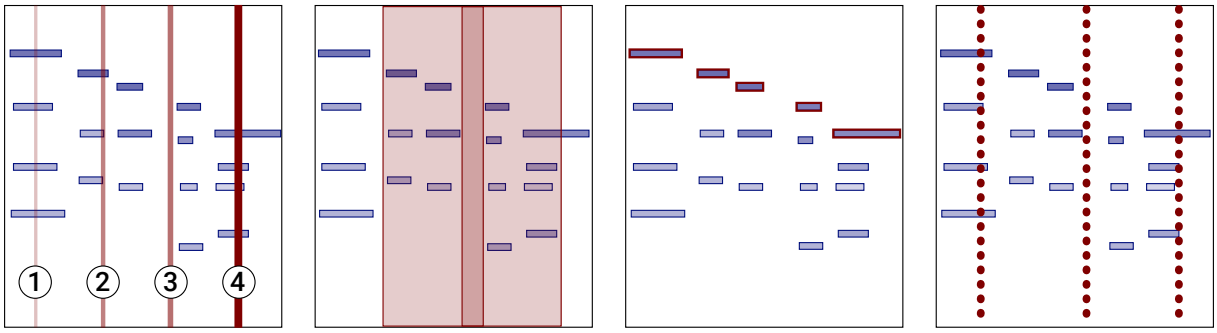
Figure 4.1: The main CosmoNote interface showing an annotated excerpt of Beethoven's Für Elise. The cyan rectangles highlight the different elements: (A) Navigation bar for the collection and account options; (B) Names of the piece, performer, composer; (C) Audio controls, annotation types, and other control buttons; (D) Visualization pane – annotations are displayed in shades of red on top of the waveform (pale green), piano roll (shades of blue), loudness (mauve), tempo (olive), and harmonic tension (shades of brown, not shown); and, (E) Contextual zoom.

Boundaries (3.5.1.1) were primarily designed to mark segmentation, so they are drawn as vertical lines that span the whole visualization pane height. Their increasing strength levels (from 1 to 4) are displayed as a function of the line's thickness and its transparency to emphasize their importance segmentating the music – see Figure 4.2a. The choice of 4 boundary strengths was made to provide sufficient granularity without overwhelming the annotators with too many options. Optionally allowing boundary placement using the keyboard while the sound is playing is essential for annotating with uninterrupted playback.

Regions (3.5.1.2) are multi-functional annotation types that highlight all the elements inside a temporal selection with defined start and end times. Regions are drawn as semi-transparent red rectangles that span the whole visualization pane height – see Figure 4.2b. Allowing one region to begin before the other ends and visualizing the selection superposition is needed, for instance, to mark nested regions or overlapping transitions in a polyphonic composition.

Note groups (3.5.1.4) are meant to single out individual salient notes, or groups of notes, that are meaningful in a given segment or prominent structure like a tipping point. Individual notes can be shared between different note groups, which may be helpful when marking a salient melody and a chord containing one of its notes. Note groups are drawn as red rectangles on top of the normal blue rectangles that represent the notes of the piano roll visualization – see Figure 4.2c. Thus, the piano roll visualization must be available and visible to use note group annotations.

Comments (3.5.1.3) are drawn as dotted lines that span the whole height of the visualization pane – see Figure 4.2d. They may be used to annotate stress without assigning a strength level or make an observation about the music for the researchers or other participants.



(a) Boundaries. (b) Regions. (c) Note groups. (d) Comments.

Figure 4.2: CosmoNote's four annotation types (shades of red) placed over the same sequence of notes (blue rectangles). (a) Four boundaries of different levels; (b) Two overlapping regions; (c) One note-group containing five notes; (d) Three comments.

4.2.2 Participants

Participants are citizen scientists interested in music and the research goals of discovering how performers create structures in performances and how listeners understand them. In order to use the interface to annotate the recorded performances, the user should have either headphones or speakers, and a computer with stable internet access. Participants first create an account using a valid email address and password to access the CosmoNote annotation platform, agreeing to the terms of the study[2] and that they are over the age of 18. At a time of their choosing, participants may answer a short questionnaire (see Figure B.1) adapted from the Goldsmiths Musical Sophistication Index (Gold-MSI, by Müllensiefen, Gingras, Musil, and Stewart (2014)) to describe their relationship to music in a more nuanced way than a simple *musician-non-musician* classification. Even though the annotation conventions are suitable for expert performers who may annotate their own work, no formal musical training is assumed or needed to contribute. An audio calibration stage allows participants to adjust their sound volume to a comfortable listening level and researchers to learn about the users' listening environments, i.e., their sound reproduction system and their hearing. It is based on a procedure by Cartwright et al. (2016) whereby participants are asked to count a number of random, equally loud, pure tones.

A training collection allowing participants to familiarize themselves with the interface and the annotations is always accessible to all annotators before engaging with the main annotation tasks. This training module currently features 3 short excerpts (around 20 s each) of the following pieces: Beethoven's "*Für Elise, Bagatelle No. 25 in A minor*", Bach's "*Minuet in G minor BWV Anh 115*", and Beethoven's "*Symphony No. 5 in C minor, Mvt II*". The excerpts were chosen for being both simple and likely to be familiar to a wide audience while exhibiting good examples of prosodic functions in music. This collection is a sandbox environment where participants can create and save annotations that will not impact their work in the actual task. In addition, the main CosmoNote YouTube channel[3] (also accessible from the CosmoNote website) has training material in the form of video examples showing how to place the different annotations.

4.2.3 Getting Feedback

After annotating a full set of performances in a collection, participants are invited to answer a questionnaire, giving feedback about their experience with the interface and with the annotation

[2] https://cosmonote.isd.kcl.ac.uk//agreement.html
[3] https://www.youtube.com/channel/UCEOzV1aQ4zOM2RLUq-AVOOg

task. For example, for a boundary annotation task, the instructions will indicate the intent of boundary strengths (as markers of degrees of disjuncture) while the feedback questionnaire will ask about the strategies annotators used, which will tell us how participants viewed/used the tools (see Figure B.2). As the experience and/or task may be different depending on the collection, custom feedback questionnaires can be shown for each collection. For example, the user survey can include questions about users' experience with marking transitions, tipping points, or salient melodies when the task involves these actions. Participants are asked to provide feedback after annotating a full collection because tasks usually involve annotating many pieces with similar properties. This also allows for a more streamlined annotating experience. If a collection involved pieces without shared properties, a simplified version of the questionnaire could be implemented and introduced after each piece or excerpt. At present, we chose to minimize the time participants spend answering questionnaires and maximize the time they spend annotating music. Incidentally, feedback at every stage of the annotation is possible as annotators may include text labels on every annotation type and the special "Comment" annotation type is designed for immediate feedback – these options allows users to mark any element they find curious, interesting, or out of the ordinary. Although these questionnaires are the main avenue for receiving feedback, other studies such as focus groups will involve direct conversations with participants. Additionally, participants can reach the CosmoNote team via email (shown on the main site) to share any other observations or comments they may have. The CosmoNote platform and our protocol have thus far been well received by the public (see Section B.2 for more details). This feedback will be used to iteratively refine the CosmoNote interface over subsequent annotation campaigns.

4.3 Musical Prosody Annotation

The performance annotation task, done with CosmoNote, is central to our method. Participants listen to audio recordings of the recorded piano performances while viewing the various music visualization layers (see Section 4.2.1, Figure 4.1) and are asked to mark segmentation and prominence in the music (as detailed in Sections 4.3.1 and 4.3.2). Participants are provided with annotations instructions that they can access at any time. Annotators currently have to read approximately one page of instructions though the exact amount of instruction/training needed may vary based on the task and the participant's musical expertise (see Chapter 6). Annotation instructions are tailored to specific collections of recorded performances, allowing us to run complementary studies as needed[4]. There is no time constraint for making annotations. This means that participants will have the option of completing a set of annotations over multiple listening sessions, in a recurrent fashion. They may also revise their annotations over separate hearings before clicking the finish button which freezes the annotations, no longer allowing further changes.

Figure 4.3 shows an organic annotation workflow highlighting the five most frequent interface interactions that listeners are faced with when marking musical prosody in CosmoNote: (1) listening/visualizing (using the sound controls and visualization options), (2) annotating (placing any of the four annotation types), (3) editing (adding labels to, moving, or deleting annotations), (4) saving (syncing to the database), and (5) finishing (concluding the process for a given piece). The following subsections describe the annotating of segmentation and prominence respectively. It should be understood that these are not necessarily separate tasks but rather are two areas of focus within the overall annotation task.

[4]See examples of different annotation instructions in Appendix C and Appendix D.

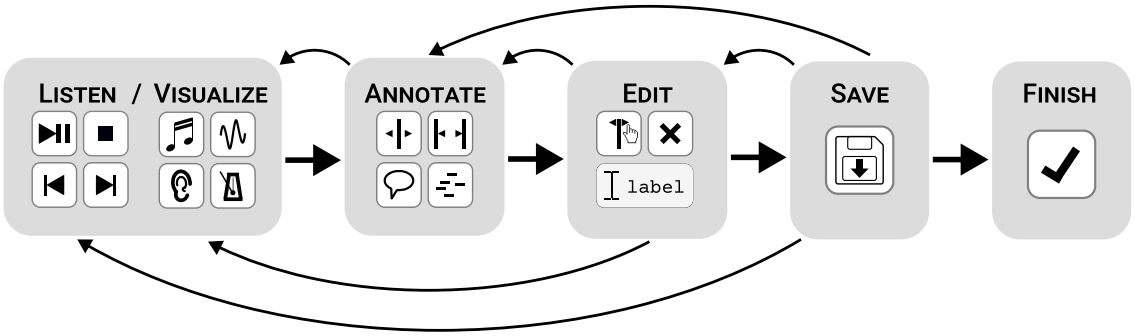


Figure 4.3: A diagram of 5 steps proposed to listeners for annotating a piece using CosmoNote, from left to right: listening/visualizing, annotating, editing, saving, and finishing. Notice that users are able to go back and forth between all intermediary stages until finishing.

4.3.1 Annotating Segmentation

Segmentation is the process of dividing something, in this case music, into meaningful units. It has been extensively studied in the fields of phonetics, speech perception, music analysis, and music information retrieval. For example, music theorists have famously studied the grouping processes that segment music into coherent chunks (Cambouropoulos, 2006; Lerdahl & Jackendoff, 1996), and how musical phrase note lengthening and intensity variations signal important segments to listeners (see Section 2.2.2).

Although prosodic cues can arise from composer-determined structures, our approach is focused on interpretive (that is performer-specific) use of musical prosody to segment musical streams. As a starting point, we examine how generalized concepts (Boundaries, Transitions, and Pauses) are traditionally defined, and in which ways they are likely to be shaped by performers.

Boundaries are well-studied structures that are used to describe segmentation (Wang et al., 2017). Even though their precise location can be ambiguous, annotated boundaries should indicate, by definition, clear points in time dividing the music stream into segments. These segments should be coherent (e.g., a complete musical idea or a musical thought) and help listeners make sense of the music. Figure 4.4 shows example boundary annotations[5].

Segmentation is mainly marked with the *Boundary* annotation type in CosmoNote. The four level *strengths* offer more granularity and provide more details for the data analysis. It is worth mentioning that, for this protocol, boundaries do not act as a nested hierarchy. This means for example that a boundary of level 1 is not contained in superior levels (2, 3, 4); and by extension, only one level is allowed at each time point. We will communicate clearly to annotators how boundaries of different strengths mark segmentation (from 1-weak to 4-strong), providing concrete examples. Boundary profiles aggregated from data of many participants will minimize effects of as blunders (e.g., an annotator meant to press 1 but pressed 3) and outliers (e.g., an annotator marked a boundary where no one else did). While the emphasis here is on prosodically demarcated boundaries, we acknowledge that experiences of boundaries may also be colored by changes in musical features such as harmony and timbre. Boundary levels can be a priori mapped to many types of segments. We describe below one of many possible mappings.

Motives, are the smallest indivisible succession of notes and/or rhythms detectable as a unit in music (Kennedy & Kennedy, 2012). They may be delineated by subtle cues like accents or micro pauses executed by the performer. In speech, letters attain meaning when they are turned

[5]Go to https://doi.org/10.6084/m9.figshare.23732664.v1 to hear the performance in Figure 4.4.

into words. Likewise, individual notes or rhythms are imbued with a meaningful context when grouped in a musical motives. Musical motives are often repeated in a piece and may represent different concepts such as: the seed of a musical idea to be developed (e.g., the main motive in Beethoven's First Movement of his 5th Symphony in C minor) or symbolize a character or an idea (e.g., in opera, a sequence of notes called *leitmotif* is repeated each time a character enters the scene). When motives have a larger priority in a piece, they are also called *figures*.

Sub-phrases are parts of a *Phrase*, which is a complete self-contained musical statement. Phrases and sub-phrases are often notated on the score using slurs, which give articulation cues to the performer. It is noticeable that these musical terms are directly linked to phrases in speech, which are governed by syntax. In that sense, the harmonic structure of a phrase (in tonal music) usually follows a set of syntactic rules (Rohrmeier & Pearce, 2018) that help to punctuate where musical phrases end; for example by using a cadence to arrive at a resting melodic or harmonic position. A common case of sub-phrases that constitute a larger phrase are antecedent and consequent sub-phrases, which resemble each other rhythmically and are complementary to each other.

Sections as defined by Spencer and Temko (1994) are major structural units made up of smaller structural phenomena. They contain, for example, phrases and motives that are related between them and function as larger parts of a whole inside a piece. Sections represent a complete, but not independent musical idea, which is why pieces are generally composed of more than one section. Since they represent important parts of a piece, composers notate sections in a score using double bar lines, and they are often labeled alphabetically in musicological analysis with capital letters (e.g., A, B, A'). Similar to what happens with phrases, the end of a section is usually demarcated with some concluding melodic or harmonic device which is perceived as being conclusive; although the resolution is often clearer with sections. For example, performers may demarcate sections using a longer pause or larger tempo change compared to what they do with other boundaries.

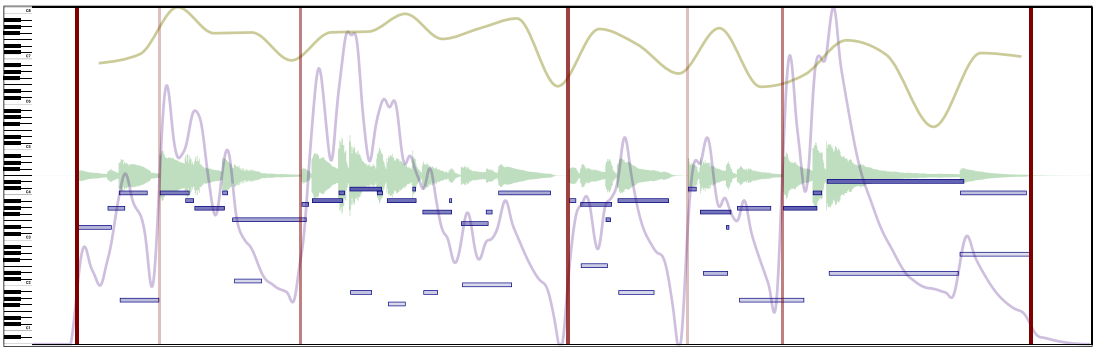


Figure 4.4: Excerpt of Beethoven's Symphony No. 5 in C minor, Mvt II. Boundaries (red vertical lines) in this example are assigned according to the strength of the change. Stronger boundaries have a thicker, less opaque red color.

Transitions may be musical passages that set up a change that is coming in the music, for example from one idea to the next. They can be seen as a link between sequential musical ideas. Transitions typically blur changes (boundaries) in the music by moving slowly through them. There are many ways to introduce a transition in music. For instance, elements from the following structure could be hinted at (e.g., a forthcoming motive is heard in a secondary voice) and/or new phrases can be introduced specifically to function as a bridge between ideas

(e.g., the last passage of the third movement of Beethoven's 5th symphony[6] shown in Figure 4.5 transitions to the fourth movement). To execute a transition musically, performers can use tempo, dynamics and articulation to prepare the listener for what is coming by intentionally making the contrast between musical ideas smoother and less rigid.

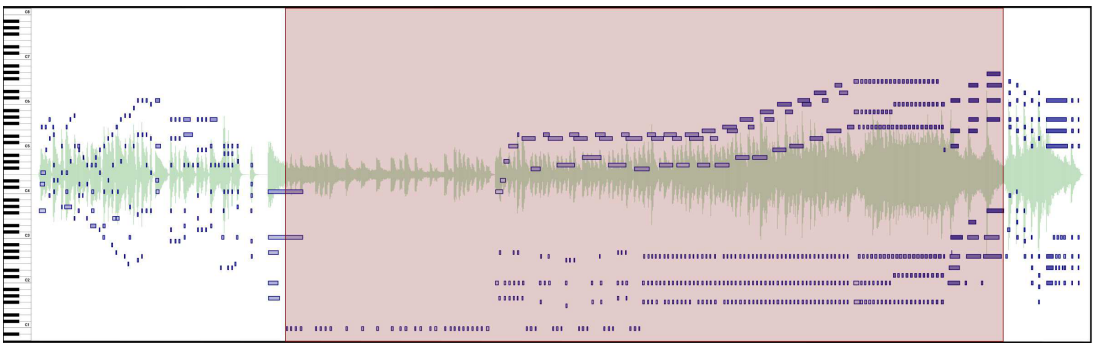


Figure 4.5: Long transition marked with a region (light red rectangle) between Movements III and IV in an excerpt of Beethoven's Symphony No. 5 in C minor.

Pauses occur analogously in speech and music as segmentation devices that add space between two adjadcent structures. In music, this concept is related to the timing of the notes in a piece and is executed by the performer via lingering on notes or by using silence. Silence is a very powerful tool in music. It is used for separating musical elements, to hold the audience in suspense, or for other expressive effect. Pause are executed before, during, or after a musical event. Then notated in the music, pauses are indicated by a comma or a *fermata* symbol placed above a note or a rest. The term *breath* is used either figuratively or literally in relation to the performer's breath when executing a musical passage (see example[7] in Figure 4.6). The duration of a pause is sometimes specified by the composer, but the performer makes the ultimate decision of how long it should be depending on the unique circumstances of a performance.

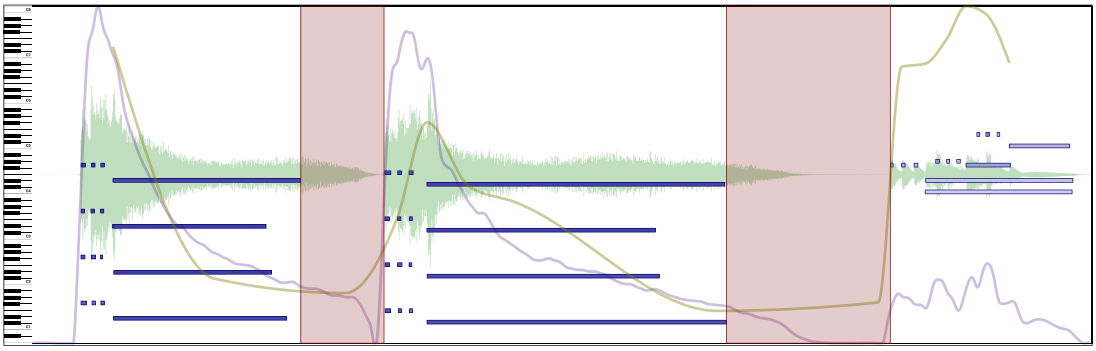


Figure 4.6: Two pauses marked with regions (light red rectangles) in an excerpt of Beethoven's Symphony No. 5 in C minor, Mvt I. Silence and timing are used to make the music breathe.

Transitions and pauses may be marked with the *Region* annotation type in CosmoNote. Participants can to mark the beginnings and ends of these structures and label them accordingly.

[6]Go to https://doi.org/10.6084/m9.figshare.23732673.v1 to hear the performance in Figure 4.5.
[7]Go to https://doi.org/10.6084/m9.figshare.23733288.v1 to hear the performance in Figure 4.6.

When the starting/ending times are not clear, especially for transitions, it is recommended to place them at the outermost value possible. Overlapping regions are allowed as needed.

4.3.2 Annotating Prominence

Prominence generally characterizes an emphasis drawn toward a certain element of a whole. Prominence in speech is important because speakers are capable of changing the meaning of an utterance by assigning more weight to a certain word or by changing their intonation. In music, experienced performers may make musical structures stand out in a way that helps to resolve ambiguities, particularly in musical meter (Sloboda, 1985). They may also introduce focal points by assigning more weight to a note or a chord.

For this description, we divide musical prominence into two sub-categories: vertical or horizontal based on their temporality. Events that may or may not segment the music, but that are easily recognized as belonging to a single moment, using timing and dynamics, are classified as vertical prominence. On the other hand, attention drawn to a particular pitch, timbre, or melodic line that cannot be pinpointed to a clear moment, is categorized as horizontal prominence. Since prominence can be viewed as either vertical or horizontal, all vertically prominent structures may be marked using *Regions* (if a preparatory stage exists) and *Boundaries* (or *Comments*) while horizontally emphasized ones may be marked with the *Note group* annotation type (e.g., notes in an important motive or a salient melody).

As was the case with segmentation (see Section 4.3.1), we will concentrate on the point of view of the performer in the following descriptions of common prominence creation techniques (stress, melodic salience, and tipping points) that listeners are likely to recognize while annotating prominence.

Stress is an emphasis on a particular element to make it more prominent than those around it. In speech, stress is used to help parse words in a language like English where syllables and words can be stressed to alter the meaning of an utterance (Ashby, 2011). In music, this category is close to how Drake and Palmer (1993) define *rhythmic grouping*, which focuses on event intensity/duration and *metric accents*, which focus on higher order regularities in a sequence. Thus, stress may be indicated by a combination of performer actions like an increase in sound intensity, duration, or even a change in timbre, as seen in the example[8] of Figure 4.7.

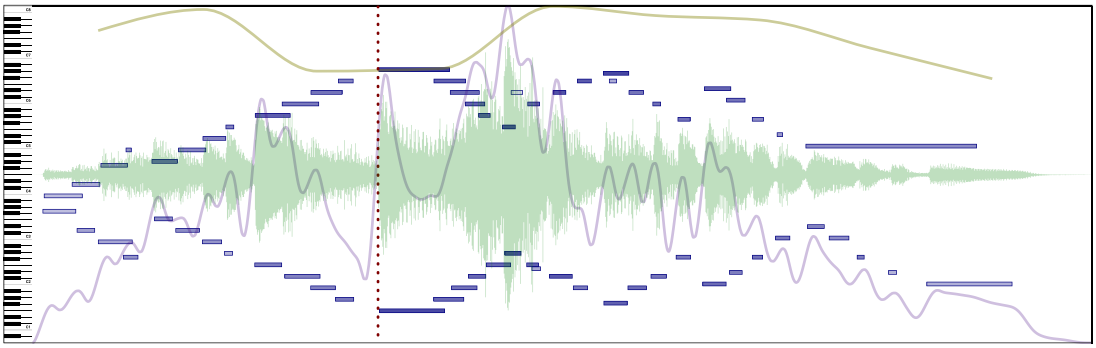


Figure 4.7: Stress is marked by a comment (red dotted line) at the pinnacle of the progression in this excerpt of Beethoven's Variation No. III from 32 Variations in C minor.

[8]Go to https://doi.org/10.6084/m9.figshare.23733276.v1 to hear the performance in Figure 4.7.

Melodic salience, as the example[9] in Figure 4.8 shows, is a special case of prominence dedicated to the melody of a piece. It relates to the concept of *melodic accents* by Drake and Palmer. Melodic salience may be recognized by an increase in loudness and duration (the notes of the main voice in a melody are usually louder and longer) or a variation in timbre (by using a different touch/technique or a different instrument altogether). Performers can indicate melodic salience in piano performances by systematic variation of intensity and duration, even within hands, to enhance the melody (Repp, 1996).

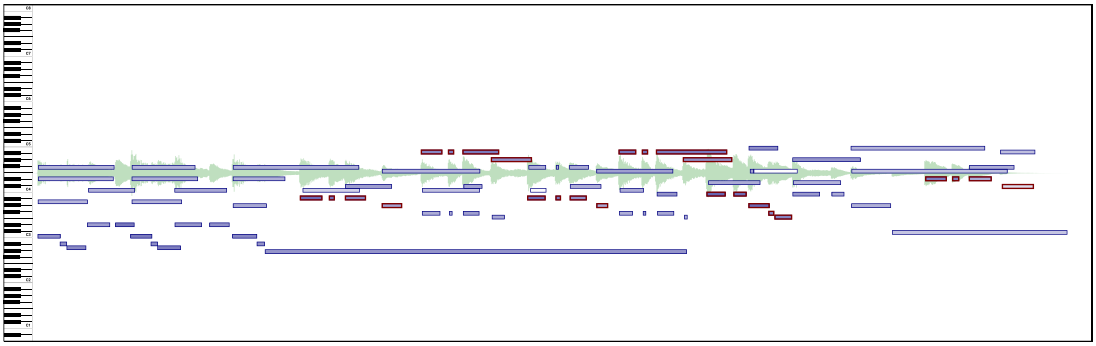


Figure 4.8: Melodic salience highlighted in note groups (light red rectangles) as the notes of the main motive are made more present by the performer on an excerpt of Chopin's Ballade No. 2 in F major.

Tipping points, as defined by Chew (2016), are cases of "extreme pulse elasticity" where musical time is suspended in an unstable state beyond which a return to the pulse is inevitable. As such, they are frequently present in musical transitions (a clear distinction exists before and after the tipping point) and musical pauses (a tipping point created at the moment a pause can no longer be stretched). Figure 4.9 shows a passage[10] with a melodic tipping point, where the performer plays with the listeners' expectations and delays closure of the sub-phrase.

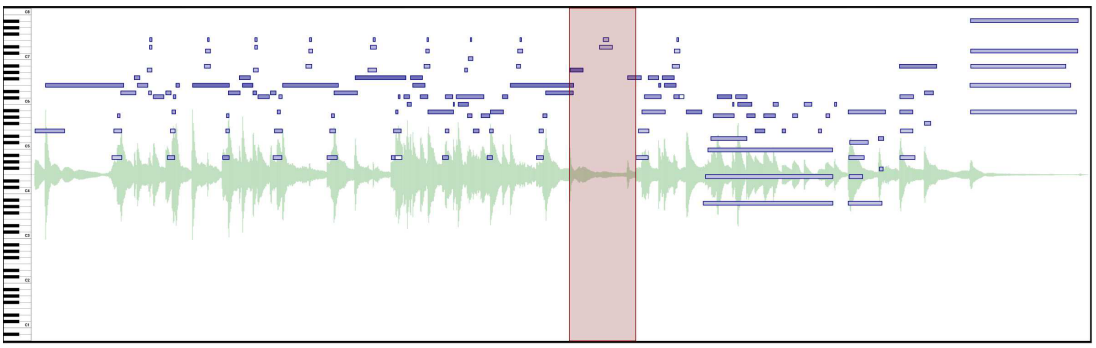


Figure 4.9: A tipping point marked with a region (light red rectangle) on an excerpt of Grieg's Solveig's Song evidences how a musical moment is stretched to its limit.

[9] Go to https://doi.org/10.6084/m9.figshare.23732679.v1 to hear the performance in Figure 4.8.
[10] Go to https://doi.org/10.6084/m9.figshare.23732676.v1 to hear the performance in Figure 4.9.

4.3.3 Potential Annotation Strategies

Participants are presented with annotation instructions to direct them in doing the annotation task. Yet the specifics of their annotation placement are driven by their individual strategies. The following is an excerpt of the general annotation instructions that participants read when they enter the main annotation campaign in CosmoNote: "*Please mark the boundaries that you hear in the music, and indicate the strength of each boundary. You may be presented with information layers such as the notes, tempo or loudness, but your ear should be your main guide*". Although there are numerous ways to reach this goal, we will center on two branching strategies that are common to all annotations, namely real-time vs. retrospective annotations and analytic vs. intuitive mindsets. These methods will be examined using actual experiences that CosmoNote participants (software testers) had during the development phase, which are used to anticipate and deal with problems that future annotators may confront.

Sometimes, participants wanted to listen to the whole piece, from start to finish, before placing any annotation. These participants found it easier to first concentrate their full attention on the music, forming their own mental model of the piece's structure and of what the performer meant to communicate, and only then proceeding to the annotation stage. This retrospective annotation process could take longer but listeners who adopted it would have to revise their work less. In contrast, some participants preferred annotating in real-time, which meant marking structures while listening to the audio, even for the first time. Since music was listened to and understood retroactively, if listeners were not already familiar with the music, they would need to go back again to correct their work after becoming more familiar with it. To maximize the benefits of real-time and retrospective annotations, we recommend marking only the biggest boundaries on the first play-through. Once the larger segments are defined, annotators may go back, focus on smaller segments and repeat the process, correcting mistakes as needed.

There is also whether to analyze the music intellectually or to annotate by intuition. An analytical approach will vary depending on the person's musical knowledge, experience or formal training. It is important to note that annotating analytically does not mean using traditional music theory or score structure analysis, nor is it about finding repeating patterns within the sounds. Annotating analytically means thinking deeply about the performer-made segments and prominent structures in the music. The counterpart of this strategy is a more intuitive, spontaneous annotation, where citizen scientists try to reduce their cognitive load, be more comfortable, focus less on being *right* and embrace the subjectivity of the task. The potential drawback of this strategy is that listeners could end up annotating the emotions conveyed by the music, which is not the purpose of this methodology. We recommend some balance between the two, where neither attentive listening nor spontaneity are privileged one over the other.

Since music is primarily an auditory stimulus, participants are discouraged from placing annotations only by relying on visual cues (see Chapter 5 for more details). However, no matter the approach, any complementary information (e.g., visuals or auditory cues like boundary sounds) should be understood more as advice than a prescribed way to perform the task since we do not wish to impose a fixed way in which annotators use the tools in CosmoNote. In fact, we wish for them to explore their possible uses, and give us feedback on how they use them. Ultimately, any instructions or strategy should be mainly used to help externalize the intuitions that are formed while listening to the music. This is why it is made clear, at various stages, that participants should trust their ears.

4.4 Outcomes from the Methodology

4.4.1 Data Structure and Analysis

As discussed in Section 3.6, data from annotations collected in CosmoNote is organized in a secure database and exported using a JSON (JavaScript Object Notation) format for its analysis. Structures of JSON arrays of paired name/value objects ensure great flexibility and modularity (see examples in Appendix A). Data for each of the four annotation types is classified by pseudonymized participant ID numbers. Objects contain primary properties (data) like timestamps and labels, and secondary properties (metadata) such as creation and modification dates. For example, note group annotations not only contain an onset time but also sub-structures like the MIDI note identifier for each note in a group. A scheme of the annotation data structures is shown in Figure 4.10. More properties may be incorporated as CosmoNote and the methodology evolve, are adopted by more users, or as the need arises. For instance, the data structure does not currently contain information on which music features the annotator chose to display and when. Such ability to track user interaction involves a non-trivial amount of development work, which can be included in future releases (see Chapter 3.7).

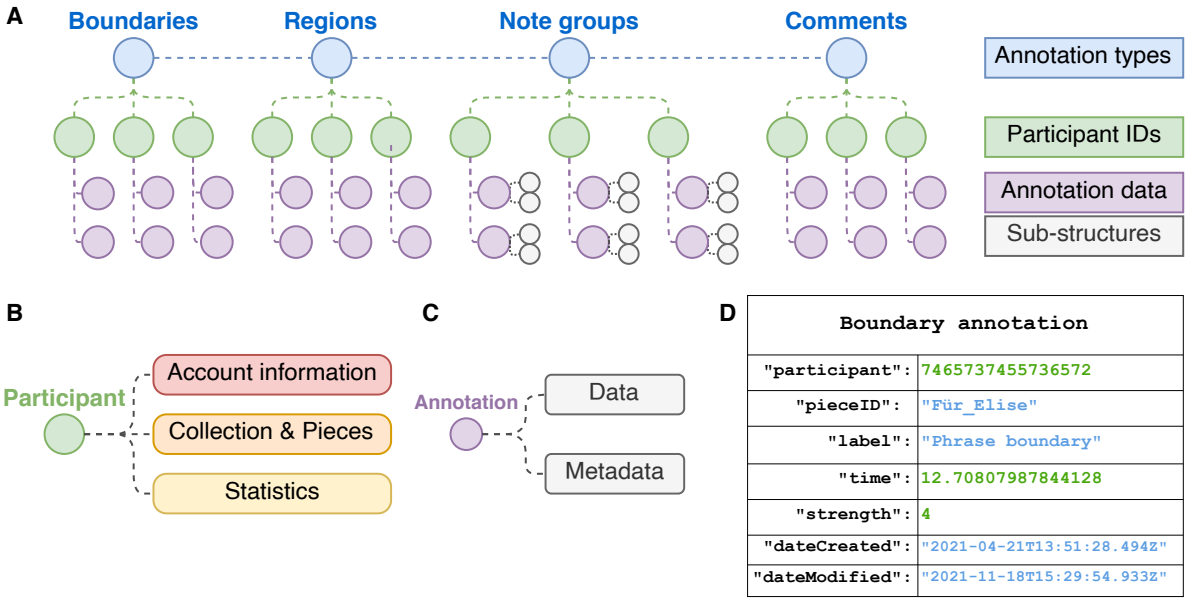


Figure 4.10: Data structure for the CosmoNote Annotations as organized in JSON objects. (A) A representation of nested data in each of the four annotation types; (B) A participant object contain information about its Account, Collection & Pieces, and aggregated Statistics (C) Annotation objects contain data (e.g., timestamps, labels) and metadata (e.g., date of creation and last modification); (D) An example boundary annotation JSON object.

For each piece, our main variables are distributed over the quantity and location (time, and pitch when applicable) of each data point from the four types of annotations, each related to a participant and containing auxiliary properties (as described above). In addition, the music features that are computed for visualizations in CosmoNote (MIDI data, loudness, tempo, and harmonic tension curves) constitute important synchronized time-based information for the analysis of the annotation data. To establish relevant relationships between aggregated annota-

tion data, high-level prosodic features, and low-level musical features and acoustic properties, we will use techniques such as dimensionality reduction, cluster analysis, and multiple regressions.

The data generated from our method will be used to inform new models for prosody perception and production both from theoretical (rule-based) and data-driven (machine learning) analyses. Theoretical models will be informed by our results, representing abstractions in performance that use musical prosody, and that are difficult for machines to recognize. Our analysis will extract a distribution of prosodic structures from which observed patterns can be described and modeled via algorithms that can be validated by expert annotators. On the other hand, data-driven models will be constructed by using human annotations to train machines to more naturally identify musical prosody. In any case, our objective is to use our models to mediate the relationship between prosodic structures and acoustic variations from a different perspective, that helps composers, performers, and music lovers in general, to enhance their understanding of the role of prosody in music performance.

4.4.2 Accuracy and Precision of Annotations

Our protocol relies on manual annotations of musical prosody. This approach is the most appropriate to explore questions about intuitions humans have that machines do not (i.e., the recognition of subtle cues of prominence in ambiguous musical structures). This means that accuracy is not strictly a problem for this method since we do not assume there is a right answer for the placement of annotations. However, precision among the annotations is an issue since the precise timing of segmentation or prominence annotations could differ across participants.

Imprecision will be accounted for by creating a large profile of annotations for a single piece. However, because of the possible deviations in timestamps, each annotation will be weighted to create this profile, where a tolerance proxy is set. For example, ± 3 s or the equivalent of 1 bar of a piece with quadruple meter at 80 bpm (Ong, 2006, p. 58). Comparing the mean profile with each individual profile will help detect outliers in an individual profile. The data will also be checked for internal consistency.

4.4.3 A Library of Examples

Another outcome of the citizen science studies is the creation of a performance-oriented library of musical prosody examples. It is an easily accessible, free resource intended for music professionals (performers, composers, musicologists, music educators), music enthusiasts (without formal musical training), and the public at large. This nascent library is composed of four main parts: The raw MIDI and audio data obtained from the pieces in the datasets (described in Section 3.3), the computed musical features (loudness, tempo, harmonic tension) as described in Section 3.4.2, individual annotations collected using the protocol detailed throughout Section 4.3, and aggregated data derived from the analysis of individual annotations, as specified in Section 4.4.2.

Anyone who accesses this library will be able to listen to the music and see the annotations as they were created. They will have total control over what representation they see, how to layer multiple visualizations, which type of annotation is shown, and which labels were assigned at a given point. Users will also be able to use see zoom controls to interact with the content at their preferred timescale, and see the aggregation of multiple users' annotations for the same piece when they are available in CosmoNote. The library already contains more than a dozen examples of musical structures and can be accessed through a playlist on YouTube[11].

[11] https://youtube.com/playlist?list=PLR0hNEZTO56Nbjc02ciW_EPxrWnmnKlp1

4.4.4 Community-Driven Musical Prosody Conventions

We believe that large scale usage and adoption of our method presents a step toward improving current practices for thinking about and annotating musical prosody in performance. Using the citizen science paradigm, we hope to build a community around the research of these structures in performed music. Each and every person that participates in our studies may come with their own motivations (e.g., passing time, learning about music, listening to their favorite artist) but they will also contribute to the same shared goal of understanding musical prosody in performed music. In the process of annotating the music, citizen scientists will refine our protocol and facilitate its wide adoption as a part of a set of conventions to annotate musical prosody.

The effort to build a community that engages with the content we propose and participates in the citizen science campaigns is a long-term endeavor. In that sense, a person (or a team) needs to be in charge of the communication in a project. Facilitators, are often in charge of being a bridge between participants and researchers. This can take the form of social media messaging, in person/distance meetings, announcements, or other types of interaction. We have already created and reached out through social media accounts (e.g., YouTube, Facebook, Twitter), communicated our work in scientific conferences and workshops, and gathered data and feedback from public and private studies. However, this is a task that must be sustained to achieve its intended effect. After the initial phases of creation and testing of the annotation platform, the COSMOS project has a roadmap for adding a person with this facilitator role. In fact, the goal is to gradually construct something which can be sustained beyond the duration of the COSMOS project.

On top of the technical capabilities of our annotation platform, social features facilitating online interactions between the community of annotators and researchers would improve its adoption (see Section 3.7). The relevance of developing a novel annotation platform has already been discussed from a technical point of view (Section 3.2). However, some reasons why we decided not to use websites with established social features, community, and continued participation such as the Zooniverse (see Section 2.4), are detailed as follows. First, at the time of CosmoNote's creation, and even at the time of this writing, the annotations and interactions with audio and visuals in CosmoNote are vaster than those possible in any citizen science platform. Second, we needed the modularity to design and run custom experiments. Third, protecting our user's data, and controlling both our tools and overall scalability are central to the project's vision. It is worth noting, that our citizen science efforts have been showcased in not for profit, public pages that aggregate many crowdsourcing projects, notably the French websites Science Ensemble[12] and Particip-Arc[13]. The promotion of CosmoNote through these types of sites are examples of a strategy that may allow us to expand to an international audience.

4.5 Discussion

In this chapter, we have presented a novel and scalable method to represent, identify, annotate and analyze musical prosody in performed music from a top-down, human-centered perspective. This method is supported by an innovative web-based platform that combines traditional and current music representations to improve the annotation experience. The results of data collection studies using this method will contribute to better understanding of how humans perceive prosodic structures in performed music. We aim to construct models that explain how these structures are created and used by performers in their real-world practice. The data we collect

[12] https://www.science-ensemble.org/projets/cosmonote
[13] https://www.participarc.net/projets/cosmonote

will also be used for the development of a library of musical prosody examples that will form the basis for formulating a musical prosody annotation convention.

Many of the advantages of our method are related to how we collect the data. Since the process is self-paced, annotators are not constrained by a fixed time limit. The instructions explain what musical prosody is and provide concrete examples of how prosody may be presented in expressive performance. Thus, annotators can afford to listen attentively, think about their choices, and even modify their annotations during a subsequent markup session. The music visualizations and training examples provide cognitive scaffolding (Yelland & Masters, 2007) for users with the various visual layers giving complementary information. For example, the piano roll visualization layer is helpful to single out the note onsets and intensities of a prominent melody but only gives vague information regarding tempo while more specific information about tempo is provided by the tempo layer. It is worth noting that piano music was chosen for the ease of aligning event-based (MIDI) information with an audio signal from a Disklavier piano, allowing for the automation of tempo extraction. However, any music audio can already be depicted on the interface and the tools work equally well for denoting expressiveness in music for other instruments, including voice. The main challenge in this case would be the difficulty of deriving precise note onset and offset information aligned with the music audio, but that is not insurmountable (e.g., with manual annotation). We anticipate that the ways in which the tool is used could evolve to encapsulate many more expressive devices appropriate for other instruments, where the ideas of segmentation and prominence still apply. With regard to the current interface, the tools available in CosmoNote are easy to use by virtue of their simplicity. The audio controls are similar to any music reproduction software and annotations are placed with a click of the mouse or a keyboard shortcut.

The challenges encountered by this methodology are typical of many citizen science projects, including ensuring data quality and sustaining community engagement. As for the data, the types of structures susceptible to being marked are potentially limited by technical constraints, difficulty, and task duration. For example, even though there is no time limit, if a specific action in CosmoNote requires sustained focus, it could be perceived as tiring or frustrating. This situation may directly impact community engagement. Solutions for these problems are generally complex; one approach is to address them methodically and iteratively at the software development level as we receive feedback on the use of the software. For example, if one prosodic marker/annotation type is rarely used, we could first investigate why that is so by getting direct feedback from the feedback questionnaires and then making changes accordingly. If annotators do not find an annotation type/visualization useful for marking musical prosody, suppose because the representation captured by this feature is not correctly communicated or is redundant, the feature can be redesigned or removed. The same process would apply if a given interaction is not consistent/intuitive, and we find that the user experience could be improved by simplifying the interaction or by progressively introducing annotation types to reduce frustration or fatigue.

Our annotation protocol, the training content, and the feedback surveys are means to bolster data validity and reliability (Balázs, Mooney, Nováková, Bastin, & Jokar Arsanjani, 2021). For instance, the first campaign was launched to gather annotations while serving the complementary purpose of providing information for improving people's interactions with CosmoNote, thus encouraging further contributions. We are in the process of building a community (see Section 4.4.4) that shares our goals of understanding how music structures are created and shaped in performance through musical prosody. We rely on members' contributions and feedback; it is with the cooperation of both expert and non-expert annotators that we will iterate and improve upon the components of our annotation method. To ensure sustained community engagement, more collaborative features that connect citizen scientists are planned for the future versions of CosmoNote.

Our method is an ambitious attempt at discerning abstract musical structures in performed music through listeners' annotations. CosmoNote is a flexible and extensible tool suitable for representing and annotating musical prosody in real-world performances. By using it to explore how humans apprehend segmentation and prominence introduced in performance, we will have the means to design models that capture the complex relationships of these structures with musical features and their acoustic properties in a comprehensive way. Subsequent phases in our studies will build iteratively upon the results of previous ones, ensuring continued progress of the annotation method. The long term goal of this research is to open new paths for the public to think about what is being communicated in expressive music through the performer's segmentation of musical ideas and creation of musical prominence, and to offer new ways to explore and talk about performed music in general.

Chapter 5

Study 1: Comparing Visual and Aural Boundary Annotations

In Chapter 4, we laid out the basis of our musical prosody annotation protocol. In this chapter, we present a study that examines the implications of using such a protocol in a platform such as CosmoNote, with visual representations enforcing the annotation process. Our results show that visuals such as piano roll and waveform visualizations contain many global segmentation structures. However, they can also show spurious patterns and be deceiving at scales of less than five seconds if not used jointly with auditory information. Throughout this chapter, we use the terms auditory and aural interchangeably to refer to stimuli related to the sense of hearing.

5.1 Introduction

Music is an auditory phenomenon by definition. However, experiencing music often involves more than one sense. When one or more sensory inputs influence what is perceived by another, we speak about cross-modal or intersensory perception (VandenBos, 2015). It may happen, for example, to an audience, a performer, or an annotator in an experiment. Audiences seek the multiple sensations of live music performances and dance or even the enhanced aspect of music videos. Performers are constantly immersed in the multisensorial act of the physical gesture, auditory feedback, and even maybe reading music notation, following a conductor or a fellow player. There is evidence that in tasks involving auditory stimuli, annotators may experience natural cross-modality links between auditory and visual information (such as going up in pitch and space), even when an experiment is only focused on one modality (Evans & Treisman, 2010). This chapter focuses on the co-occurrence of auditory and visual stimulation while annotating music structures.

Much has been written about the effect of the visual component on music learning, appreciation, and understanding. Firstly, many musicians learn to read and interpret music through notation. Indeed, Stein (1979) argues that even in a short eight-measure unit, the eye comprehends musical form better than the ear. Secondly, concerning music appreciation, Tsay (2013) shows the prevalence of visual over aural information when musicians and non-musicians judge the outcomes of music performance competitions, despite reporting that sound is the most important source of information for this task. Indeed, a meta-analysis by Platz and Kopiez (2012), explains that the visual component in music perception is essential in communicating meaning, influencing an audience's evaluation of music performance. Visual cues, such as the physical gestures of a percussionist, can also alter note duration perception (Schutz & Lipscomb, 2007). Thirdly, several approaches use visualizations to promote a better understanding of music struc-

tures. For instance, Cruz, Rolla, Kestenberg, and Velho (2018) proposed a visual language that uses geometrical shapes of different sizes and colors to identify pitch and timbre. Malandrino, Pirozzi, and Zaccagnino (2019) created software that maps chord changes in traditional four-part harmony to colored rectangles. In parallel, Miyazaki, Fujishiro, and Hiraga (2003) developed a system that uses 3-dimensional, multicolored explorations of MIDI data as a means to provide different perspectives of musical structures.

To our knowledge, no study has compared the differences between boundary annotations in unimodal (aural or visual) vs. cross-modal (aural and visual) conditions. This study aims to explore that question using CosmoNote and our annotation protocol to annotate segmentation boundaries of four levels. We designed an experiment where participants were exposed to combinations of visual/aural cues and could freely place or remove boundaries. We used unbalanced optimal transport to compute distances between the aggregated boundaries, thus detecting which boundaries determine their value.

Because CosmoNote was conceived to offer visual layers synchronized with music, we are interested in how aggregated segmentation annotations differ when participants mainly use their eyes rather than their ears. The rest of this chapter is structured as follows to engage with this question. Section 5.2 explains the choices for gathering annotations in this study, practical aspects of our experimental design, conditions, and statistical methods. Next, a necessary overview of structural, historical, and technical aspects of the musical stimuli is delineated in Section 5.3. Subsequently, the results of our analysis are reported and interpreted in Section 5.4. Finally, Section 5.5 is dedicated to a broad discussion about the results in the context of the Musical prosody annotation protocol (Chapter 4) and future work.

5.2 Materials and Methods

5.2.1 Participants

Boundary annotations were collected from 56 participants (31 female, 25 male) at INSEAD-Sorbonne Université Behavioural Lab. Participants were classified by age group: 58% between 18-24 years old, 40% between 25-34 years old, and 2% between 35-44 years old. The group was equally divided into musicians and non-musicians. Musicians reported at least five years of musical practice and one year of formal music theory training.

5.2.2 Musical Stimuli

We selected a performance by Elaine Chew of the 32 variations in C minor, composed by Ludwig van Beethoven, played on the Bosendorfer ENSPIRE at Ircam. The performance was split into 33 individual pieces, a theme (the name Tema, in Italian, is used throughout this chapter), and 32 variations on that theme. The music was chosen because it has a variety of tempo and dynamics while preserving a similar musical style and ideas. The music's basic structure and main characteristics are described in Section 5.3 in context and more detail since they are essential to understanding the results.

5.2.3 Experimental Task

There were five main stages for the experiment. Participants first completed a short questionnaire about their musical abilities (self-reported). Next, they calibrated their audio levels and did a hearing environment test. Then, they got familiarized with the interface using the Training

Collection in CosmoNote, while guided by the experimenter. The main task consisted in marking the boundaries heard in the music and indicating the strength of each boundary. Boundaries were defined as: "time points that separate a music stream into segments representing meaningful chunks of music, e.g., a musical idea or a musical thought. Boundaries not only separate a larger piece of music into smaller, coherent units, they also help listeners make sense of the music." Participants could choose between four strength levels, defined from 1 (weakest) to 4 (strongest). See Figure C.1 in Appendix C for more details. After completing the annotations, they were asked to complete a feedback survey. The experiment was designed to be completed in approximately 100 minutes.

5.2.3.1 Experiment Duration

Participants took different amounts of time to complete the experiment. The experiment duration was tracked using the first and last annotation timestamps, recorded by collection. We did not record each participant's time completing the questionnaires or annotating the Training collection. The experiment's mean duration was 56 minutes, with a standard deviation of 24 minutes. The median value was of 53 minutes. The shortest time was 10.5 minutes, and the longest was 2 hours.

5.2.3.2 Experimental Conditions

Participants were presented with different conditions combining two visual information layers, a waveform or a piano roll (notes without pedal data) and the audio information (the actual sound of the music). Each piece was presented in a forced condition, preventing participants from changing the visualization/audio options in CosmoNote. Three global experimental conditions were designed, from which seven total conditions were extracted, as seen in Figure 5.1 and described below:

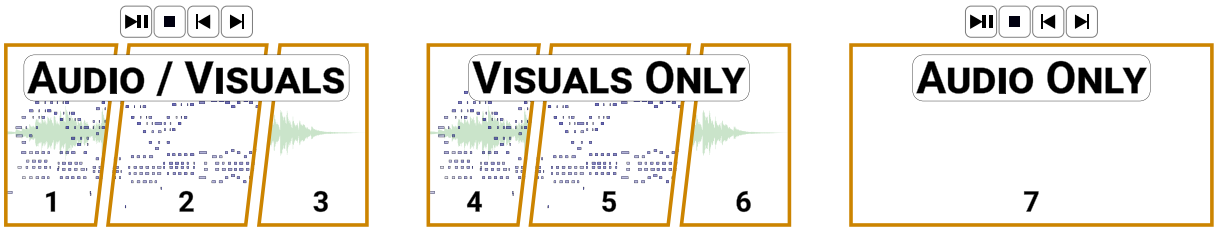


Figure 5.1: Simplified interface representation of the seven experimental conditions. 1: APW, 2: AP, 3: AW, 4: PW, 5: P, 6: W, 7: A. Audio controls only show when the music can be played. Subdivisions are highlighted in orange.

i. **Audio and visuals:** Cross-modal conditions where both audio and visual information is presented. Participants were asked to listen to the music to annotate and were told they might use the visual representations to help them annotate. Three conditions fit these criteria.

   1: APW → Audio + Piano roll + Waveform

   2: AP → Audio + Piano roll

   3: AW → Audio + Waveform

ii. **Only visuals:** Unimodal conditions where only the audio information is not presented. Participants were reminded that they would see the representations but not hear the sound and asked to annotate visually. Three conditions fit these criteria.

4: PW → Piano roll + Waveform

5: P → Piano roll

6: W → Waveform

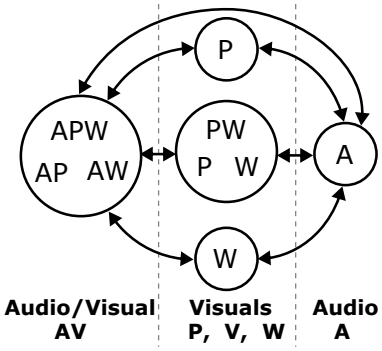
iii. **Audio only:** Unimodal condition where only the audio information is presented; the visual information is absent. Participants were asked to listen to the music to annotate and were instructed that they could not see any music representations. One condition fits the criteria.

7: Audio (A)

Each of the seven conditions was assigned to the pieces as a Latin Square. The order of the pieces participants annotated was shuffled. Thus, participants were divided into seven groups of eight people so that all the pieces would be annotated by a group in each of the seven conditions. All participants were exposed to all the conditions for the same number of pieces. Additionally, since experimental conditions are derived from the same musical source, there is no conflicting information between conditions; participants are only presented with more or fewer components. The PW visuals for all stimuli (Piano roll and Waveform overlapped) are shown in Figure 5.5, as displayed in CosmoNote during the experiment.

5.2.4 Grouping Experimental Conditions

To highlight the contrast between annotations placed using audio/visuals, we combined the seven experimental conditions detailed in Section 5.2.3.2, i.e., we pooled all boundary annotations from each condition, into two groups with common elements, then split into subgroups: (1) Conditions with audio, with two subgroups, labeled 'AV'and 'A', and (2) Conditions without audio, with three subgroups, labeled 'V', 'P', and 'W'. Annotations of each subgroup within 'Conditions with audio' were compared to annotations of each subgroup within 'Conditions without audio'. We do not focus on comparisons between visuals only because they do not provide helpful information for the central question of this chapter. Thus, we choose 7 out of 10 possible combinations, as explained in Figure 5.2.



(a) Comparison scheme

	V {PW, P, W}	**P**	**W**	**A**
AV {APW, AP, AW}	① AV vs. V	③ AV vs. P	⑤ AV vs. W	⑦ AV vs. A
A	② A vs. V	④ A vs. P	⑥ A vs. W	

(b) Groups of conditions to compare and their labels

Figure 5.2: Visual scheme and labels for comparing combined conditions.

It is essential to remember that when grouping and comparing annotations with seemingly repeating conditions (e.g., APW vs. PW), we are not copying the same annotations into different groups because different people annotated each condition. We are comparing the support used to create each group of annotations: visual/aural to visual to aural annotations.

We are particularly interested in comparisons 1, 2, and 7 in Table 5.2b. Comparisons 1 and 7 evaluate how close cross-modal annotations (visual/aural) are to unimodal annotations (visual or aural). Comparison 2 evaluates how close to each other are the two unimodal annotations (visual and aural).

5.2.5 Statistical Methods

The annotation data contained, among other information, the number of boundaries participants placed for each level by condition and piece. This number allowed us to calculate statistical descriptors such as the mean and standard deviation of the number of boundaries placed by piece. We identified, for example, that some participants marked an atypical amount of boundaries compared to the mean. Outliers were defined by boundary level as the number of boundaries that fell outside the upper outer fence, that is, more than three times the interquartile range (IQR) above the upper quartile. These annotations may be caused by participants not using all four levels (all annotations were done using only one level) or not understanding/respecting the annotation instructions. If data for a participant/level pair were considered outliers, all boundary data for that level were removed from the rest.

We applied four different techniques to analyze the differences between boundary annotations before choosing only one. (1) the complement of the correlation coefficient on continuous Gaussian KDE boundary profiles, (2) the complement of the weighted F-measure on individual boundaries, (3) the optimal transport distance on discretized boundary annotations, and (4) the unbalanced optimal transport distance on discretized boundary annotations. The following text describes the advantages and disadvantages of using any individual technique. Then, Section 5.2.6 presents a comparison between all distance metrics on our data, and Section 5.2.7 argues why the unbalanced optimal transport is the most convenient distance for this study.

5.2.5.1 Kendall's Tau Correlation

One technique to analyze the differences between annotation boundary profiles is to compute correlation coefficients between KDE curves that aggregate boundary annotations (Hartmann, 2017). However, since KDE boundary profiles do not follow normal distributions, we used the non-parametric correlation coefficient Kendall's τ instead of the Pearson correlation coefficient used by Hartmann. To express this metric as a distance, we use the complement of the Kendall's τ coefficient $(1 - \tau)$.

Because KDE profiles depend on a scale (bandwidth) parameter, the value of the correlation coefficient changes for different scales. Figure 5.3 displays curves of the complement of Kendall's τ computed on all comparisons in Table 5.2b with different KDE scales (see Section C.2 for more details on the KDE technique, and the bandwidth parameter in particular). The markers show the value of the correlation complement at 1 second, chosen as the best compromise to aggregate KDE boundary profiles in our studies.

Correlation coefficients between KDE boundary profiles can be modified and used as a distance metric to compare annotation profiles but are dependent on the scale parameter of the KDE curve.

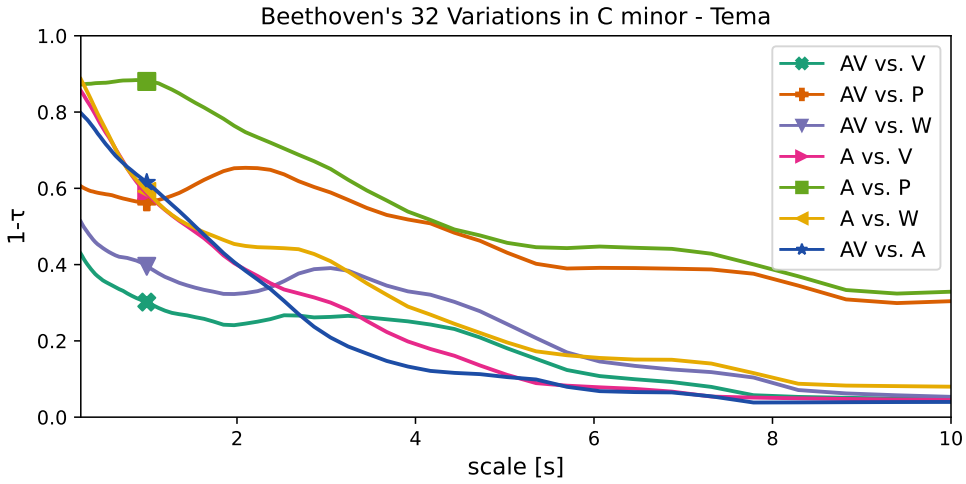


Figure 5.3: Comparing grouped experimental conditions (audio and visual) for the Theme of Beethoven's 32 variations in C minor. The y-axis represents the complement of Kendall's τ correlation $(1-\tau)$. The x-axis, in seconds, represents the KDE profile scale parameter. Markers for all comparisons are placed at the scale chosen for our studies, 1 second.

5.2.5.2 F-score

The F-score, also called F-measure or F_1–score, is a common technique used for comparing flat (as opposed to nested) boundary annotations (Nieto et al., 2014, 2020). It measures how close estimated boundary annotations (within a specific time window) are to reference boundary annotations. Annotations that fall within the window are called hits or positive results. F-scores are computed using two parameters: Precision (P) and Recall (R). Precision measures positive results compared to the number of estimated boundaries, and recall measure positive results compared to the reference boundaries. Nieto et al. (2014) introduced a variation accounting for perceptive bias favoring precision over recall using an α parameter smaller than 1 (we keep the default value $\alpha = 0.58$). F_1 and F_α are computed using the following equations:

$$F_1 = 2\frac{P \cdot R}{P + R} \qquad F_\alpha = (1 + \alpha^2)\frac{P \cdot R}{\alpha^2 P + R}$$

To accommodate this technique to our data, we use a weighted variation of the F_α–score, computing precision and recall values separately by strength levels (1–4), then combining them using a weighted sum, with scaled weights ranging from 0.25 to 1, and finally computing the F_α–score using the weighted versions of precision and recall.

The F-score metrics were designed for situations where ground truth annotations exist. They have the advantage of providing values for precision and recall but do not give insight into the importance of individual boundaries in each annotation set.

5.2.5.3 Optimal Transport

Boundary annotations over a time axis placed by humans are prone to imprecision errors within and between annotators (see Section 4.4.2). This is the temporal nuance that we would like to capture when comparing two boundary annotation curves. The problem is, however, that usual distance metrics, e.g., Euclidean distance, calculate point-by-point vertical comparisons between each point of the curves. Computing a distance using optimal transport alleviates this issue because it calculates a horizontal comparison. The distance d_p associated with the optimal

transport between two curves f and g, having the same area under their curves, is also called the Wasserstein distance, or even 'Earth moving distance' (Kolouri, Park, Thorpe, Slepčev, & Rohde, 2017). In the particular one-dimensional case (where $p = 1$), this can be interpreted as "pushing" the area under the first curve towards the area under the second curve. By defining F and G as the cumulative distribution functions of the curves f and g, respectively, the optimal transport distance is equal to:

$$d_1(f, g) = \sum_i |F(i) - G(i)|$$

We use this type of movement for comparing musical annotations on a timescale. However, there are two problems with this technique: (1) The value of the optimal transport distance is proportional to the size of the area under both curves, meaning longer pieces yield larger distances. To be able to compare the optimal transport distance of pieces with different lengths, we scale the distance by the duration of each piece. (2) Isolated boundaries in one curve may distort the distance between the two. For example, an isolated boundary near the end of the curve f must be pushed towards at least one boundary of the curve g, which would inflate the optimal transport distance between the curves even if there is only one boundary that is different between them. This is why we use a slightly modified version of this distance, described in the following subsection.

5.2.5.4 Unbalanced Optimal Transport

While unidimensional optimal transport is already a better representation of the horizontal difference between two curves, it does not reveal how or where the two samples differ the most. The unbalanced version of the optimal transport distance solves this problem, introducing a cost element that allows us to remove individual boundaries, one by one, to minimize the distance between the two samples (Guichaoua, Lascabettes, & Chew, 2023).

The cost determines the likelihood of removing a boundary in one of the two distributions. Higher costs yield fewer boundaries removed because the distance between the distributions without such boundaries will be higher than between the original distributions. Thus, if the cost is too high, the unbalanced optimal transport distance will equal the optimal transport distance. Conversely, if it is too low, the algorithm will remove boundaries until both distributions have the same number of non-zero boundaries. We have chosen a conservative cost equivalent to a window of ±20 ms to capture subtle differences between distributions. To compute the unbalanced optimal transport distance, we adapted an algorithm provided by a member of the COSMOS team, Paul Lascabettes.

The unbalanced optimal transport technique improves the results of the optimal transport distance and reveals which boundaries are causing the largest discrepancies between any two curves. This insight is helpful at the stage of results interpretation. As mentioned above, distances computed with this technique are also scaled to be comparable between pieces with different durations.

5.2.6 Comparing Between Distance Metrics

To visualize the results of all the distance metrics presented in the previous section, we computed seven comparisons between annotations of grouped conditions (Section 5.2.4) and observed their results.

Figure 5.4 aims to compare all distance metrics at a glance. Each box plot shows a distribution with 33 points (all pieces) computed from a specific distance metric (panels) on a given

comparison (rows). Panels display, from left to right, Kendall's correlation coefficient complement (first), F_α–score complement (second), optimal transport distance (third), and unbalanced optimal transport distance (fourth). Distances in individual panels are displayed from lowest to highest across the x-axis. The y-axis is divided into four rows. The first three rows in every panel contain two box plots. They compare either audio/visual (AV) or audio (A) annotations to one condition with visuals only (V, P, W). The last row contains only one box plot. It compares audio/visual (AV) to audio (A) annotations.

The same tendency is observed on the three top rows of all panels. First, compared to solely visual annotations, distributions of audio/visual conditions (AV) have a lower median than distributions where only audio was presented (A). Second, box plots between panels are visually similar, suggesting that the four metrics generate comparable measured outcomes when analyzing boundary annotations. In other words, rows in Figure 5.4 show that hearing the music *and* looking at the visuals (AV) results in annotations more similar to those placed solely with visuals (V, P, W) than solely with audio (A). Access to the notes and the waveform simultaneously (V) seems to enhance this similarity, compared to having only one visual at a time (P or W). We break down these relationships in more depth in Section 5.4.

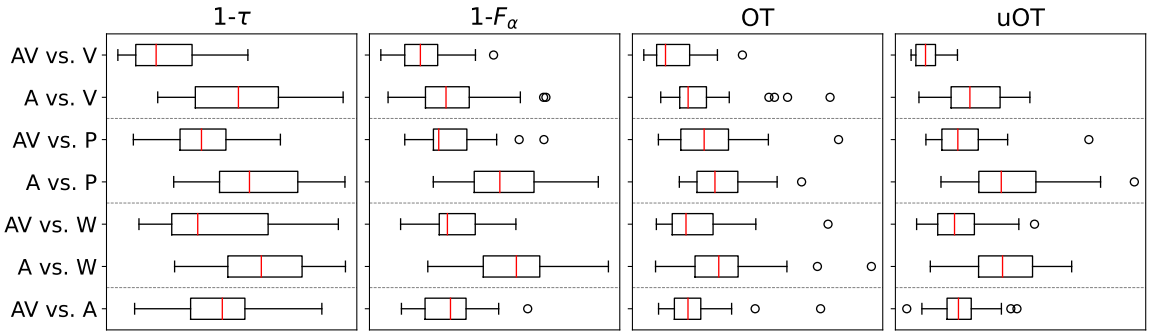


Figure 5.4: Box plot showing distances comparing conditions using audio and/or visuals. The panels show from left to right: the Kendall's τ complement (first panel), the F_α–score's complement (second panel), the optimal transport distance (third panel), and the unbalanced optimal transport distance (fourth panel). The distances are computed on all 33 pieces and their distribution is shown as a box plot.

5.2.7 Choosing a Distance Metric

Because of the high variability of correlation coefficients compared to their scaling factor, and the inability to distinguish important F-score boundaries, we will not be using those metrics to analyze our results. However, we will continue using the KDE boundary profile technique for visualization purposes because of its convenience for aggregating annotations and illustrating trends. The optimal transport distance, as already stated, has a better weighing of the boundaries of our annotation profiles, but it has the same problem as F-scores. In contrast, unbalanced optimal transport distances give insight into which individual boundaries make profiles more distant. Consequently, for the rest of this chapter, we will focus solely on the results produced by the one-dimensional unbalanced optimal transport distance and use its abbreviation uOT.

5.3 Brief Analysis of Beethoven's WoO 80

Before looking at the results of comparing segmentation annotations, we must dive into the fundamental aspects of the music we used for this experiment.

Published in 1807 and later labeled WoO 80 in the Kinsky-Halm Catalogue[1], the "Thirty-Two Variations on an Original Theme" by Ludwig van Beethoven (1770-1827) are a solo piano set usually performed as a whole[2]. The recorded performance of the set used for this study[3] was no exception, with a total duration of approximately 12 minutes. The median duration of individual pieces is 18 seconds. The shortest piece (Variation XX) is 12 seconds long, while the longest piece (Variation XXXII) is approximately 2 minutes long. Most pieces are under 30 seconds long.

5.3.1 Variation as a Form

As stated in Section 5.2.2, the choice of musical stimuli was informed by the musical similarity of the pieces between them, derived from the concept of variation. A variation is a commonly used compositional technique. Variations can be simple or complex. They often spawn from changes to the melody, harmony, rhythm, mode, among other musical elements, which can also be combined to create a new version of the original concept and is recognizable in the new idea (Kennedy & Kennedy, 2012). Beethoven's 32 Variations in C minor is an archetype composition of the *theme and variation* form. According to Abromont and de Montalembert (2010), this type of composition is usually a set that presents a principal theme and then cycles through transformations of such theme, preferring technical and aesthetic display over depth.

5.3.2 Structural Components

The theme is composed in the key of C minor, and all variations stay in the same key, though they alternate from minor to major mode from Var XII through Var XVI. The theme and most variations maintain an 8-bar structure with a 3/4 time signature, and a descending baseline, which are typical characteristics of the Chaconne, as defined by Kennedy and Kennedy. In addition, rhythmic accents on the weak beats of the main melody are reminiscent of the French Sarabande, as defined by Apel (1969, p. 750).

A recurring phrasing of ascending and descending notes is a dominant prominent figure throughout the composition. A similar form of this figure (not necessarily preserving the same intervals) is reiterated four times on most pieces with increasing stress, usually co-occurring with an increasing pitch, followed by a slight pause and a sudden accent (*sforzando*), and ending with a slower conclusion, with decreased intensity, marked harmonically with a cadence to the tonic.

Most pieces also preserve roughly the same harmonic progression with the following functional moments: i-V-IV-i-iv-V-i, which are extensively varied (e.g., using different chord inversions or substitutions). In contrast, Var XXXII is a unique case that functions as an extended epilogue with a coda, notably departing from the main structure in its duration, number of bars, and harmonic progression.

Despite the recurrent harmonic progression, lack of modulation, or time signature changes, numerous cues for dynamic changes give the performer a variety of intensities to be expressive. Dynamics of the set range from pianissimo to fortissimo, with many instances of *diminuendo*,

[1] The label WoO means "work without opus number" (Kennedy & Kennedy, 2012).

[2] Go to https://doi.org/10.6084/m9.figshare.24635568 to see the full score of the set.

[3] Go to https://doi.org/10.6084/m9.figshare.c.6755340.v1 to hear the individual performances in Figure 5.5.

crescendo, and *sforzando*. The recorded performance used in the study registered MIDI velocity values consistent with these ranges, with a lower average of 13 and a higher average of 92 out of 127.

Explicit indications to change tempo are rarely given. The set defines an *Alegretto* tempo at the start, and offers no other straightforward tempo change. However, performers can map expressive indications (e.g., *dolce*, *leggiermente*, *semplice*) to corresponding temporal variations. In our recording, the performer used the expressive resources at her disposal via many tempo changes throughout the set. Tempo curves were computed for each piece; the mean is approximately 90 BPM across all pieces. The median tempo on individual pieces varies from 47 to 123 BPM. The slowest passages of a given piece oscillate around 30 BPM, which usually occurs at a ritardando (e.g., at the end of Variation XII), while the fastest virtuosic sections of the set go up to 145 BPM (e.g., towards the end of Var XXXII)[4].

The WoO 80 set presents virtuosic passages, which are indeed a technical challenge for a performer, not only because of the fast tempo in which they are executed but also because of the piano techniques required, of which García Stan (2015) makes an inventory. Among many possible cases, two examples of contrasting tempo but considerable difficulty are provided: (1) Though Var IX is played at a slow tempo, the performer must be able to make the *legato* melody salient while managing the polyphony and polyrhythms of the piece. (2) In contrast, Var XXIX needs the same texture control to make the melody salient, this time upon an energetic, virtuosic execution and *fortissimo* dynamic, where fast contrary motion arpeggios with triplets (on both hands) are combined with octave jumps.

One interesting historical aspect is that the 32 Variations were composed on a 5½ octave Erard piano, which ranges from F1 to C7. Beethoven makes deliberate use of the full range of the piano in crucial moments of the piece (see Var III and Var XXXII in Figure 5.5), as noticed by Ferraguto (2019). Although the lowest note is used many times, we can highlight its first use in Var III, where the right and left hands arrive at their most distant position on the keyboard, of five octaves and a minor third. The highest note is reserved for the set's climax on Var XXXII, after a series of ascending and descending diatonic scales on the right hand and ascending arpeggios on the left, which marks the end of the first section of this last variation.

5.3.3 Groups of Variations

Individual pieces on the set can be grouped by different criteria, such as their mode (major vs. minor) or technical motivic patterns (the same ascending arpeggio on the right hand vs. the left hand). The following list details five large structural groups, described by Jost (1994), and eight smaller groups of what García Stan calls sibling variations. When the piano roll visualizations are used, noticeable visual patterns emerge from these similarities between sibling variations, as highlighted in Table 5.1 and Figure 5.5.

[4] A tempo jump to 184 BPM at one point in Var XI is disregarded because it occurs at a skipped beat.

Groups	Variations				
Group 1	Tema	V1 V2 V3		V4	V5
	V6	V7 V8	V9	V10 V11	
Group 2	V12	V13 V14		V15 V16	
Group 3	V17	V18	V19	V20 V21	V22
Group 4	V23	V24	V25	V26 V27	
	V28	V29	V30		
Group 5	V31	V32			

Table 5.1: Global organization of the Beethoven's 32 Variations in C minor as five large groups. Sibling variations within each group are highlighted in orange. Roman numerals are omitted to improve readability

- Five-part structure:
  1. Tema, and Var I through XI.
  2. Var XII through Var XVI. Var XII establishes this subset of variations in major mode with a simplified rendition of the Tema (e.g., no quintuplets, slower rhythm, and quieter dynamics).
  3. Var XVII through Var XXII. The group, which returns to minor mode, starts on piece 18 out of 33, the second half of the set.
  4. Var XXIII through Var XXX.
  5. Var XXXI and Var XXXII. Here, the second-to-last variation prepares the last variation, which itself is divided into multiple sections.
- Sibling variations:
  - Var I, Var II, and Var III. They share the same arpeggiated motive played by the right hand, left hand, and both in the first, second, and third variations, respectively.
  - Var VII, and Var VIII. These variations share an implicit melody played in octaves on the left hand.
  - Var X, and Var XI. They share ascending patterns that alternate between hands in both variations.
  - Var XIII, and Var XIV. These variations are linked by their baroque voice counterpoints.
  - Var XV, and Var XVI. These variations differ rhythmically in their triplet motives that transform into sixteenth-note motives on the right hand.
  - Var XX, and Var XXI. They are once again grouped by the mirroring of a pattern that alternates between the right and left hands.
  - Var XXVI, and Var XXVII. Variations that are related by double notes and octave jumps on both hands.

Figure 5.5: Beethoven's 32 variations in C minor WoO 80. Main visual panes showing the waveform in pale green, and notes of the piano roll in shades of blue, as presented in CosmoNote. The theme and Var XXXII are enlarged. Similar variations are highlighted in orange. Pieces in major mode are highlighted in cyan.

5.4 Results

This section compares uOT distances between boundaries from the seven conditions described in Section 5.2.4, that pooled annotations with/without audio information. We first examine the results by computing the mean and median distances from the WoO 80 set, using representative pieces as examples of the main tendencies. We then inspect the results from both ends of the uOT distances distribution, going deeper in specific cases. Section 5.3 already addressed the musical aspect of the pieces. Thus, descriptions mentioning musical pieces are referenced primarily in the context of their segmentation structure and the presence or absence of visual and aural cues.

Global results show how annotations with cross-modal (aural and visual) components rank above those with only one component, meaning there is a clear advantage to receiving multiple representations of the same information instead of only one. Results recounting individual differences highlight the prevalence of patterns derived from visual information and that auditory information prevails mostly to resolve specific contradictions. In most other cases, auditory information is best used to identify the music's lower-level, subtle segmentation structures.

5.4.1 Global Results

We computed distances for seven comarisons over 33 pieces of the set summarized in Table 5.2. Calculating the mean and median of these uOT distances produces rankings of grouped conditions. As shown in Table 5.2a and Table 5.2b, the smallest distance (where annotations are the most similar) occupies the first position of the ranking. In contrast, the largest distance (where annotations are the most different) occupies the last position. For context, the global mean of uOT distances is 1.43 units with a standard deviation of 0.58 units; see Table 5.2c.

Ranking	Comparisons	uOT distance
1º	AV vs. V	0.79
2º	AV vs. W	1.27
3º	AV vs. A	1.27
4º	AV vs. P	1.33
5º	A vs. V	1.47
6º	A vs. W	1.91
7º	A vs. P	1.97

(a) mean

Ranking	Comparisons	uOT distance
1º	AV vs. V	0.77
2º	AV vs. W	1.19
3º	AV vs. P	1.24
4º	AV vs. A	1.25
5º	A vs. V	1.41
6º	A vs. P	1.87
7º	A vs. W	1.88

(b) median

Descriptor	uOT distance
mean	1.43
std	0.58
min	0.50
25%	1.00
median	1.29
75%	1.77
max	3.79

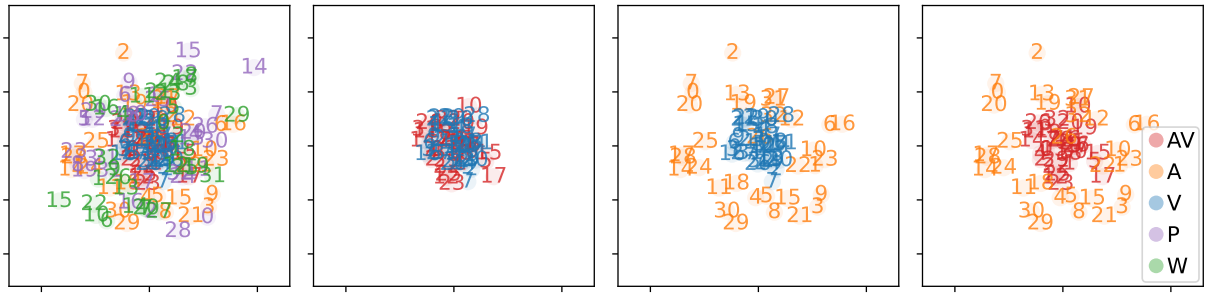
(c) descriptive statistics

Table 5.2: uOT distances rankings and descriptive statistics

First on the ranking is AV vs. V. Visual and aural (cross-modal) annotations are the closest to solely visual (unimodal) annotations than to other combinations. On the one hand, annotations between AV vs. V share the most attributes. On the other hand, annotations comparing A vs. P and A vs. W seem to share the least. The ranking obtained from calculating the median

across all pieces is the same, inverting the third and fourth and the last two positions. This ranking can also be seen on the right panel of Figure 5.4, as the red median lines of the box plots. However, the box plot representation hides how individual pieces relate to each distribution.

Figure 5.6 is a two-dimensional representation of the uOT distances between pieces in each comparison, filtered by grouped conditions in each panel to highlight the clustering of annotations with smaller distances. We focus on the relative distance of points in the two-dimensional space rather than on the individual position of specific pieces. However, indicative labels show pieces numbered from 0 (Tema) to 32 (Var XXXII). Figure 5.6a shows all pieces and grouped conditions, where outliers of visual conditions P and W are most evident. Two clusters, AV and V, are hidden in the first panel but highlighted at the center of Figures 5.6b, 5.6c, and 5.6d, which show only two conditions at a time to illustrate how they compare spatially. Figures 5.6b and 5.6d depict how visual/aural annotations are closer to solely visual than solely aural annotations. Visual information seems to provide cues that auditory information lacks, making pieces cluster (we will explore this claim by describing examples). The same interpretation can be made from Figures 5.6b and 5.6c, which show that adding visuals to the audio and not audio to the visuals makes pieces cluster.



(a) All conditions. (b) AV vs. V. (c) A vs. V. (d) AV vs. A.

Figure 5.6: The pairwise uOT distances between grouped conditions (colored legend) are mapped into a two-dimensional space to visualize their relationship using multidimensional scaling (MDS) (VanderPlas, 2016, p. 449). Each number represents a piece. Each panel shows the same data, filtered by how many conditions are shown: (a) All conditions. (b) Audio/Visual vs. Visual. (c) Audio vs. Visual. (d) Audio/Visual vs. Audio.

Comparisons between conditions exemplified in Figures 5.7, 5.8, and 5.9 examine how boundary annotations differ within a given piece. In all subsequent figures of this type, KDE boundary profiles of conditions with audio (AV and A) are drawn as solid lines, while shaded areas show boundary profiles of visual conditions (V, P, and W). Vertical dashed lines represent the boundaries removed by the uOT distance computation; they emphasize the timestamps where the two overlaid profiles differ the most. The rest of this chapter only details specific examples of the WoO 80 set to illustrate the differences in annotations for relevant comparisons in each piece. A compendium of figures for all the pieces in the study, displaying the seven comparisons simultaneously, can be consulted online[5].

[5]Go to https://doi.org/10.6084/m9.figshare.23937924.v3 to see the additional figures.

Three pieces (Tema, Var XII, and Var XXXII) are plotted to highlight the main tendencies of the median uOT distance ranking in Table 5.2b. The similarity in profiles between AV and V is evident in how the contours of their boundary profiles superpose compared to other profiles; see Figure 5.7a and Figure 5.7b. Even though the largest peaks are gathered around identical timestamps, subtle variations in the position of peaks indicate that distinct conditions (that may vary per piece) give the most information about the segmentation structure of a piece. The overall distance between two conditions increases as differences in the position of peaks increase. Figure 5.7c contrasts the first and last conditions of the ranking using an excerpt of Var XXXII. In more detail:

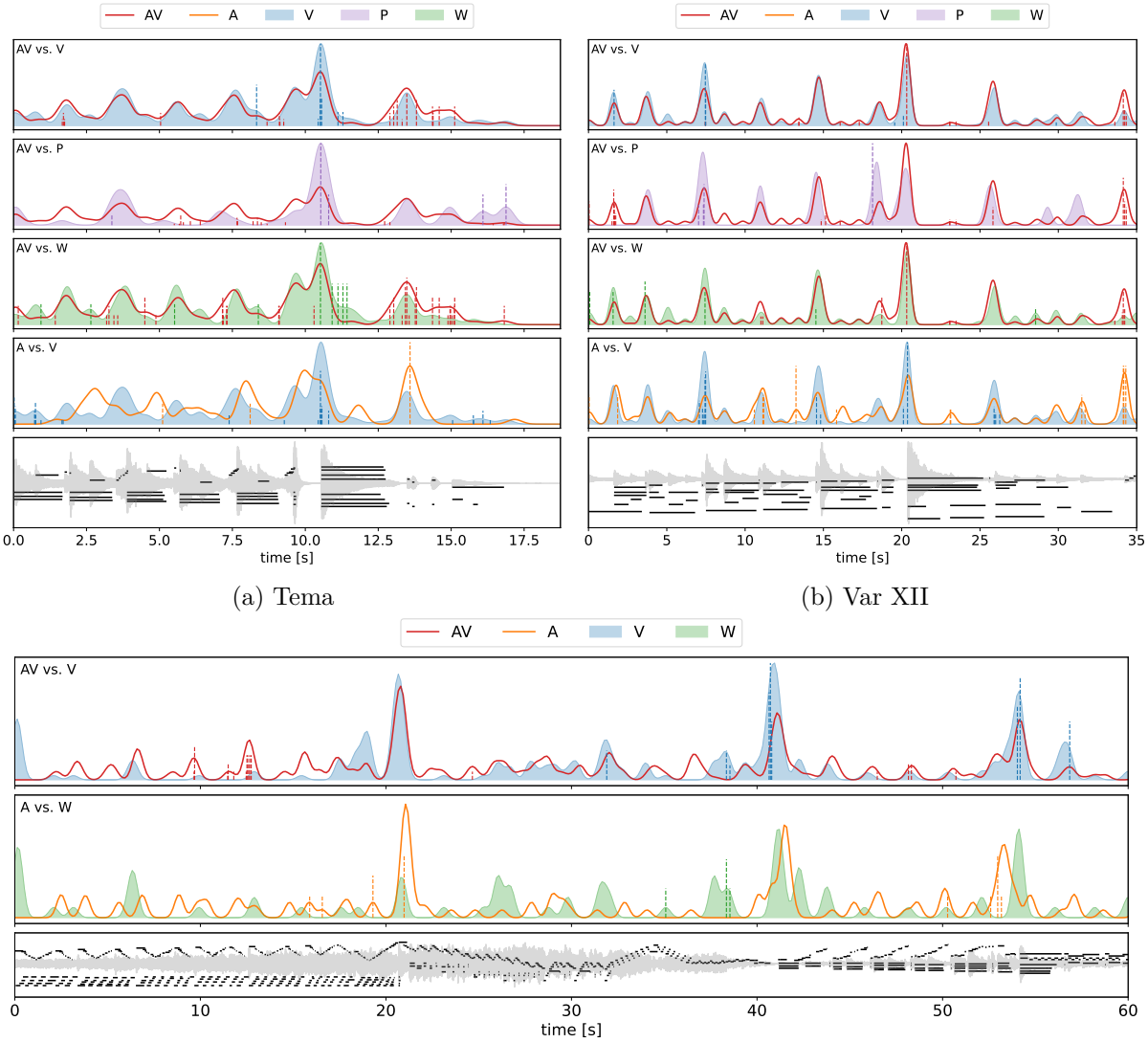
- Tema: Figure 5.7a shows the largest segmentation peak in profiles for all conditions is the *sforzando* at 10.5 s. The pause at 9.1 s, the change in dynamics to *piano* at 13.4 s, and the first quintuplet at 3.5 s are also worth noting. The main driver of the segments for visual annotations seems to be the chords on the left hand and the quintuplets. Segments annotated aurally shift the distribution peaks. For instance, they are focused on melody notes (on the right hand) and, lacking a visual cue, the sforzando is marked early.

- Var XII: Boundary profiles in Figure 5.7b appear roughly similar for all conditions. As before, the pause at 19.5 s and *sforzando* at 20.3 s are the most salient characteristics of the profiles annotated with the music. Visual annotations primarily, but also audio/visual annotations, assign more importance to the accent on the reiteration of the main motive at 7.5 s than audio annotations. As before, segmentation is driven by the first notes of the arpeggiated chords on the left hand.

- Var XXXII: Figure 5.7c zooms in a one-minute excerpt showing three sections of this piece. First, comparing the shaded blue and green areas in this example, the piano roll is predominant in visual annotations. The section change at 20.7 s is less evident with only the waveform representation than with the piano roll. This could be explained, for example, by the abundance of notes and overall crescendo making this change difficult to see in the waveform. Second, the pause and dynamic change to *pianissimo* at 42.7 s are evident in all representations. However, annotations with audio shift the placement of the boundary after the first C major chord, to the B3 eight note at 42.1 s. Third, even though the *sforzando* at 54.2 s on the left hand was visually more salient, annotations with audio placed a boundary on the harmonic motive on the left hand at 52.7 s that concludes with the *sforzando* in question.

5.4.2 Smallest Distances

For this ranking, we start by considering all pieces and comparisons equally. That is, 231 points from 33 pieces and 7 comparisons. Then, we keep only the smallest distances in each piece to see which pieces have a ranking that is different from the global ranking in Table 5.2. The smallest uOT distance is the comparison AV vs. A for Var XXXII, the set's longest piece. For this variation, the annotation profiles indicate that participants could obtain the most similar segmentation structure of the piece from cross-modal (visual and aural) or unimodal aural, instead unimodal visual cues (see Table C.1 in Appendix C for more details).



(a) Tema

(b) Var XII

(c) Var XXXII

Figure 5.7: Boundary annotations for the Tema (left), Var XII (right), and an excerpt of Var XXXII (bottom) of Beethoven's 32 Variations in C minor. KDE boundary profiles are drawn as either solid lines (conditions with audio) or shaded areas (conditions without audio). Boundaries removed by the uOT distance are shown as vertical dashed lines. Bottom panels of each subfigure display waveform and piano roll visuals.

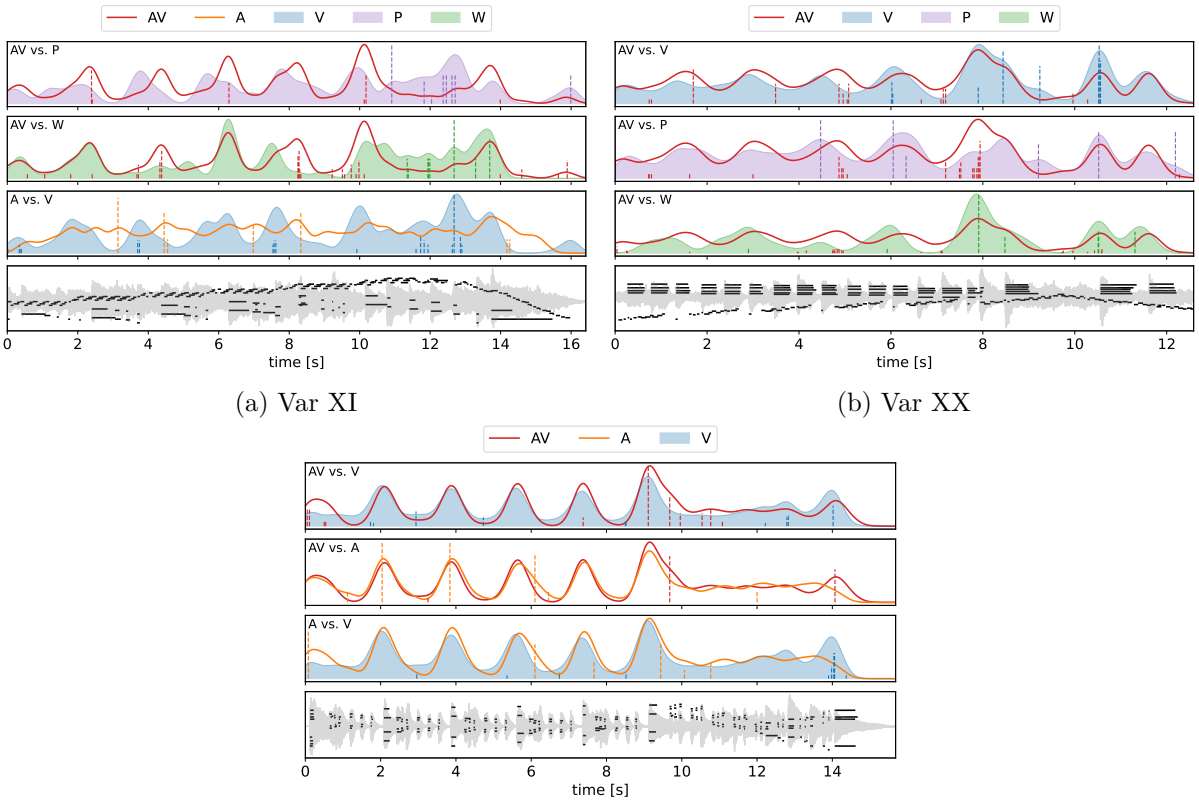
When grouping the pieces before ranking comparisons, i.e., ranking 7 comparisons for 33 pieces, the smallest uOT distances are between AV vs. V (27 pieces), AV vs. A (3 pieces), AV vs. W (1 piece), AV vs. P (1 piece), and A vs. V (1 piece). Here, one could argue that 32 out of 33 pieces are consistent with the results reported in Table 5.2. Figure 5.8 highlights the subtlety of these three exceptions, where AV vs. V is still ranked second or third. In more detail:

- Var XI: Annotations made using the waveform only noticeably resemble those placed with audio and visuals more than the other conditions (Figure 5.8a). The green shaded area shows that the absence of the music misleads annotators into missing the accent at 4.3 s and placing the accent at 8.2 s too early. The presence of the piano roll visualization and the music seem to shift the segmentation toward the right-hand patterns at the bar level. Additionally, the note visuals enhance a beat the performer skipped at 12.6 s, which may

have been visually mistaken for a pause (with a peak in the distribution only when the notes were present) but was aurally less striking in context, with the left hand continuing the descent.

- Var XX: As seen in Figure 5.8b with the green shaded area, the waveform favors the accented chords on the right hand. It is only with the addition of the piano roll visuals that the segmentation on the melodic motifs of the left hand is more apparent to the participants.

- Var XXVI: As with Var XII, all conditions have similar boundary profiles. Figure 5.8c shows that unimodal visual conditions did not lead to segmentations capturing the performance. The performer chooses to linger on the first notes of the gallop figures at each bar, creating a difference in segmentation that is only resolved by listening to the music. Both the waveform and the piano roll visuals privileged the accented chords with octaves on both hands at the bar level, seen as similar transients on the waveform and vertical lines on the piano roll.



(a) Var XI

(b) Var XX

(c) Var XXVI

Figure 5.8: Boundary annotation profiles for Var XI (left), Var XX (right), and Var XXVI (bottom) of Beethoven's 32 Variations in C minor. KDE boundary profiles are drawn as either solid lines (conditions with audio) or shaded areas (conditions without audio). Boundaries removed by the uOT distance are shown as vertical dashed lines. Bottom panels of each subfigure display waveform and piano roll visuals.

5.4.3 Largest Distances

As before, considering all pieces and all comparisons equally, the largest uOT distance is the comparison A vs. P for Var XIV (see Table C.2 in Appendix C for more details). This result is consistent with Table 5.2a. Annotations for this piece, placed with audio only compared to those with the piano roll only, present the largest differences among all annotations. Additionally, the result suggests that the visual pattern created by the notes only led to annotations that conflicted with annotations done with audio only (see Section 5.4.1).

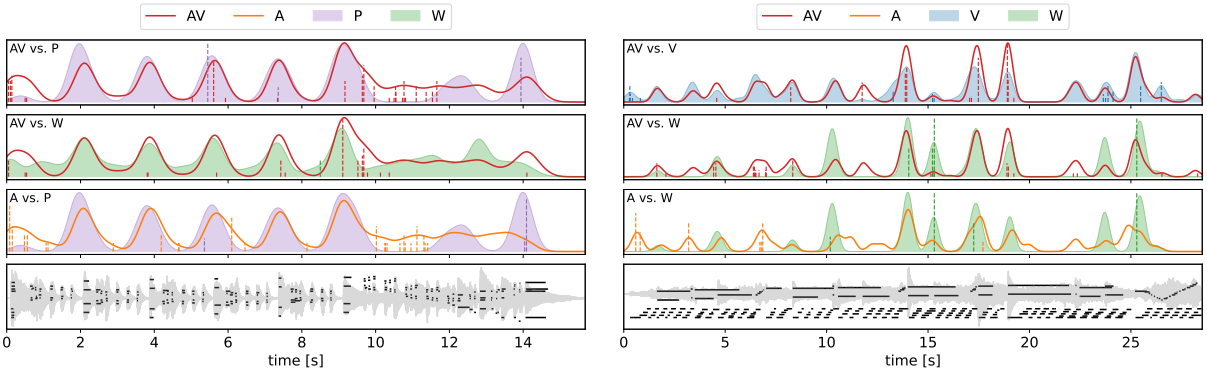
When grouping the pieces before ranking comparisons, the largest uOT distances are shared between three comparisons: A vs. W (16 pieces), A vs. P (15 pieces), AV vs. P (1 piece), and AV vs. W (1 piece). Here, the tendency for 31 out of 33 pieces is expected (Tables 5.2a and 5.2b), i.e., annotations placed only with audio are less similar to those placed only with one visual layer (either P or W). Two pieces, Var XXVI and Var XXXI, can be considered exceptions although they are not too far from the global trend. For Var XXVI, A vs. P is the second-largest distance after AV vs. P, while for Var XXXI, A vs. W is the second-largest distance after AV vs. W. For these outliers in Figure 5.9, annotations made from a particular visual layer (P or W) were closer to audio alone than to audio paired with visuals. More specifically:

- Var XXVI: This time, Figure 5.9a focuses on comparing unimodal visual conditions (P and W) to cross-modal and unimodal aural conditions (AV and A). In this case, the piano roll visualization favors two peaks (12.2 s and 14 s) over the waveform. As mentioned before, this segmentation is best understood when hearing the music. Accents during the descending motives between 10 s and 12 s seem to be missed by listeners, as evidenced by the concentration of destroyed boundaries (dashed lines) in that zone. The transients on the waveform convey more of the segmentation here than the piano roll, which is insufficient to emphasize prominence.

- Var XXXI: Differences between the green shaded area to the red and orange solid lines in Figure 5.9b show that annotations created solely with the waveform visual lack boundaries marked when audio and the notes were present. We see that, for example, the syncopated sixteenth notes (at 3.3 s or 22.2 s) and the thirty-second-note quintuplet (at 6.3 s) are not noticeable with the waveform alone. Adding the notes makes these notes more salient, possibly competing with the accents that participants heard with the music alone, which are compatible with the more evident transients of the waveform.

5.5 Discussion

Visual representations (piano roll and waveform) contain, by definition, less information than their audio source. In other words, visual representations are a reduction or a simplification of more complex data. They are (as explained in Section 2.2.1) purposefully conceived to emphasize certain aspects of the sound while neglecting others. For example, the music's tempo, the instrument's timbre, and the room acoustics are only accessible through hearing. (1) Piano roll representations may assist in quickly spotting visual patterns like groups of notes that form a chord or an ascending scale but do not provide cues to the duration and intensity of the attack/release of notes. (2) Waveform representations may help to estimate properties like transients, loudness, and pauses more accurately, but they lack information about specific notes. Visual layers thus accentuate crucial information and divert the annotator's attention from music features that may not be necessary for a specific analysis but constitute crucial components of a performance. In this sense, adding visuals to the audio gives people more support, a form



(a) Var XXVI (b) Var XXXI

Figure 5.9: Boundary annotation profiles for Var XXVI (left) and Var XXXI (right) of Beethoven's 32 Variations in C minor. KDE boundary profiles are drawn as either solid lines (conditions with audio) or shaded areas (conditions without audio). Boundaries removed by the uOT distance are shown as vertical dashed lines. Bottom panels of each sub-figure display waveform and piano roll visuals.

of cognitive scaffolding for accurately identifying otherwise less salient patterns and marking a given boundary's start and end times. However, as expected, removing the audio component leads to frustration and complicates the task by obscuring subtle cues that are only available through hearing, as reported by some participants. They reported that annotating uniquely visual representations of the music made them feel lost, frustrated, and less confident about the cues they used to identify boundaries. This uncertainty would explain why annotations created with only one visual layer seem the most limited, as observed in the aggregated profiles.

To insist on the advantages of adding visual layers for structure annotations, it would seem that annotations created with more information are more accurate than those where less information was available. Gingras, Rowland, and Stein (2009) argue that performance enhancement in tasks with cross-modal stimuli is not due to simple target redundancy (two stimuli rather than one) but is likely due to increased error reduction. For instance, boundary placement after a pause is made more accurate by following visual cues that indicate the start of notes, which are missing in unimodal audio conditions.

We have seen that the first position of the rankings in Table 5.2 suggest cross-modal stimulation produces annotations sharing most of their properties with unimodal annotations created by combining visual representations. This result indicates a prevalence of visual stimulation over aural stimulation across pieces, consistent with the cognitive scaffolding paradigm and the work of Platz & Kopiez and Tsay on cross-modal stimulation in music perception. However, the second, third, and fourth positions in that ranking are more dependent on individual characteristics of pieces, i.e., there are some pieces where the annotations profiles improve by focusing on different visual components or even by removing the visuals altogether. How can we interpret this?

Results, where the distances AV vs. W are small, are (on average) the second most similar comparison, followed by AV vs. A and AV vs. P, with individual fluctuations (see Figure 5.6d). For certain pieces, such as Var XI, the information that the waveform contributed to the segmentation structure is predominant over that derived from the notes or the audio. For others, conflicts between audio and visuals (in the cross-modal condition) are resolved to favor the audio. Since AV vs. A is ranked consistently after AV vs. V, the tendency of visual over aural sensory preference is not contradicted.

Some of the challenges we faced are inherent to the design choices assumed by this study. We will discuss the generalization of our results, the role of musical abilities in our sample, and the choice of using the boundary annotation type only.

First, generalizing musical annotation results is complex because of the many unique characteristics of each stimulus. Instead of presenting various musical styles and practices, our study privileges the differences between modalities. It minimized the differences in stimuli, employing music with a consistent structural framework and short durations so that participants could annotate many pieces relatively quickly. Moreover, we counterbalanced the possibility of a perceived global structure emerging from the theme and variation form by presenting the 33 pieces in a shuffled order.

Second, we did not examine the differences between musician and non-musician participants, though we recognize the influence of specialization and familiarity with music notation and representation. For example, Møller et al. (2021) show how the specialization of brain structures in professional musicians makes their segmentation approach different from non-musicians. Nevertheless, there is evidence that musicians improve less with the help of visual cues than non-musicians in objectively quantifiable tasks, such as pitch detection (Møller et al., 2021). Participants with previous experience using waveform/piano roll representations may have had an advantage in our task over those without it. Such an advantage is not exclusive to musicians, who would be ahead in working with music notation. Indeed, both groups may have expressed similar learned cross-modal correspondences (e.g., increase in pitch and height) when looking at the notes of the piano roll (Spence, 2011). Ultimately, we chose to analyze annotations assembled from a balanced group with varying musical abilities.

Third, we intentionally restricted the interface to allow only boundary placement and not other structural annotations such as regions or groups. We wanted participants to focus on a more straightforward task. Likewise, we allowed the use of labels, though we did not take them into account for our results. Chapter 6 presents a study that includes boundaries, regions, note groups, comments, and their labels.

This chapter has focused on boundary annotations created using either unimodal (visual or aural) or cross-modal (visual/aural) stimuli. The objective of this study was not to evaluate the validity of the structures obtained in the results, even though we have examined several structural aspects of the music as separate entities and as a whole. Instead, we have quantified the distance between annotations using different modalities with an indicator that shows which samples differ the most between any two distributions. In addition, we have shown that visuals play a considerable role in shaping boundary annotations of a recorded performance. These results are vital for our citizen science project because they make apparent the underlying consequences of presenting (on the side of experimenters) and using (on the side of participants) visualizations in CosmoNote.

Even if visual cues are sometimes overemphasized or somewhat misleading, they often seem to have improved the overall result of annotation profiles. Thus, we ought to consider biases in cross-modality while designing annotation campaigns, but it is preferable to have visuals than not to have them. We can then think about ways to tackle these potential obstacles. For instance, in free annotation tasks, participants may deliberately choose to show or hide a visual component without being aware of the subtle changes that this action may produce in their annotations. To counterbalance this effect, we can take a few approaches. Beginner annotators may need to be shown specific visualizations in a defined order to minimize bias. Segmentation tasks can be started by a group annotating with restricted visual representations, and these annotations would then be given to a second group to refine without restricted visuals. Annotators' actions could also be tracked more precisely so that annotations created with different modalities are analyzed separately.

Work by Guichaoua et al. illustrates how loudness and tempo curves are used to estimate the segmentation structure of performed music. Designing experiments to study the same differences in cross-modal annotations with loudness and tempo data through our musical prosody annotation protocol would be interesting, considering the advantages and disadvantages of these different visual representations. For example, single curves are simpler to read because they encode information in only one dimension but, at the same time, they limit the possibilities of interpreting the musical meaning to that dimension. We also need to be aware of the possibility of saturation by using too many visual cues simultaneously, which may be studied in the future.

Chapter 6

Study 2: Analyzing Free-Form Musical Prosody Annotations

The study in Chapter 5 was centered around boundary annotation data. This chapter features a study that carries on with the theme of annotating musical prosody while broadening the visualization and annotation capabilities at the user's disposal. This study presents an analysis framework based on scalable annotation classification and novel data aggregation techniques.

6.1 Introduction

Annotations created in CosmoNote contain specific information about the annotation task, the music (or any other time series being annotated), and the annotator. However, the reliability of the resulting data is determined by its quality. For the citizen science approach, data quality depends on, among other factors, how well the instructions are planned and followed, the expectations and qualifications of annotators, and successfully analyzing the resulting data with validation and quality assurance processes (Balázs et al., 2021).

We evaluate how instructions are planned and followed through the experimental conditions of contrasting annotations when providing either minimal or detailed protocol instructions in an experimental design characterized as free-form, i.e., participants engaged with every CosmoNote visualization (piano roll, pedals, waveform, loudness, and tempo) and annotation type (boundaries, regions, comments, and note groups) at will to mark musical prosody. They could create text labels to complement their data. This experiment assessed how the core concept of musical prosody and our annotation task, defined in Chapter 4, were understood and applied in a first encounter with the complete protocol.

The participant's expectations and qualifications are considered by including annotators with diverse levels of musical expertise, defined using the criteria of our short musical questionnaire (see Section B.1).

Data validation and quality assurance are contemplated firstly via detailed aggregation methods for region and note group annotations and secondly through analysis methods comparing similar annotations according to their type, labels, categories, and properties. For example, from one participant to another, the same prosodic structure may have been marked by placing a boundary or note group, assigned the same or a different label, or created within a fifty-millisecond window discrepancy.

Text labels included in CosmoNote annotations contain complementary information about the motivations, reasoning, and details behind a particular marking on the music. Annotators may, consciously or unconsciously, use metaphorical (as opposed to technical) language referring

to music structure and expressiveness (Spitzer, 2015). For instance, these shortcuts are used by performers to facilitate sound control and production (Leech-Wilkinson & Prior, 2014). It is worth noting that even when transcribing chord labels, an arguably objective task, text labels differ between annotators due to an inherent annotator subjectivity (Koops et al., 2019). Even though they may carry meaningful information, we have not focused on their content until this point because they need a more involved analysis, as we will later illustrate.

Our study shows that annotations in CosmoNote can be analogous to rich annotations (X. Li et al., 2012), which can be used for data-driven analysis, that their aggregation generates similar profiles for novice and expert participants, and that even with minimal instructions, participants understood the concept of musical prosody and its functions, allowing them to annotate expressiveness in performed music. We limit the scope to the aggregation and comparison of annotation results, analyzing how they fit together without diving into the details of individual or collective observations.

The rest of this chapter is organized in the following manner: Section 6.2 describes the design and methodological considerations for the study. Section 6.3 inspects the two musical pieces within their historical context and examines their essential musical structure elements. Section 6.4 presents the results of the annotation analysis as a framework that can be implemented in other studies and scaled in a citizen science approach. Finally, Section 6.5 interprets the results, presents conclusions, and discusses improvements and future work.

6.2 Materials and Methods

6.2.1 Participants

We collected annotations from 116 participants (69 female, 46 male, 1 unspecified) at INSEAD-Sorbonne Université Behavioural Lab. The following age groups were represented: 55.17% between 18-24 years old, 34.48% between 25-34 years old, 2.59% between 35-44 years old, 5.17% between 45-54 years old, and 2.59% between 55-64 years old. The group was divided into musicians (57%) and non-musicians (43%). Musicians reported at least 5 years of musical practice and 1 year of formal music theory training. We use the musician and non-musician categories for simplicity while acknowledging that differences between the two groups for processing complex auditory information go beyond a simple dichotomy and are more specific, nuanced, and task-dependent (Susini, Houix, Wenzel, & Ponsot, 2022).

6.2.2 Musical Stimuli

In contrast to the 33 pieces presented for the study in Chapter 5, only 2 pieces, performed by Elaine Chew, were presented in this study. Participants were tasked to annotate an excerpt of Edvard Grieg's piano arrangement of Solveig's Song[1], and Pierre Boulez's *Fragment d'une ébauche*[2]. A detailed description of the main structural characteristics of the music stimuli is presented in Section 6.3, in context, to convey how results are interpreted in this study.

6.2.3 Experimental Task

The study used the annotation protocol described in Section 4.3 and the same primary stages presented in Section 5.2.3: Before the experiment, participants completed a self-reported ques-

[1]Go to `https://doi.org/10.6084/m9.figshare.23732643.v1` to hear the performance (Figure 6.1).
[2]Go to `https://doi.org/10.6084/m9.figshare.23732640.v1` to hear the performance (Figure 6.2).

tionnaire on musical abilities, calibrated the audio levels, and received basic training on the interface and task. Then, they annotated musical prosody in performances according to instructions (see Section D.1) defined by the experimental conditions. Finally, they filled out a feedback questionnaire.

6.2.3.1 Experimental Conditions

As previously mentioned, the 116 participants were divided equally based on their musical abilities into two large groups of experts (33 musicians) and novices (25 non-musicians) per piece. These groups were then subdivided around the description of musical prosody concepts:

1. *Less detailed*, where participants have minimal information about the definitions used in our protocol. No explicit prosody label names or definitions were given. See Figure D.1 in Appendix D for more details. The rationale for creating this experimental condition is to evaluate our data's internal robustness and consistency if participants receive less information, do not read instructions, or fail to follow them properly during an experiment. To prevent this behavior during an active citizen science campaign, other annotators or researchers would intervene to prevent deviations from the protocol.

2. *More detailed*, where participants can access information about individual concepts in our protocol and examples of these structures. Explicit structure names, prosody examples, and explanations were given according to the definitions in Chapter 4. See Figure D.2 in Appendix D for more details. This condition is the standard for the citizen science applications of our protocol.

6.2.3.2 Interface Configuration

According to the experimental protocol, the interface was configured in the following manner: No restrictions to the playback, visuals, or navigation controls were applied. However, to counterbalance order effects, the two stimuli were shuffled. In the same way, to reduce possible biases, information about the piece, such as title, performer, and composer, was always hidden. Participants could freely control waveform, piano roll, pedals (when available), loudness, and tempo visuals. While only the waveform and notes were visible by default, participants could toggle the remaining visualizations and were shown how to do so. Participants were free to use boundaries, regions, note groups, and comments to fulfill their task and could choose to include text labels with their annotations in either English or French. Participants in both experimental conditions had access to a printed guide summarizing the most common actions for placing annotations in CosmoNote (see Figure D.3 in Appendix D).

6.2.3.3 Experiment Duration

From what we learned in Pilot Study 2 (see Section 3.5.3.2), it would take participants 45 minutes on average to fully annotate (using the four annotation types) a 2-minute piece. Because of the increased number of annotation types available, we compensated participants for 70 minutes of annotation time among both pieces while allowing a few annotators to finish outside the allotted time.

Participants took different amounts of time to complete the experiment. As in Chapter 5, the experiment duration was tracked using annotation timestamps, recorded by collection. This study calculates the total duration from the earliest to the latest dates among all annotation

types. Again, questionnaire completion times were not recorded. The experiment's mean duration was 35 minutes, with a standard deviation of 10 minutes. The median value was of 34 minutes. The shortest time was 10 minutes, and the longest was approximately 1 hour.

6.2.4 Statistical Methods

The main data collected from participants are the four annotation types (boundaries, regions, note groups, and comments), each with specific properties (see Section 3.5) and optional text labels. In this experiment, the four annotation types are interconnected to mark prosodic structures in performance. The data cleaning process consisted of pre-processing labels and detecting outliers. The following paragraphs describe this process and its implications.

The pre-processing stage started with raw annotations. While participants were allowed to write in their native language (French for most participants), some created labels in English, sometimes including technical musical terms in Italian on their own or mixed with plain French or English. We corrected spelling errors, removed abbreviations and contractions, and then translated labels and comments to English using the `deepl` Python API[3]. Musical terms in Italian, when used, were preserved. The `deepl` library offers a good compromise between accuracy and computation time. However, the resulting translated labels had to be revised manually and rectified when the translation or the context was not respected. For example, the French word *accord* was incorrectly translated many times as 'agreement' instead of 'chord' (more on this in Section 6.2.4.2).

Outlier analysis was handled differently than in Chapter 5 because users could access all annotation types. We used a ±25 ms time window from the timestamps stored with boundaries, regions, and comments to detect potentially repeated annotations. For groups, repeated annotations were detected by comparing whether the same number of notes (pitch and durations) with two different group identifications (IDs) were annotated by the same participant. All repeated annotations were marked as outliers and removed from the data except those with distinct labels (duplicated on purpose).

Even without outliers, the diversity of structures marked by participants obscured global tendencies. For example, Figure D.11 represents all note group annotations for both experimental conditions in Grieg's romantic piece. The opacity of each note is analogous to its frequency in the dataset. Although some notes appeared fewer times than others, finer details about most note groups were lost[4]. While it is possible to focus on the annotations of an individual to analyze their data, this defeats the purpose of large-scale data collection. Thus, ancillary filtering of annotations was needed to obtain a cohesive characterization of pooled annotations and their role in the music.

We split the data analysis of all annotation types in CosmoNote (boundaries, regions, comments, and note groups), focusing on three approaches: common labels, common categories, and common properties. Section 6.4 presents the results using the same separation, while the techniques used to obtain the data for each approach are detailed as follows:

6.2.4.1 Common Labels

This subdivision of the data is based on the presence of a given label in the dataset for any annotation type, for example, the word "pause". It is the easiest to implement as it only requires

[3] https://pypi.org/project/deepl/

[4] Appendix D contains other figure examples of all annotations in the dataset (boundaries, regions, comments, and note groups) plotted at once per piece.

filtering out all annotations that lack the searched term and, by definition, annotations without a label.

6.2.4.2 Common Categories

This subdivision is also based on the annotation labels and therefore filters out annotations without labels. Moreover, this technique extends the previous one by creating annotation categories and classifying individual labels into one or many categories.

In CosmoNote, annotations with labels are analogous to rich annotations, defined by X. Li et al. as additional information provided by human annotators. Rich annotations are often used for classification problems. They are subdivided into three levels: Level 1 provides simple evidence (highlighting a relevant concept), Level 2 provides generalized evidence (indicating links to broader concepts), and Level 3 provides annotation metadata (making judgments and detailed observations). Likewise, we can divide labels in our experiment based on how much information is contained within them: (1) simple labels consisting of one concept, word, or symbolic identification, e.g., "A1" or "tipping point", (2) title labels composed of a sentence indicating a theme, e.g., "Change of rhythm and tone", and (3) complex labels with context and a deeper understanding of the material, e.g., "The tempo is set in motion again a little more driven after a pause on the half-cadence and the nuance is also reinforced, in a fine overdrive". The analogy between rich annotation levels and our labels is not perfect. However, it allows us to appreciate the value of text labels and the possibility of classifying them to gain insights applicable to our research.

We applied elementary Natural Language Processing (NLP) tools to perform automatic multi-label classification on annotations with labels. We describe two approaches: The first is unsupervised, meaning there was no human intervention in the categorization step. The second, supervised approach, is outlined but remains out of the scope of this work.

6.2.4.2.1 Basic text classification

The following steps describe how we pre-processed labels using the `NLTK`[5] Python library. We then proceeded to classify them into predefined categories and subcategories. The example in Table 6.1 shows how each step works:

Step	Result
Original	`"Variations of theme A'."`
Tokenized	`["Variations", "of", "theme", "A'", "."]`
Without punctuation	`["Variations", "of", "theme", "A'"]`
Without stop words	`["Variations", "theme", "A'"]`
Lemmatized	`["Variation", "theme", "A'"]`
Subcategories	`prosody_segmentation, melody, form_and_genre`
Categories	`segmentation, structure, music_descriptors`

Table 6.1: Example of text classification pre-processing steps and results.

1. Tokenization and punctuation removal. This stage consists of separating individual words into tokens. Some tokenizers assign a part-of-speech (POS) to each token, e.g., noun, verb, adverb. However, a bespoke tokenizer was used because automatic tokenizers separate non-alphabetic characters into different tokens. For example, the apostrophe on the label A' usually denotes a prime version of label A and must be kept (see Section 2.1.2). Punctuation characters were removed.

[5] https://www.nltk.org/

2. Stop word removal. English words that are frequent but not meaningful in speech, e.g., articles ('the', 'a'), prepositions ('of', 'over'), or single letters separated from contractions ('s', 't') are removed. As in the previous step, the indefinite article 'a' often denotes music structures, so we encoded structural labels with uppercase letters and kept them.

3. Lemmatization. This stage reduces a word into its lemma, i.e., a word keeping the original meaning in its basic form, without gender or number alterations. With this step, words such as "melodies" and "melody" would have the same token, "melody".

4. Categorization. Labels are first assigned to a subcategory if a token is present in the list of terms of that subcategory. Then, a label is put in a larger category if its subcategory is part of it. For this study, we focused only on the *segmentation* and *prominence* categories.

Tables D.1 and D.2 in Appendix D detail the custom dictionary with 22 subcategories (keys) and terms (values) we created for this analysis. The second custom dictionary in Table D.3 contains the 7 larger categories (keys) gathering the 22 subcategories (values) from the previous dictionary, that would be assigned and kept with this technique. While subcategories (from the first dictionary) are specific and valuable for refined analysis, they contain many overlapping terms. Fewer, larger categories (from the second dictionary) were used to facilitate visual inspection of the categorization results and could be more easily verified manually for a supervised classification task.

The classification obtained with this technique provided an acceptable first approximation of the categories for the analysis. However, it was limited by hard-coded terms in the category dictionaries and therefore ignored the labels' context and semantic information. An option to incorporate semantic information is to use synonym sets (*synsets*). Another is to represent words as vectors in a multidimensional space. Once a word was converted to an alternative representation, a similarity comparison between tokens in our dataset and terms in a subcategory was added at the last phase of the categorization method. Despite adding a few more correct classifications, synonymy and word vectorization were unreliable. The algorithm cannot process words with different POS, while vectorized words are too abstract outside their original phrases. Other limitation examples of unsupervised text classification approaches comparing word by word, without context, can be seen in Table 6.2.

6.2.4.2.2 Supervised Classification

Supervised classification techniques can be used to consider the annotation's context. A greater contextual understanding can be achieved by training a language model to generate vector embeddings for complete annotations rather than individual words independently. These embeddings will be representative of our categories and their context in a multidimensional space. Instead of training a model from scratch, fine-tuning a pre-trained language model such as BERT[6] from the `transformers`[7] Python library is becoming standard practice (Tenney, Das, & Pavlick, 2019). However, a prerequisite of the supervised approach is having already classified a portion of the dataset that will be used for training, which is why it is outside the scope of this work. In any case, we outline the steps to perform supervised classification on text annotations (Section 6.5 discusses leveraging rich annotations to perform supervised classification):

1. Create three subsets of the original dataset by applying a training-validation-test split (Géron, 2022), where annotations are distributed evenly between each partition, typically 80% for training/validation and 20% for testing.

[6] Bidirectional Encoder Representations from Transformers.

[7] `https://pypi.org/project/transformers/`

Phenomenon	Importance of having context	Examples
Translation	Understanding foreign language translations	*accord* (French) ≡ chord chord ≠ agreement
Mispellings	Discriminating between a typographical error and a distinct word	tone ≠ tune ≠ one
Synonymy	Identifying if different words have the same meaning	boundary ≈ frontier ≈ limit
Polysemy	Recognizing when one word has multiple meanings and choosing the correct one	piano (instrument) ≠ piano (dynamic marking)
Quantification and negation	Noticing how added words can change what an annotation means	not happy, less pretty
Loan words	Keeping the original foreign word's meaning instead of translating it	*allegro* (fast tempo) ≠ allegro (Italian for "cheerful")
False cognates	Telling apart words that look similar but have a different meaning	outro (closing section) ≠ outro (Portuguese for "other")
Sarcasm	Realizing when something means the opposite of what is written	This is just great
Figurative language	Judging what is meant without it being said directly	Electronic meteorites with high-pitched embers

Table 6.2: Examples showing how context can change the meaning of text annotations.

2. Manually classify the training data into the categories of Table D.3.

3. Fine-tune the pre-trained model using the manually-classified subset on the validation subset and iterate, if needed, by changing the model's parameters.

4. Classify the test subset (annotations the model has not been trained on). The performance on the test subset should be equivalent to those of the validation subset.

6.2.4.3 Common Properties

This subdivision of the data is not based on labels but on other properties captured in the metadata (see Section 4.4.1) by the less detailed and more detailed experimental conditions. Annotations with common properties are extracted by annotation type according to the following steps (see Section D.2 for more on these techniques):

1. Compare annotations by pairs. Similar annotations are defined based on time similarity (for boundaries, regions, and comments) or their Jaccard similarity coefficient (for note groups).

2. Create a graph of similar annotations.

3. Extract clusters of similar annotations from the annotation graph using the Louvain community detection algorithm from the `networkx`[8] Python library.

4. Similar annotation clusters are aggregated by musical ability. For example, similar aggregated regions will be represented by tuples with the clustered annotations' mean start/end times. In contrast, similar aggregated note groups will be represented by sets of unique note tuples with a note's start/end times and MIDI number.

[8] https://networkx.org/

6.3 Structural Analysis in Pieces by Grieg & Boulez

The musical stimuli presented in this study come from periods with contrasting intentions, tempo, dynamics, techniques, and musical framework. Combined with our annotation protocol, they were chosen to maximize the number of annotations that capture expressiveness over score structures. For example, the romantic piece was shortened to facilitate the annotation task in the context of the experiment's limited time. The contemporary piece, in particular, controls for annotations with a heavy reliance on the harmonic cues participants may get from their implicit knowledge of Western tonal harmony.

6.3.1 Solveig's Song

Solveig's Song (*Solvejgs Sang*, in Norwegian) Op. 23 No.19 is a romantic era lied composed in 1875 by Edvard Grieg[9] (1843-1907) and premiered as a part of the incidental music for Henrik Ibsen's play Peer Gynt in 1876. It was a lament for the character of Solveig, who tragically longs for Peer Gynt's return. The recorded performance used for this study is the Op. 52, No.4 piano arrangement by the composer[10].

6.3.1.1 The Song Form

Songs are musical forms that span from the use of voice and music. The word song portrays many musical forms that have evolved from medieval compositions to modern pop songs. The piece used in this study has more in common with the German denomination of *lied*, which is often used to designate songs of the romantic period with emphasis on emotional and dramatic expression (Kennedy & Kennedy, 2012).

Songs of this type are usually composed around the text of a poem. They can be either strophic, repeating the same musical ideas with different text, or through-composed, presenting original musical ideas for each stanza (arrangement of lines) of a poem (Abromont & de Montalembert, 2010). According to Jarrett (2003b), Grieg's songs are composed as strophic or through-composed, depending on the length of the poems he utilized.

6.3.1.2 Structural Components

Solveig's Song is composed of a typical ABA form with an introduction and a coda. The sections of the verse and refrain will be referred to as the A and B sections, respectively. The piece is written in A minor, modulating to A major through the B section. The accompaniment, played by the left hand, is characterized throughout the piece by open fifth chords that evoke Norwegian folk music. The same can be said for the melody, which was possibly inspired by the Norwegian folk song "I lay down so late" (*Jeg lagde mig saa sildig*) (Foster, 2007).

For this study, we chose a 2-minute excerpt cut by the performer to maintain musical sense while being significantly shorter than Grieg's approximately 5-minute piano arrangement of the piece. The excerpt used in this study skips the short introduction of the original piece, beginning at the start of the A section (measure 8 of the score) and ending after the first instance of the B section (measure 39 of the score). All possible visualization options for this excerpt, as displayed in CosmoNote, can be seen in Figure 6.1.

[9] Grieg is considered a prominent figure of romanticism as well as Norwegian nationalism.

[10] Go to https://doi.org/10.6084/m9.figshare.24635589 to see the score of the excerpt used in this study.

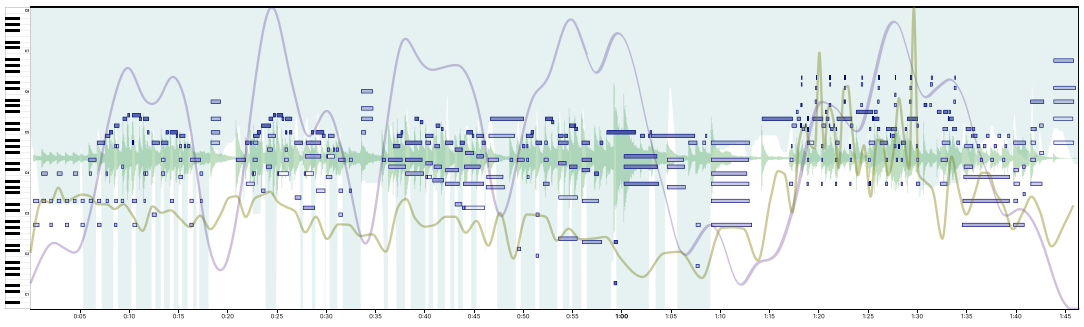


Figure 6.1: Edvard Grieg's Solveig's Song excerpt, as presented CosmoNote; all possible visuals are overlaid. Waveforms (pale green), notes (blue rectangles), pedal data (light blue area), loudness (mauve), and tempo (olive).

The A section starts with a 4/4 meter, an *Andante* tempo, and *piano* dynamics. Several pauses are marked after each line of the verses in the melody, which lay out an ascending and descending pattern with each phrase. The melody was composed so that the last four words of every verse were repeated throughout the song using similar motives but often different pitches. The piece modulates to its relative major (C major) for the poem's second line and goes back to the original tonality for the remaining portion of this section. Since Solveig's song is part of a set of incidental music, the melody of the A section is heard outside this piece, in the prelude of Act I '*I brudlaupsgarden*' (At the wedding) of Op. 23.

Although the B section was written to be sung on the interjection 'Ah', it is often called the humming section (Jarrett, 2003a). It is characterized by a change to the parallel major mode (A major), an arpeggiated drone accompaniment alternating between the tonic and the dominant, a change in meter to 3/4, and a change to a faster *Allegretto con moto* tempo. Jarrett argues that this section adds an element of optimism to the loyalty and devotion shown by Solveig as a character. The B section (and the excerpt with it) ends on an A major chord, two bars after the return to the original tempo, concluding the refrain. This chord marks the return to A minor, contains the highest note in the excerpt (an A6), and ends on a *fermata*, indicating a pause before the next section when listening to the entire piece.

The tempo curve computed from the recorded performance remains steady in the A section and is faster in the B section. At the same time, the dynamics delineate a slow fluctuation between soft and loud passages throughout the excerpt. Indeed, the performance's computed median tempo was approximately 68 BPM, which corresponds mainly to the overall slower tempo of the A section. The slowest passage of the piece (approximately 21 BPM) occurs at the closure of the A section, along with a *diminuendo*. This point is prepared by two salient events: (1) an accented C5 note –the loudest single note in the excerpt– with a recorded velocity of 88 out of 127, and (2) an E1 –the lowest note in the excerpt– played on the left hand. The contrast with the B section is highlighted by a faster execution, with high notes arpeggiated around the melody, which has to be made prominent by the performer. The fastest passage of the piece, approximately 203 BPM, occurs at an early hastening of the tempo, followed by a compensatory *ritardando*. This point marks the second-highest note in the melody, after the A6 on the final A major chord mentioned above.

6.3.2 Fragment d'une ébauche

The contemporary piece, *Fragment d'une ébauche*, translated as "Fragment of a draft" from the original French title, was composed by Pierre Boulez (1925-2016) in 1987 as a friendly token to Jean-Marie Lehn to commemorate his Nobel Prize in chemistry. Its first performance, in 2013, was part of a concert at 'Musica Festival' in Strasbourg. The piece was initially planned to be developed into a composition for piano and instrumental ensemble, without ever being released in that format, as many unfinished works in progress left by Boulez (O'Hagan, 2016). However, *Fragment d'une ébauche* may have served, after all, as the seed to other works by the author (Manoury, 2013). The piece was chosen for our study because of its short duration (approximately 30 seconds long), diversity of dynamic changes, and lack of tonal harmony.

6.3.2.1 Form in Boulez's Compositions

Because of the vast quantity of styles emerging in this period, music of the Western tradition composed around the beginning of the twentieth century falls under the umbrella term of contemporary music. Compositions of this period are defined by one or more of the following characteristics: breaking free of the limits of tonality, focusing on timbre and electronic sounds, developing algorithmic composition, conceiving complex rhythmic structures, and exploring conceptual art in music, among others (Ammer, 2004).

Pierre Boulez is one of the most significant figures of the contemporary music period. He was highly influenced by the rigorous and systematic approach of serialism techniques, which he applied in his early compositions. However, the composer eventually developed a personal style that incorporated poetic and emotional aspects of music. We are interested in Boulez's work after 1970, where his music transformed as he reconciled the strict and unfettered aspects of his compositional methods. Here, Boulez's music arises from the play of oppositions between strict, rigid, structural listening and open or 'aleatoric', resonant listening. These oppositions come from the acoustic characteristics of the instruments present in a piece and inform his work's structures, often making them difficult to understand. Nonetheless, in compositions of the last quarter of the twentieth century, Boulez is more concerned with how listeners perceive his music and thus combines clear and obscure structures. In the years surrounding his intervention at *Collège de France*, he explains how form can be thought of as a recomposition process by the listener's perception and memory (Ammer, 2004; Goldman, 2011).

Since Boulez aimed to redefine musical language through his compositions, analysis of his work should come from consulting sources characterizing specific pieces and considering each piece in isolation rather than applying generic analytic tools (O'Hagan, 2016). Regardless, as Ammer noted, we can mention how soft-loud dynamics, rapid figurations, drones, trills, and heavy percussion elements are commonly featured in Boulez's writing.

6.3.2.2 Structural Components

There is not a concrete form to which *Fragment d'une ébauche* adheres. Instead, it is a short, virtuosic, atonal composition from the late twentieth century. It seems to be written as a large arc or gesture building up to a climax and then descending to its conclusion. Visualization options for this piece, as shown in CosmoNote, can be seen in Figure 6.2.

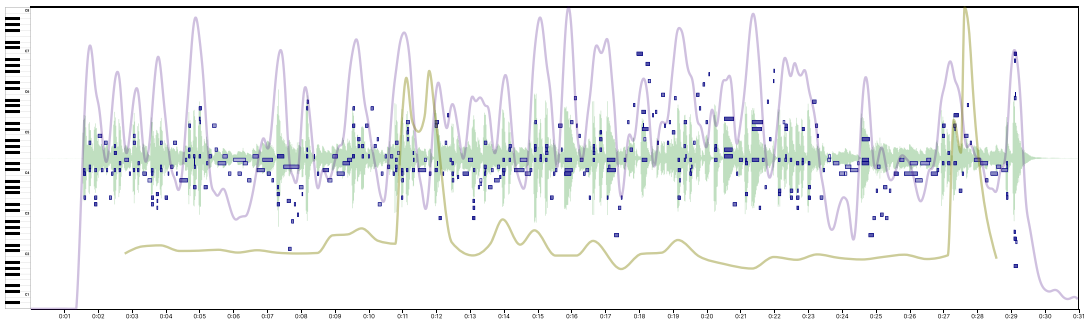


Figure 6.2: Pierre Boulez's *Fragment d'une ébauche*, as presented CosmoNote; all possible visuals are overlaid. Waveforms (pale green), notes (blue rectangles), loudness (mauve), and tempo (olive). Peaks in the tempo curve at 11 s, 12 s, and 28 s are outliers from the tempo extraction algorithm.

Analyzing the sketches that would reveal more information on this piece's compositional structure is beyond the scope of this study. However, we can comment on three main elements (here referred to as materials *a*, *b*, and *c*) used throughout the piece and shown in Figure 6.3: (*a*) accented sixteenth-note chord pairs, (*b*) ascending/descending sixteenth-note staccato triplets, and (*c*) legato sixteenth-note melodies. We will explore the relevant characteristics of these three materials while looking at the piece's dynamics, tempo, and pitch.

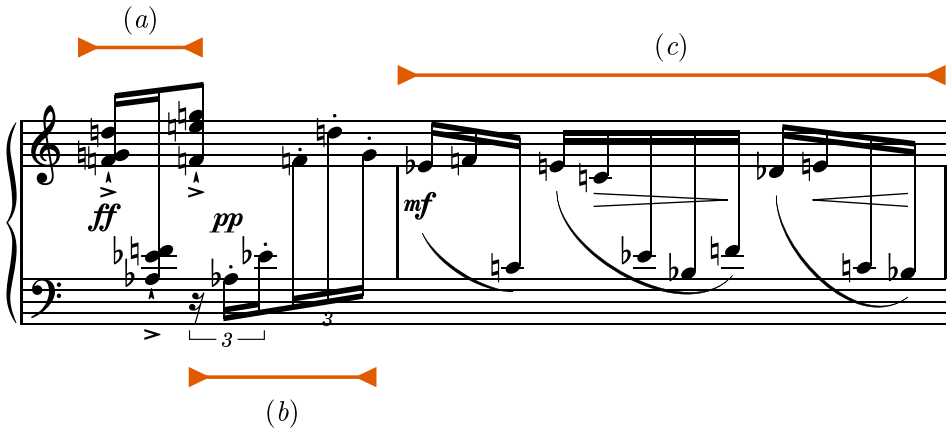


Figure 6.3: Score excerpt of Pierre Boulez's *Fragment d'une ébauche* showing materials *a*, *b*, and *c*. © Copyright 2013 by Universal Edition A.G., Wien / UE36098.

The piece alternates rapidly between distant dynamics, going almost instantly from *pianissimo* to *fortissimo* and back, challenging the performer. Material *a* features *fortissimo* dynamics, while materials *b* and *c* are *pianissimo* and *mezzo forte*. The mean and maximum velocities recorded for this performance are 74 and 102 out of 127, respectively. The last chord is notably one of the loudest instants in the piece.

An energetic 118 BPM per quarter note tempo, marked as *Extrêmement vif*, is maintained for the whole piece. The recorded performance has a median tempo of 106 BPM. In this context, sixteenth-note ternary patterns in material *b* are heavily contrasted with binary rhythmic patterns in materials *a* and *c*. Notably, most notes in material *a* chords are often expanded to form materials *b* and *c*.

The lowest note is an A1 and the highest is a B6, giving the piece a range of 5 octaves and a major second. The median note is E♭4 (D♯4). The music frequently uses pitches close to the median (D♭4 to F4) as central anchor notes, a concept similar to an axis of symmetry so that other notes expand from or converge to them. This oscillating motion is characteristic of all materials. Mirror images (symmetrical intervals to the left and right of anchor notes) in material *a* climbing and falling pitch motions in material *b*, and sparse use of interval inversions are also noteworthy.

6.4 Results

This section presents results from our free-form annotation experiment, showcasing a selection of notable annotations. We begin by doing an overview of all annotations, followed by a summarized description of the subsets of the data obtained using the techniques described in Section 6.2 for both experimental conditions, one musical piece at a time, drawing connections to the structures described in Section 6.3. This section does not attempt to comprehensively analyze the music structures through their annotations, which would need a deep dive into one specific technique and individual annotators' contributions. It is instead an exploration and characterization of the analysis techniques and the results that are possible when collecting and processing musical prosody annotations in CosmoNote.

6.4.1 All Annotations

Table 6.3 tallies all annotations in this study by their type. It also divides annotations according to participants' use of the label field after translating the text to English: We see that participants placed more boundaries (41.3%) than regions (27.5%) and note groups (25.3%); also that comments (5.9%) were used only sporadically. Close to one-third of all annotations had unique labels, i.e., they appear only once in the set containing all text labels, while one-sixth of all annotations had non-unique labels, i.e., the same label is used more than once. Unique and non-unique labels on regions and note groups amount to one-third of the annotation data. Another third of the data (31.7%) are boundaries, the largest portion of annotations without labels. In contrast, the smallest portion of unlabeled data (1.7%) corresponds to comment annotations, which should, by definition, contain text. Indeed, most comments had unique labels since they were defined to contain unrestricted text.

Annotation types	Without labels	(%)	Unique labels	(%)	Non-unique labels	(%)	TOTAL	(%)
Boundaries	1634	(31.7)	317	(6.2)	177	(3.4)	2128	(41.3)
Regions	631	(12.3)	475	(9.2)	311	(6.0)	1417	(27.5)
Note groups	609	(11.8)	466	(9.1)	225	(4.4)	1300	(25.3)
Comments	88	(1.7)	175	(3.4)	40	(0.8)	303	(5.9)
TOTAL	2962	(57.5)	1433⋆	(27.9)	753⋆	(14.6)	5148	(100)

Table 6.3: Translated annotation label counts by annotation type. Each column is computed from individual annotation types. ⋆When all annotations types are combined there are 1375 (26.7%) unique labels and 811 (15.8%) non-unique labels.

Figures displaying annotation results throughout this section are plotted separated by condition (less detailed or more detailed) and differentiated by musical ability (musicians shown in

red and non-musicians in blue), musical excerpt, and annotation type. Boundaries and comments are plotted as KDE profiles with solid and dashed lines. Regions are drawn as overlapping rectangles across the visualization pane. Note groups are drawn as squares or diamonds on top of the piano roll visualization.

Figures with all annotations are shown in Appendix D: Figures D.5 and D.6 for boundaries, Figures D.7, and D.8 for regions, Figures D.9 and D.10 for comments, and Figures D.11 and D.12 for note groups. However, since looking at all annotations without filtering decreases the interpretability of the results, we will focus on a sample of the results created from their common labels, categories, and properties.

6.4.2 By Common Labels

Among all annotation types, pieces, and experimental conditions, only four terms: "pause", "stress", "transition", and "tipping point" are used repeatedly as a complete label. This behavior was expected on the more detailed condition, where the terms (defined in Section 4.3) were mentioned as part of the instructions. However, some of these words, already part of common expressions in music, do appear in the less detailed condition.

In this study, users could use any annotation type to mark musical prosody. However, some annotation types were preferred over others. This behavior was expected because of our recommendations on the more detailed condition instructions (see Figure D.2), and therefore, we will focus only on showing results for the more detailed condition in this section. For example, we recommended regions (time selections) for pauses and transitions, boundaries and comments (timestamps) for stress or tipping points, and note groups for pauses, stress, and tipping points.

The following subsections show examples of annotations of the words "pause" and "transition" because they contain the two most representative samples of users using the same labels.

6.4.2.1 Pause

All annotation labels of any type containing the word "pause" are described as follows:

Solveig's Song

Annotations for the more detailed condition are shown in Figure 6.4. Boundary, comment, and note group annotations represent around 10% of participants, while region annotations represent around 40%. It is worth highlighting the convergence of regions in the same few moments in the piece. Notably, the concluding Am chords at each structural segment. Moreover, 7% of participants (distributed among all annotation types) also used this label on the same chords in the less detailed condition.

Fragment d'une ébauche

Figure 6.5 shows annotations in the more detailed condition. Regions represent 19% of participants, while the rest of the annotations represent 5%. The most consistent pause annotations are marked on repeated occurrences of the melody in material *c*. We remark that two participants in the less detailed condition marked the final three instances of material *c* as pauses.

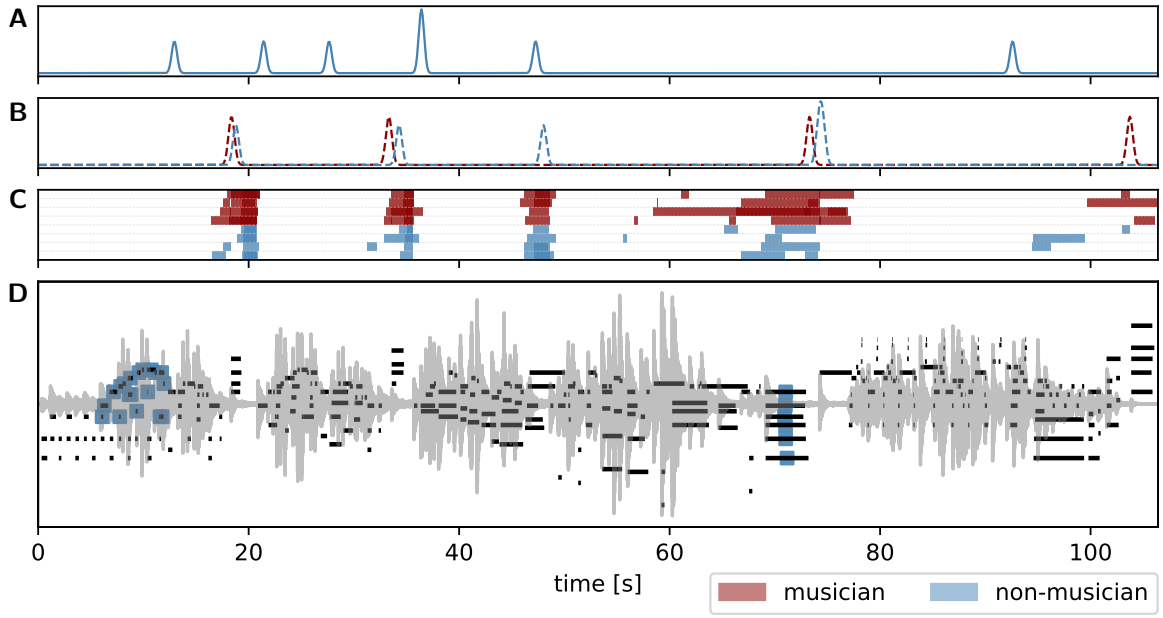


Figure 6.4: All annotations containing the label "pause" in Edvard Grieg's Solveig's Song for the more detailed condition. (A) Boundaries (solid lines). (B) Comments (dashed lines). (C) Regions (stacked rectangles). (D) Note groups (squares) overlaid on top of waveform and notes (gray curve and black lines). In all panels musicians are displayed in red and non-musicians in blue.

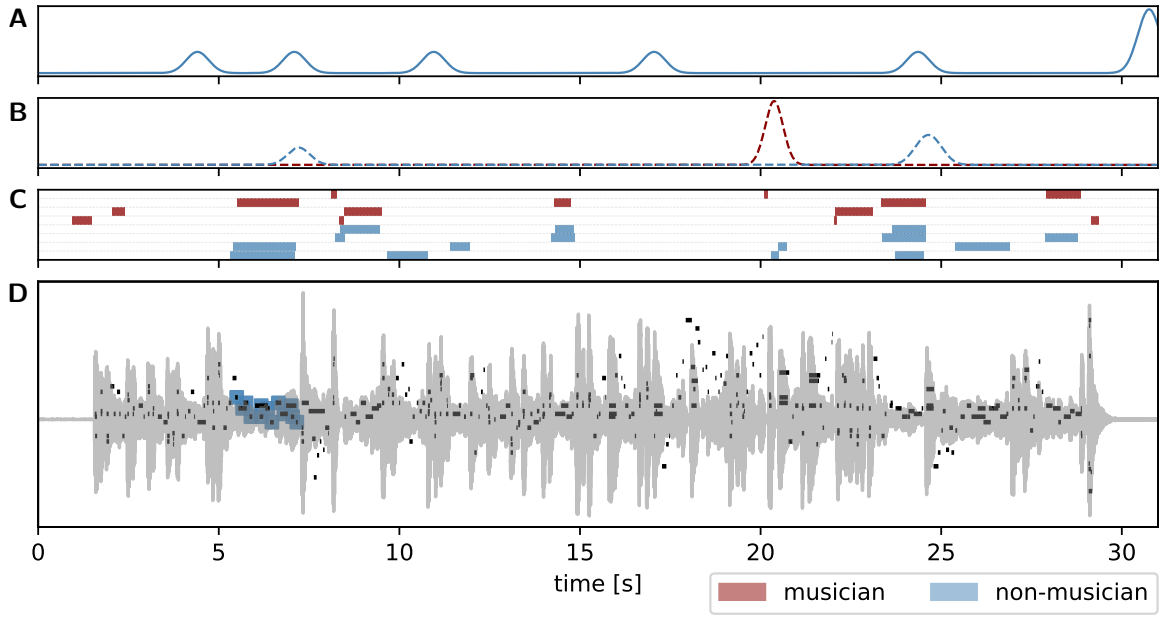


Figure 6.5: All annotations containing the label "pause" in Pierre Boulez's *Fragment d'une ébauche* for the more detailed condition. (A) Boundaries (solid lines). (B) Comments (dashed lines). (C) Regions (stacked rectangles). (D) Note groups (squares) overlaid on top of waveform and notes (gray curve and black lines). In all panels musicians are displayed in red and non-musicians in blue.

6.4.2.2 Transition

All annotation labels of any type containing the word “transition” are described as follows:

Solveig’s Song

Figure 6.6 shows annotations where participants marked transitions in the more detailed condition. Here, regions represent 31% of participants, while boundaries, comments, and note groups represent up to 10%. Most annotation types coincide with the connections between structural segments in the excerpt, emphasizing the conclusion of the A section. The same passages are marked in the less detailed condition, with regions representing 10% of participants. It is also worth noticing how different participants’ perception of the same transition starting and ending times varies up to 7 seconds.

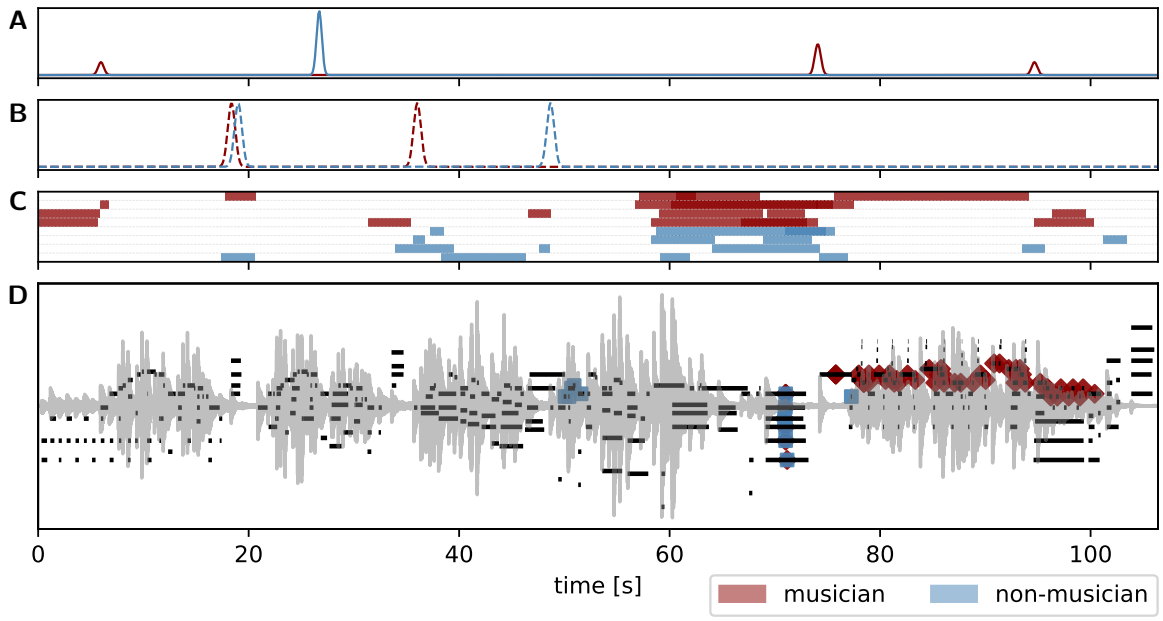


Figure 6.6: All annotations containing the label “transition” in Edvard Grieg’s Solveig’s Song for the more detailed condition. (A) Boundaries (solid lines). (B) Comments (dashed lines). (C) Regions (stacked rectangles). (D) Note groups (diamonds and squares) overlaid on top of waveform and notes (gray curve and black lines). In all panels musicians are displayed in red and non-musicians in blue.

Fragment d'une ébauche

Annotations shown in Figure 6.7 correspond to transitions in the more detailed condition. Here, 31% of participants are represented in the region annotations; up to 7% are represented in the rest. Many participants mark the first and third to last occurrences of material *c* as transitions, dividing the piece roughly into three parts.

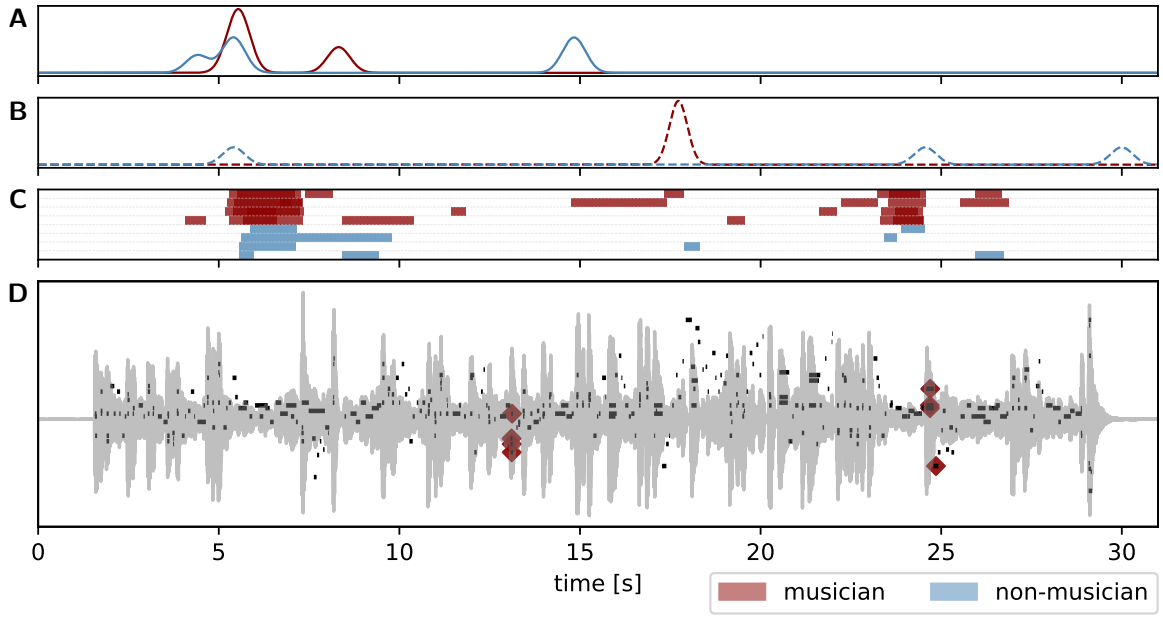


Figure 6.7: All annotations containing the label "transition" in Pierre Boulez's *Fragment d'une ébauche* for the more detailed condition. (A) Boundaries (solid lines). (B) Comments (dashed lines). (C) Regions (stacked rectangles). (D) Note groups (diamonds) overlaid on top of waveform and notes (gray curve and black lines). In all panels musicians are displayed in red and non-musicians in blue.

6.4.3 By Common Categories

Results by common categories include annotations automatically classified to one of the broad categories in Table D.3. Here, we will focus on boundary annotations belonging to the segmentation and prosody categories since they are the most comprehensive, comparing the two experimental conditions.

As we will see, differences between experimental conditions and musical abilities are mainly found in the frequency of boundaries placed around a given structure and the precision of KDE peak locations.

6.4.3.1 Segmentation

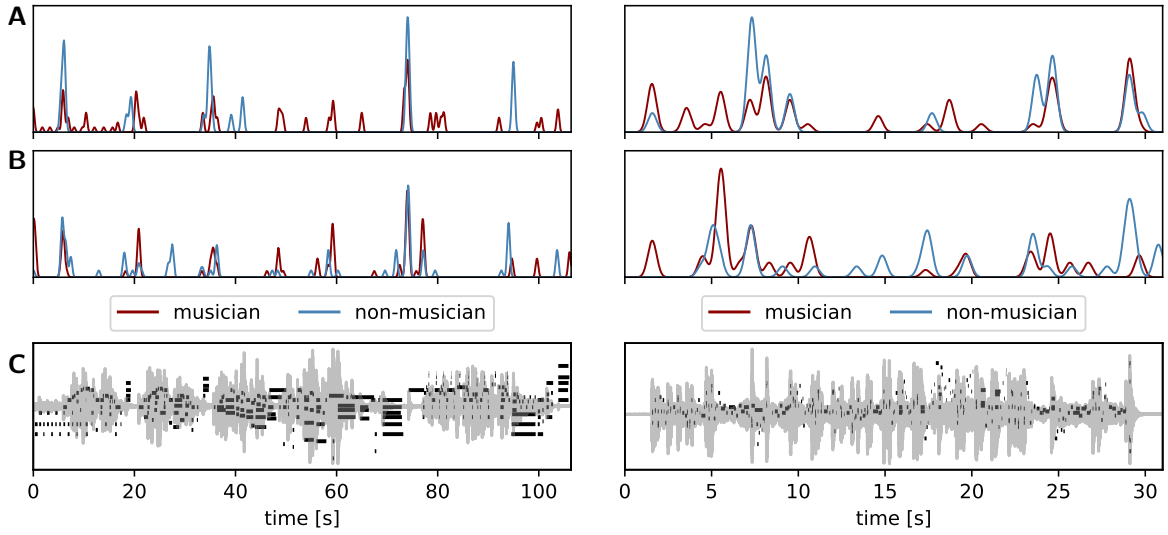
Solveig's Song

Figure 6.8a shows boundary KDE profiles for the less detailed and more detailed conditions, drawn as solid lines for musicians (red) and non-musicians (blue). Participants in both experimental conditions and musical abilities distinctly marked the main segments of the excerpt, representing up to 47% of participants. Since the height of the KDE peaks is analogous to the frequency of annotations around a specific time, we can appreciate recurrent segments. The piece is most frequently divided into its A and B sections, after which participants marked phrase

boundaries for each verse and the introductory chords. The repeating motif in the last verse, closing the A section, was consistently marked as a segment due to a *ritardando* and *diminuendo* execution that contrasts with its surroundings.

Fragment d'une ébauche

Figure 6.8b shows segmentation marks for both experimental conditions, representing up to 28% of participants. The segments are less clear than in the romantic piece. However, participants in both conditions marked reasonable segmentation boundaries, driven mainly by reappearances of material *c*. The piece is divided roughly into four segments marked by the presentation of materials *a*, *b*, and *c*, their reappearance, a climax, and the last reinstatement of the materials before the end. The last chord, marked frequently in both conditions, is described as bringing a 'sudden' or 'abrupt' end to the music by participants in the less detailed condition.



(a) Edvard Grieg's Solveig's Song. (b) Pierre Boulez's *Fragment d'une ébauche*.

Figure 6.8: KDE boundary profiles for the "Segmentation" category are drawn as solid lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

6.4.3.2 Prominence

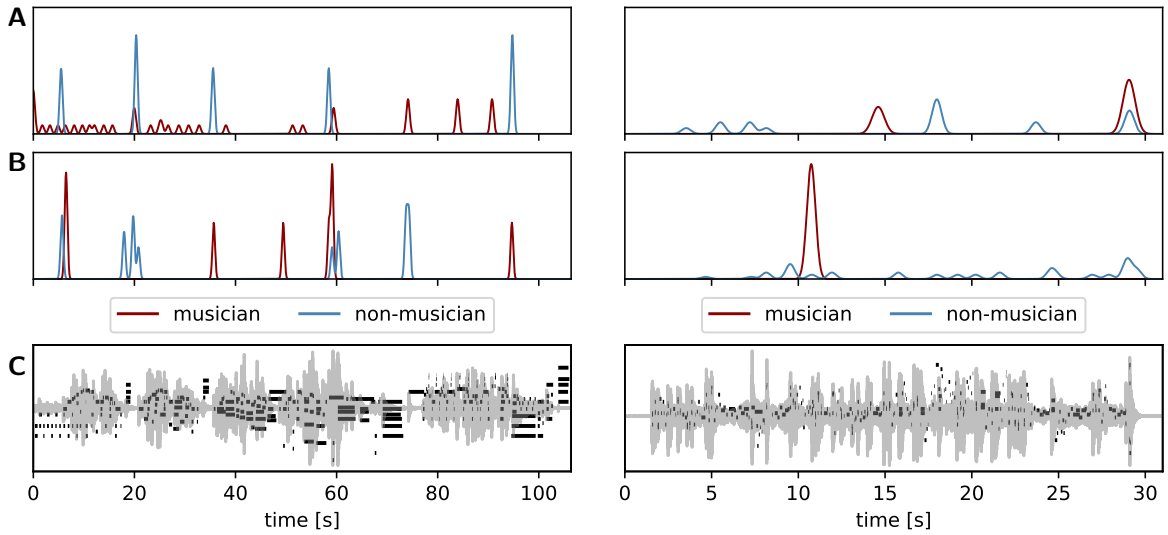
Solveig's Song

Figure 6.9a shows boundary KDE profiles for both experimental conditions, representing up to 16% of participants. The peaks of KDE profiles resemble those from the Segmentation category, dividing the piece into roughly the same segments. In the more detailed condition, the largest KDE peaks coincide with the lowest E octave played loudly by the left hand at the A section's conclusion, followed by the central theme's presentation at 6 s, and the humming section's melody introduction with the E5. This note and other specific notes, such as an accented G5, the highest note in the B section's melody, are marked as prominent in the less detailed condition.

Fragment d'une ébauche

Figure 6.9b shows boundary annotations for both experimental conditions, representing up to 10% of participants. Despite not containing many annotations after the automatic classification process, this category highlights prominent aspects of the piece. They are punctuated by the

stress of a modified version of material *a* at 10.7 s, the climax's highest note (B6) at 17.9 s, and, again, the final chord.



(a) Edvard Grieg's Solveig's Song.

(b) Pierre Boulez's *Fragment d'une ébauche*.

Figure 6.9: KDE boundary profiles for the "Prominence" category are drawn as solid lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

6.4.4 By Common Properties

Communities with common properties by experimental condition were extracted using the steps described in Section 6.2.4.3 from the entire dataset by annotation type, regardless of their label content. Since KDE profiles, which can be used for boundary and comment annotations, are already a form of data aggregation, we will focus on describing the aggregation of common regions and note groups that have remained unexplored up to this point.

6.4.4.1 Common Regions

Figure 6.10 and Figure 6.11 shows common regions by experimental condition with musicians (red) and non-musicians (blue) split. Figures display which regions were most commonly annotated per experimental condition, as explained in Section 6.2.4.3. Individual regions (colored rectangles) have been aggregated from communities containing at least four nodes (see Section D.2). They are displayed vertically stacked (in no particular order) to facilitate their visual inspection.

Tables D.4, D.5, D.6, and D.7 detail the specific start/end times, subset separations extracted from communities, and individual region counts in each subset for the 10 most common regions annotated per piece and experimental condition. These results are described as follows:

Solveig's Song

For the condition with minimal instructions, participants marked similar region segments dividing the excerpt into its phrases. This data represents up to 24% of participants. Both musicians and non-musicians used regions similarly. Common regions include the introductory chords of the left hand and the melodies for each of the verses in the A section, as well as the B section

split in two halves at the highest note of the melody. Smaller common regions highlight the tonic chords at the end of the second verse, the end of the A and B sections, and the high E that transitions to the B section.

Common region annotations in the condition with more detailed instructions represent up to 17% of participants. They focus on the A section's introduction, concluding chords, and pauses rather than melodies. Larger regions in the B section have the same subdivision as the one found in the less detailed condition (see Figure 6.10). Shorter common regions highlight pauses and the repeated arpeggiated E minor chord at the end of the B section.

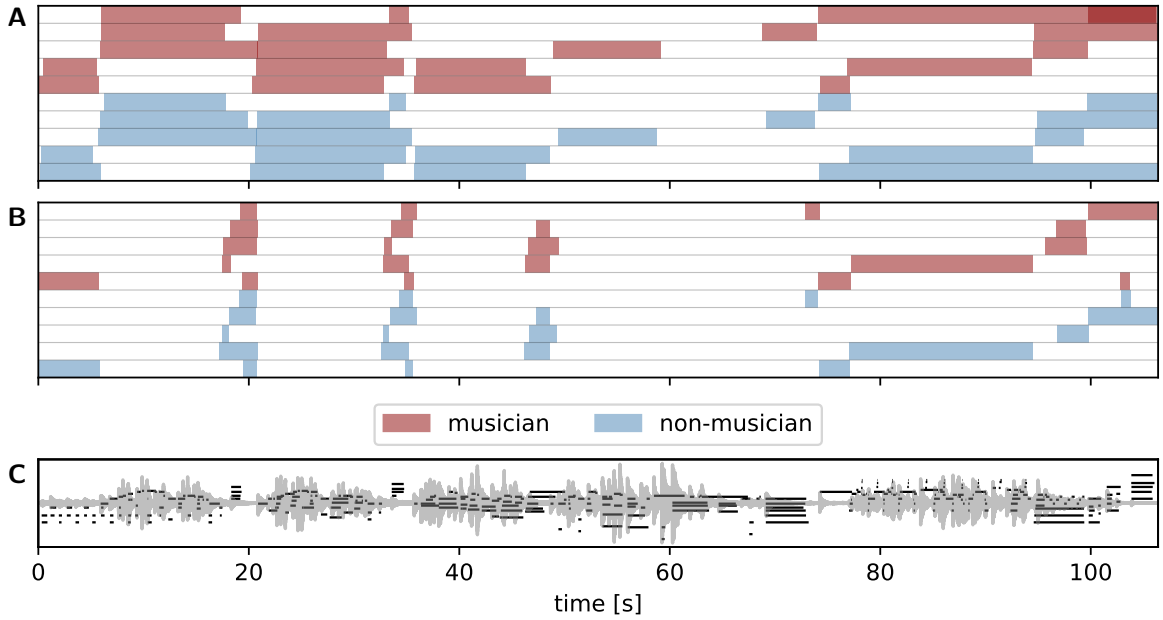


Figure 6.10: Common regions in Edvard Grieg's Solveig's Song. Regions for each condition are vertically stacked to improve visual inspection. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

Fragment d'une ébauche

For musicians and non-musicians, common regions in both experimental conditions follow the same patterns (see Figure 6.11), with region selections reminiscent of the segmentation structure outlined in Section 6.4.3.1. Aggregated regions in the less detailed condition represent up to 29% of participants, while those in the more detailed condition represent up to 33%. Smaller common regions highlight melodic motives, chords, or individual materials, whereas larger aggregated regions encompass segments.

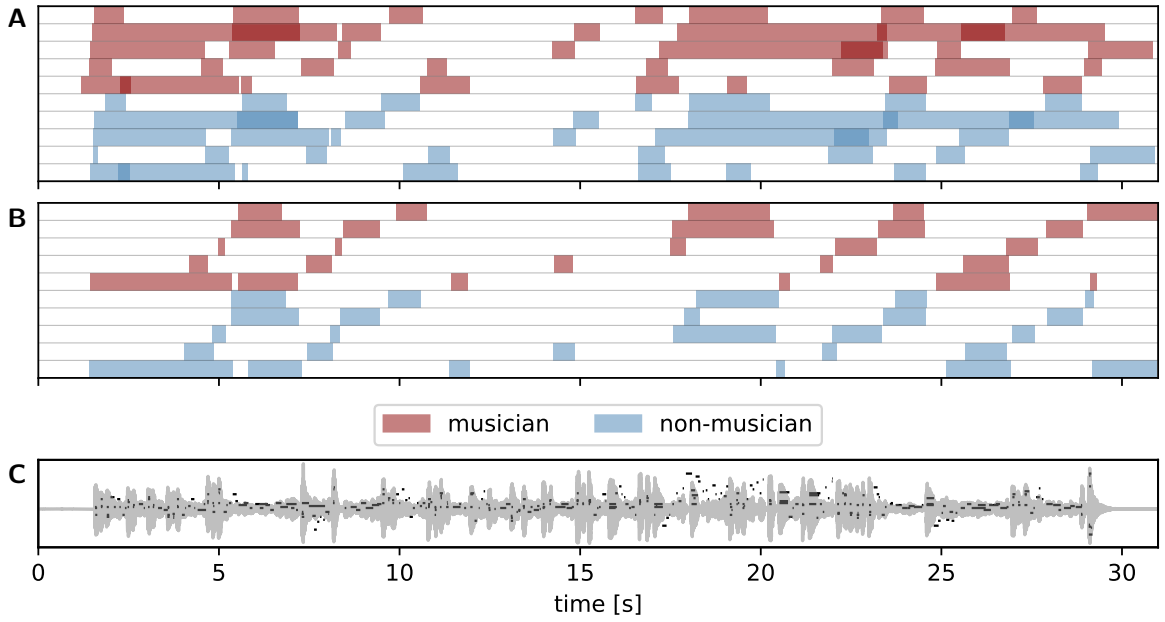


Figure 6.11: Common regions in Pierre Boulez's *Fragment d'une ébauche*. Regions for each condition are vertically stacked to improve visual inspection. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

6.4.4.2 Common Note Groups

Figures 6.12 and 6.13 show common note groups by experimental condition with musicians and non-musicians split. Common note groups in communities containing at least four similar groups are shown and aggregated into their common notes (colored diamonds and squares) in a piano roll. The note's opacity is analogous to how often it was used in a common group. Tables D.8–D.15 in Section D.9 detail the specifics of each community subset. The mean number of notes contained in each subset and the frequency count of individual note tuples (note start/end times and MIDI note number) are given for the 10 most common groups annotated per piece and experimental condition. Results are described below:

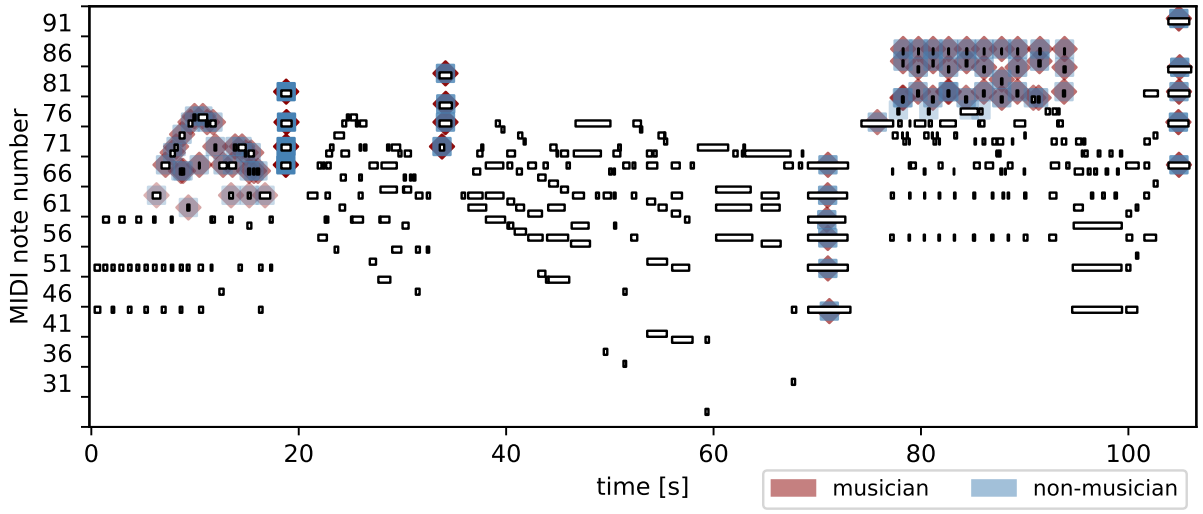
Solveig's Song

Common groups in the less detailed condition represents up to 38% of participants. They include the main melody at the start of the A section and the arpeggiated ornaments above the main melody of the B section.

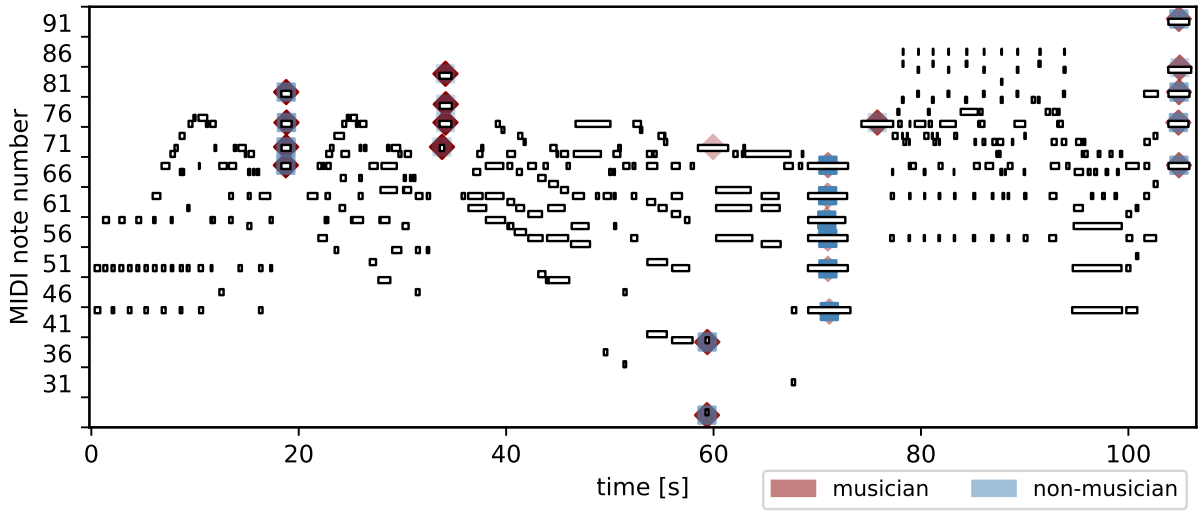
Common groups represent up to 19% of participants in the more detailed condition. Here, musicians and non-musicians repeatedly marked the lowest notes played by the left hand, concluding the A section (see Figure 6.12b). In both experimental conditions, participants grouped the prolonged concluding Am chords at the end of each phrase/section of the piece.

For common note groups, it is also worth considering the mean number of notes inside similar groups in a community, which can be higher for less common note groups. An example in Tables D.10 and D.11 of common groups helps to illustrate this point: The most common community of similar groups (Set 1 in Table D.10) is conformed by the first A minor chord at 18 s. It had 4 notes on average and was created independently by 11 users. In contrast, the eighth most common community of similar groups (Set 8 in Table D.11), most of the notes in the B section, had 107 notes on average and was created independently by 5 users. Significant

chords or single relevant notes tend to be annotated more often than groups with a higher note count. In this extreme case, however, the note group included almost an entire section (see Figure 6.12a). Thus, although fewer participants may have marked them, common groups with high average note counts (e.g., a melody) should be considered salient.



(a) Less detailed.



(b) More detailed.

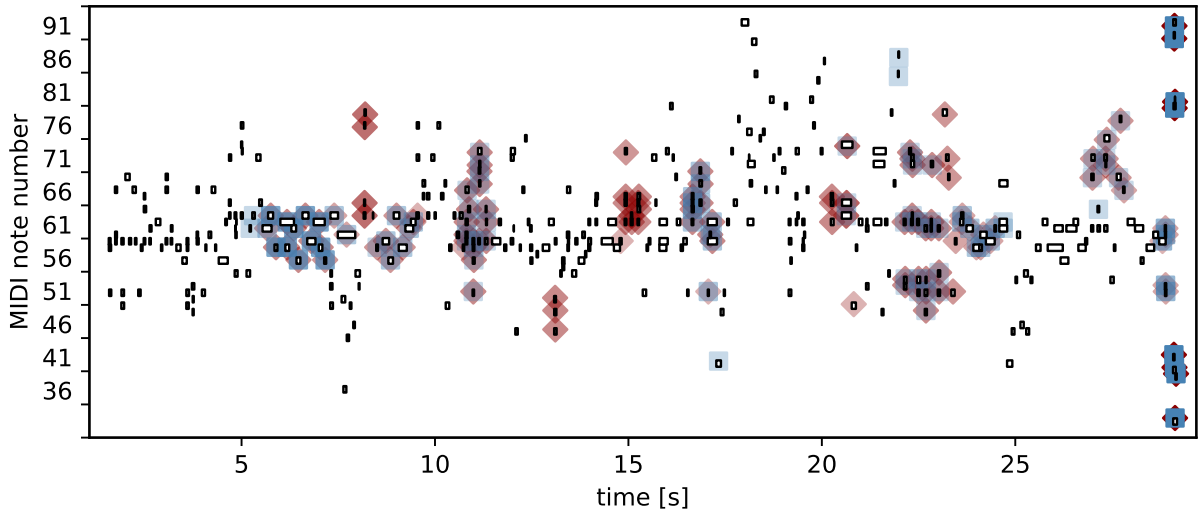
Figure 6.12: Common note groups in Edvard Grieg's Solveig's Song. Note groups (diamonds and squares) overlaid on top of notes (outlined rectangles). Musicians are displayed in red and non-musicians in blue. Opacity represents how often a note appeared in a common group.

Fragment d'une ébauche

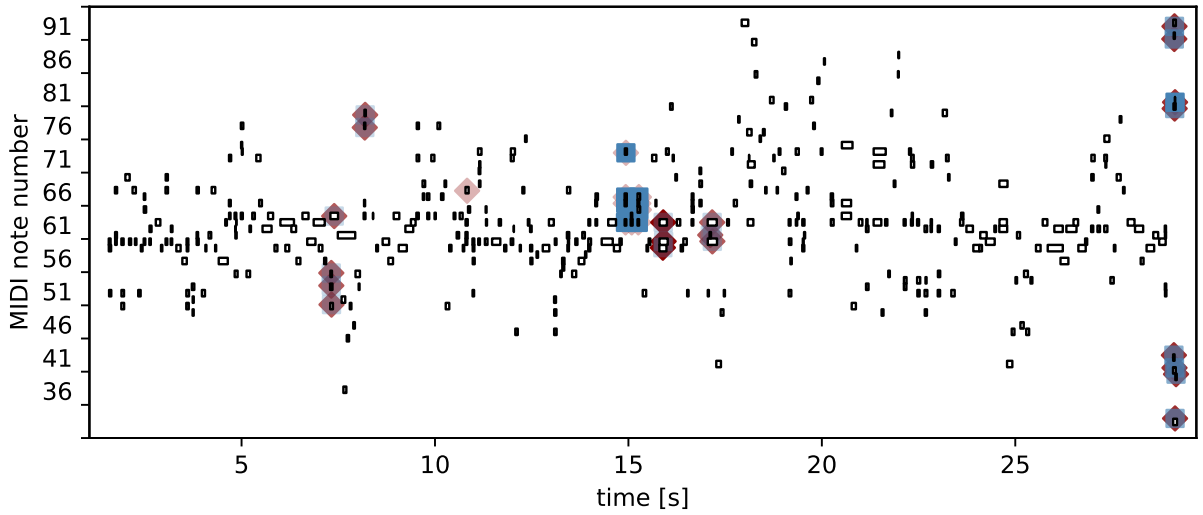
Data in the less detailed condition represent up to 38% of participants. Figure 6.13a shows that commonly annotated groups by experts and novices included examples of materials *a*, *b*, and *c*, sporadic loud chords during the piece, and the final chords.

Annotations in the more detailed condition represent up to 10% of participants. There are fewer common groups in this experimental condition (see Figure 6.13b). Interestingly, there are no examples of melodies from material *c*, only accented chords from materials *a* and *b* (at 8 s, 15 s, and 16 s), and the final chords.

The community of note groups with the most similar groups for both experimental conditions is the last chords of the piece, which are one of the loudest moments of the entire performance, as mentioned in Section 6.3.2.2.



(a) Less detailed.



(b) More detailed.

Figure 6.13: Common note groups in Pierre Boulez's *Fragment d'une ébauche*. Note groups (diamonds and squares) overlaid on top of notes (outlined rectangles). Musicians are displayed in red and non-musicians in blue. Opacity represents how often a note appeared in a common group.

6.5 Discussion

We have presented an experiment demonstrating a free-form application of the four CosmoNote annotation types enhanced by text descriptions that mark recurring musical elements for two musical pieces across two experimental conditions and musical abilities. Data collected from 58 participants per experimental condition, analyzed with different methods, show a substantial convergence of segmentation and prominence annotations across musical expertise, highlighting

our protocol's robustness. The following paragraphs reflect upon choices of the experimental design and their implications, open challenges, and perspectives for future work.

The stimuli used for the study were chosen to contrast the conspicuous melody, familiar Western harmony, and more predictable tempo and dynamic structures of the romantic musical excerpt against more subtle lines, the atonal nature, and the intentional breach of popular performance practices of the contemporary piece. Because of the length, difficulty, and subjectivity of the task, this choice of music was also designed to prompt a new perspective and change of pace for listeners, despite the order in which the pieces were presented to them.

Conducting this study was an essential part of testing the reliability of CosmoNote for a larger audience, the effectiveness of our instructions, and their outcomes. The experiment was deployed in an ecological setting where participants were fully autonomous in a task where time, attention, and patience were valued. The overall reaction of participants was positive, both from feedback questionnaires (see Section B.2) and recounted experiences. However, there was some expected constructive criticism about the platform and its first use. Most commonly, participants spent an average of fifteen minutes familiarizing themselves with the interface controls, others being at ease with them only by the end of the experiment. We know that CosmoNote is a complex tool, and users must spend varying amounts of time learning how to use it before being comfortable and focusing uniquely on the annotation task. Reduced, homogenized training time was the consequence and compromise of a laboratory study. At this stage, the experiment may have been quite demanding for participants without the researcher's presence. For instance, though frequent, autonomous use of the printed guide proved effective, many participants had lingering questions throughout the experiment. This adjustment phase is why improved, dynamic, and shorter tutorials for less-steep software learning are apparent, as Section 3.7 mentions.

Learning how to use the software is only the first part of the process. Comprehensive instructions by annotation task must complement suitable tutorials. In our case, minimal protocol instructions were sufficient for obtaining consistent global prosodic annotations. Our results show that participants can understand musical prosody and, particularly, what segmentation and prominence represent in music performance. An elementary version of the annotation instructions may be adequate for inviting beginners to annotate high-level segmentation or the most salient passages in a piece without requiring previous specialized knowledge. However, in a citizen science application, more detailed, specific instructions should be used in advanced stages for marking subtle structures. Moreover, at this stage, automatically detecting deviation from protocol is an open challenge to be resolved with an agreement between novice annotators, experts in the community, and researchers.

Data aggregation is essential in achieving consensus on widely accepted representations of prosodic structures in annotated music. By aggregating annotations based on their shared characteristics, we can overcome the limitations of diverse labeling expressions or the absence of labels altogether. The aggregation techniques presented in this chapter are effective as standalone procedures but can complement the methods discussed in Section 6.2 or serve as tools for more in-depth analysis. For example, note groups could be consolidated to illustrate common chords by solely considering MIDI note numbers without considering the notes' start and end times. Throughout the data validation stage, all aggregation techniques prove valuable as they offer simplified visualizations that summarize the markings made by a group of annotators.

Expert and novice participants could identify structures serving segmentation and prominence functions in the two musical excerpts and experimental conditions. Non-musicians were likely more open to our instructions and adopting new concepts about music structures than musicians. Musically trained participants applied their specialized auditory expertise and were sometimes biased toward analyzing theoretical elements in music rather than its expressiveness. For instance, musicians tend to mention technical details as evidence for their annotations, such

as the type of chords or scales being used in a specific passage. However, tasks involving musical structures rather than pitch or chord identification rely less on musical expertise (Bigand, 2003). For example, region annotations in Figure 6.4 show that the "pause" concept was well understood and annotated in mostly equivalent portions. It is also relevant to consider that the expert (musician) vs. novice (non-musician) division was determined using participant responses to our musical questionnaire (see Section B.1) and be aware that musical abilities are a multidimensional concept along a spectrum. In a citizen science context, we aim to bring multiple perspectives to understand the musical experience. Interestingly, there was a more noticeable difference between experimental conditions than between expert and novice participants within a condition.

One meaningful difference between annotations in each experimental condition was how participants used labels. The more detailed condition intentionally caused participants to match the words in the instructions to designate prosodic functions in segmentation and prominence. However, it carries the undesired effect of limiting the number of concepts in participants' minds. Indeed, as previously stated (see Section 4.4.4), we wish annotators to be free to explore a variety of notions and terms to come up with shared conventions in a community. A solution to balancing between excessive specificity and inadequate formalization may be accentuating the difference between simple labels, title labels, and complex labels (Section 6.2.4.2).

Even though many labels in our data contain elaborate information supporting a particular choice of timing or pitch, the inclusion of a tag (mentioned in Section 3.7) associated with larger categories (such as the ones presented in this chapter and the ones in Table D.3) would substantially improve the quality of our data. This action represents only a marginal increase in participants' effort (X. Li et al., 2012) toward bridging the gap to Level 2 or Level 3 rich annotations.

Future research may also be focused on improving automatic label classification through fine-tuning a pre-trained language model using a ground truth categorization of annotation data. Although classification is necessary, descriptive labels' subtlety, complexity, and subjectivity escape an introductory analysis. A possible avenue to explore in a scalable, data-driven approach is using Large Language Models (LLMs) for data analysis, preferably local (offline) alternatives when they become widely available. Although LLMs have proven helpful in summarizing and extracting meaning from large quantities of text, at the time of writing, their ethical, safe, and reliable use is still being studied by the scientific community (Watkins, 2023).

This experiment is crucial in recognizing future development milestones and understanding how to exploit CosmoNote's versatility and assets. Annotating performed music using boundaries, regions, comments, note groups, and their respective detailed labels illustrates how collecting rich annotations in large-scale applications can paint a fuller picture of how humans perceive and represent expressiveness.

Chapter 7

Conclusion and Future Work

This chapter concludes this dissertation by recapitulating the research. First, we summarize the previous chapters' core results aligning to the aim of capturing music expressiveness in performed music within the scope presented in Chapter 1. Next, we will outline the contributions of this work. Then, we will examine future work. Finally, we will discuss the research as a whole.

7.1 Conclusion

In this section, we will revisit each chapter (2–6) and highlight the most important ideas and research findings in each one:

Chapter 2

This chapter contextualized the foundational research regarding the three main intersections of this work: music structure analysis, music performance, and citizen science.

We introduced computational music structure analysis as a multidisciplinary field studying music structures with the help of computational tools. We were particularly interested in techniques based on human perception, such as the GTTM segmentation rules, LBDM, and Melisma. We outlined techniques for boundary segmentation identification and evaluation used in MIR. Even though ambiguity and subjectivity are inherent factors in annotating music structures, sensibly addressing them in analysis algorithms remains an open problem. We thus commented on how scalable and interpretive frameworks like the one we proposed are necessary within the current machine learning and data-driven models.

In studying music performance, it is crucial to understand the role of expressive performance research. Thus, we explored how this field evolved from analyzing music as writing (notation) to music as performance. We described expressive variations communicating structure and expressiveness through expressive devices such as dynamics, tempo, and articulation. Finally, we explored the role of modeling music expressiveness. We saw how this field embraces data-driven techniques for which there is a prevailing need for large-scale, adequately labeled data.

This chapter also recounted the origins and the current state of citizen science. It explored the challenges of classifying it and two reasonable alternatives of dividing citizen science projects by their involvement levels or epistemic practices (based on participant actions). Finally, the chapter delineated how citizen science is used in Computational Music Structure Analysis projects and its nascent use in research for performance expressiveness, where it is underexplored.

Chapter 3

This chapter introduced the interactive, customizable, web-based annotation platform CosmoNote. CosmoNote was created for annotating musical structures created by performers and felt by listeners. We detailed the step-by-step process from obtaining performance data to its analysis.

CosmoNote is a web application using the Web Audio API with the D3 visualization library, managing its data with JSON format files through the PouchDB in-browser database. CosmoNote organizes its documents into users with distinctive roles, music pieces with various data types, and collections gathering pieces with thematic or research-directed topics. The central components of the platform are:

1. The ability to individually toggle multiple superimposed visual layers synchronized to the audio content: the audio waveform, recorded piano roll (discrete note pitch, start/end time, velocity, and pedal data), extracted audio features (loudness and tempo), score features (harmonic tension), custom markers (score sections defined by the composer), and custom time series (physiological data).

2. Four independent annotation types with optional text labels: boundaries (vertical lines of different strength levels), regions (superimposable time selections with start and end times), comments (dashed vertical markers without level differentiation), and note groups (sets of notes selected from the MIDI data).

3. An intuitive visual interface inspired by audio production software, where annotators can access structural information, annotation data, and playback options at a glance. Annotators can control the interface using their preferred method using mouse or keyboard shortcuts.

4. Interactive audio/visual controls allow listeners to manipulate the music and annotations. For example, visually zooming in on a specific part of the performance, listening to a particular selection's audio, displaying multiple visual representations, and editing (adding, removing, moving) all annotation types.

The CosmoNote interface is entirely modular, i.e., it can be configured to show/hide informative text about the piece, visual layers, annotation types, and interface controls. For example, multiple collections are accessible to different user roles concurrently, so more than one annotation campaign can be active simultaneously.

The chapter concluded by outlining our contributions and ideas for improving the platform and its user interaction. For instance, though the note group annotation type brings new possibilities for annotating music, the simultaneous visualization of multiple note groups can be ameliorated. Current research uses CosmoNote to study relationships between physiological data and music performances. Applications for the platform in domains such as music education and musicology have also been considered and applied in workshops but have yet to be tested in a real-world scenario.

Chapter 4

This chapter described the annotation protocol that is core to our research, based on the principles of citizen science and the paradigm of musical prosody. We elaborated on the theoretical framework of the protocol by defining musical prosody, characterized by the acoustic variations that make music expressive, with four main functions: segmentation, prominence, emotional

reaction and coordination; our method focalizes only on the first two. We remarked on how few systematic and scalable studies exist on prosodic functions in music performance and how CosmoNote is an effective tool for executing such studies.

To detail the specific steps of the annotation protocol, we first specified the criteria for selecting the music and the participants and how we collect data on their demographics, musical abilities, and experience with the CosmoNote platform.

Next, we delineated the annotation task as a top-down, human-centered method relying on listeners' annotations of prosodic functions to analyze the link between these functions and acoustic properties. The task was designed to work with the four annotation types in CosmoNote (boundaries, regions, note groups, and comments). We described the two prosodic functions used in our protocol as follows:

Segmentation

Segmentation is defined as the process of dividing music into meaningful units. We focused on three generalized concepts as a starting point:

- Boundaries: time points separating a music stream into segments representing meaningful chunks of music. CosmoNote's boundaries are divided from 1 (weakest) to 4 (strongest). For example, these levels can mark motives, sub-phrases, phrases, and sections.
- Transitions: musical passages setting up a coming change in the music, linking musical ideas and blurring the change between them.
- Pauses: spaces added between two adjacent structures by lingering on notes or using silence.

Prominence

Prominence characterizes an emphasis drawn towards a particular whole in the music. As before, we denoted three generalized starting concepts:

- Stress: An emphasis on a particular element to make it more prominent than those around it.
- Melodic salience: A salient melody may be recognized by an increase in loudness and duration or a variation in the melody notes timbre.
- Tipping points: Moments where musical time is suspended/stretched before an inevitable return to the pulse.

To help participants, researchers, and music enthusiasts familiarize themselves with these concepts, we presented our contributions to a growing library of prosodic examples. We discussed three annotation strategies: intuitive vs. analytical, real-time vs. retrospective, and audio vs. visual. While each annotation strategy has its value, we recommend that beginners adopt a real-time, intuitive strategy guided by any cues at their disposal, and then refine their work retrospectively and analytically. We learned that time variable is crucial to the overall enjoyment of the experience and that a the balance between training, data collection, and engagement is delicate, specially for a citizen science project where there is no time-limit. These topics motivated the studies in Chapters 5, Chapter 6, and part of the perspectives mentioned in Section 7.2.

The chapter ends by describing how the data collection process is improved when annotators are trained to use the interface and develop internalized personal strategies, while respecting the annotation task. All of the processes presented in this chapter were designed and have been continuously revised accordingly, to obtain reliable and consistent annotations for theoretical and data-driven models of musical prosody.

Chapter 5

This chapter investigated the cooccurrence of auditory and visual stimuli during the annotation task. When one or more sensory inputs influence what is perceived by another, we speak of cross-modal or intersensory perception. The study examined the differences between four-level segmentation boundary annotations in unimodal (auditory or visual) conditions and cross-modal (auditory and visual) conditions using CosmoNote and our annotation protocol. We designed an experiment in which participants were exposed to combinations of visual/auditory cues and could freely place or remove boundaries. We used the technique of unbalanced optimal transport to calculate a distance between boundaries grouped by condition, which as opposed to conventional distance metrics, allowed us to detect which individual boundaries most contributed to their similarity or were less impactful and could thus be removed.

The overall results show that annotations with multimodal components (auditory and visual) rank higher than those with a single component, meaning there is a clear advantage in receiving several representations of the same information instead of just one. However, interpreting these results also requires considering the characteristics of the music being annotated. The results concerning individual differences highlight the predominance of models derived from visual information and that auditory information prevails mainly for resolving specific contradictions. In most other cases, auditory information is better used to identify subtle, low-level segmentation structures in music.

The chapter concluded by stating that multimodal components enhance segmentation annotations and that visual stimuli are a powerful tool, but may mislead annotators if the full range of information is not considered. Unsurprisingly, removing the audio component may lead to a loss of confidence and frustration. Adding visuals to the audio provides additional support, a form of cognitive scaffolding to accurately identify otherwise less salient patterns and mark a given boundary's start and end times.

Chapter 6

This chapter depicted a study closer to the actual conditions of the citizen science project. This experiment was built as a free-form task where participants had access to all the annotation types, visualizations, and interactions of CosmoNote to annotate structures of musical prosody. Participants could create text labels to complement their data. The experiment focused on three factors determining data quality: planning and adherence to instructions, annotator expectations and qualifications, and successful data analysis and validation. This study assessed how the central concept of musical prosody and our annotation task was understood and applied in a first encounter with the entire task.

We contrasted the annotations by providing minimal or detailed instructions while considering the participants' musical abilities. In addition, we detailed aggregation methods for region and note group annotations. We presented analysis methods for comparing similar annotations based on (1) their type, (2) their labels, (3) their categories, and (4) their properties. From one participant to another, for example, the same prosodic structure may have been marked by the placement of a boundary or note group, with the same or a different label, referring to

prominence or emotions, or just placed close to another annotation with a 50 ms margin. The study showed that annotations in CosmoNote can be analogous to rich annotations, which can be used for data-driven analysis, that their aggregation generates similar profiles for novice and expert participants, and that even with minimal instructions, the concepts of musical prosody enable participants to annotate expressiveness in performed music.

The data showed substantial convergence of segmentation and prominence annotations across all levels of musical expertise. We also underlined the experimental protocol's robustness to potential protocol non-compliance or neglect. However, at this stage, automatic control of protocol deviations is an open challenge that needs to be resolved by an agreement between novice annotators, community experts, and researchers. The overall results for the four annotation approaches are as follows: (1) boundaries are the most widely used annotation type, followed by regions, note groups, and comments. (2) There are fewer annotations with common labels, but they are more consistent with each other. (3) Annotations with common categories give an overview of annotated themes even if the individual labels differ. A supervised model could improve the automatic classification of common categories. (4) Annotations with common properties can be aggregated for easier comparison with modifiable parameters and show local convergence of annotations independently of their labels.

7.2 Future work

This section reflects on the perspectives for this research, issued from its connected threads. Some of this future work has already been evoked in previous chapters.

Development work improving the CosmoNote platform is naturally considered a primary next step for this research. Because specific development goals were already suggested in Section 3.7, it suffices to say that active development on the software is needed to ensure the tool's longevity. On the same subject, further controlled studies may be carried out to test more ambitious changes, such as including interactive tutorials and gamification elements or adding visual representations of the music (see Section 5.5). Once the platform can track the annotator's actions, we could, for example, measure the time participants take to master specific steps in the tutorial or get accurate data on visualization and interaction usage.

Along with CosmoNote, the musical prosody annotation protocol may be expanded to include music outside the Western canon. Vocal or instrumental music performances from soloists or ensembles can be studied using the same methodology. The interface is already compatible with any MIDI data, so the challenge would be to design intuitive visualizations for more than one MIDI or audio track. Extending the annotation task and its definitions to embrace the prosodic functions of emotional reaction and coordination is also contemplated in the perspectives of this research. Annotating emotional reaction (emotions that are experienced by listeners) or communicated emotion (emotions that are recognized but not necessarily experienced by listeners) is currently possible through CosmoNote, although it was outside the scope of this thesis. Annotating coordination depends on how the question of using data from musical ensembles is approached.

Similarly, future efforts with a specialized focus on supporting the citizen science initiative, for which we laid the groundwork, are required to accomplish the goals set in Chapter 4. Improving and expanding the example library and the music collections to be annotated are relevant steps in this process. These aspects of the project are essential for more direct contact between the community (e.g., citizens, music experts, performers) and the researchers.

A critical theme mentioned throughout the manuscript is a constant improvement of the data analysis and validation phase, dependent on the procedures implemented in the data collection

phase. Applying advanced data analysis and aggregation techniques, as discussed in Chapter 6, is one of the principles of our protocol. For example, the function of the comment annotation type as a marker without strength levels could be integrated to the boundary annotation type to facilitate analysis and aggregation, so comments can function as auxiliary annotations for interaction between annotators and researchers, as they were intended originally. We reiterate our interest in maintaining a solid data processing protocol bridging the gap to collecting rich annotations in a large-scale database used in theoretical and data-driven models to understand music performance.

Chapter 4 states that the musical prosody annotation protocol is iterative by design, meaning its progress is measured in development cycles. As the current cycle comes to a close, we have shown the efficacy of the annotation platform and the protocol to represent and annotate expressive structures in performed music, prepared all the steps necessary for its fruitful continuation, and contributed to the expressive music performance research field.

7.3 Final thoughts

This work focused on capturing music expressiveness and, specifically, musical prosody. My colleagues and I wanted to provide the tools anyone with or without a technical or musical background would need to annotate expressive structures in performed music. With that end in mind, this manuscript described: (1) The design process of a functional browser-based annotation platform, requiring only an internet connection to access our collections and campaigns and place interactive audio/visual annotations. (2) A complete annotation protocol for representing and annotating musical prosody in performed music. (3) Two experiments applying the platform and the protocol in a wide array of cases and observing the entire process.

The CosmoNote platform is a powerful software annotation tool freely accessible through a web browser. Among its many assets, its modularity, interactivity, and intuitiveness make it appealing for numerous applications in music research. I recommend opening the platform's code to the public for these reasons. CosmoNote's development cycle could also be improved by making the tool open source, following the same principles of collaboration of citizen science. Some steps towards this goal have already been achieved by CosmoNote's developer, who has refactored parts of the code to render it more modular. This transition may be done by the end of the active research of the COSMOS project so that people interested in using the platform may contribute to and maintain it.

The citizen science component of this protocol has been emphasized many times throughout the document. Moreover, I believe that it is possible, and even necessary, to achieve a level of a participatory science project–where participants engage in data collection and analysis–to fulfill the potential of the annotation methodology of advancing knowledge in expressiveness in performed music through high-quality, large-scale, adequately labeled data collection. As previously stated, this challenge will require active community engagement and support.

This research has directly interacted with the human component in expressiveness and music performance by capturing musical prosody annotations. The balancing act between rigorously collecting, analyzing, and aggregating data and embracing the subjectivity in individual human annotations is one I have tried to carry out diligently during these years of work, hoping, at the same time, to convey my deep appreciation for music as an art and as a science.

Appendix A

CosmoNote

A.1 Documents for Conducting Experiments


```
{
  "_id": "Bach Minuet in g, BWV Anh 115",
  "_rev": "23-9bcd00df09b4cf533764931a94a652e0",
  "name": "Minuet in G minor, BWV Anh 115",
  "composer": "Christian Petzold",
  "compositionYear": "1725",
  "performer": "Elaine Chew",
  "performanceDate": "2021-04-27",
  "events":[
    {
      "StartTime": 0.5315104166666668,
      "EndTime": 1.0461979166666668,
      "Type": "note_on",
      "Note": 82,
      "Velocity": 51
    }
  ],
  "loudness":[
    {
      "Time": 0.011609977324263,
      "Loudness": 0.0389273783711402
    }
  ],
  "tempo":[
    {
      "Time": 0.7853968255,
      "Tempo": 117.245657739161
    }
  ],
  "featureData":[
    {
      "type": "tension",
      "name": "Cloud Momentum",
      "color": "#795C34",
      "style": "solid",
      "width": "3px",
      "data":[
        {
          "time": 1.0452600000000003,
          "value": 1.4605934866804429
        }
      ]
    },
```


(a) piece - part 1


```
    {
      "type": "tension",
      "name": "Cloud Diameter",
      "color": "#993404",
      "style": "solid",
      "width": "3px",
      "data":[ ...]
    },
    {
      "type": "tension",
      "name": "Tensile Strain",
      "color": "#FEC44F",
      "style": "solid",
      "width": "3px",
      "data":[ ...]
    }
  ],
  "instantsData":[
    {
      "type": "instants",
      "name": "Tempo Values",
      "color": "#303030",
      "style": "solid",
      "width": 2,
      "opacity": 0.5,
      "data":[
        {
          "time": 0.5315104166666668,
          "value": "mezzo forte"
        }
      ]
    }
  ],
  "_attachments":{
    "Bach Minuet in g, BWV Anh 115":{
      "content_type": "audio/mpeg",
      "revpos": 22,
      "digest": "md5-5yF8OySGVdB6ZQ9VkXbOww==",
      "length": 848948,
      "stub": true
    }
  }
}
```


(b) piece - part 2

Figure A.1: Example of a piece JSON file in CosmoNote's database split into two columns. Only one datapoint per property is shown. The ellipsis represents truncated data.

```
{
  "_id": "Beethoven_B32_2_c4",
  "name": "Experiment c4",
  "description": "Please mark the boundaries...",
  "image": "images/visuals-vs-audio.png",
  "active": true,
  "shufflePieces": true,
  "interfaceMode": 2,
  "showPieceInfo": false,
  "alwaysShowNext": false,
  "showPieceDataControls": true,
  "showWaveformControl": false,
  "showNoteControl": false,
  "showPedalsControl": false,
  "showInstantsControl": false,
  "showLoudnessControl": false,
  "showTempoControl": false,
  "showTensionControl": false,
  "waveformOn": true,
  "notesOn": true,
  "pedalsOn": false,
  "instantsOn": false,
  "loudnessOn": false,
  "tempoOn": false,
  "tensionOn": false,
  "allowBoundaries": true,
  "allowRegions": false,
  "allowComments": false,
  "allowGroups": false,
  "showAnnotationModeControls": true,
  "allowSaving": true,
  "allowFinishing": true,
  "notesStartHidden": false,
  "showPlayLine": true,
  "highlightNotes": false,
  "playBoundarySounds": true,
  "showContext": true,
  "showFeedbackQuestionnaire": true,
```


(a) collection - part 1


```
  "pieces":[
    "Beethoven_B32_2_Tema",
    "Beethoven_B32_2_Var_I",
    "Beethoven_B32_2_Var_II",
    "Beethoven_B32_2_Var_III",
    "Beethoven_B32_2_Var_IV"
  ],
  "pieceOptions":[
    {
      "pieceID": "Beethoven_B32_2_Tema",
      "interfaceMode": 1,
      "waveformOn": true,
      "notesOn": true
    },
    {
      "pieceID": "Beethoven_B32_2_Var_I",
      "interfaceMode": 1,
      "waveformOn": false,
      "notesOn": true
    },
    {
      "pieceID": "Beethoven_B32_2_Var_II",
      "interfaceMode": 1,
      "waveformOn": true,
      "notesOn": false
    },
    {
      "pieceID": "Beethoven_B32_2_Var_III",
      "interfaceMode": 0,
      "waveformOn": false,
      "notesOn": false
    },
    {
      "pieceID": "Beethoven_B32_2_Var_IV",
      "interfaceMode": 2,
      "waveformOn": true,
      "notesOn": true
    }
  ]
}
```


(b) collection - part 2

Figure A.2: Example of a collection JSON file in CosmoNote's database split into two columns. The text after the ellipsis truncated. Global properties for the collection are set first as boolean variables. Pieces included in a collection can be listed by ID. Interface properties for individual pieces can be set using their ID.

```
{
  "user": "494e534541445f45322d3031",
  "enabled": true,
  "agreed": true,
  "eighteen": true,
  "musicalQuestionnaireFinished": true,
  "hearingTestFinished": true,
  "date": "2022-06-17T16:24:16.891178Z",
  "roles":[
    "experiment_prosody_training"
  ],
  "currentCollectionID": "prosody_training_experiment",
  "musicalQuestionnaireResults":[
    {
      "question": "Q1",
      "answer": "7"
    },
    {...}
  ],
  "hearingTestResults":[
    {
      "attemptNumber": 1,
      "correctToneCount": 4,
      "userToneCount": 4,
      "toneOrder":[ ...],
      "tonePlayed":[ ...]
    }
  ],
  "collections":[
    {
      "_id": "prosody_training_experiment",
      "collectionName": "Experiment INSEAD",
      "options":{
        "active": true,
        "shufflePieces": true,
        "interfaceMode": 2,
        "showPieceInfo": false,
        "alwaysShowNext": true
      },
      "pieces":[
        {
          "pieceID": "Grieg_Solveigs_Song_split",
          "interfaceMode": 2,
          "started": true,
          "finished": true,
          "startDate": "2022-06-22T07:52:20.742Z",
          "finishDate": "2022-06-22T08:34:06.009Z"
        },
        {...}
      ],
```


(a) user - part 1


```
        "currentPiece": 1,
        "startDate": "2022-06-22T07:52:20.742Z",
        "finished": true,
        "finishDate": "2022-06-22T09:03:00.041Z",
        "feedbackQuestionnaireFinished": true,
        "feedbackQuestionnaireResults":[
          {
            "question":{
              "number": 1,
              "type": "agreementScale",
              "category": "userExperience",
              "question": "Annotating music in...",
              "optional": false,
            },
            "answer": "5"
          },
          {
            "question":{ ...},
            "answer": "detecting smaller..."
          }
        ]
      }
    ],
    "stats":{
      "annotations":{
        "total": 54,
        "boundaries": 13,
        "regions": 8,
        "comments": 8,
        "groups": 25
      },
      "pieces":{
        "total": 2,
        "started": 2,
        "finished": 2
      },
      "collections":{
        "total": 1,
        "started": 1,
        "finished": 1
      }
    }
  }
}
```


(b) user - part 2

Figure A.3: Example of a user JSON file in CosmoNote's database split into two columns. Only one datapoint per property is shown. The ellipsis represents truncated data.

A.2 Data structure examples


```
[
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "A1",
    "time": 5.807065796617684,
    "strength": 4,
    "dateCreated": "2022-06-22T08:06:04.953Z"
  },
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "B1",
    "time": 35.7200000000001,
    "strength": 4,
    "dateCreated": "2022-06-22T08:21:13.696Z"
  },
  {
    "user": "494e534541445f45322d3032",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "",
    "time": 35.48654708520189,
    "strength": 2,
    "dateCreated": "2022-06-22T08:05:07.293Z",
    "dateUpdated": "2022-06-22T08:38:55.597Z"
  },
```


(a) Boundaries - part 1


```
  {
    "user": "494e534541445f45322d3032",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "",
    "time": 73.47802690582981,
    "strength": 3,
    "dateCreated": "2022-06-22T08:06:30.188Z",
    "dateUpdated": "2022-06-22T08:38:55.597Z"
  },
  {
    "user": "494e534541445f45332d3238",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "court arrêt avant l'accord",
    "time": 53.510610153699886,
    "strength": 1,
    "dateCreated": "2022-06-30T14:59:29.597Z"
  },
  {
    "user": "494e534541445f45332d3332",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "label": "Stress",
    "time": 20.222113228699552,
    "strength": 2,
    "dateCreated": "2022-07-01T09:01:19.870Z"
  }
]
```


(b) Boundaries - part 2

Figure A.4: Example of a boundary JSON file in CosmoNote's database split into two columns.


```
[
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "A",
    "startTime": 5.999484304932753,
    "endTime": 35.371950672645845,
    "dateCreated": "2022-06-22T08:05:48.521Z"
  },
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "intro",
    "startTime": 0.16555506905981934,
    "endTime": 5.538457729417853,
    "dateCreated": "2022-06-22T08:11:05.412Z"
  },
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "B",
    "startTime": 35.667750000000105,
    "endTime": 73.8245000000002,
    "dateCreated": "2022-06-22T08:20:25.663Z"
  },
```


(a) Regions - part 1


```
  {
    "user": "494e534541445f45322d3033",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "label": "",
    "startTime": 1.5639013452914798,
    "endTime": 7.228699551569507,
    "dateCreated": "2022-06-22T08:17:53.986Z"
  },
  {
    "user": "494e534541445f45322d3035",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "Introduction",
    "startTime": 0.05964125560538133,
    "endTime": 5.964125560538134,
    "dateCreated": "2022-06-22T08:03:44.936Z",
    "dateUpdated": "2022-06-22T08:18:03.798Z"
  },
  {
    "user": "494e534541445f45332d3435",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "melody major left hand",
    "startTime": 34.769247320477085,
    "endTime": 35.70117537050826,
    "dateCreated": "2022-07-01T13:26:20.608Z",
    "dateUpdated": "2022-07-01T13:37:49.099Z"
  }
]
```


(b) Regions - part 2

Figure A.5: Example of a region JSON file in CosmoNote's database split into two columns.

```
[
  {
    "user": "494e534541445f45322d3031",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "comment": "melange du chant d'oiseau ...",
    "time": 8.406130420918366,
    "dateCreated": "2022-06-22T08:51:36.605Z"
  },
  {
    "user": "494e534541445f45322d3132",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "comment": "Rupture ",
    "time": 23.545403587443943,
    "dateCreated": "2022-06-23T08:45:44.716Z"
  },
  {
    "user": "494e534541445f45322d3239",
    "pieceID": "Grieg_Solveigs_Song_split",
    "comment": "",
    "time": 30.178475336322954,
    "dateCreated": "2022-06-24T08:29:10.399Z"
  },
  {
    "user": "494e534541445f45322d3536",
    "pieceID": "Grieg_Solveigs_Song_split",
    "comment": "accord parfait",
    "time": 99.82455156950701,
    "dateCreated": "2022-06-29T10:32:10.504Z"
  },
```


(a) Comments - part 1


```
  {
    "user": "494e534541445f45332d3032",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "comment": "Baisse en intensité",
    "time": 23.67572869955157,
    "dateCreated": "2022-06-29T12:08:24.836Z",
    "dateUpdated": "2022-06-29T12:15:47.450Z"
  },
  {
    "user": "494e534541445f45332d3533",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "comment": "stress - high pitch",
    "time": 18.1130070623407,
    "dateCreated": "2022-07-01T14:58:43.051Z"
  },
  {
    "user": "494e534541445f45332d3533",
    "pieceID": "Grieg_Solveigs_Song_split",
    "comment": "third part",
    "time": 35.48841688368065,
    "dateCreated": "2022-07-01T15:13:07.909Z"
  }
]
```


(b) Comments - part 2

Figure A.6: Example of a comment JSON file in CosmoNote's database split into two columns.


```
[
  {
    "user": "494e534541445f45322d3032",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "label": "",
    "notes":[
      {
        "Note": 60,
        "StartTime": 20.87921874999996,
        "EndTime": 20.914166666666627
      },
      {
        "Note": 63,
        "StartTime": 21.14046874999996,
        "EndTime": 21.20578124999996
      },
      {
        "Note": 64,
        "StartTime": 21.14921874999996,
        "EndTime": 21.181145833333293
      },
      {
        "Note": 70,
        "StartTime": 21.85880208333329,
        "EndTime": 21.901041666666625
      }
    ],
    "dateCreated": "2022-06-22T08:12:02.021Z",
    "dateUpdated": "2022-06-22T08:43:41.527Z"
  },
```


(a) Note groups - part 1


```
  {
    "user": "494e534541445f45322d3032",
    "pieceID": "Grieg_Solveigs_Song_split",
    "label": "",
    "notes":[
      {
        "Note": 72,
        "StartTime": 33.6021354166665,
        "EndTime": 34.02583333333317
      }
    ],
    "dateCreated": "2022-06-22T08:37:42.052Z"
  },
  {
    "user": "494e534541445f45322d3431",
    "pieceID": "2020-12-23_EC_Boulez_Fragment",
    "label": "mélodie interrompue",
    "notes":[ ...],
    "dateCreated": "2022-06-27T12:39:49.482Z"
  }
]
```


(b) Note groups - part 2

Figure A.7: Example of a note group JSON file in CosmoNote's database split into two columns.

Appendix B

Annotation Protocol

B.1 Questionnaires

CosmoNote

Musical Questionnaire

The purpose of this short self-report questionnaire is:
a) to quantify your amount of musical engagement and behaviour in its many possible facets
b) to record your self-assessed level amongst various musical skills.

Please answer these questions by selecting the most appropriate option for you.

Agreement questions

Question	1	2	3	4	5	6	7
I spend a lot of my free time doing music related activities	○	○	○	○	○	○	○
I can reproduce (e.g., sing, whistle, hum) music from memory	○	○	○	○	○	○	○
I am able to hit the right notes when I reproduce a melody (e.g., sing, whistle, hum) along with a recording	○	○	○	○	○	○	○
I can compare and discuss differences between two performances or versions of the same piece of music	○	○	○	○	○	○	○
I am not able to reproduce a melody (e.g., sing, whistle, hum) in harmony when somebody is singing a familiar tune	○	○	○	○	○	○	○
I am able to identify what is special about a given musical piece	○	○	○	○	○	○	○
When I reproduce a melody (e.g., sing, whistle, hum), I have no idea whether I'm in tune or not	○	○	○	○	○	○	○
Music is kind of an addiction for me - I couldn't live without it	○	○	○	○	○	○	○
After hearing a new song two or three times, I can usually reproduce its melody (e.g., sing, whistle, hum) it by myself	○	○	○	○	○	○	○

(a) Musical questionnaire - part 1. Reformatted to fit the page. The numbered options for 'Agreement questions' are: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree.

Figure B.1: Musical questionnaire.

Multiple choice

I engaged in regular, daily practice of a musical instrument (including voice) for _____ years

At the peak of my interest, I practised _____ hours on my primary instrument (including voice)

I have had formal training in music theory for _____ years

I can play the following number of musical instruments (including voice)

Demographic data

What is your current occupational status

- Still at School
- At University
- In Full-time employment
- In Part-time employment
- Self-employed
- Homemaker/full time parent
- Unemployed
- Retired

What is the musical genre you mainly listen to?

- Rock/Pop
- Jazz
- Classical Music
- Hip-Hop/Rap
- Electronica
- Other

What is the Highest educational qualification you have attained?

- Did not complete any school qualification
- Completed first school qualification (e.g GSCE/Junior High School)
- Completed second school qualification (e.g A levels / High School
- Undergraduate degree of professional qualification
- Postgraduate degree
- I am still in education

Please select your age group

Please select your gender

Please select your current country of residence

Finish

(b) Musical questionnaire - part 2. Reformatted to fit the page.

Figure B.1: Musical questionnaire (continued).

Feedback Questionnaire

Thank you for having finished annotating this collection!

Please take a minute to answer the following questions by selecting the most appropriate option for you.
This will help us to better understand your results.

User Experience

We would like to know how was your overall experience using CosmoNote.
Using a scale from 1 (strongly disagree) to 7 (strongly agree) answer the following questions:

Question	1	2	3	4	5	6	7
Annotating music in CosmoNote was enjoyable	○	○	○	○	○	○	○
Recognizing and marking boundaries was easy	○	○	○	○	○	○	○
CosmoNote was easy to use	○	○	○	○	○	○	○
The music's sound quality was good	○	○	○	○	○	○	○

Please elaborate on your experience using our platform *(optional)*

Suggestions for improving the CosmoNote User Experience *(optional)*

Annotation Strategies

We are also interested in how you approached the annotation task(s).

What were your main strategies for marking boundaries?

Some visualization options are on/off by default, but you can also toggle between them.

We would like you to rate how much you used the available visualizations to actually inform your annotations.
Using a scale from 1 (never used it) to 7 (always used it) answer the following questions:
(if the visualization wasn't shown, then use N/A)

(a) Feedback questionnaire - part 1.

Figure B.2: Feedback questionnaire for a boundary annotation task. Reformatted to fit the page. The numbered options for 'User Experience' are: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree.

Question	1	2	3	4	5	N/A
Audio waveform (pale-green)	○	○	○	○	○	○
Piano roll (notes - blue rectangles)	○	○	○	○	○	○
Pedals (pale areas)	○	○	○	○	○	○
Loudness (purple line)	○	○	○	○	○	○
Tempo (green line)	○	○	○	○	○	○
Tension	○	○	○	○	○	○

Please elaborate on your choices e.g., why you preferred a visualization over another *(optional)*

Familiarity and Preference

We are interested in knowing if you had heard this music before and how much you enjoyed it. Using a scale from 1 (strongly disagree) to 7 (strongly agree) answer the following questions:

Question	1	2	3	4	5	6	7
The music was new to me	○	○	○	○	○	○	○
I like the music I annotated	○	○	○	○	○	○	○

Please elaborate on your previous knowledge of the music you just annotated *(optional)*

Further Comments

Tell us if you have any additional comments about your experience annotating this collection *(optional)*

Submit Reset

We hope to see you again soon!

(b) Feedback questionnaire - part 2.

Figure B.2: Feedback questionnaire for a boundary annotation task (continued). Reformatted to fit the page. The numbered options for 'Annotation Strategies' are: (1) Never, (2) Rarely, (3) Sometimes, (4) Often, (5) Always. The numbered options for 'Familiarity and Preference' are: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree.

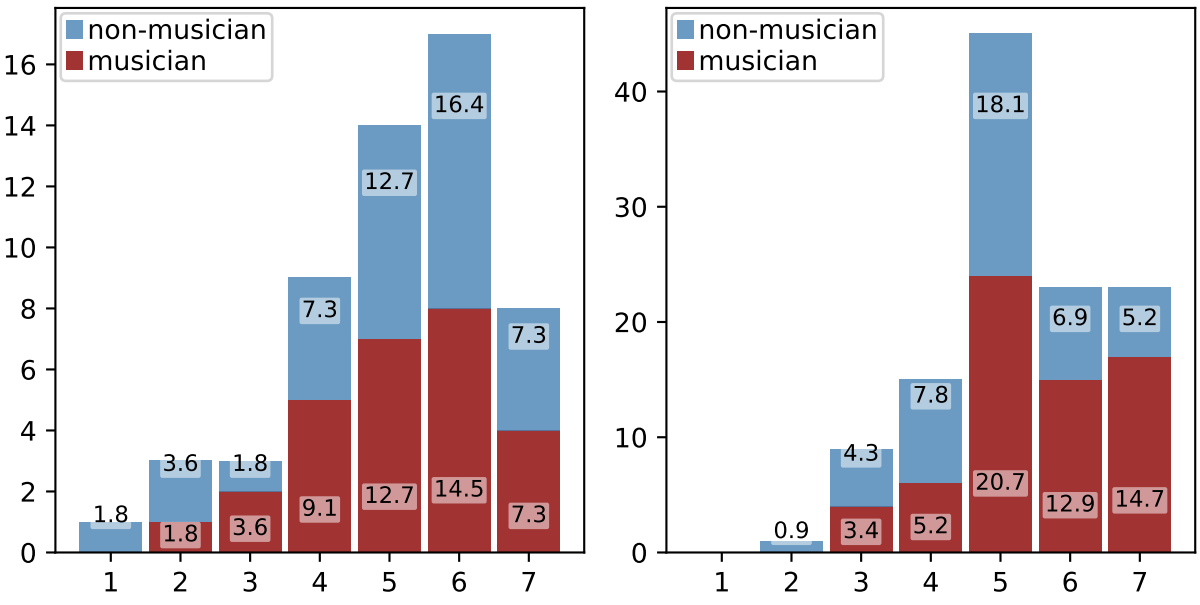
B.2 User Experience Feedback

In this section, we present results from the answers that participants of the studies in Chapter 5 (Study 1) and Chapter 6 (Study 2) gave to the Feedback questionnaires they received after having finished annotating. We show how participants responded to the "User Experience" questions described in Figure B.2.

For all figures in this section, the *y-axis* shows the number of participants that selected a given rating on the Agreement scale of the *x-axis* from 1 to 7. Rating counts are displayed as bars where ratings for non-musicians (blue) are stacked over those of musicians (red). Percentages from the total number of participants that took the questionnaire (55 for Study 1 and 116 for Study 2) are shown with one decimal place over each bar, except for cases where no participants used a particular rating.

Figure B.3 shows results to a question that evaluated the overall enjoyment of participants with the interface. The majority of participants report having enjoyed their time annotating with CosmoNote. In Study 1, Figure B.3a shows that 70.9% of participants marked a rating of 4 or higher while Figure B.3b shows that 78.5% marked a rating of 5 or higher.

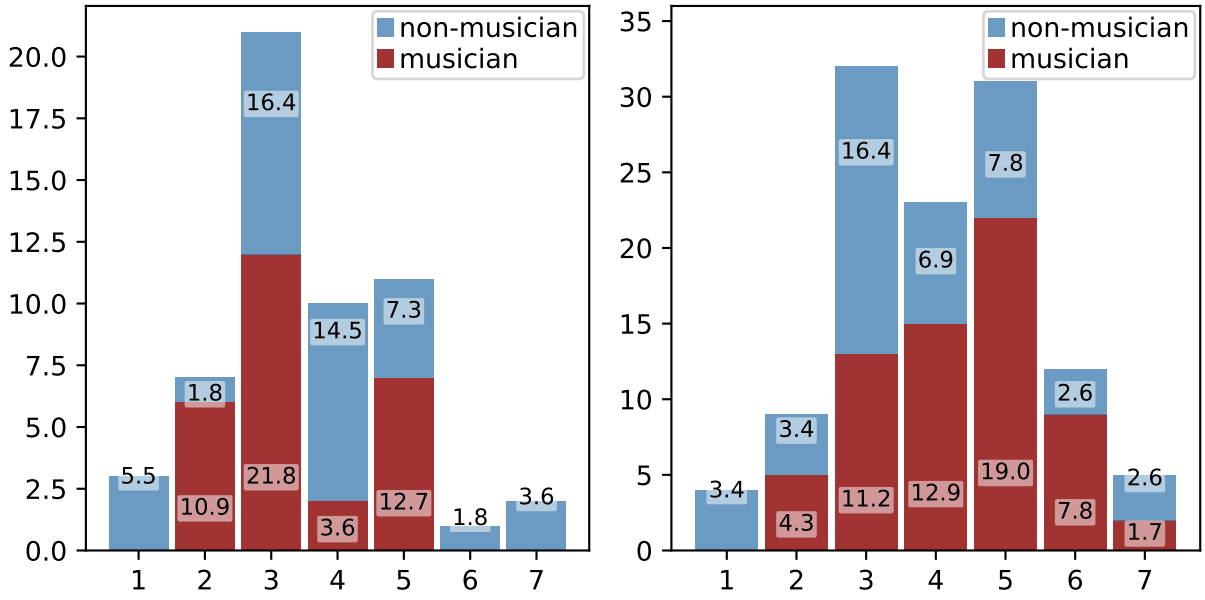
Figure B.4 describes responses related to the respective annotation tasks of Study 1 (boundaries) and Study 2 (all annotations). We asked if the tasks were easy. For both studies, the majority of participants rated the task with more neutrality. In Study 1, Figure B.4a shows a rather negative rating of 38.2% of participants marked a rating of 3. This can be explained by the nature of visual annotations the task in Study 1 (all participants annotated only visually approximately half of the pieces), which caused some frustration. Figure B.4b shows that 27.6% marked a rating of 3 while 26.8% marked a rating of 4. This experiment was more complex, thus requiring better knowledge of CosmoNote and the task, as discussed in Chapter 6.



(a) Study 1 - Count

(b) Study 2 - Count

Figure B.3: Results for the question "Annotating in CosmoNote was enjoyable". The numbered options were: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree. Percentages shown over bars.



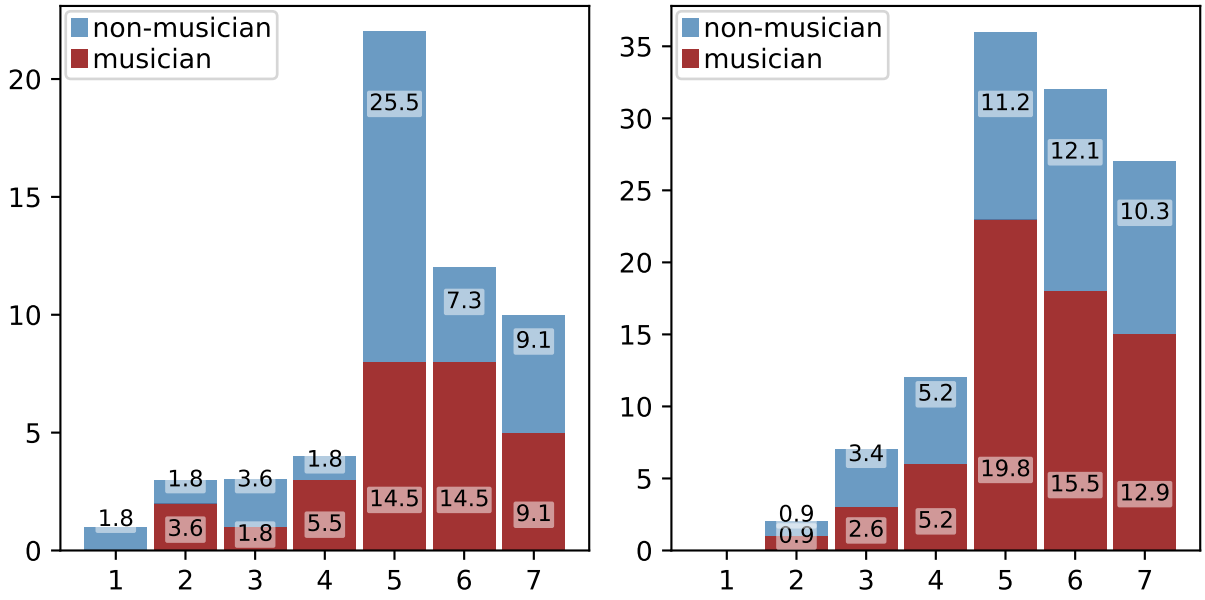
(a) Study 1 - Count

(b) Study 2 - Count

Figure B.4: Results for the questions (a) "Recognizing and marking boundaries was easy" and (b) "Recognizing and marking musical prosody was easy". The numbered options were: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree. Percentages shown over bars.

Figure B.5 presents the evaluation results of CosmoNote's ease of use. The interface and its features were changed between the data collection stages in Study 1 and Study 2, not only because the specific tasks required a different configuration but also because we improved the interaction and fixed minor computer bugs between the studies. Consequently, CosmoNote's usability rating in Study 2 was higher than that of participants in Study 1. Figure B.5a shows that from 80% of users declaring the platform was easy to use with a rating of 5 or higher, 40% of those responses are concentrated on a rating of 5 and 40% on ratings 6 and 7. In contrast, Figure B.5b shows that among 81.8% of participants who rated the platform's usability positively, only 31% is on rating 5, while 50.8% is distributed on ratings 6 and 7.

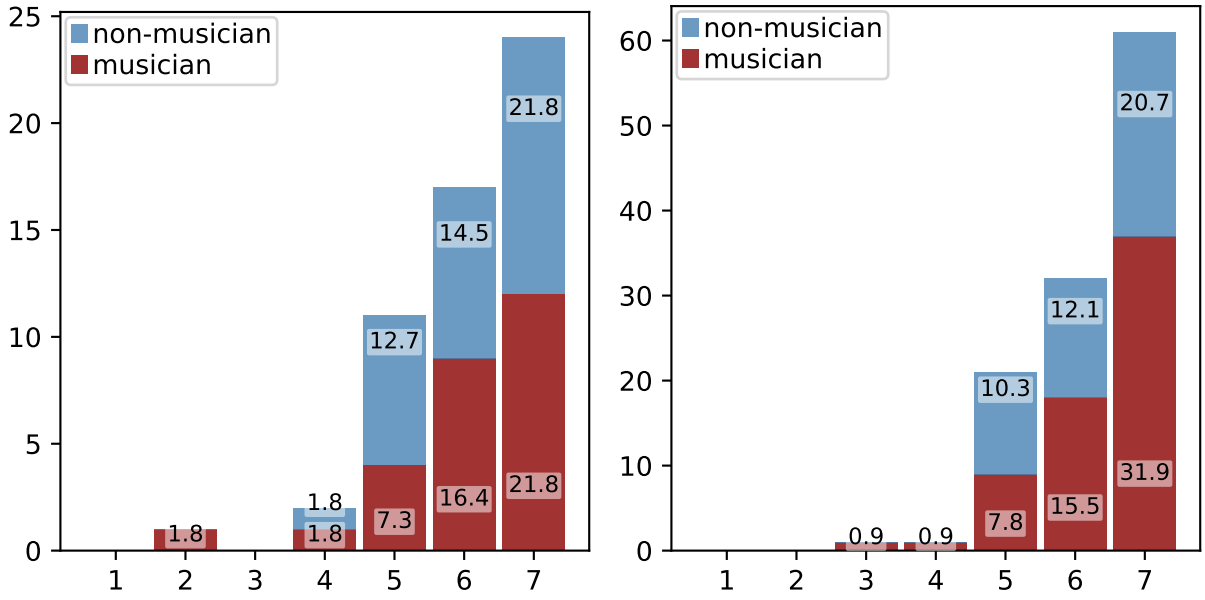
Figure B.6 evaluated the sound quality of the recorded performances. Most participants, 94.5 % for Study 1 and 98.3% for Study 2, rated the sound quality positively. The outliers are 1.8% of participants (1 person) in Figure B.6a, who chose a rating of 2, and 0.9% of participants (1 person) in Figure B.6b, who chose a rating of 3. Since these participants did not elaborate on these choices, the exceptions could be explained by rare instances where the audio playback stopped working or became sluggish, as verbally reported by participants during the experiments.



(a) Study 1 - Count

(b) Study 2 - Count

Figure B.5: Results for the question "CosmoNote was easy to use". The numbered options were: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree. Percentages shown over bars.



(a) Study 1 - Count

(b) Study 2 - Count

Figure B.6: Results for the question "The music's sound quality was good". The numbered options were: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree. Percentages shown over bars.

Appendix C

Cross-modal Study

C.1 Annotation Instructions

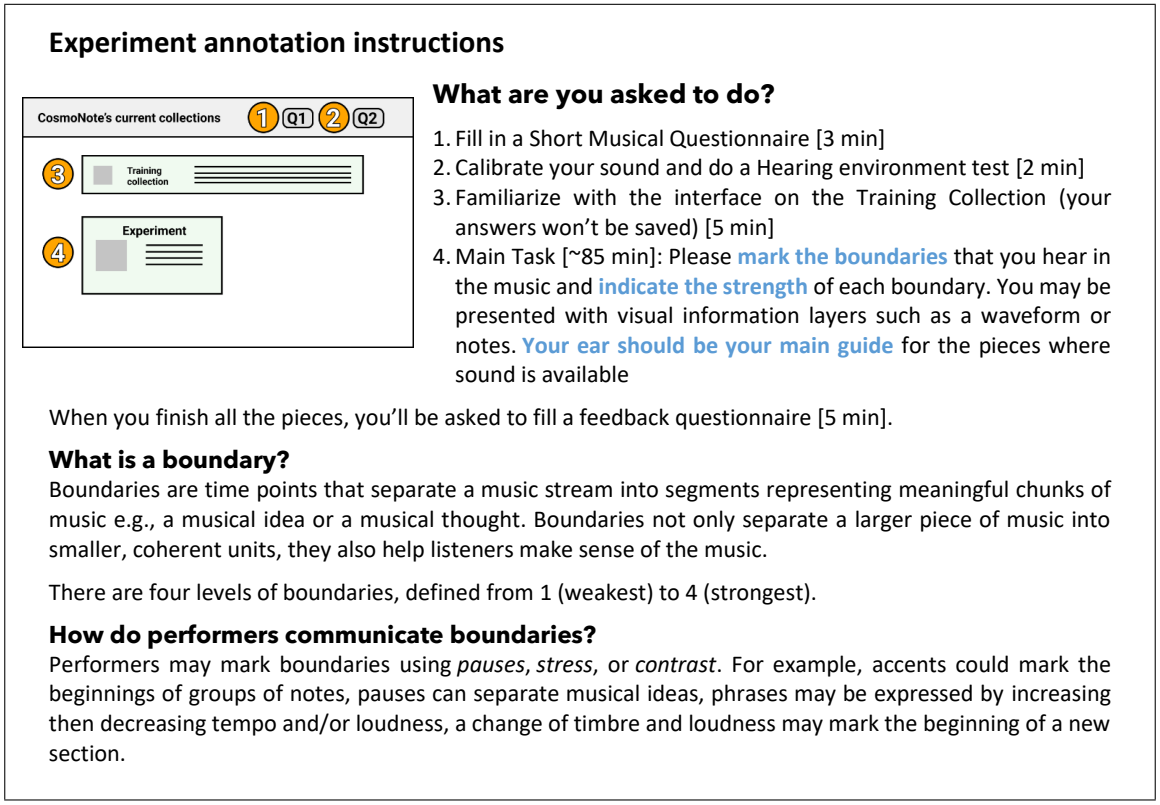
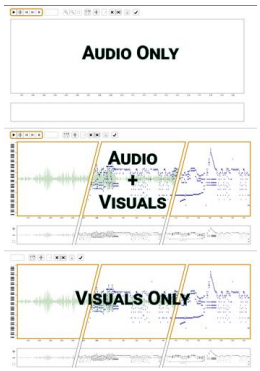


Figure C.1: English version of the instructions for the cross-modal annotations experiment (reformatted to fit the page). First part of the instructions explaining the task and the definition of boundaries that is used for the experiment.



What you will see/hear

In this experiment, you will annotate 33 pieces of music, shown in different configurations of the CosmoNote interface.

You'll be presented with one of three cases:

1. *Audio only*: Listen to the music to annotate. When you press play (or spacebar) you'll see the playhead position in time but won't see any other visual representations
2. *Audio + visuals*: Listen to the music to annotate. When you press play (or spacebar) you'll see the playhead position in time. You may use visual representations to help you annotate
3. *Visuals only*: Annotate visually. You'll see visual representations of the music but won't be able to play/hear any sound

How to place boundaries

There are two ways to place a boundary:

1. While the audio is stopped (press) : Select the boundary button and click anywhere on the screen; you can select the boundary strength level from the dropdown list next to the boundary button
2. While the audio is playing ( ): Press [1], [2], [3] or [4] on your keyboard (the strength is defined by the key you press)

You can click on boundaries to adjust them (level/position) if necessary.

Remember to save often and push on the "Finish" button in order to advance to the next piece.

Tips

- Don't dwell on the same piece for too long
- Feel free to take small breaks every 10 pieces

The following image shows a common way of annotating in CosmoNote:

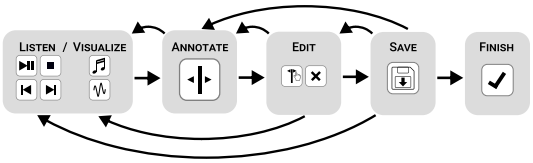


Figure C.1: (continued) English version of the instructions for the cross-modal annotations experiment (reformatted to fit the page). Second part of the instructions explaining the three possible interface configurations, instructions to place boundaries, and annotation tips.

C.2 KDE Profiles

Boundary annotations are often aggregated into continuous profile curves using Gaussian Kernel Density Estimation (KDE) curves (Silverman, 2017) to more easily visualize the contribution of multiple boundaries occurring around the same time than using one vertical bar per boundary. Each boundary contributing to the KDE representation can be weighted using its strength level so that its contribution to the segmentation of a piece is represented accordingly. The density estimator $\hat{f}$ for x_i samples drawn from a distribution, at a point x is defined by Silverman (2017) as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Where K is the kernel function, in our case a Gaussian kernel $K(x;h) \propto \exp(-\frac{x^2}{2h^2})$. These

curves create continuous distributions of annotations, concentrating the impact of annotations close to each other, depending on a parameter h, often called scale, bandwidth, or smoothing parameter. This parameter represents the horizontal scale or the standard deviation of the Gaussian kernel. The choice of h has to be reflected upon because it influences the result of each curve.

We present an example of the study described in Chapter 5. The three vertical panels in Figure C.2 show multiple KDE profile curves computed from boundary annotations of three pieces used in the study. The KDE profiles were computed using an evenly spaced array of 40 bandwidth values (on a logarithmic scale), ranging from 0.5 to 3.5 seconds (from yellow to violet in the color bar). A thick red line represents the bandwidth of 1 second, the same value used in all examples of this document. On a piece with a quadruple meter at 90 BPM, this bandwidth represents at least 1.5 beats.

Because stronger boundary levels (e.g., level 4) mark a more noticeable change, there are often fewer, more spaced-out strong boundaries throughout the music, most people identify them, and their contribution to the profile is substantial even in small numbers. On the contrary, lower level boundaries (e.g., level 1) may often occur in a smaller time scale, tend to occur more often, are more challenging to identify, and their contribution to the profile is minor unless many boundaries are placed around the same point.

The proposed bandwidth parameter of 1 second allows the profile curve to capture these boundaries with weak and strong levels, as shown by the red line in Figure C.2. Using KDE curves with this scale parameter, time differences in boundary annotations of any strength level produced by precision errors (e.g., placing a boundary 0.5 seconds to the left or the right of a note's onset) are less impactful on the aggregated curve. The most prominent peaks on boundary profiles aggregate many individual boundaries of higher levels.

The KDE representation helps interpret multiple annotations with unique strength levels at a glance. Higher peaks represent the segmentation agreed upon by the majority of annotators. Less frequent or weaker annotations produce peaks (with smaller heights) when concentrated around the same timestamp.

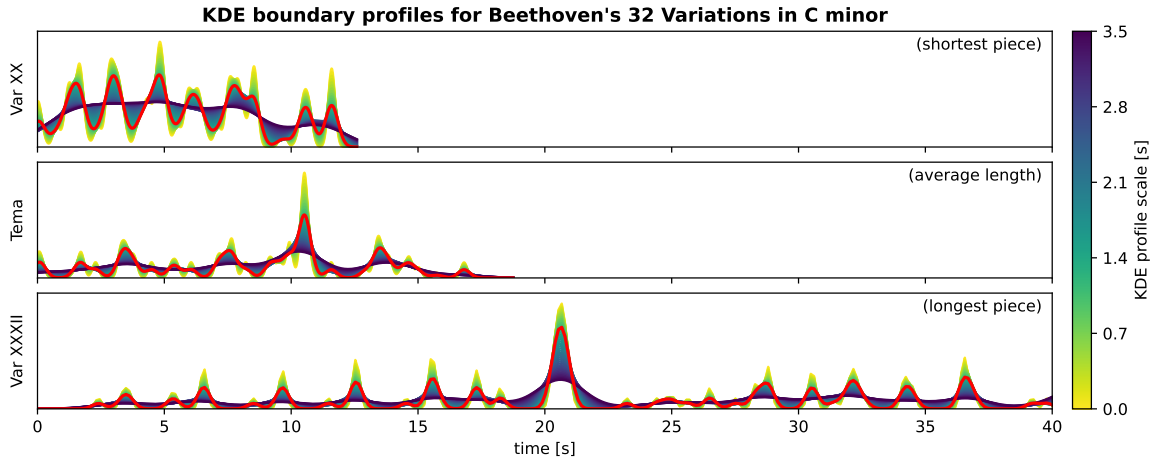


Figure C.2: Examples of scale differences when creating KDE boundary profiles for pieces with different lengths (the longest piece is truncated). The color bar represents an array of different scales in seconds. A scale of 1 s is displayed as a solid, red line.

C.3 Comparison Between Aural/Visual Annotations

Title	Duration	uOT	uOT normalized	Comparison
Var XXXII	129.04	64.04	0.5	AV vs. A
Var XIII	20.75	11.6	0.56	AV vs. V
Var XXX	25.45	14.7	0.58	AV vs. V
Var XXXI	28.52	16.79	0.59	AV vs. V
Var XXVII	18.76	11.25	0.6	AV vs. V
Tema	18.74	11.36	0.61	AV vs. V
Var III	14.55	8.91	0.61	AV vs. V
Var XXIV	14.25	8.82	0.62	AV vs. V
Var XVIII	24.75	15.49	0.63	AV vs. V
Var XVI	18.74	11.77	0.63	AV vs. V
Var XXI	13.26	8.72	0.66	AV vs. V
Var XXVI	15.68	10.55	0.67	V vs. A
Var XII	35.01	23.64	0.68	AV vs. V
Var XIX	13.56	9.32	0.69	AV vs. V
Var V	18.03	12.84	0.71	AV vs. V
Var IX	24.27	17.71	0.73	AV vs. V
Var VIII	19.47	14.76	0.76	AV vs. V
Var XXV	13.86	10.65	0.77	AV vs. V
Var XX	12.59	9.74	0.77	AV vs. P
Var IV	16.97	13.26	0.78	AV vs. V
Var X	16.84	13.2	0.78	AV vs. V
Var XXVIII	22.35	17.56	0.79	AV vs. V
Var II	13.22	10.57	0.8	AV vs. V
Var XI	16.42	14.5	0.88	AV vs. W
Var XIV	21.89	19.43	0.89	AV vs. V
Var XXIX	16.1	14.7	0.91	AV vs. V
Var I	13.79	12.6	0.91	AV vs. V
Var XXIII	16.92	15.9	0.94	AV vs. V
Var VI	15.4	14.56	0.95	AV vs. V
Var XVII	22.22	21.69	0.98	AV vs. V
Var XV	18.91	18.97	1	AV vs. A
Var VII	20.82	20.93	1.01	AV vs. V
Var XXII	15.37	17.2	1.12	AV vs. A

Table C.1: Smallest distances between visual and aural annotations. Exceptions to the trend are shown in red.

Title	Duration	uOT	uOT normalized	Comparison
Var XXXII	129.04	162.05	1.26	W vs. A
Var XVIII	24.75	31.92	1.29	P vs. A
Var XXVI	15.68	22.5	1.43	AV vs. P
Var IV	16.97	26.13	1.54	P vs. A
Var XXV	13.86	21.95	1.58	W vs. A
Var XIX	13.56	22.03	1.62	P vs. A
Var XIII	20.75	34.08	1.64	W vs. A
Var XXII	15.37	25.64	1.67	W vs. A
Var XI	16.42	28.24	1.72	P vs. A
Var XII	35.01	60.56	1.73	P vs. A
Var I	13.79	24.44	1.77	W vs. A
Var XXXI	28.52	51.92	1.82	AV vs. W
Var V	18.03	36.18	2.01	P vs. A
Var XXIV	14.25	29.79	2.09	W vs. A
Var XXVII	18.76	40.49	2.16	W vs. A
Var XXX	25.45	56.6	2.22	P vs. A
Var XXVIII	22.35	50.92	2.28	W vs. A
Var XX	12.59	29.18	2.32	W vs. A
Var XXI	13.26	31.25	2.36	W vs. A
Var XVI	18.74	44.36	2.37	P vs. A
Var III	14.55	34.51	2.37	P vs. A
Var X	16.84	40.57	2.41	W vs. A
Var IX	24.27	62.36	2.57	P vs. A
Var XVII	22.22	57.26	2.58	W vs. A
Var VII	20.82	53.77	2.58	W vs. A
Var VIII	19.47	50.33	2.59	W vs. A
Var XV	18.91	50.46	2.67	P vs. A
Var XXIII	16.92	46.54	2.75	P vs. A
Var VI	15.4	43.58	2.83	W vs. A
Var II	13.22	37.77	2.86	P vs. A
Var XXIX	16.1	46.46	2.89	W vs. A
Tema	18.74	61.93	3.3	P vs. A
Var XIV	21.89	83.04	3.79	P vs. A

Table C.2: Largest distances between visual and aural annotations. Exceptions to the trend are shown in red.

Appendix D

Free-Form Study

D.1 Annotation Instructions

Experiment annotation instructions

Thank you for agreeing to participate in this annotation experiment. The goal of this study is to understand how people perceive structures in performed music.

What are you asked to do?

1. Fill in a Short Musical Questionnaire [2 min]
2. Calibrate your sound and do a Hearing environment test [2 min]
3. Familiarize with the interface on the Training Collection (your answers won't be saved) [4 min]
4. Main task [~50 min]: Please annotate the performances of the music you hear. Use Boundaries (specific points in time), Regions (time selections), Groups (ensembles of notes) and Comments (something you could not annotate using the other tools). **Your ear should be your main guide**.

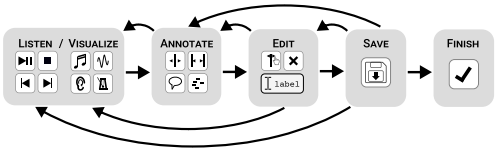
When you finish annotating the pieces, please fill our feedback questionnaire [2 min].

What does annotating a performance mean?

Annotating a performance means focusing on the performer's actions rather than the score. Use CosmoNote to mark how sound is manipulated to make the music expressive, this is called Musical Prosody. We are trusting your best judgement to annotate the music relying only on this information.

Recommendations

- You have around 45 minutes to annotate 2 pieces. If you've reached half that time, please proceed to the next piece. You can come back to finish your annotations of the first piece if you have time
- We recommend a balance between spontaneity and attentive listening
- Embrace the subjectivity of the task. Do not focus on finding a "correct answer"
- You may use labels to be more specific about your annotations
- Visual cues should be used as complementary information. Trust your ears!
- The following image shows a common strategy for annotating in CosmoNote:



- Remember you can use all four annotation types in CosmoNote: Boundaries with their strength levels from 1 (weakest) to 4 (strongest), Regions, Groups and Comments. See "Placing annotations in CosmoNote" document.

Figure D.1: English version of the instructions for the free-form annotations experiment (reformatted to fit the page), less detailed condition.

Experiment annotation instructions

Thank you for agreeing to participate in this annotation experiment. The goal of this study is to understand how people perceive structures in performed music.

What are you asked to do?

1. Fill in a short Musical Questionnaire [2 min]
2. Calibrate your sound a do a hearing environment test [2 min]
3. Familiarize with the interface on the Training Collection (your answers won't be saved) [4 min]
4. Main task [~50 min]: Please annotate the performances of the music you hear. Use Boundaries (specific points in time), Regions (time selections), Groups (ensembles of notes) and Comments (something you could not annotate using the other tools). **Your ear should be your main guide**.

When you finish annotating the piece, please fill our feedback questionnaire [2 min].

What does annotating a performance mean?

Annotating a performance means focusing on the performer's actions rather than the score. Use CosmoNote to mark how sound is manipulated to make the music expressive, this is called Musical Prosody.

What are segmentation and prominence?

Segmentation

Segmentation is the process of dividing something, in this case music, into meaningful units. These are some examples of segmentation you may annotate:

- Boundaries: time points that separate a music stream into segments representing meaningful chunks of music e.g., a musical idea or a musical thought. Boundaries not only separate a larger piece of music into smaller, coherent units, they also help listeners make sense of the music. There is a boundary annotation type in CosmoNote, with four levels of boundaries you may choose from, defined from 1 (weakest) to 4 (strongest).
- Transitions: musical passages that set up a change that is coming in the music. They can be seen as a link between musical ideas. Transitions blur changes in the music by moving slowly through them. You may use regions to annotate transitions.
- Pauses: devices that add space between two adjacent structures. Executed by the performer via lingering on notes or using silence. You may use regions to annotate pauses.

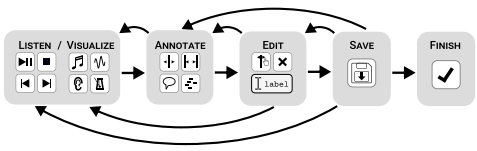
Prominence

Prominence characterizes an emphasis drawn towards a certain of a whole in the music. A non-exhaustive list of examples of prominence in music are:

- Stress: An emphasis of a particular element to make it more prominent than those around it. Stress may be indicated by a combination of performer actions like an increase in sound intensity, duration, or a change in timbre. You may use boundaries, comments or groups to mark stress.
- Melodic salience: a special case of prominence dedicated to the melody of a piece. It may be recognized by an increase in loudness and duration or a variation in timbre of melody-notes. You may use groups to annotate melodic salience.
- Tipping points: Moments where musical time is suspended/stretched in a state beyond which a return to the pulse is inevitable. You may use regions, boundaries or groups to mark tipping points.

Recommendations

- You have around 45 minutes to annotate 2 pieces. If you've reached half that time, please proceed to the next piece. You can come back to finish your annotations of the first piece if you have time
- We recommend a balance between spontaneity and attentive listening
- Embrace the subjectivity of the task. Do not focus on finding a "correct answer"
- You may use labels to be more specific about your annotations
- Visual cues should be used as complementary information. Trust your ears!
- The following image shows a common strategy for annotating in CosmoNote:



- Remember you can use all four annotation types in CosmoNote: Boundaries, Regions, Groups and Comments. See "Placing annotations in CosmoNote" document.

Figure D.2: English version of the instructions for the free-form annotations experiment, more detailed condition.

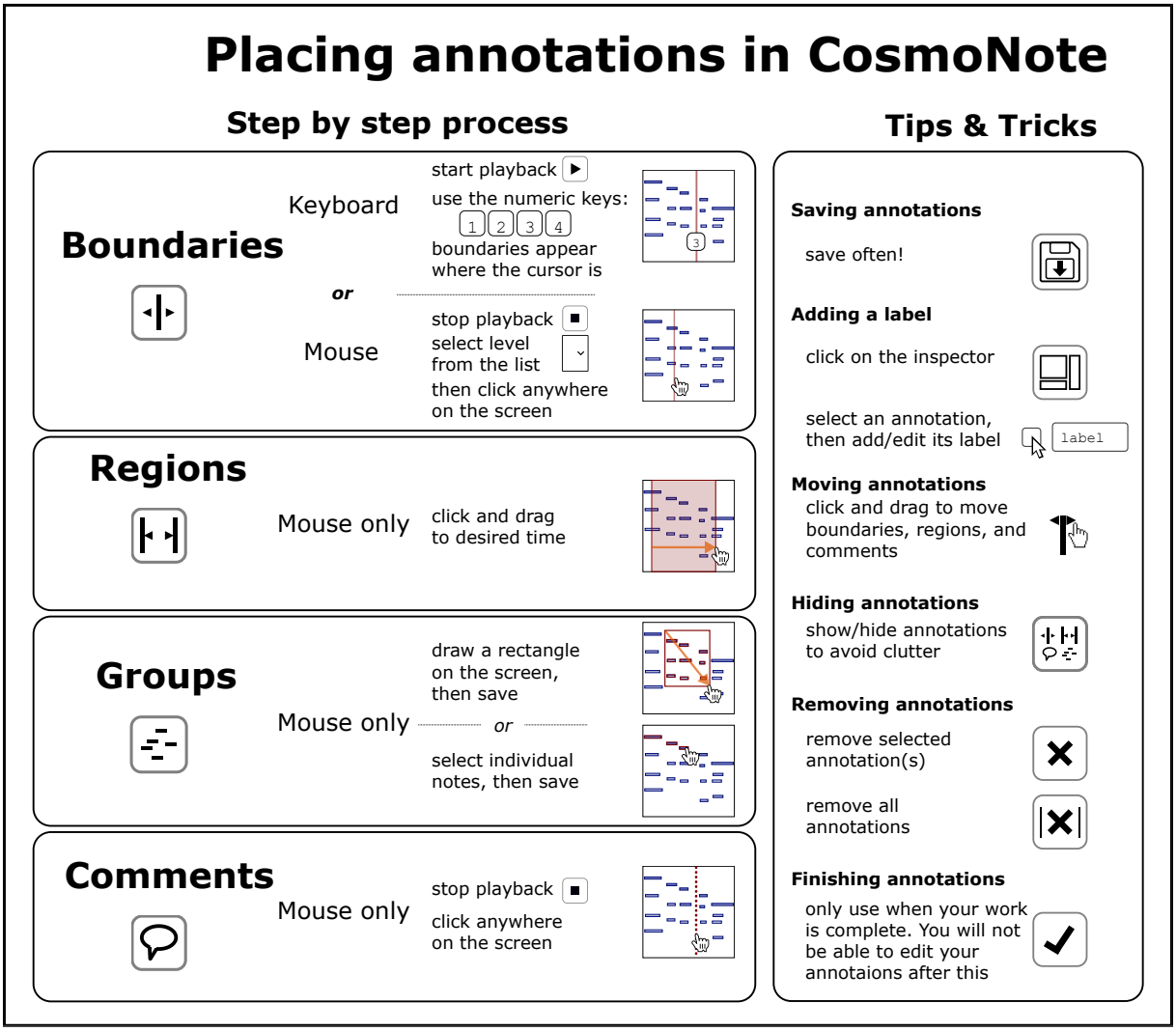


Figure D.3: English version of the printed annotation placement guide for the free-form annotations experiment, for both experimental conditions.

D.2 Comparing Common Properties

D.2.1 Time Similarity

The simplest way to discriminate between annotations is to compare their corresponding time properties by annotation type. We can compare the time similarity between boundaries, regions, and comments but not note groups because of their complexity. For example, in the case of two annotations, A and B, annotation A is considered to be similar to annotation B if its timestamp (a single point for boundaries and comments, but two points start/end for regions) is within a window (we defined by default a window of 0.5 seconds) of the timestamp of annotation B.

D.2.2 Jaccard Similarity

The Jaccard similarity coefficient or Jaccard index $J(A, B)$ is defined as the intersection of two sets A and B over its union:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard coefficient can be used to measure the similarity between sets of patterns because of its conceptual and computational simplicity, interpretability and applicability (Fletcher & Islam, 2018). We use this coefficient to measure note group similarity. Each note group a user creates has a unique ID and can be reduced to a list of tuples containing three note properties: start time, end time, and MIDI note number. Jaccard similarity coefficients are computed between note groups with distinct IDs to find similar note groups. Group pairs with a similarity coefficient over the threshold value of 0.8 are considered similar groups and kept for further analysis.

D.2.3 Graphs and Communities

Once we have identified similar annotations using time similarity or the Jaccard similarity coefficient, we evaluate the relationship of similar annotation pairs by representing them as graphs. In a graph, nodes (its elements) that are related in some way (in our case, their similarity) are connected by edges (lines). Pairs of annotations having common properties are the rows of a two-column edge list table. Connections between nodes are extracted from the edge list table, building a graph. Our analysis considers graphs created by similar annotations with highly interconnected nodes called communities. Figure D.4 shows an example where the `louvain_communities` function in the `networkx` Python package is applied to automatically find the best community partition in our graphs with the Louvain Community Detection Algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

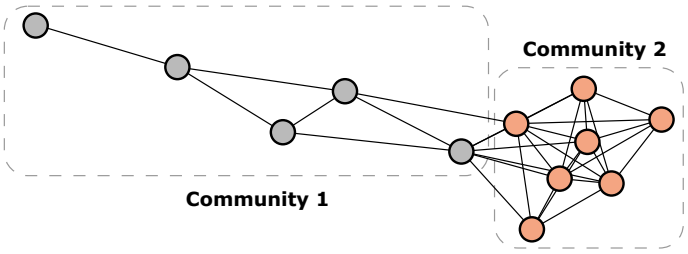


Figure D.4: Example of a graph created from a portion similar region annotations. The graph (interconnected nodes and edges) is split into two communities using the Louvain community detection algorithm. Nodes filled in gray represent Community 1, and nodes filled in tangerine represent Community 2.

D.3 Category Terms

Category	Associated terms
prosody segmentation	A, A', A1, A2, AA, answer, B, B', B1, B2, BB, beginning, binary, boundaries, boundary, break, breath, bridge, C, C', C1, C2, call, chorus, coda, codetta, conclusion, D, D', D1, D2, development, end, ending, epilogue, episode, exposition, figure, form, idea, intro, introduction, motif, motive, movement, outro, parallel period, part A, part A', part B, passage, pattern, pause, period, phrase, postlude, prelude, punctuate, punctuation, question, recall, recapitulation, refrain, repetition, response, ritornello, section, segment, segmentation, start, subpart, subphrase, subsection, transition, verse
prosody prominence	accent, accentuate, accentuated, accentuation, emphasis, emphasize, emphasized, highlight, melodic salience, melody, prominence, prominent, salient, stress, stretched, suspended, suspension, tipping point
dynamics	amplitude, crescendo, decrescendo, diminuendo, dynamic, dynamics, forte, fortissimo, inaudible, intensity, loud, louder, loudest, loudness, mezzo-forte, mezzo-piano, morendo, pianissimo, piano, quiet, sforzando, silence, soft, softer, softest, strong, stronger, strongest, subito, volume
tempo	accelerando, accelerate, accelerated, acceleration, adagio, allargando, allegretto, allegro, andante, andantino, fast, faster, fastest, grave, larghetto, largo, lento, meno, moderato, molto, pace, piu, poco, prestissimo, presto, quick, rallentando, ritardando, rubato, slow, slowdown, slower, slowest, slowing, stringendo, tempi, tempo, troppo, vivace, vivo
tuning	temperament, microtonal, microtone, tune, tuning
harmony	accidental, augmentation, augmented, basso, bitonality, cadence, cadenza, chord, chromatic, chromaticism, cluster, consonance, diatonic, diminished, disjunct, dissonance, dominant, enharmonic, flat, half step, harmonization, harmony, home key, homophonic, interval, inversion, key, major, minor, mode, modulation, natural, octave, pedal, polytonality, progression, reharmonization, retrograde, scale, seventh, sharp, sixth, step, suspended, tension, texture, tonality, tone, tonic, transposition, triad, unison, whole
rhythm	bar, beat, dotted, downbeat, duple, measure, meter, metric, metrical, polyrhythm, pulse, quadruple, rhythm, rhythmic, syncopation, triple, triplet, upbeat
melody	bass, contrary motion, countermelody, counterpoint, diminution, idée fixe, imitation, Leitmotif, melisma, melodious, melody, monophonic, note, pentatonic, polyphonic, polyphony, range, register, sequence, subject, theme, tone, voice

Table D.1: Categories for basic text classification of annotations - part 1. Eight arbitrary keywords created from preliminary data inspection are shown on the left column while associated terms (tokens) with a few variations are shown on the right column. These categories are used in the first stage basic text classification.

Category	Associated terms
musical periods	atonal, atonality, Baroque, classic, classical, classicism, contemporary, expressionism, Medieval, minimalism, modern, nationalism, neo-classicism, Renaissance, romantic, romanticism, serialism
form and genre	aria, art, avant-garde, ballet, bebop, blues, buffa, canon, cantata, capella, chamber, chant, chorale, computer, concert, concerto, concrète, Dies Irae, etude, film, form, fugue, fusion, genre, Gregorian chant, habanera, hip-hop, hot jazz, Impressionism, improvisation, incidental, jazz, libretto, Lied, madrigal, mass, mazurka, motet, Musikdrama, nocturne, opera, operetta, organum, overture, poem, polka, polonaise, pop, popular, postlude, prelude, quotation, R&B, raga, ragtime, rap, Requiem, ritornello, RnB, rock, rondo, scherzo, serenade, Singspiel, sonata, sonata-rondo, song, soundtrack, Sprechstimme, strophic, suite, swing, symphony, tala, ternary, through-composed, trio, variation, verismo, waltz, Western
musical instruments	alto, band, baritone, bass, bassoon, bells, brass, cello, chimes, clarinet, conductor, contrabassoon, contralto, cornet, cymbals, drum, ensemble, flute, fortepiano, gamba, gamelan, glockenspiel, gong, guitar, harp, harpsicord, hi-hat, horn, instrument, instrumentation, kettledrums, koto, lute, marimba, metronome, mezzo-soprano, mute, oboe, orchestra, orchestration, organ, percussion, piano, pianoforte, quartet, recorder, reed, sackbut, saxophone, shakuhachi, shawm, sitar, soprano, sousaphone, string, synthesizer, tabla, tenor, timpani, trombone, tuba, Ud, viol, viola, violin, violoncello, voice, woodwind, xylophone
performer	author, artist, composer, composition, instrumentalist, musician, performer, pianist, player, soloist, virtuoso
piano terms	corda, key, keyboard, pedal, piano, sostenuto, sustain
pitch and timbre	attack, bright, chroma, color, decay, flatter, high, high-pitched, higher, highest, low, lower, lowest, pitch, release, rough, roughness, round, sharper, sustain, timbre, timing, warm
performance and technique	accompaniment, arpeggiate, arpeggiated, arpeggio, arpeggios, comma, da capo, execution, expressive, expressiveness, expressivity, falsetto, fermata, glissando, hand, legato, ornament, ornamented, ostinato, notation, performance, pizzicato, play, recitative, recitativo, scat, score, staccato, tremolo, trill, tutti, vibrato, virtuosic
sound and recording	audio, cable, compression, delay, echo, equalization, mic, microphone, MIDI, mp3, reverberation, signal, noise, sound, stereo
positive	acceptance, admiration, affect, amazement, amorous, amusement, arousal, awe, contentment, delight, ecstasy, elated, enjoy, enthusiastic, euphoria, excited, fun, glad, glee, good, happier, happiness, happy, joy, joyful, joyous, love, nice, overexcite, overexcited, overjoy, overjoyed, pleasure, positive, pride, relief, satisfied, trust, wonderful
negative	afraid, anger, annoyance, anxiety, apprehension, bad, bitter, bitterness, boredom, boring, contempt, discontent, disgust, dislike, disappointment, distraction, enraged, fear, fearful, frustration, furious, fury, gloom, grief, guilt, hate, hopeless, irritation, loathing, mad, miserable, negative, nervous, rage, regret, sad, sadness, shame, startle, startled, tense, terror, ugly, worried
neutral-ambiguous	anticipation, calm, calmer, compassion, emotion, interest, mood, neutral, nostalgia, peace, pensiveness, relax, relaxed, serenity, surprise, surprised, vigilance
ID category	g1, g2, g3, g4, g5, g6, g7, g8, g9, R1, R2, R3, R4, R5
useless	N/A
miscellaneous	sci-fi

Table D.2: Categories for basic text classification of annotations - part 2. Fourteen arbitrary keywords created from preliminary data inspection are shown on the left column while associated terms (tokens) with a few variations are shown on the right column. These categories are used in the first stage basic text classification.

Larger category	Smaller categories
segmentation	prosody segmentation
prominence	prosody prominence
structure	form and genre musical periods
music descriptors	dynamics harmony melody pitch and timbre rhythm tempo tuning
performance and sound	musical instruments performance and technique performer piano terms sound and recording
emotions	negative neutral-ambiguous positive
miscellaneous	ID category miscellaneous useless

Table D.3: Seven larger categories (keys) gathering the twenty-two subcategories (values) for classifying annotation labels into common categories.

D.4 Boundaries

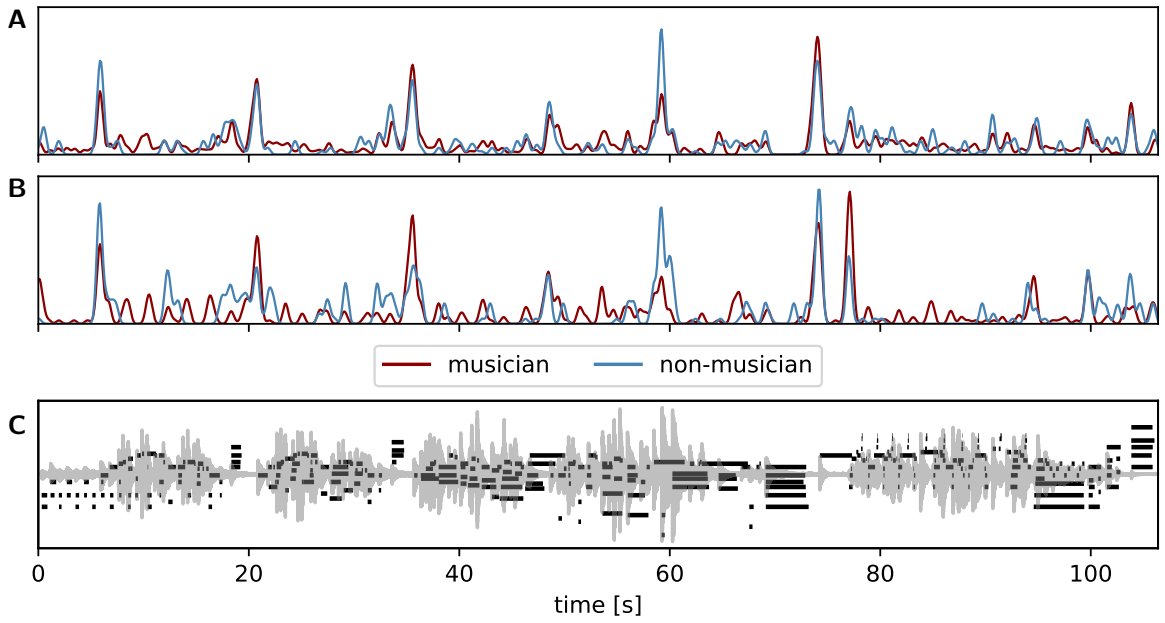


Figure D.5: KDE boundary profiles for all boundary annotations in Edvard Grieg's Solveig's Song are drawn as solid lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

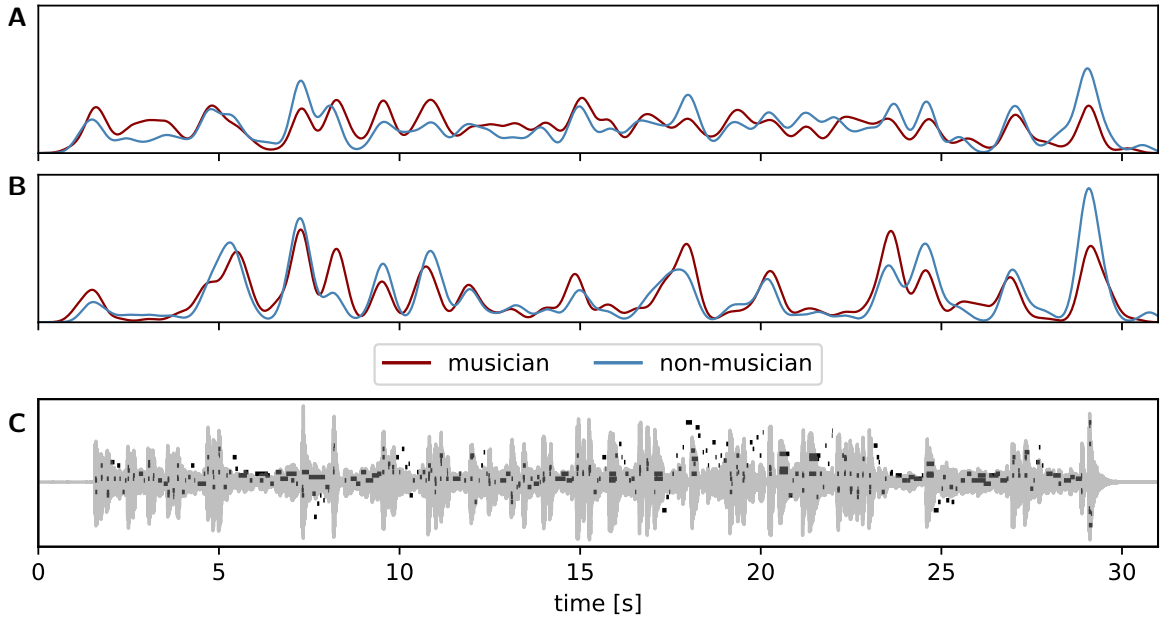


Figure D.6: KDE boundary profiles for all boundary annotations in Pierre Boulez's *Fragment d'une ébauche* are drawn as solid lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

D.5 Regions

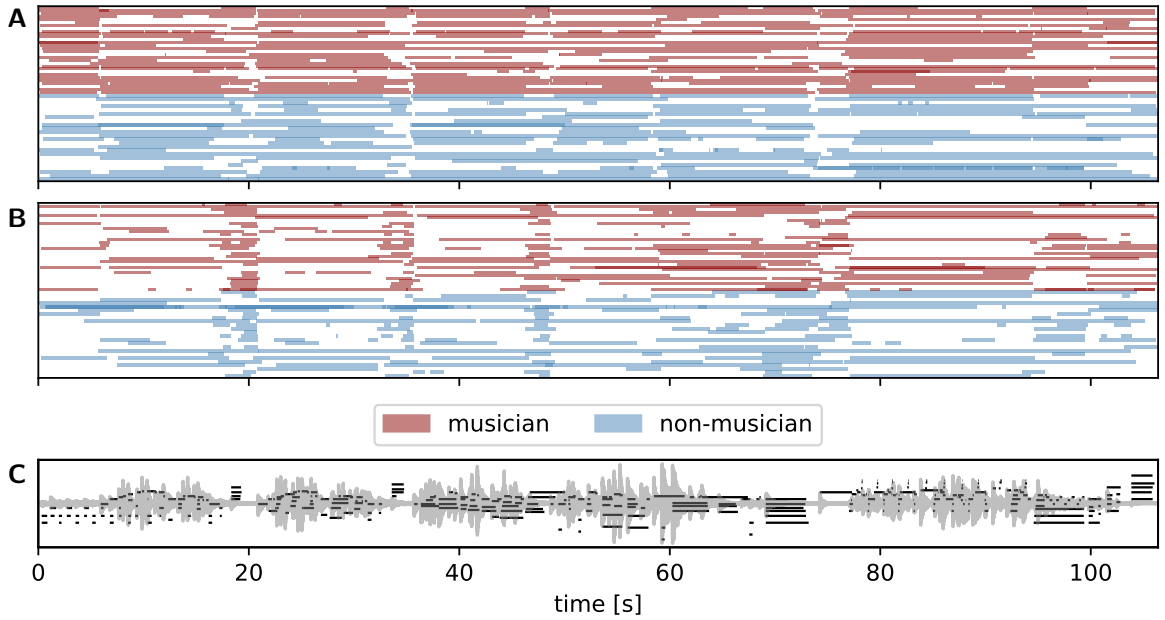


Figure D.7: All region annotations in Edvard Grieg's Solveig's Song. Regions are vertically stacked by user ID for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

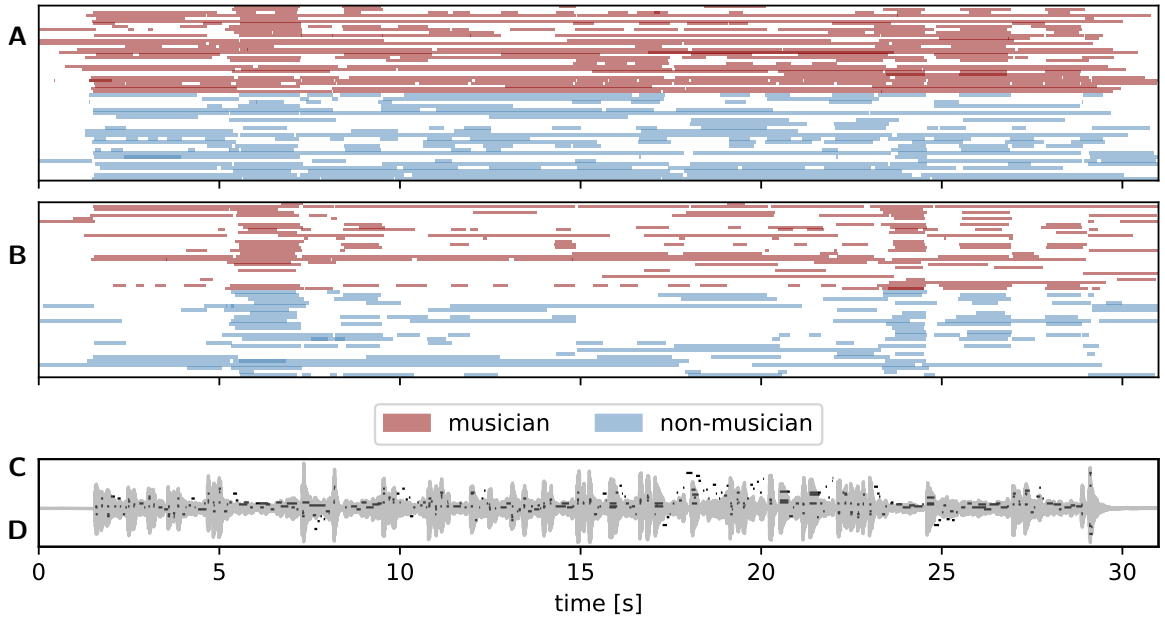


Figure D.8: All region annotations in Pierre Boulez's *Fragment d'une ébauche*. Regions are vertically stacked by user ID for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

D.6 Common Regions

Tables in this section detail the first ten communities found by the time similarity on region annotations. Each community (set) is split by musical ability into two subsets (musician, non-musician). Each subset contains a number of similar regions given by the "Subset count" column of each table. The start and end time values are averaged from all common regions in the subset.

Index	Musical abilities	Start time	End time	Subset count	Total count
1	musician	23.76	24.54	11	18
	non-musician	23.83	24.58	7	
2	musician	23.32	24.55	9	17
	non-musician	23.42	24.59	8	
3	musician	5.33	7.23	7	16
	non-musician	5.34	7.20	9	
4	musician	25.60	26.86	6	14
	non-musician	25.66	26.82	8	
5	musician	8.53	9.49	5	12
	non-musician	8.36	9.46	7	
6	musician	27.91	28.92	7	12
	non-musician	27.92	28.90	5	
7	musician	5.55	7.19	8	12
	non-musician	5.80	7.27	4	
8	musician	7.38	7.91	6	9
	non-musician	7.39	7.94	3	
9	musician	22.06	23.19	5	9
	non-musician	21.98	23.34	4	
10	musician	20.51	20.78	5	8
	non-musician	20.43	20.68	3	

Table D.4: Common regions Grieg, less detailed

Index	Musical abilities	Start time	End time	Subset count	Total count
1	musician	47.29	48.56	4	10
	non-musician	47.29	48.59	6	
2	musician	74.14	77.20	7	10
	non-musician	74.21	77.14	3	
3	musician	46.24	48.61	6	9
	non-musician	46.18	48.58	3	
4	musician	18.23	20.82	7	9
	non-musician	18.19	20.70	2	
5	musician	19.40	20.84	4	9
	non-musician	19.49	20.79	5	
6	musician	34.78	35.69	4	8
	non-musician	34.86	35.62	4	
7	musician	72.92	74.27	5	7
	non-musician	72.90	74.12	2	
8	musician	46.51	49.42	3	7
	non-musician	46.66	49.24	4	
9	musician	33.49	35.56	5	7
	non-musician	33.42	35.96	2	
10	musician	19.22	20.78	3	7
	non-musician	19.09	20.79	4	

Table D.5: Common regions Grieg, more detailed condition.

Index	Musical abilities	Start time	End time	Subset count	Total count
1	musician	5.40	7.22	10	17
	non-musician	5.49	7.18	7	
2	musician	27.83	28.90	11	12
	non-musician	27.89	28.89	1	
3	musician	1.19	5.56	5	11
	non-musician	1.43	5.42	6	
4	musician	25.55	26.76	7	10
	non-musician	25.50	26.87	3	
5	musician	28.95	29.44	4	9
	non-musician	28.83	29.33	5	
6	musician	26.96	27.62	4	9
	non-musician	26.88	27.55	5	
7	musician	23.58	24.62	5	9
	non-musician	23.67	24.57	4	
8	musician	23.34	24.50	7	9
	non-musician	23.38	24.54	2	
9	musician	7.28	8.19	5	9
	non-musician	7.42	7.99	4	
10	musician	8.41	9.49	5	9
	non-musician	8.48	9.58	4	

Table D.6: Common regions Boulez, less detailed

Index	Musical abilities	Start time	End time	Subset count	Total count
1	musician	23.66	24.51	12	19
	non-musician	23.71	24.59	7	
2	musician	5.33	7.23	7	16
	non-musician	5.34	7.20	9	
3	musician	25.60	26.86	6	14
	non-musician	25.66	26.82	8	
4	musician	23.26	24.55	7	14
	non-musician	23.38	24.55	7	
5	musician	27.91	28.92	7	12
	non-musician	27.92	28.90	5	
6	musician	5.55	7.19	8	12
	non-musician	5.80	7.27	4	
7	musician	8.44	9.46	4	11
	non-musician	8.36	9.46	7	
8	musician	22.06	23.19	5	9
	non-musician	21.98	23.34	4	
9	musician	20.51	20.78	5	8
	non-musician	20.43	20.68	3	
10	musician	1.44	5.35	5	7
	non-musician	1.42	5.39	2	

Table D.7: Common regions Boulez, more detailed

D.7 Comments

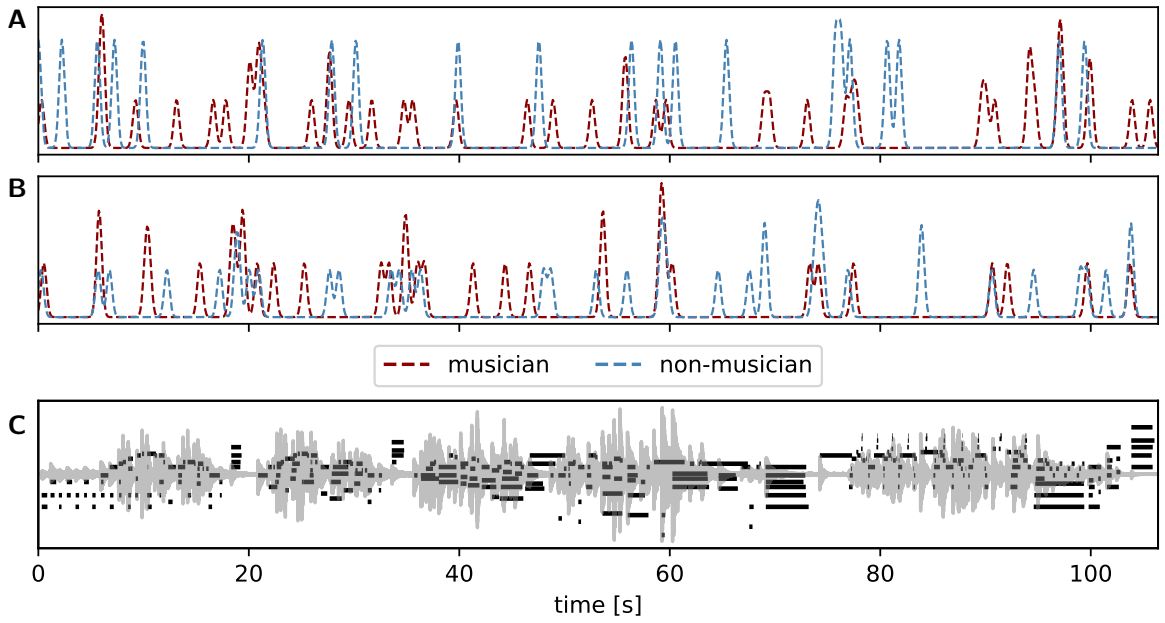


Figure D.9: KDE profiles for all comment annotations in Edvard Grieg's Solveig's Song are drawn as dashed lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

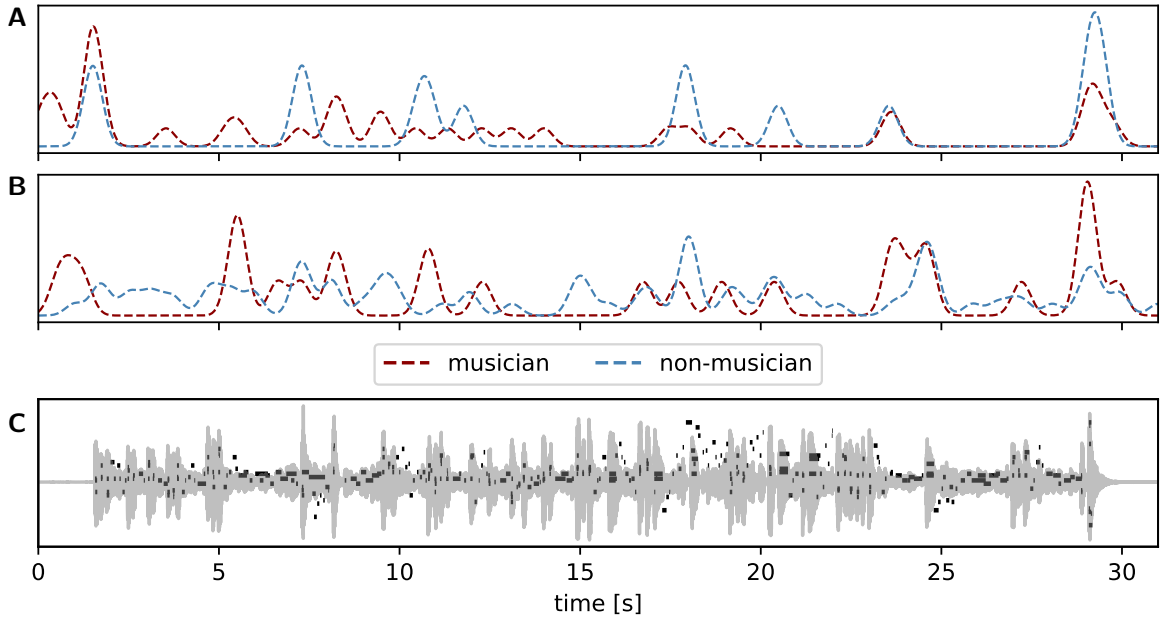
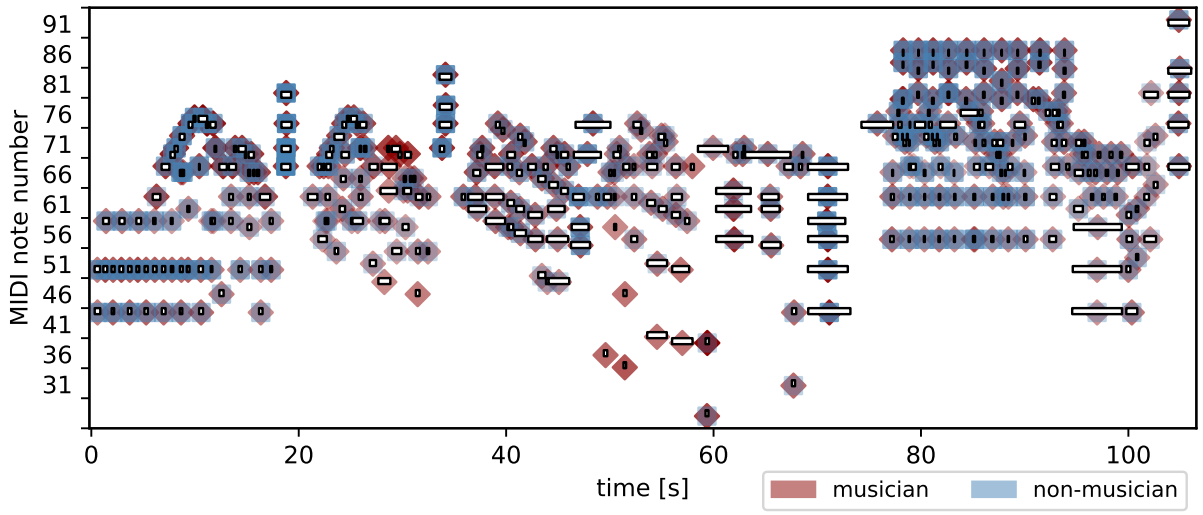
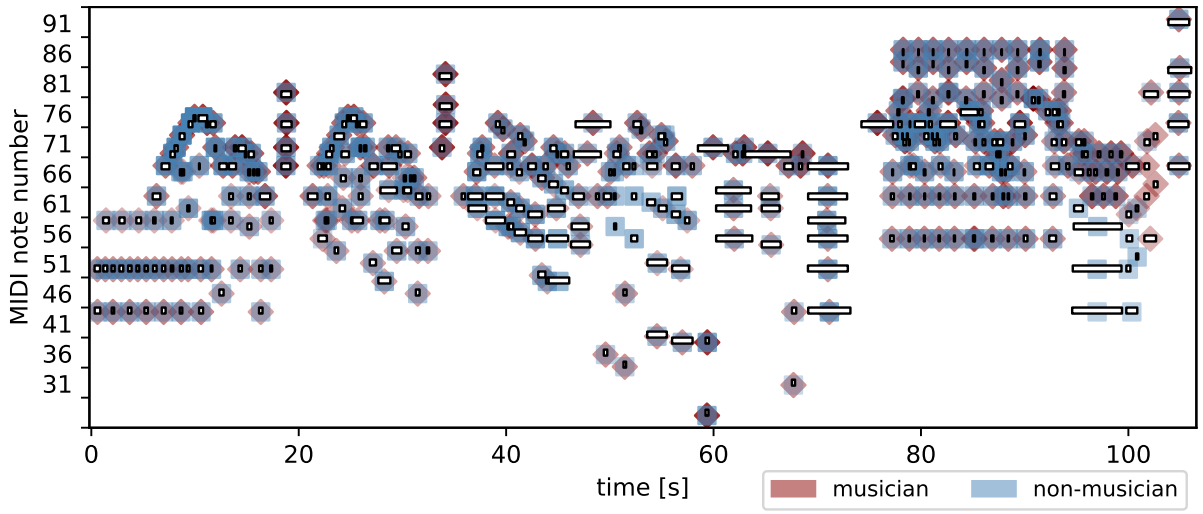


Figure D.10: KDE profiles for all comment annotations in Pierre Boulez's *Fragment d'une ébauche* are drawn as dashed lines for each condition. (A) Less detailed (B) More detailed. (C) Waveform and notes (gray curve and black lines). The first two rows display musicians in red and non-musicians in blue.

D.8 Note Groups

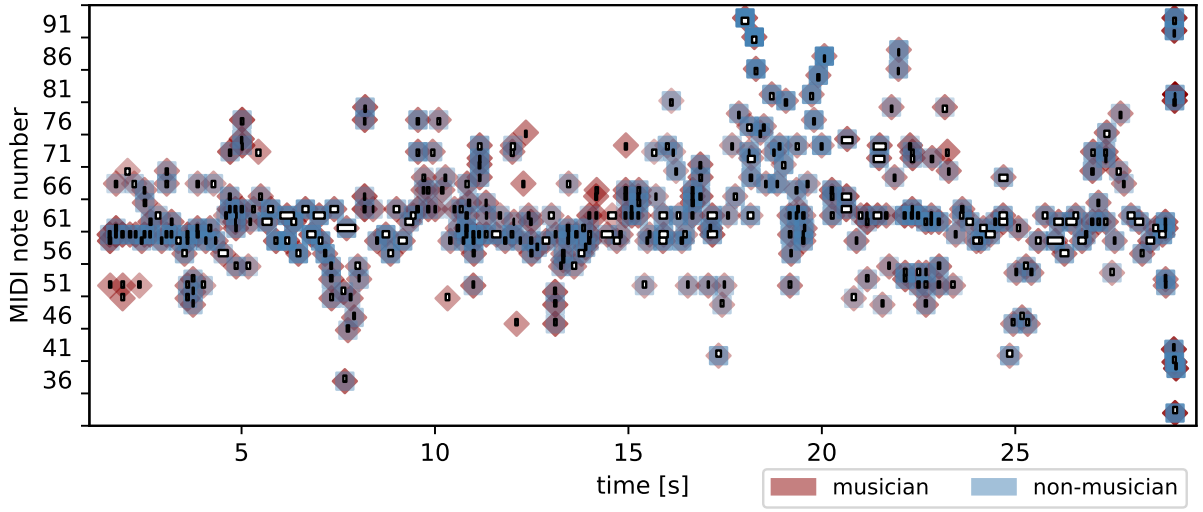


(a) Less detailed.

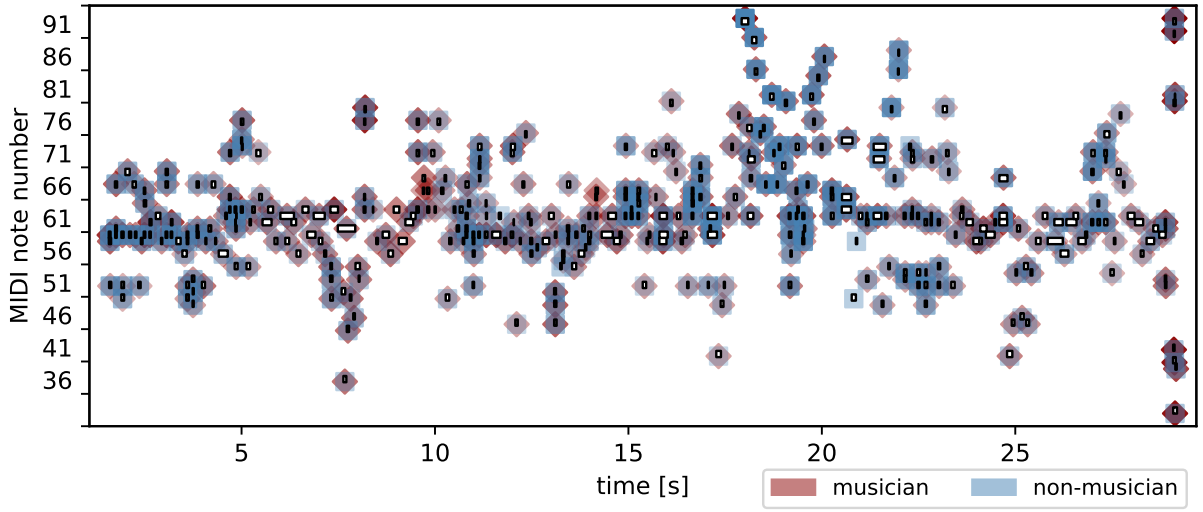


(b) More detailed.

Figure D.11: All note group annotations in Edvard Grieg's Solveig's Song. Note groups (diamonds and squares) overlaid on top of notes (outlined rectangles). Musicians are displayed in red and non-musicians in blue. The color's opacity represents how often a note was selected.



(a) Less detailed.



(b) More detailed.

Figure D.12: All note group annotations in Pierre Boulez's *Fragment d'une ébauche*. Note groups (diamonds and squares) overlaid on top of notes (outlined rectangles). Musicians are displayed in red and non-musicians in blue. The color's opacity represents how often a note was selected.

D.9 Common Note Groups

Tables in this section detail the first ten communities found by the Jaccard similarity on note group annotations. Each community (set) is split by musical ability into two subsets (musician, non-musician). Each subset contains note tuples (with start time, end time, and note number) detailing how common annotations relate. To better illustrate the data, let us analyze one annotation in Table D.8: Set 4 ("Index" column) has five note group annotations (shown in the "Total count" column), and is formed by two subsets, Subset 4–musician containing three unique annotations (shown in the "Subset count" column) and Subset 4–non-musician containing two unique annotations (also shown in the "Subset count" column). Looking at the "Subset detail" column of Subset 4–musician, the first note (8.48, 8.53, 60) appears in all three annotations while the last note (9.52, 9.58, 65) appears only in one annotation. Similarly, looking at the

detail of Subset 4–non-musician, the first note (8.65, 8.81, 61) appears in its two annotations while the last note (8.48, 8.53, 60) appears only in one. The value in the "Mean notes" column is a weighted mean of number of note tuples in each subset, weighted by their frequency. The subset detail for subsets containing more than ten note tuples is truncated to fit the tables.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
1	musician	8	{(29.08, 29.15, 41): 12, (29.08, 29.12, 43): 12, (29.08, 29.18, 33): 12, (29.08, 29.15, 95): 12, (29.09, 29.15, 82): 12, (29.09, 29.13, 93): 12, (29.13, 29.13, 83): 12, (29.14, 29.17, 40): 12}	12	15
	non-musician	8	{(29.08, 29.15, 95): 3, (29.09, 29.13, 93): 3, (29.09, 29.15, 82): 3, (29.13, 29.13, 83): 3, (29.08, 29.15, 41): 3, (29.08, 29.12, 43): 3, (29.08, 29.18, 33): 3, (29.14, 29.17, 40): 3}	3	
2	musician	10.25	{(16.63, 16.68, 64): 4, (16.63, 16.69, 66): 4, (16.64, 16.68, 67): 4, (16.64, 16.68, 68): 4, (16.83, 16.9, 70): 4, (16.84, 16.89, 67): 4, (16.84, 16.88, 72): 4, (17.05, 17.28, 64): 4, (17.05, 17.29, 61): 4, (17.1, 17.13, 62): 4, ...}	4	6
	non-musician	11.5	{(16.63, 16.68, 64): 2, (16.63, 16.69, 66): 2, (16.64, 16.68, 67): 2, (16.64, 16.68, 68): 2, (16.83, 16.9, 70): 2, (16.84, 16.89, 67): 2, (16.84, 16.88, 72): 2, (17.05, 17.28, 64): 2, (17.05, 17.08, 53): 2, (17.05, 17.29, 61): 2, ...}	2	
3	musician	4	{(8.16, 8.2, 65): 6, (8.17, 8.19, 67): 6, (8.17, 8.2, 79): 6, (8.18, 8.21, 81): 6}	6	6
4	musician	7.33	{(8.48, 8.53, 60): 3, (8.65, 8.81, 61): 3, (8.79, 8.91, 58): 3, (8.93, 9.08, 65): 3, (9.06, 9.27, 60): 3, (9.23, 9.44, 63): 3, (9.38, 9.5, 64): 3, (9.52, 9.58, 65): 1}	3	5
	non-musician	6.5	{(8.65, 8.81, 61): 2, (8.79, 8.91, 58): 2, (8.93, 9.08, 65): 2, (9.06, 9.27, 60): 2, (9.23, 9.44, 63): 2, (9.38, 9.5, 64): 2, (8.48, 8.53, 60): 1}	2	
5	musician	17.25	{(10.79, 10.84, 61): 4, (10.79, 10.84, 62): 4, (10.79, 10.87, 69): 4, (10.79, 10.85, 64): 4, (10.85, 10.91, 66): 4, (10.97, 11.02, 61): 4, (10.97, 11.01, 53): 4, (10.97, 11.02, 60): 4, (10.98, 11.03, 58): 4, (11.12, 11.16, 61): 4, ...}	4	5
	non-musician	17	{(10.79, 10.87, 69): 1, (10.85, 10.91, 66): 1, (10.79, 10.85, 64): 1, (10.79, 10.84, 62): 1, (10.79, 10.84, 61): 1, (10.97, 11.02, 61): 1, (10.97, 11.02, 60): 1, (11.11, 11.2, 75): 1, (11.14, 11.17, 73): 1, (11.13, 11.17, 72): 1, ...}	1	

Table D.8: Common note groups (1–5). Grieg, less detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
6	musician	24	{(22.11, 22.17, 64): 4, (22.12, 22.19, 54): 4, (22.12, 22.19, 55): 4, (22.25, 22.31, 75): 4, (22.3, 22.35, 64): 4, (22.3, 22.39, 73): 4, (22.31, 22.37, 74): 4, (22.32, 22.35, 65): 4, (22.46, 22.51, 64): 4, (22.46, 22.54, 55): 4, ...}	4	5
	non-musician	23	{(21.97, 21.99, 87): 1, (21.98, 22.0, 90): 1, (22.11, 22.17, 64): 1, (22.12, 22.19, 54): 1, (22.12, 22.19, 55): 1, (22.25, 22.31, 75): 1, (22.3, 22.35, 64): 1, (22.3, 22.39, 73): 1, (22.31, 22.37, 74): 1, (22.32, 22.35, 65): 1, ...}	1	
7	musician	12.5	{(28.72, 28.89, 61): 1, (28.86, 28.9, 63): 2, (28.87, 28.9, 54): 2, (28.87, 28.9, 62): 2, (28.87, 28.91, 53): 2, (29.08, 29.15, 41): 2, (29.08, 29.12, 43): 2, (29.08, 29.18, 33): 2, (29.08, 29.15, 95): 2, (29.09, 29.15, 82): 2, ...}	2	5
	non-musician	13	{(28.72, 28.89, 61): 3, (28.86, 28.9, 63): 3, (28.87, 28.9, 54): 3, (28.87, 28.9, 62): 3, (28.87, 28.91, 53): 3, (29.08, 29.15, 41): 3, (29.08, 29.12, 43): 3, (29.08, 29.18, 33): 3, (29.08, 29.15, 95): 3, (29.09, 29.15, 82): 3, ...}	3	
8	musician	10	{(14.9, 14.96, 67): 4, (14.9, 14.96, 68): 4, (14.9, 14.95, 64): 4, (14.91, 14.96, 75): 3, (15.06, 15.11, 64): 4, (15.06, 15.1, 65): 4, (15.23, 15.28, 67): 4, (15.23, 15.28, 64): 4, (15.24, 15.3, 68): 4, (15.24, 15.3, 66): 4, ...}	4	4
9	musician	9	{(23.44, 23.48, 61): 1, (23.57, 23.66, 64): 2, (23.6, 23.66, 65): 2, (23.74, 23.91, 63): 2, (23.92, 24.07, 60): 2, (24.07, 24.14, 60): 2, (24.09, 24.26, 62): 2, (24.24, 24.47, 61): 2, (24.46, 24.5, 63): 2, (23.35, 23.43, 53): 1}	2	4
	non-musician	9	{(23.57, 23.66, 64): 2, (23.6, 23.66, 65): 2, (23.74, 23.91, 63): 2, (23.92, 24.07, 60): 2, (24.07, 24.14, 60): 2, (24.09, 24.26, 62): 2, (24.24, 24.47, 61): 2, (24.46, 24.5, 63): 2, (24.59, 24.79, 63): 1, (24.59, 24.78, 64): 1}	2	
10	musician	6	{(20.23, 20.29, 68): 4, (20.23, 20.3, 64): 4, (20.23, 20.3, 67): 4, (20.51, 20.75, 65): 4, (20.52, 20.76, 67): 4, (20.52, 20.79, 76): 3, (20.77, 20.88, 51): 1}	4	4

Table D.9: Common note groups (6–10). Grieg, less detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
1	musician	4.14	{(18.32, 19.26, 81): 7, (18.32, 19.19, 69): 7, (18.33, 19.31, 76): 7, (18.34, 19.22, 72): 7, (74.26, 77.32, 76): 1}	7	11
	non-musician	4	{(18.32, 19.26, 81): 4, (18.32, 19.19, 69): 4, (18.33, 19.31, 76): 4, (18.34, 19.22, 72): 4}	4	11
2	musician	4.14	{(33.58, 34.72, 84): 7, (33.58, 34.72, 76): 7, (33.59, 34.76, 79): 7, (33.6, 34.03, 72): 7, (58.49, 61.41, 72): 1}	7	10
	non-musician	4	{(33.58, 34.72, 84): 3, (33.58, 34.72, 76): 3, (33.59, 34.76, 79): 3, (33.6, 34.03, 72): 3}	3	10
3	musician	2	{(59.22, 59.51, 28): 6, (59.23, 59.56, 40): 6}	6	10
	non-musician	2	{(59.23, 59.56, 40): 4, (59.22, 59.51, 28): 4}	4	10
4	musician	5	{(103.85, 106.05, 85): 4, (103.86, 105.87, 93): 4, (103.86, 105.79, 76): 4, (103.87, 105.89, 81): 4, (103.87, 105.88, 69): 4}	4	8
	non-musician	4.75	{(103.86, 105.87, 93): 4, (103.86, 105.79, 76): 4, (103.87, 105.89, 81): 4, (103.87, 105.88, 69): 4, (103.85, 106.05, 85): 3}	4	8
5	musician	7	{(69.09, 72.97, 69): 3, (69.1, 72.95, 52): 2, (69.11, 72.92, 57): 2, (69.12, 72.88, 64): 2, (69.13, 72.7, 60): 2, (69.13, 73.2, 45): 2, (74.26, 77.32, 76): 1}	2	7
	non-musician	6	{(69.09, 72.97, 69): 5, (69.1, 72.95, 52): 5, (69.11, 72.92, 57): 5, (69.12, 72.88, 64): 5, (69.13, 72.7, 60): 5, (69.13, 73.2, 45): 5}	5	7

Table D.10: Common note groups (1–5). Grieg, more detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
6	musician	1	{(74.26, 77.32, 76): 3}	3	6
	non-musician	1	{(74.26, 77.32, 76): 3}	3	6
7	musician	20	{(78.22, 78.31, 86): 3, (78.25, 78.31, 88): 3, (79.68, 79.81, 85): 3, (79.74, 79.78, 88): 3, (81.12, 81.22, 86): 3, (81.14, 81.2, 88): 3, (82.62, 82.72, 85): 3, (82.65, 82.69, 88): 3, (84.32, 84.47, 86): 3, (84.37, 84.46, 88): 3, ...}	3	5
	non-musician	17.5	{(79.68, 79.81, 85): 2, (79.74, 79.78, 88): 2, (81.12, 81.22, 86): 2, (81.14, 81.2, 88): 2, (82.62, 82.72, 85): 2, (82.65, 82.69, 88): 2, (84.32, 84.47, 86): 2, (84.37, 84.46, 88): 2, (86.02, 86.15, 85): 2, (86.06, 86.11, 88): 2, ...}	2	5
8	musician	107	{(77.03, 77.37, 57): 1, (77.09, 77.39, 64): 1, (77.22, 77.31, 68): 1, (77.28, 77.75, 74): 1, (77.73, 77.92, 78): 1, (77.81, 78.13, 76): 1, (78.17, 78.26, 80): 1, (78.19, 78.24, 74): 1, (78.22, 78.31, 86): 1, (78.25, 78.31, 88): 1, ...}	1	5
	non-musician	106.75	{(77.03, 77.37, 57): 2, (77.09, 77.39, 64): 2, (77.22, 77.31, 68): 3, (77.28, 77.75, 74): 3, (77.73, 77.92, 78): 4, (77.81, 78.13, 76): 4, (78.17, 78.26, 80): 4, (78.19, 78.24, 74): 4, (78.22, 78.31, 86): 4, (78.25, 78.31, 88): 4, ...}	4	5
9	musician	1	{(90.69, 90.98, 80): 4}	4	5
	non-musician	1	{(90.69, 90.98, 80): 1}	1	5
10	musician	1	{(58.49, 61.41, 72): 1}	1	4
	non-musician	1	{(58.49, 61.41, 72): 3}	3	4

Table D.11: Common note groups (6–10). Grieg, more detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
1	musician	8	{(29.08, 29.15, 41): 12, (29.08, 29.12, 43): 12, (29.08, 29.18, 33): 12, (29.08, 29.15, 95): 12, (29.09, 29.15, 82): 12, (29.09, 29.13, 93): 12, (29.13, 29.13, 83): 12, (29.14, 29.17, 40): 12}	12	15
	non-musician	8	{(29.08, 29.15, 95): 3, (29.09, 29.13, 93): 3, (29.09, 29.15, 82): 3, (29.13, 29.13, 83): 3, (29.08, 29.15, 41): 3, (29.08, 29.12, 43): 3, (29.08, 29.18, 33): 3, (29.14, 29.17, 40): 3}	3	
2	musician	10.25	{(16.63, 16.68, 64): 4, (16.63, 16.69, 66): 4, (16.64, 16.68, 67): 4, (16.64, 16.68, 68): 4, (16.83, 16.9, 70): 4, (16.84, 16.89, 67): 4, (16.84, 16.88, 72): 4, (17.05, 17.28, 64): 4, (17.05, 17.29, 61): 4, (17.1, 17.13, 62): 4, ...}	4	6
	non-musician	11.5	{(16.63, 16.68, 64): 2, (16.63, 16.69, 66): 2, (16.64, 16.68, 67): 2, (16.64, 16.68, 68): 2, (16.83, 16.9, 70): 2, (16.84, 16.89, 67): 2, (16.84, 16.88, 72): 2, (17.05, 17.28, 64): 2, (17.05, 17.08, 53): 2, (17.05, 17.29, 61): 2, ...}	2	
3	musician	4	{(8.16, 8.2, 65): 6, (8.17, 8.19, 67): 6, (8.17, 8.2, 79): 6, (8.18, 8.21, 81): 6}	6	6
4	musician	7.33	{(8.48, 8.53, 60): 3, (8.65, 8.81, 61): 3, (8.79, 8.91, 58): 3, (8.93, 9.08, 65): 3, (9.06, 9.27, 60): 3, (9.23, 9.44, 63): 3, (9.38, 9.5, 64): 3, (9.52, 9.58, 65): 1}	3	5
	non-musician	6.5	{(8.65, 8.81, 61): 2, (8.79, 8.91, 58): 2, (8.93, 9.08, 65): 2, (9.06, 9.27, 60): 2, (9.23, 9.44, 63): 2, (9.38, 9.5, 64): 2, (8.48, 8.53, 60): 1}	2	
5	musician	17.25	{(10.79, 10.84, 61): 4, (10.79, 10.84, 62): 4, (10.79, 10.87, 69): 4, (10.79, 10.85, 64): 4, (10.85, 10.91, 66): 4, (10.97, 11.02, 61): 4, (10.97, 11.01, 53): 4, (10.97, 11.02, 60): 4, (10.98, 11.03, 58): 4, (11.12, 11.16, 61): 4, ...}	4	5
	non-musician	17	{(10.79, 10.87, 69): 1, (10.85, 10.91, 66): 1, (10.79, 10.85, 64): 1, (10.79, 10.84, 62): 1, (10.79, 10.84, 61): 1, (10.97, 11.02, 61): 1, (10.97, 11.02, 60): 1, (11.11, 11.2, 75): 1, (11.14, 11.17, 73): 1, (11.13, 11.17, 72): 1, ...}	1	

Table D.12: Common note groups (1–5). Boulez, less detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Note tuples (start, end, note number) : frequency	Subset count	Total count
6	musician	24	{(22.11, 22.17, 64): 4, (22.12, 22.19, 54): 4, (22.12, 22.19, 55): 4, (22.25, 22.31, 75): 4, (22.3, 22.35, 64): 4, (22.3, 22.39, 73): 4, (22.31, 22.37, 74): 4, (22.32, 22.35, 65): 4, (22.46, 22.51, 64): 4, (22.46, 22.54, 55): 4, ...}	4	5
	non-musician	23	{(21.97, 21.99, 87): 1, (21.98, 22.0, 90): 1, (22.11, 22.17, 64): 1, (22.12, 22.19, 54): 1, (22.12, 22.19, 55): 1, (22.25, 22.31, 75): 1, (22.3, 22.35, 64): 1, (22.3, 22.39, 73): 1, (22.31, 22.37, 74): 1, (22.32, 22.35, 65): 1, ...}	1	
7	musician	12.5	{(28.72, 28.89, 61): 1, (28.86, 28.9, 63): 2, (28.87, 28.9, 54): 2, (28.87, 28.9, 62): 2, (28.87, 28.91, 53): 2, (29.08, 29.15, 41): 2, (29.08, 29.12, 43): 2, (29.08, 29.18, 33): 2, (29.08, 29.15, 95): 2, (29.09, 29.15, 82): 2, ...}	2	5
	non-musician	13	{(28.72, 28.89, 61): 3, (28.86, 28.9, 63): 3, (28.87, 28.9, 54): 3, (28.87, 28.9, 62): 3, (28.87, 28.91, 53): 3, (29.08, 29.15, 41): 3, (29.08, 29.12, 43): 3, (29.08, 29.18, 33): 3, (29.08, 29.15, 95): 3, (29.09, 29.15, 82): 3, ...}	3	
8	musician	10	{(14.9, 14.96, 67): 4, (14.9, 14.96, 68): 4, (14.9, 14.95, 64): 4, (14.91, 14.96, 75): 3, (15.06, 15.11, 64): 4, (15.06, 15.1, 65): 4, (15.23, 15.28, 67): 4, (15.23, 15.28, 64): 4, (15.24, 15.3, 68): 4, (15.24, 15.3, 66): 4, ...}	4	4
9	musician	9	{(23.44, 23.48, 61): 1, (23.57, 23.66, 64): 2, (23.6, 23.66, 65): 2, (23.74, 23.91, 63): 2, (23.92, 24.07, 60): 2, (24.07, 24.14, 60): 2, (24.09, 24.26, 62): 2, (24.24, 24.47, 61): 2, (24.46, 24.5, 63): 2, (23.35, 23.43, 53): 1}	2	4
	non-musician	9	{(23.57, 23.66, 64): 2, (23.6, 23.66, 65): 2, (23.74, 23.91, 63): 2, (23.92, 24.07, 60): 2, (24.07, 24.14, 60): 2, (24.09, 24.26, 62): 2, (24.24, 24.47, 61): 2, (24.46, 24.5, 63): 2, (24.59, 24.79, 63): 1, (24.59, 24.78, 64): 1}	2	
10	musician	6	{(20.23, 20.29, 68): 4, (20.23, 20.3, 64): 4, (20.23, 20.3, 67): 4, (20.51, 20.75, 65): 4, (20.52, 20.76, 67): 4, (20.52, 20.79, 76): 3, (20.77, 20.88, 51): 1}	4	4

Table D.13: Common note groups (6–10). Boulez, less detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
1	musician	8	{(29.08, 29.15, 41): 4, (29.08, 29.12, 43): 4, (29.08, 29.18, 33): 4, (29.08, 29.15, 95): 4, (29.09, 29.15, 82): 4, (29.09, 29.13, 93): 4, (29.13, 29.13, 83): 4, (29.14, 29.17, 40): 4}	4	6
	non-musician	8.5	{(29.08, 29.15, 95): 2, (29.09, 29.13, 93): 2, (29.13, 29.13, 83): 3, (29.09, 29.15, 82): 2, (29.08, 29.12, 43): 2, (29.08, 29.15, 41): 2, (29.14, 29.17, 40): 2, (29.08, 29.18, 33): 2}	2	
2	musician	3	{(15.8, 15.98, 60): 5, (15.8, 16.02, 61): 5, (15.8, 16.0, 64): 5}	5	6
	non-musician	3	{(15.8, 15.98, 60): 1, (15.8, 16.02, 61): 1, (15.8, 16.0, 64): 1}	1	
3	musician	3	{(17.05, 17.28, 64): 3, (17.05, 17.29, 61): 3, (17.1, 17.13, 62): 3}	3	4
	non-musician	3	{(17.05, 17.28, 64): 1, (17.05, 17.29, 61): 1, (17.1, 17.13, 62): 1}	1	
4	musician	2	{(8.17, 8.2, 79): 3, (8.18, 8.21, 81): 3}	3	4
	non-musician	2	{(8.17, 8.2, 79): 1, (8.18, 8.21, 81): 1}	1	
5	musician	4.33	{(7.29, 7.35, 54): 3, (7.29, 7.36, 51): 3, (7.3, 7.34, 56): 3, (7.3, 7.49, 65): 3, (10.79, 10.87, 69): 1}	3	4
	non-musician	4	{(7.29, 7.35, 54): 1, (7.29, 7.36, 51): 1, (7.3, 7.34, 56): 1, (7.3, 7.49, 65): 1}	1	

Table D.14: Common note groups (1–5). Boulez, more detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

Index	Musical ability	Mean notes	Subset detail {(Note tuples) : frequency in subset}	Subset count	Total count
6	musician	10	{(14.91, 14.96, 75): 1, (14.9, 14.96, 68): 1, (14.9, 14.96, 67): 1, (14.9, 14.95, 64): 1, (15.06, 15.11, 64): 1, (15.24, 15.3, 66): 1, (15.24, 15.3, 68): 1, (15.23, 15.28, 67): 1, (15.06, 15.1, 65): 1, (15.23, 15.28, 64): 1}	1	4
	non-musician	10	{(14.9, 14.96, 68): 3, (14.9, 14.96, 67): 3, (14.9, 14.95, 64): 3, (15.06, 15.11, 64): 3, (15.06, 15.1, 65): 3, (15.24, 15.3, 66): 3, (15.23, 15.28, 67): 3, (15.24, 15.3, 68): 3, (15.23, 15.28, 64): 3, (14.91, 14.96, 75): 3}	3	
7	musician	3	{(13.08, 13.13, 52): 3, (13.08, 13.14, 47): 3, (13.08, 13.14, 50): 3}	3	3
8	musician	3	{(24.59, 24.8, 70): 1, (24.59, 24.78, 64): 1, (24.59, 24.79, 63): 1}	1	3
	non-musician	3	{(24.59, 24.79, 63): 2, (24.59, 24.8, 70): 2, (24.59, 24.78, 64): 2}	2	
9	non-musician	54	{(1.57, 1.61, 61): 3, (1.57, 1.63, 60): 3, (1.58, 1.63, 53): 3, (1.73, 1.78, 69): 3, (1.73, 1.78, 62): 3, (1.73, 1.77, 61): 3, (1.89, 1.96, 51): 3, (1.89, 1.95, 61): 3, (1.9, 1.96, 53): 3, (2.0, 2.11, 71): 3, ...}	3	3
10	musician	12	{(28.86, 28.9, 63): 3, (28.87, 28.9, 54): 3, (28.87, 28.9, 62): 3, (28.87, 28.91, 53): 3, (29.08, 29.15, 41): 3, (29.08, 29.12, 43): 3, (29.08, 29.18, 33): 3, (29.08, 29.15, 95): 3, (29.09, 29.15, 82): 3, (29.09, 29.13, 93): 3, ...}	3	3

Table D.15: Common note groups (6–10). Boulez, more detailed condition. Each numbered row represents a common note group set, divided into subsets for musicians and non-musicians. The mean number of notes per subset is computed. Subsets contain note tuples where each tuple contains three elements (start time, end time, note number). The number next to each tuple counts how many times that note was selected as part of a note group in the subset. The number of unique group IDs (unique note group annotations saved into CosmoNote) per subset is counted and added into a total count.

References

Abromont, C. (2001). *"Guide de la théorie de la musique", Claude Abromont, Eugène de Montalembert* (1st ed.). Clamecy, France. Retrieved from https://www.fayard.fr/musique/guide-de-la-theorie-de-la-musique-9782213609775

Abromont, C., & de Montalembert, E. (2010). *Guide des formes de la musique occidentale.*

Amatriain, X., Arumí, P., & Ramírez, M. (2002, November). CLAM, yet another library for audio and music processing? In *Companion of the 17th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications* (pp. 46–47). New York, NY, USA. doi: 10.1145/985072.985097

Amatriain, X., Massaguer, J., Garcia, D., & Mosquera, I. (2005). The CLAM Annotator: A Cross-platform Audio Descriptors Editing Tool. In *Proceedings of the 6th International Conference on Music Information Retrieval* (pp. 426–429). London. Retrieved from https://ismir2005.ismir.net/proceedings/3017.pdf

Ammer, C. (2004). *The Facts on File Dictionary of Music* (4th ed.). New York, NY, USA.

Apel, W. (1969). *Harvard Dictionary of Music: Second Edition, Revised and Enlarged.* Cambridge, MA.

Ashby, P. (2011). Beyond the segment. In *Understanding Phonetics.*

Balázs, B., Mooney, P., Nováková, E., Bastin, L., & Jokar Arsanjani, J. (2021). Data Quality in Citizen Science. In K. Vohland et al. (Eds.), *The Science of Citizen Science* (pp. 139–157). Cham. doi: 10.1007/978-3-030-58278-4_8

Bauer, W. R. (2014, July). Expressiveness in Jazz Performance: Prosody and Rhythm. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 133–153). doi: 10.1093/acprof:oso/9780199659647.003.0008

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, *3*, 30. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.663&rep=rep1&type=pdf

Bedoya, D., Fyfe, L., & Chew, E. (2022a, June). Creating Experiments with Cosmonote: Advancing Web-Based Annotations for Performed Music. In *Proceedings of the 19th Sound and Music Computing Conference* (pp. 683–684). Saint-Étienne, France. doi: 10.5281/zenodo.6576284

Bedoya, D., Fyfe, L., & Chew, E. (2022b). A Perceiver-Centered Approach for Representing and Annotating Prosodic Functions in Performed Music. *Frontiers in Psychology*, *13*. doi: 10.3389/fpsyg.2022.886570

Bigand, E. (2003). More About the Musical Expertise of Musically Untrained Listeners. *Annals of the New York Academy of Sciences*, *999*(1), 304–312. doi: 10.1196/

annals.1284.041

Bigo, L., Ghisi, D., Spicher, A., & Andreatta, M. (2015, September). Representation of Musical Structures and Processes in Simplicial Chord Spaces. *Computer Music Journal*, *39*(3), 9–24. doi: 10.1162/COMJ_a_00312

Bjørndalen, O. M. (2023, February). *Mido - MIDI Objects for Python.* mido. Retrieved from https://github.com/mido/mido

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008, October). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Boersma, P., & van Heuven, V. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–347. Retrieved from https://hdl.handle.net/11245/1.200596

Bostock, M., Ogievetsky, V., & Heer, J. (2011, December). D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. doi: 10.1109/TVCG.2011.185

Bray, T. (2017, December). *The JavaScript Object Notation (JSON) Data Interchange Format* (Request for Comments No. RFC 8259). Internet Engineering Task Force. doi: 10.17487/RFC8259

Bresin, R., & Friberg, A. (2012, May). Evaluation of Computer Systems for Expressive Music Performance. In A. Kirke & E. R. Miranda (Eds.), *Guide to Computing for Expressive Music Performance* (pp. 181–203). London. doi: 10.1007/978-1-4471-4123-5_7

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2009). The perception of structural boundaries in melody lines of Western popular music. *Musicae Scientiae*, *13*(2), 273–313. doi: 10.1177%2F102986490901300204

Brugman, H., & Russel, A. (2004). Annotating Multi-media / Multi-modal resources with ELAN. In *LREC 2004* (p. 4). Lisbon, Portugal. Retrieved from http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf

Cambouropoulos, E. (2006, February). Musical Parallelism and Melodic Segmentation: A Computational Approach. *Music Perception*, *23*(3), 249–268. doi: 10.1525/mp.2006.23.3.249

Cancino-Chacón, C. E., Grachten, M., Goebl, W., & Widmer, G. (2018). Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities*, *5*. Retrieved from https://www.frontiersin.org/article/10.3389/fdigh.2018.00025

Cannam, C., Landone, C., & Sandler, M. (2010, October). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1467–1468). New York, NY, USA. doi: 10.1145/1873951.1874248

Cartwright, M., Pardo, B., Mysore, G. J., & Hoffman, M. (2016, March). Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 619–623). Shanghai, China. doi: 10.1109/ICASSP.2016.7471749

Cartwright, M., Seals, A., Salamon, J., Williams, A., Mikloska, S., MacConnell, D., ... Nov, O. (2017, December). Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. In *Proceedings of the ACM*

on Human-Computer Interaction (Vol. 1, pp. 29:1–29:21). New York, NY, USA. doi: 10.1145/3134664

Catalán, B. M. (2022, May). *BAT - BMAT Annotation Tool.* Retrieved from https://github.com/BlaiMelendezCatalan/BAT

Chew, E. (2016, February). Playing with the Edge: Tipping Points and the Role of Tonality. *Music Perception: An Interdisciplinary Journal*, *33*(3), 344–366. doi: 10.1525/mp.2016.33.3.344

Chew, E. (2017). From Sound to Structure: Synchronizing Prosodic and Structural Information to Reveal the Thinking Behind Performance Decisions. In C. Mackie (Ed.), *New Thoughts on Piano Performance: Research at the Interface between Science and the Art of Piano Performance* (pp. 123–150). London.

Chew, E. (2018, July). Notating Disfluencies and Temporal Deviations in Music and Arrhythmia. *Music & Science*, *1*. doi: 10.1177/2059204318795159

CNRS. (2020, November). *Le piano virtuose | Reportage CNRS.* Retrieved from https://youtu.be/yXkwusNyte4

Cook, N. (1994). *A Guide to Musical Analysis.*

Cook, N. (2013). *Beyond the Score: Music as Performance.*

Couprie, P. (2012, May). EAnalysis : Aide à l'analyse de la musique électroacoustique. In *Journées d'Informatique Musicale* (pp. 183–189). Mons, Belgium. Retrieved from https://hal.archives-ouvertes.fr/hal-00823848

Couprie, P. (2018). Approches audionumériques pour l'analyse musicale. *Musicologies nouvelles*, *5*, 120–132. Retrieved from https://hal.archives-ouvertes.fr/hal-02084867

Couprie, P. (2022, June). Designing Sound Representations For Musicology. In *Proceedings of the 19th Sound and Music Computing Conference* (pp. 570–576). Saint-Étienne (France). doi: 10.5281/zenodo.6798315

Couturier, L., Bigo, L., & Levé, F. (2022, June). Annotating Symbolic Texture in Piano Music: A Formal Syntax. In *Proceedings of the 19th Sound and Music Computing Conference* (pp. 570–577). Saint-Étienne (France). doi: 10.5281/zenodo.6573655

Cruz, L., Rolla, V., Kestenberg, J., & Velho, L. (2018). Visual Representations for Music Understanding Improvement. In M. Aramaki, M. E. P. Davies, R. Kronland-Martinet, & S. Ystad (Eds.), *Music Technology with Swing* (pp. 468–476). Cham. doi: 10.1007/978-3-030-01692-0_31

De Roure, D., Downie, J. S., & Fujinaga, I. (2010, May). SALAMI: Structural Analysis of Large Amounts of Music Information. In *UK e-Science All Hands Meeting 2010 (12/09/10 - 15/09/10).* Retrieved from https://eprints.soton.ac.uk/271171/

Doğantan-Dack, M. (2014, July). Philosophical Reflections on Expressive Music Performance. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 3–21). doi: 10.1093/acprof:oso/9780199659647.003.0001

Drake, C., & Palmer, C. (1993, April). Accent Structures in Music Performance. *Music Perception*, *10*(3), 343–378. doi: 10.2307/40285574

Evans, K. K., & Treisman, A. (2010, January). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6. doi: 10.1167/10.1.6

Fabian, D., Timmers, R., & Schubert, E. (2014). *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures.*

Ferraguto, M. (2019, October). Music for a French Piano: WoO 80. In M. Ferraguto (Ed.), *Beethoven 1806* (p. 0). doi: 10.1093/oso/9780190947187.003.0006

Fillon, T., Simonnot, J., Mifune, M.-F., Khoury, S., Pellerin, G., & Le Coz, M. (2014, September). Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology* (pp. 1–8). New York, NY, USA. doi: 10.1145/2660168.2660169

Fletcher, S., & Islam, M. Z. (2018, March). Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*, *22*. doi: 10.3127/ajis.v22i0.1538

Foster, B. (2007). *The Songs of Edvard Grieg*.

Friberg, A. (2006). pDM: An Expressive Sequencer with Real-Time Control of the KTH Music-Performance Rules. *Computer Music Journal*, *30*(1), 37–48. Retrieved from https://www.jstor.org/stable/3682025

Friberg, A., & Battel, G. U. (2011). Structural Communication. In *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning* (pp. 199–218). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-176261

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, *2*(2-3), 145–161. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-51912

Fyfe, L., Bedoya, D., & Chew, E. (2022, November). Annotation and Analysis of Recorded Piano Performances on the Web. *Journal of the Audio Engineering Society*, *70*(11), 962–978. doi: 10.17743/jaes.2022.0057

Fyfe, L., Bedoya, D., Guichaoua, C., & Chew, E. (2021, July). CosmoNote: A Web-based Citizen Science Tool for Annotating Music Performances. In *Proceedings of the International Web Audio Conference* (pp. 1–6). Barcelona, Spain. Retrieved from http://webaudioconf.com/posts/2021_25/

García Stan, C. A. (2015). *Análisis técnico interpretativo de las 32 variaciones para piano en DO menor (WoO. 80) de Ludwig Van Beethoven* (Doctoral dissertation). Retrieved from http://repository.udistrital.edu.co/handle/11349/23093

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*.

Gingras, G., Rowland, B. A., & Stein, B. E. (2009, April). The Differing Impact of Multisensory and Unisensory Integration on Behavior. *Journal of Neuroscience*, *29*(15), 4897–4902. doi: 10.1523/JNEUROSCI.4120-08.2009

Giraud, M., Groult, R., & Leguy, E. (2018, May). Dezrann, a Web Framework to Share Music Analysis. In *International Conference on Technologies for Music Notation and Representation (TENOR 2018)* (p. 104). Retrieved from https://hal.science/hal-01796787

Goldman, J. (2011). *The Musical Language of Pierre Boulez: Writings and Compositions*.

Gómez-Cañón, J. S., Gutiérrez-Páez, N., Porcaro, L., Porter, A., Cano, E., Herrera-Boyer, P., ... Gómez, E. (2023, April). TROMPA-MER: An open dataset for personalized music emotion recognition. *Journal of Intelligent Information Systems*, *60*(2), 549–570. doi: 10.1007/s10844-022-00746-0

Goto, M. (2006). AIST annotation for RWC music database. In *In ISMIR*.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003, October). RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In *Proceedings of the 3rd International Conference on Music Information Retrieval.* Paris, France. doi: 10.5281/zenodo.1416474

Goto, M., Yoshii, K., Fujihara, H., Mauch, M., & Nakano, T. (2011). SONGLE: A WEB SERVICE FOR ACTIVE MUSIC LISTENING IMPROVED BY USER CONTRIBUTIONS. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 311–316). Miami, FL, USA. Retrieved from https://archives.ismir.net/ismir2011/paper/000014.pdf

Guichaoua, C., Lascabettes, P., & Chew, E. (2023). *End-to-end Bayesian segmentation and similarity assessment of performed music tempo and dynamics without score information* [Manuscript Submitted for Publication].

Guo, R., Herremans, D., & Magnusson, T. (2019, October). Midi Miner – A Python library for tonal tension and track classification. *arXiv:1910.02049 [cs, eess]*. Retrieved from http://arxiv.org/abs/1910.02049

Gutiérrez Páez, N. F., Gómez-Cañón, J. S., Porcaro, L., Santos, P., Hernández-Leo, D., & Gómez, E. (2021). Emotion Annotation of Music: A Citizen Science Approach. In D. Hernández-Leo, R. Hishiyama, G. Zurita, B. Weyers, A. Nolte, & H. Ogata (Eds.), *Collaboration Technologies and Social Computing* (pp. 51–66). Cham. doi: 10.1007/978-3-030-85071-5_4

Haklay, M. (2013). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (pp. 105–122). Dordrecht. doi: 10.1007/978-94-007-4587-2_7

Haklay, M. M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., & Vohland, K. (2021). What Is Citizen Science? The Challenges of Definition. In K. Vohland et al. (Eds.), *The Science of Citizen Science* (pp. 13–33). Cham. doi: 10.1007/978-3-030-58278-4_2

Hartmann, M. A. (2017). *Modelling and prediction of perceptual segmentation* (Doctoral dissertation, University of Jyväskylä, Finland). Retrieved from https://jyx.jyu.fi/handle/123456789/52654

Herremans, D. (2016). *MorpheuS Tension Visualiser.* Retrieved from https://dorienherremans.com/tension

Herremans, D., & Chew, E. (2016). Tension ribbons: Quantifying and visualising tonal tension. In *Second International Conference on Technologies for Music Notation and Representation (TENOR)* (Vol. 2, p. 10). Cambridge, UK.

Herrera, P., Celma, Ò., Massaguer, J., Cano, P., Gómez, E., Gouyon, F., ... Wack, N. (2005). MUCOSA: A Music Content Semantic Annotator. In *Proceedings of the 6th International Conference on Music Information Retrieval* (pp. 77–83). Barcelona, Spain.

Huron, D. (2001, September). Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*, *19*(1), 1–64. doi: 10.1525/mp.2001.19.1.1

Jarrett, S. (2003a). The First Ibsen Songs: Opus 23, from Peer Gynt. In *Edvard Grieg and His Songs.*

Jarrett, S. (2003b). Writer of Songs. In *Edvard Grieg and His Songs.*

Jost, C. (1994). 32 Variationen c-Mollfür Klavier WoO 80. In A. Riethmüller, C. Dahlhaus, & A. L. Ringer (Eds.), *Beethoven. Interpretationen seiner Werke* (Vol. 2, pp. 481–185). Retrieved from https://laaber-verlag.de/detailview?no=01107

Juslin, P. N., Friberg, A., & Bresin, R. (2001, September). Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, *5*(1_suppl), 63–122. doi: 10.1177/10298649020050S104

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, *129*(5), 770–814. doi: 10.1037/0033-2909.129.5.770

Kennedy, M., & Kennedy, J. (2012). *The Oxford Dictionary of Music*. Oxford, England.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., ... Players, F. (2011, November). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, *108*(47), 18949–18953. doi: 10.1073/pnas.1115898108

Kirke, A., & Miranda, E. R. (2012, May). An Overview of Computer Systems for Expressive Music Performance. In A. Kirke & E. R. Miranda (Eds.), *Guide to Computing for Expressive Music Performance* (pp. 1–47). London. doi: 10.1007/978-1-4471-4123-5_1

Kolouri, S., Park, S., Thorpe, M., Slepčev, D., & Rohde, G. K. (2017, July). Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, *34*(4), 43–59. doi: 10.1109/MSP.2017.2695801

Kong, Q., Li, B., Chen, J., & Wang, Y. (2022, May). GiantMIDI-Piano: A Large-Scale MIDI Dataset for Classical Piano Music. *Transactions of the International Society for Music Information Retrieval*, *5*(1), 87–98. doi: 10.5334/tismir.80

Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A., & Volk, A. (2019, May). Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, *48*(3), 232–252. doi: 10.1080/09298215.2019.1613436

Lalitte, P. (2011). Du son au sens : vers une approche sub-symbolique de l'analyse musicale assistée par ordinateur. *Musurgia*, *XVIII*(1-2), 99–116. doi: 10.3917/musur.111.0099

Lartillot, O. (2021). Computational Musicological Analysis of Notated Music: A Brief Overview. *Nota Bene*, *15*, 142–161. Retrieved from https://www.duo.uio.no/handle/10852/85647

Lee, D. (2023, April). Organising music's structures: The classification of musical forms in Western art music. *Journal of Information Science*, 01655515231167384. doi: 10.1177/01655515231167384

Leech-Wilkinson, D. (2009). *The Changing Sound of Music: Approaches to Studying Recorded Musical Performance*. Retrieved from http://www.charm.kcl.ac.uk/studies/chapters/intro.html

Leech-Wilkinson, D., & Prior, H. (2014, July). Heuristics for expressive performance. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 34–57). doi: 10.1093/acprof:oso/9780199659647.003.0003

Lerch, A., Arthur, C., Pati, A., & Gururani, S. (2019). Music Performance Analysis: A Survey. In *Proceedings of the 20th International Society for Music Information*

Retrieval Conference (pp. 33–43). Delft, The Netherlands. doi: 10.5281/zenodo.3527735

Lerdahl, F. (2004). *Tonal Pitch Space*.

Lerdahl, F., & Jackendoff, R. (1996). *A Generative Theory of Tonal Music*. Cambridge, MA, USA. Retrieved from https://mitpress.mit.edu/books/generative-theory-tonal-music-reissue-new-preface

Li, B., Burgoyne, J. A., & Fujinaga, I. (2006). Extending Audacity for Audio Annotation. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)* (pp. 1–2). Victoria, Canada. Retrieved from https://archives.ismir.net/ismir2006/paper/000116.pdf

Li, X., Ji, H., Farooq, F., Li, H., Lin, W.-P., & Yu, S. (2012, December). Rich Annotation Guided Learning. *International Journal On Advances in Intelligent Systems*, *5*(3 and 4), 261–277. Retrieved from http://personales.upv.es/thinkmind/IntSys/IntSys_v5_n34_2012/intsys_v5_n34_2012_4.html

Lima, H. B., Santos, C. G. R. D., & Meiguins, B. S. (2021, July). A Survey of Music Visualization Techniques. *ACM Computing Surveys*, *54*(7), 143:1–143:29. doi: 10.1145/3461835

Lorenzo de Reizábal, A., & Lorenzo de Reizábal, M. (2009). *Análisis musical: claves para entender e interpretar la música* (2nd ed.). Barcelona, Spain. Retrieved from https://boileau-music.com/es/obras/analisis-musical-b.3261

Malandrino, D., Pirozzi, D., & Zaccagnino, R. (2019, December). Learning the harmonic analysis: Is visualization an effective approach? *Multimedia Tools and Applications*, *78*(23), 32967–32998. doi: 10.1007/s11042-019-07879-5

Manoury, P. (2013, September). *Mario Caroli / Wilhem Latchoumia: Récital flûte & piano* [Program notes]. Strasbourg, Fr. Retrieved from https://festivalmusica.fr/telecharger/1542

Marsden, A. (2016). Music Analysis by Computer: Ontology and Epistemology. In D. Meredith (Ed.), *Computational Music Analysis* (pp. 3–28). Cham. doi: 10.1007/978-3-319-25931-4_1

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating Hierarchical Structure in Music Annotations. *Frontiers in Psychology*, *8*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01337

Meléndez Catalán, B., Molina, E., & Gómez Gutiérrez, E. (2017). BAT: An open-source, web-based audio events annotation tool. In *3rd Web Audio Conference*. London, UK. Retrieved from http://repositori.upf.edu/handle/10230/43406

Migliore, O., & Obin, N. (2018, June). At the Interface of Speech and Music: A Study of Prosody and Musical Prosody in Rap Music. In *Proceedings of the 9th International Conference on Speech Prosody* (pp. 557–561). Poznan, Poland. doi: 10.21437/SpeechProsody.2018-113

Mikloska, S. (2017). *Audio-annotator*. CrowdCurio. Retrieved from https://github.com/CrowdCurio/audio-annotator

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285–290. doi: 10.1890/110278

Miyazaki, R., Fujishiro, I., & Hiraga, R. (2003, July). Exploring MIDI datasets. In *ACM SIGGRAPH 2003 Sketches & Applications* (p. 1). New York, NY, USA. doi:

10.1145/965400.965453

Møller, C., Garza-Villarreal, E. A., Hansen, N. C., Højlund, A., Bærentsen, K. B., Chakravarty, M. M., & Vuust, P. (2021, February). Audiovisual structural connectivity in musicians and non-musicians: A cortical thickness and diffusion tensor imaging study. *Scientific Reports*, *11*(1), 4324. doi: 10.1038/s41598-021-83135-x

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2).

Nakamura, E., Yoshii, K., & Katayose, H. (2017). Performance Error Detection and Post-Processing for Fast and Accurate Symbolic Music Alignment. In *Proceedings of the 18th ISMIR Conference* (p. 7). Suzhou, China.

Nieto, O., Farbood, M. M., Jehan, T., & Bello, J. P. (2014). Perceptual analysis of the f-measure for evaluating section boundaries in music: 15th International Society for Music Information Retrieval Conference, ISMIR 2014. In (pp. 265–270). Retrieved from http://www.scopus.com/inward/record.url?scp=85066072277&partnerID=8YFLogxK

Nieto, O., Mysore, G. J., Wang, C.-i., Smith, J. B. L., Schlüter, J., Grill, T., & McFee, B. (2020, December). Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications. *Transactions of the International Society for Music Information Retrieval*, *3*(1), 246–263. doi: 10.5334/tismir.54

Notess, M., & Swan, M. (2004). Timeliner: Building a Learning Tool into a Digital Music Library. In *EdMedia + Innovate Learning* (pp. 603–609). Retrieved from https://www.learntechlib.org/primary/p/12995/

O'Hagan, P. (2016). *Pierre Boulez and the Piano: A Study in Style and Technique*. Retrieved from https://doi.org/10.4324/9781315517858

Ong, B. S. (2006). *Structural Analysis and Segmentation of Music Signals* (Doctoral dissertation, Universitat Pompeu Fabra, Barcelona, Spain). doi: http://hdl.handle.net/10803/7544

Palmer, C. (1997). Music Performance. *Annual Review of Psychology*, *48*, 115–138. doi: 10.1146/annurev.psych.48.1.115

Palmer, C., & Hutchins, S. (2006, January). What Is Musical Prosody? In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 46, pp. 245–278). Urbana, IL, USA. doi: 10.1016/S0079-7421(06)46007-2

Pampalk, E. (2004). A Matlab Toolbox to compute music similarity from audio. In *In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04), Universitat Pompeu Fabra* (pp. 254–257). Barcelona, Spain. doi: 10.5281/zenodo.1418077

Platz, F., & Kopiez, R. (2012, September). When the Eye Listens: A Meta-analysis of How Audio-visual Presentation Enhances the Appreciation of Music Performance. *Music Perception*, *30*(1), 71–83. doi: 10.1525/mp.2012.30.1.71

Repp, B. H. (1996, December). Patterns of note onset asynchronies in expressive piano performance. *The Journal of the Acoustical Society of America*, *100*(6), 3917–3932. doi: 10.1121/1.417245

Rohrmeier, M., & Pearce, M. (2018). Musical Syntax I: Theoretical Perspectives. In R. Bader (Ed.), *Springer Handbook of Systematic Musicology* (pp. 473–486). Heidelberg, Germany. doi: 10.1007/978-3-662-55004-5_25

Santana, I. A. P., Pinhelli, F., Donini, J., Catharin, L., Mangolin, R. B., da Costa, Y. M. e. G., ... Domingues, M. A. (2020, July). Music4All: A New Music Database and Its Applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 399–404). doi: 10.1109/IWSSIP48289.2020.9145170

Sauda, A., Giraud, M., & Leguy, E. (2022, June). Soutenir en Classe L'écoute Active, L'Autonomie Et L'échange en Analyse Musicale Avec la Plateforme Web Dezrann. In *Proceedings of the 19th Sound and Music Computing Conference* (pp. 617–623). Saint-Étienne (France). doi: 10.5281/zenodo.6800855

Schubert, E., Canazza, S., De Poli, G., & Rodà, A. (2017, April). Algorithms can Mimic Human Piano Performance: The Deep Blues of Music. *Journal of New Music Research*, *46*(2), 175–186. doi: 10.1080/09298215.2016.1264976

Schubert, E., & Fabian, D. (2014, July). A Taxonomy of Listeners' Judgements of Expressiveness in Music Performance. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 283–303). doi: 10.1093/acprof:oso/9780199659647.003.0016

Schutz, M., & Lipscomb, S. (2007, June). Hearing Gestures, Seeing Music: Vision Influences Perceived Tone Duration. *Perception*, *36*(6), 888–897. doi: 10.1068/p5635

Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B. L., & Cristia, A. (2021, July). Describing Vocalizations in Young Children: A Big Data Approach Through Citizen Science Annotation. *Journal of Speech, Language, and Hearing Research : JSLHR*, *64*(7), 2401–2416. doi: 10.1044/2021_JSLHR-20-00661

Senabre Hidalgo, E., Perelló, J., Becker, F., Bonhoure, I., Legris, M., & Cigarini, A. (2021). Participation and Co-creation in Citizen Science. In K. Vohland et al. (Eds.), *The Science of Citizen Science* (pp. 199–218). Cham. doi: 10.1007/978-3-030-58278-4_11

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., ... Wallin, J. (2014, February). Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America*, *135*(2), 953–962. doi: 10.1121/1.4861348

Silverman, B. W. (2017). *Density Estimation for Statistics and Data Analysis*. New York. doi: 10.1201/9781315140919

Sjölander, K., & Beskow, J. (2000, October). Wavesurfer - an open source speech tool. In *6th International Conference on Spoken Language Processing (ICSLP 2000)* (Vol. 4, pp. 464–467). Beijing, China. doi: 10.21437/ICSLP.2000-849

Sloboda, J. A. (1985). Expressive skill in two pianists: Metrical communication in real and simulated performances. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 273–293. doi: 10.1037/h0080062

Smith, J. B., & Chew, E. (2013). Using Quadratic Programming to Estimate Feature Relevance in Structural Analyses of Music. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 113–122). New York, NY, USA. doi: 10.1145/2502081.2502124

Smith, J. B. L., Burgoyne, J., Fujinaga, I., Roure, D. D., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 555–560). Miami, FL, USA. Retrieved from https://ismir2011.ismir.net/papers/

PS4-14.pdf

Smith, J. B. L., Schankler, I., & Chew, E. (2014, September). Listening as a Creative Act: Meaningful Differences in Structural Annotations of Improvised Performances. *Music Theory Online*, *20*(3). Retrieved from http://www.mtosmt.org/issues/mto.14.20.3/mto.14.20.3.smith_schankler_chew.html

Speer, S., & Blodgett, A. (2006, January). Prosody. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics (Second Edition)* (pp. 505–537). London. doi: 10.1016/B978-012369374-7/50014-6

Spence, C. (2011, May). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. doi: 10.3758/s13414-010-0073-7

Spencer, P., & Temko, P. M. (1994). *A Practical Approach to the Study of Form in Music*. Long Grove, Illinois, USA. Retrieved from https://books.google.fr/books?id=IXUfAAAAQBAJ

Spitzer, M. (2015, December). Metaphor and Musical Thought. In *Metaphor and Musical Thought*. doi: 10.7208/9780226279435

Stein, L. (1979). *Structure & Style: The Study and Analysis of Musical Forms*.

Stowell, D., & Chew, E. (2013). Maximum a Posteriori Estimation of Piecewise Arcs in Tempo Time-Series. In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad (Eds.), *From Sounds to Music and Emotions* (pp. 387–399). Berlin, Heidelberg. doi: 10.1007/978-3-642-41248-6_22

Strasser, B., Baudry, J., Mahr, D., Sanchez, G., & Tancoigne, É. (2019, May). Citizen Science: Rethinking Science and Public Participation. *Science & Technology Studies*. doi: 10.23987/sts.60425

Susini, P., Houix, O., Wenzel, N., & Ponsot, E. (2022, April). Beyond the musician vs. non-musician dichotomy: Evidence for a multi-step reorganization of auditory processing with musical learning. In *16ème Congrès Français d'Acoustique, CFA2022.* Marseille, France. Retrieved from https://hal.science/hal-03848181

Temperley, D. (2004). *The Cognition of Basic Musical Structures*.

Temperley, D. (2009). A Unified Probabilistic Model for Polyphonic Music Analysis. *Journal of New Music Research*, *38*(1), 3–18. doi: 10.1080/09298210902928495

Tenney, I., Das, D., & Pavlick, E. (2019, July). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Florence, Italy. doi: 10.18653/v1/P19-1452

Tsay, C.-J. (2013, September). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, *110*(36), 14580–14585. doi: 10.1073/pnas.1221454110

Tzanetakis, G., & Cook, F. (1999, September). A framework for audio analysis based on classification and temporal segmentation. In *Proceedings 25th EUROMICRO Conference. Informatics: Theory and Practice for the New Millennium* (Vol. 2, p. 61-67 vol.2). doi: 10.1109/EURMIC.1999.794763

Tzanetakis, G., & Cook, P. R. (2000, April). Experiments in computer-assisted annotation of audio. In *Proceeding of International Conference on Auditory Display* (pp. 1–5). Georgia, USA. Retrieved from https://smartech.gatech.edu/handle/1853/50671

Valière, J.-C., Lefèvre, C., Colloud, F., & Villard, A. (2019, July). Identification of the consequences of a pianist's motions on his/her sound by means of psychoacoustic

tests and motion capture. In *Conference Proceedings of 26 th International Congress on Sound and Vibration (ICSV 2019)* (pp. 1–8). Montréal.

Valin, J.-M., Maxwell, G., Terriberry, T. B., & Vos, K. (2013, October). High-Quality, Low-Delay Music Coding in the Opus Codec. In *135th AES Convention.* New York, NY, USA. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=16992

VandenBos, G. R. (2015). *APA dictionary of psychology* (Second Edition ed.). Washington, DC.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data.*

Wang, C.-i., Mysore, G. J., & Dubnov, S. (2017). Re-Visiting the Music Segmentation Problem with Crowdsourcing. In *Proceedings of the 18th ISMIR Conference* (p. 7). Suzhou, China.

Watkins, R. (2023, May). Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics.* doi: 10.1007/s43681-023-00294-5

White, J. D. (1994). *Comprehensive Musical Analysis.*

Wilson, C. (2013, January). A tale of two clocks. *web.dev.* Retrieved from https://web.dev/audio-scheduling/

Yelland, N., & Masters, J. (2007, April). Rethinking scaffolding in the information age. *Computers & Education*, *48*(3), 362–382. doi: 10.1016/j.compedu.2005.01.010