REAL-TIME SOFTWARE FOR CREATING MULTIMODAL CORPORA IN MUSIC AND HEALTH SCIENCE CONTEXTS

A Thesis Presented

by

OGUZHAN TUGRAL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF ARTS

August 2025

Music Theory

ABSTRACT

REAL-TIME SOFTWARE FOR CREATING MULTIMODAL CORPORA IN MUSIC AND HEALTH SCIENCE CONTEXTS

AUGUST 2025

OGUZHAN TUGRAL
M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

This literature review examines current research at the intersection of music theory, machine learning, and cognitive science, emphasizing the limitations of existing systems for real-time multimodal analysis [1] . Contemporary models for Roman numeral analysis rely heavily on expert-annotated symbolic datasets and operate exclusively in offline, batch-processing environments, preventing dynamic alignment with physiological or behavioral data. Studies in neurophysiology, cognition, and performance further reveal that musical structures are typically analyzed prior to experimentation, resulting in limited temporal precision when interpreting memory, expectation, or affective responses. Together, these findings highlight a critical gap in technological infrastructures capable of synchronizing symbolic musical information with EEG, behavioral, or expressive data. This review motivates the need for an integrated, real-time multimodal system to advance interdisciplinary music and memory research.

[1] This thesis is scheduled for defense in May 2026; therefore, certain sections of the literature review may be revised or expanded in accordance with feedback and guidance from the thesis committee.

TABLE OF CONTENTS

Page

List of Figures

Figure **Page**

Chapter 1

LITERATURE REVIEW

This chapter reviews the existing scholarship at the intersection of music theory, machine learning, and cognitive science, with particular emphasis on the limitations of current systems for real-time, multimodal musical analysis in therapeutic contexts. It highlights significant gaps in the integration of symbolic music-analytical methods with physiological and behavioral data, underscoring the challenges that impede interdisciplinary research in music and memory. By outlining these deficiencies, the chapter establishes the conceptual and methodological foundation for the system proposed in this thesis.

1.1 Introduction

This chapter begins by outlining the organizational structure of the thesis and introducing the central research problem. Section 1.2, titled Contemporary Machine Learning Systems for Roman Numeral Analysis, identifies a fundamental challenge: the construction of multimodal datasets that achieve precise temporal alignment across heterogeneous data streams, including symbolic musical data (such as MIDI), physiological measurements (e.g., EEG), and behavioral observations. This task remains technically complex due to longstanding difficulties in music tokenization and, more specifically, in generating reliable Roman numeral analyses.

The absence of technological infrastructures capable of integrating these data modalities in real time has resulted in a scarcity of standardized, well-synchronized corpora. This gap not only constrains current research practices but also impedes the development of systematic approaches within music and memory studies. Following this problem framing, before the chapter introduces the system developed in this thesis, which collects and synchronizes musical, physiological, and expressive data in real time to enable fine-grained, multimodal analysis, I would like to clarify two concepts which is repeatedly emphasized in the present work: music and memory, multimodal data. In addition sequential data, real time analysis and batch-processing.

In this thesis, the term memory refers to both its internal computational properties and building upon this, its formation with personalized learned experiences. For example, sensory memory operates on the order of milliseconds, whereas short-term memory typically maintains a capacity of approximately 4±3 items, a characteristic that does not vary dramatically from one individual to another. If so, it becomes challenging to learn everyday experiences. In most cases severe memory problems requires professional medical assistance e.g. alzheimer's and dementia. As a natural consequence of this fundamental and relatively simple precondition, as memory then well encompasses the inevitable integration of conscious cognitive processes, emotional responses, and individual learned experiences which are able to reshape internal mechanisms of brain e.g. brain plasticity. Overall, the central motivation is to venture looking at the whole picture how music acquires meaning within human nature and experience, which requires examining musical data with particular emphasis on neurocognitive indicators and their resulting physiological and behavioral outcomes. Regarding the focus on multimodal corpora, the intention is to draw on the technical parameters and analytical value of at least two interdisciplinary domains, rather than merely extending existing parameters within a single discipline. So, when the term is used in this present work, it represents an interdisciplinary emphasis in corpus studies.

There is another important point to clarify that distinction between sequential data, real-time analysis, and batch or offline processing, as these concepts are often conflated despite referring to fundamentally different aspects of system behavior. To begin with, sequential data simply denotes information that unfolds in a temporally ordered stream—such as EEG signals, acoustic waveforms, and other time-series modalities—and this temporal structure does not, in itself, require the analysis to occur in real time. Real-time analysis, by contrast, introduces an explicit timing constraint: the system must process incoming data at the same rate at which it is produced, thereby matching human perceptual timescales or the sampling frequency of the sensors. This requirement imposes strict limits on latency and computational complexity, making real-time processing substantially more demanding than merely handling sequential data. Finally, batch or offline processing represents a different analytical paradigm altogether, one in which models can revisit the entire dataset repeatedly—often through computationally intensive architectures such as CNNs, RNNs, or Transformers—because no real-time responsiveness is required.

1.2 Problem Domain

The development of robust multimodal corpora that integrate parameters from multiple disciplinary domains remains a fundamental challenge in computational music theory and music cognition. In particular, the construction of datasets that align symbolic musical representations with physiological and behavioral data continues to introduce methodological limitations for research on music and memory. These limitations further constrain progress in therapeutic and health-science applications, as articulated through three interconnected problems summarized in Figure 1.1 and detailed in Appendix 2.[1] . The first two problems form a causal sequence, in which each problem provides the foundational explanation for the emergence of the next. As illustrated in the figure, the dependence of current machine learning systems on manually annotated datasets represents the root cause from which subsequent technological and methodological constraints arise. Although this thesis chapter was completed in August 2025, it is noteworthy that a significant development appeared shortly thereafter in the form of the Traces.js library [11], presented at the Web Audio Conference in November 2025. While Traces.js constitutes an important advancement for

[1] For more information on research in music, health and data analysis I conducted, see: `https://oguzhantugral.com/research/musicTheory/dataAnalysisMusicHealth.html`.

presenting and interacting with musical, physiological, and other time-series data on the web, its design philosophy remains oriented toward facilitating user-driven, browser-based annotation rather than reducing the reliance on manual annotations. In practice, this approach has the potential to expand the volume of human-generated labels, thereby further entrenching the field's dependence on manually curated datasets rather than advancing the automated analytical capabilities—such as Roman numeral analysis—that are fundamental to computational music theory. Moreover, the ability of such systems to retrospectively modify or reorganize time-series signals introduces significant ethical concerns. Post hoc manipulation of physiological or musical data obscures the boundary between raw experimental recordings and data shaped by later interpretation, creating ambiguity around reliability, reproducibility, and safety. In interdisciplinary settings where musical structure is examined alongside neurophysiological or behavioral responses, even minor retrospective edits can compromise the authenticity of the recorded event and weaken the evidentiary standards expected in cognitive, therapeutic, and music-analytic research. These risks reinforce the central claim of this thesis: without automated and verifiable real-time multimodal analysis, the field remains vulnerable both to methodological constraints and to emerging ethical uncertainties regarding the integrity and trustworthiness of research data.

Against this background, the first of the structural issues underpinning these limitations is that contemporary machine learning systems rely extensively on costly, manually produced expert annotations [19, 5, 21, 20, 9, 16]. This dependence is especially pronounced in music tokenization, chord recognition, and Roman numeral analysis, where expert intervention remains essential for generating reliable ground truth. This reliance directly contributes to Problem 2: the absence of technological infrastructures capable of supporting real-time, synchronized multimodal processing. Existing systems are not equipped to automatically align symbolic musical parameters with physiological or behavioral data streams during live or unfolding musical events. As a result, no current framework can integrate these modalities in a therapeutically or cognitively meaningful manner.

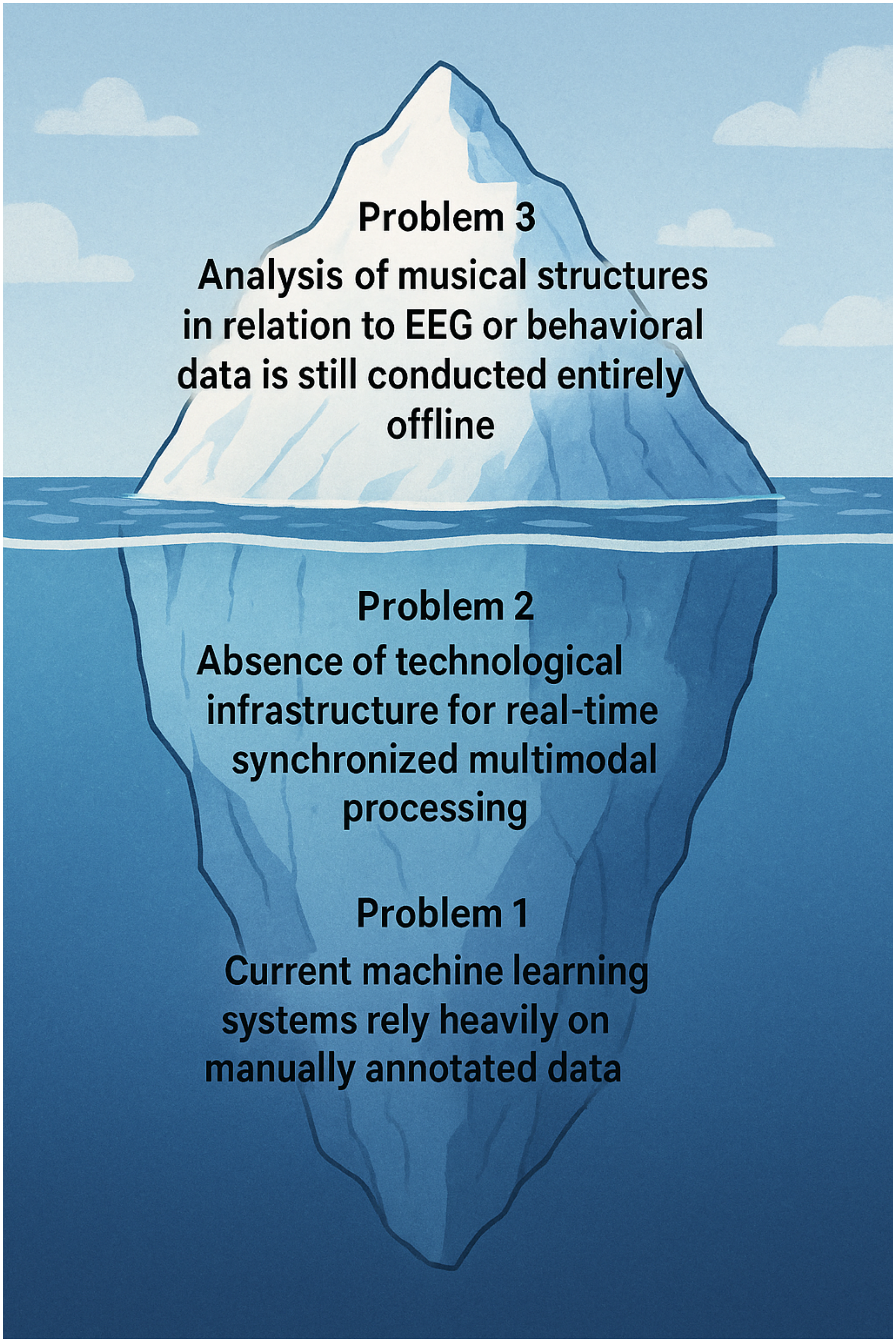
Problem 3
Analysis of musical structures in relation to EEG or behavioral data is still conducted entirely offline
Problem 2
Absence of technological infrastructure for real-time synchronized multimodal processing
Problem 1
Current machine learning systems rely heavily on manually annotated data

Figure 1.1: Problem Domain

These two foundational issues culminate in Problem 3, the surface-level manifestation of a deeper structural deficit in the field. Despite substantial advances in neuroscience, behavioral science, and music cognition, the analysis of musical structure—such as sectional boundaries, expectation violations, harmonic deviations, or formal cues—in relation to EEG or behavioral responses remains almost entirely offline [18, 1, 12, 4, 25, 13, 17, 28]. In practice, this means that musical events are analyzed retrospectively and subsequently aligned with neural or behavioral data, rather than being processed in real time.

Taken together, Problems 1 and 2 generate a systemic lack of integration, synchronization, and temporal precision across musical, physiological, and behavioral modalities during experimental contexts. This absence of real-time multimodal alignment limits the field's capacity to capture moment-to-moment cognitive, affective, and performance-related dynamics—an essential requirement for advancing research on musical expectation, memory, learning, and therapeutic intervention.

1.2.1 Thesis Proposal

To address the central problem identified above, this thesis proposes the development of a software system that targets the underlying structural limitations of current methodologies by generating a multimodal corpus situated at the intersection of music theory, psychology, and brain sciences. The core function of this system is to automatically map symbolic musical parameters onto facial-expression data and electroencephalography (EEG) signals. Through this integration, the system produces temporally precise and analytically meaningful datasets that directly support advanced research in music cognition, particularly in studies examining the relationship between music and memory.

The multimodal corpus itself is not the primary objective but rather the natural outcome of a broader technological framework designed to achieve automated, expert-level Roman numeral analysis and to align this musical tokenization with behavioral and physiological data in real time. This approach responds to a critical gap in current music research:

the need for technological infrastructures that move beyond engineering-centered solutions and incorporate the experiential, stylistic, and interpretive dimensions central to music-theoretical inquiry. The proposed system is therefore designed to accommodate a wide range of musical practices—composed or improvised—and to operate flexibly across symbolic formats such as MIDI and MusicXML. In summary, the principal aim of this thesis is to design and implement a real-time system for advanced musical data analytics capable of automatically generating multimodal corpora from unfolding musical input, thereby contributing to interdisciplinary research on musical processing, memory, and experience.

1.2.2 Route Map

In this first chapter, I aim to demonstrate how the central research problem mentioned earlier emerges from an examination of eight recent studies in Roman numeral analysis, chord recognition, and music tokenization. These studies represent state-of-the-art machine learning systems and were selected based on the independent data analysis conducted as a complementary component of the present research [26]. Having established the first part of Problem namely, that current systems remain dependent on costly, manually produced expert annotations—I will then turn to the second part: the lack of technological capacity for real-time, synchronized multimodal processing. This limitation constitutes the major bottleneck preventing the integration of symbolic musical parameters with data streams from health sciences, thereby restricting their applicability in therapeutic and cognitive research contexts.

To substantiate this claim, I will examine the structure of pretraining corpora and datasets employed in recent machine learning studies that generate Roman numeral analyses. I will then broaden the discussion by situating the problem within selected studies from music cognition, behavioral science, and neurophysiology—domains in which the system proposed in this thesis may offer a methodological contributions, particularly to ongoing research in music and memory where synchronized multimodal corpora are becoming increasingly essential.

1.3 Contemporary Machine Learning Systems for Roman Numeral Analysis

Recent developments in computational harmonic analysis demonstrate substantial progress in the automatic prediction of Roman numeral (RN) labels, yet they also reveal consistent methodological patterns that motivate the contributions of this thesis. The following section synthesizes representative state-of-the-art systems, focusing on how each model is engineered, whether it depends on expert-annotated data, and whether it operates in batch or real-time settings. These characteristics form the technical and conceptual contrast against which the present work positions its contributions.

Micchi, Gotham, and Giraud (2020 [19]) develop a hybrid neural architecture combining convolutional and recurrent layers to generate RN labels from symbolic scores. The model introduces improved pitch-spelling features and leverages expert-annotated corpora to enhance functional harmonic prediction. Despite these innovations, the system processes complete symbolic files offline in batch mode and remains dependent on curated RN annotations. As such, it typifies the dominant paradigm this thesis seeks to expand beyond: supervised, symbolic, and non-real-time analysis.

Chen and Su (2021 [5]) extend this paradigm through Transformer-based symbolic chord recognition. Their Harmony Transformer uses self-attention to capture both local and long-distance harmonic relationships, improving segmentation and RN prediction over earlier CNN- or RNN-based approaches. However, the model relies heavily on large labeled symbolic datasets and processes entire sequences simultaneously, remaining firmly batch-oriented rather than real-time or incremental.

Nápoles López et al. (2021 [21]) propose AugmentedNet, a convolutional–recurrent system that combines chroma- and bass-based feature modules with multitask learning across several tonal dimensions. By augmenting expert-annotated RN corpora and distributing learning across related subtasks, the model achieves state-of-the-art accuracy on several benchmark datasets. Nonetheless, it remains restricted to symbolic inputs, operates exclusively offline, and relies on extensive annotation. Nápoles López's subsequent study (2022 [20]) further develops this multitask framework, integrating predictions of key, inversion, and chord quality. This work strengthens the broader pattern observed across the field: increasingly sophisticated architectures continue to depend on fixed symbolic data and batch-mode training.

Moving to audio-based approaches, Donahue et al. (2022 [9]) introduce a system that transcribes melodies and chord progressions from full audio recordings. The method relies on features extracted from Jukebox, a large generative audio model, and trains a Transformer to produce lead sheets. Although this represents an expansion beyond symbolic corpora, the system still processes audio in large chunks and indirectly depends on expert-annotated harmonic data through resources such as HookTheory. It does not attempt real-time inference and remains grounded in offline processing.

Karystinaios and Widmer (2023 [16]) approach the harmonic analysis problem by modeling each note as a node in a graph structure. Their ChordGNN architecture uses note-wise relationships and edge contraction to derive onset-wise chord labels without slicing music into time windows. This design preserves more musical nuance than frame-based approaches, yet the system still depends on expert-annotated symbolic datasets and analyzes entire works offline rather than during playback. Its single-modality focus also limits its capacity for expansion into multimodal domains.

Gotham et al. (2023 [14]) take a corpus-building approach, assembling and normalizing over 2,000 symbolic harmonic analyses into the RNTXT format. This work supplies a crucial resource for machine learning research and supports reproducible, large-scale harmonic analysis. At the same time, it reinforces the field's reliance on expert-annotated symbolic corpora: the dataset itself and the tools it provides remain tied to static offline analysis rather than dynamic or multimodal contexts.

Finally, Uehara (2024 [27]) presents a notable exception to the supervised-learning norm by proposing an unsupervised system based on a neural hidden semi-Markov model (HSMM). The model incorporates both explicitly defined chord-quality templates and deep latent-variable components that infer many harmonic structures directly from symbolic pitch-class sequences. While this approach reduces dependence on annotated corpora, it still operates exclusively on symbolic inputs and performs offline batch analysis.

Across these studies, the current state of automatic Roman numeral analysis is defined by three broad characteristics: (1) a predominant reliance on supervised or annotation-dependent architectures, with only isolated attempts at unsupervised learning; (2) batch-based symbolic processing, with no real-time or incremental inference; and (3) single-modality designs that treat harmony as a static symbolic problem rather than an experiential or multimodal one. These patterns collectively motivate the present thesis, which approaches harmonic understanding as a process unfolding in time and potentially integrated with multimodal indicators of human memory and experience.

1.3.1 System Benchmark Assessment

The purpose of this benchmark assessment is not simply to compare machine learning architectures, but to evaluate how effectively existing systems support the specific aims outlined at the beginning of this chapter. Rather than prioritizing large-scale batch processing or the ability to label thousands of symbolic files automatically, the central focus of this thesis is the capacity to align Roman numeral (RN) analyses with moment-to-moment changes in multimodal data streams, particularly EEG and facial and behavioral expression measurements. In other words, the goal is high temporal precision rather than high throughput: a system that operates in real time and synchronizes symbolic musical features with neurophysiological and behavioral indicators offers fundamentally different analytical possibilities than systems designed for offline corpus processing.

As demonstrated in my earlier summer 2025 research [26][2] , current state-of-the-art

[2] This research is currently unpublished; a revised and extended version will appear in the appendix of

RN analysis systems are not designed for real-time integration with domains such as neuroscience or affective computing. Their architectures operate exclusively in offline, batch-processing modes, preventing any meaningful synchronization with physiological or behavioral data streams as they unfold. The contribution of the present work, therefore, lies in evaluating how effectively the proposed system performs on core symbolic benchmarks while simultaneously establishing the technical basis for real-time, multimodal alignment. Success is evaluated not only through conventional accuracy metrics but also through the system's ability to maintain temporal precision and responsiveness in synchronizing harmonic analysis with moment-to-moment changes in neurophysiological and behavioral signals.

Most reviewed studies share a common structure: they advance harmonic or melodic analysis through machine learning applied to symbolic inputs. Systems such as Micchi et al. (2020 [19]), Chen and Su (2021 [5]), Nápoles López (2021, 2022 [21, 20]), Karystinaios and Widmer (2023 [16]), and Gotham et al. (2023 [14]) all rely on expert-annotated corpora for supervised learning. Donahue et al. (2022 [9]) work directly with audio, but their system also incorporates harmonic information derived from annotated datasets. Only Uehara (2024 [27]) attempts an unsupervised approach through a neural hidden semi-Markov model, reducing reliance on manual labels.

Performance reporting varies widely. Karystinaios and Widmer claim improvements over AugmentedNet but depend on a post-processing step to achieve superior results. In contrast, Uehara provides explicit accuracy metrics (61–66%), offering clearer grounds for comparison and underscoring the need for standardized evaluation criteria across the field.

A defining characteristic of existing systems is their reliance on batch-mode symbolic analysis. With the exception of Donahue's audio-based model, all reviewed systems process entire pieces offline. None operate incrementally in real time or synchronize output with multimodal data such as EEG, which limits their applicability in interactive, therapeu-

this thesis. A link to the full document is provided here for reference.

tic, or perceptual studies.

These limitations align with White's observation that high-quality musical datasets remain scarce due to the specialized expertise required for annotation (White 2025 [29]). Manual entry and expert labeling continue to shape what is possible in supervised learning. While MIDI files are more accessible than engraved scores, expert-annotated RN datasets remain costly to produce.

This context motivates the system proposed in this thesis: a symbolic RN analysis engine capable of operating in real time and coordinating its predictions with multimodal data streams. By addressing both annotation scarcity and temporal alignment, the system expands the methodological foundations for future work in music cognition, pedagogy, and therapeutic applications.

To address the second part of the identified problem—namely, the technological absence of real-time, synchronized multimodal processing that links musical parameters with health-science data for therapeutic purposes—this chapter proceeds by examining the general features of the datasets and corpora used in current machine learning systems which perform Roman numeral analysis. In doing so, it lays the groundwork for motivating the system proposed in this thesis, which operates on symbolic input in real time while synchronizing with behavioral and physiological data.

1.4 Pretraining Corpora and Datasets in Contemporary Machine Learning Systems for Roman Numeral Analysis

This section examines the corpora most frequently used as pretraining datasets in contemporary machine learning systems for Roman numeral analysis (Neuwirth, 2018 [22]; Gotham, 2023 [14]; Burgoyne, 2011 [3]; DeClercq, 2011 [7]; Simonetta, 2018 [24]; Granroth, 2014 [15]; Broze, 2011 [2]; Cuthbert, 2010 [6]; Eremenko, 2018 [10]; Pfleiderer, 2017 [23]). While it is important to understand which corpora are practically implemented in music-related machine learning tasks, this study also highlights a critical gap: the absence of multimodal data that links musical parameters with behavioral, physiological, or affective signals. Such data is essential for advancing real-time, cognitively aware music systems that contribute meaningfully to fields like music cognition and health sciences.

(Neuwirth, 2018 [22]), as an Annotated Beethoven Corpus (ABC), provides expert-level functional harmonic analyses of Beethoven's complete string quartets. Each work is annotated with Roman numerals indicating key, root, inversion, and cadences. MuseScore and TSV files offer both machine- and human-readable formats. This dataset supports symbolic music processing and machine learning pretraining with a high analytical resolution.

The "When in Rome" corpus [14] is the largest meta-collection of Roman numeral annotations, aggregating sub-corpora like RomanText and TAVERN. It covers over 2,000 pieces across 1,500 works. Its unified RNTXT format and annotation tools make it the primary benchmark for supervised learning in harmonic analysis, offering symbolic files labeled by expert theorists.

The McGill Billboard Project [3] blends symbolic chord labels for popular songs with audio descriptors such as timbre, rhythm, and key extracted from Spotify and AcousticBrainz. This dual-modality format supports audio-based chord estimation as well as symbolic harmonic studies, particularly in the MIR (Music Information Retrieval) community.

De Clercq and Temperley's rock corpus [7] contains hand-annotated Roman numeral analyses for 100 iconic rock songs spanning five decades. The data includes repeated harmonic patterns and transitions. Its CSV and JSON-based files facilitate statistical analysis of rock harmony trends like plagal motion and IV–I resolution.

The Enhanced Wikifonia Dataset [24] represents a novel graph-based approach to lead sheet encoding. MusicXML files are converted into graph structures that map phrase transformations and hierarchical reductions. These advanced representations enable structural similarity queries and deepen symbolic modeling capabilities beyond flat chord lists.

Granroth-Wilding and Steedman [15] created a jazz corpus to evaluate grammar-based harmonic parsing. Annotated chord sequences are parsed using NLP-inspired techniques, showing improvements over simpler models like Hidden Markov Models. The dataset provides tonal grammar templates that mirror Roman numeral interpretations.

The iRb Jazz Corpus [2] consists of over 1,100 jazz lead sheets encoded in Humdrum format. With over 47,000 chord tokens, this dataset spans jazz styles from swing to modal. Its validity is reinforced by comparisons with fake books, making it ideal for diachronic studies of jazz harmony.

Music21 [6], a widely-used Python toolkit, offers built-in symbolic corpora such as the JSBChorales371 dataset. These Bach chorales are a standard resource for tasks including key finding, voice leading, and supervised Roman numeral prediction. The toolkit's programmability makes it essential for academic and pedagogical applications.

The MTG/JAAH dataset [10] is a jazz-specific chord transcription dataset with beat-level annotations for 113 recordings. It provides meter, form, and chord progression labels derived from Smithsonian collections. This corpus enables the development and evaluation of audio-aligned jazz chord transcription algorithms.

The Weimar Jazz Database [23] is a detailed symbolic and audio-derived dataset of 299 jazz solos. Alongside transcription data (onset, pitch, duration), it includes intensity contours, modulation ranges, and metadata like performer and tempo. The dataset is ideal for large-scale performance analysis and stylistic study.

1.4.1 Dataset and Corpora Benchmark Assessment

While these corpora have substantially advanced machine learning applications in music—from classical harmonic analysis to jazz improvisation—they remain oriented toward offline, single-domain workflows rather than the multimodal specifications outlined earlier in this chapter. None of these datasets integrate symbolic musical information with the kinds of technical descriptors necessary for interdisciplinary applications, nor do they incorporate real-time affective or physiological measurements, which is a central concern of the present work. This absence leaves a critical gap for research seeking to model cognitive and emotional processes as they unfold during listening.

Moreover, nearly all reviewed systems rely on expert-annotated Roman numeral labels for supervised pretraining. Only Uehara (2024) departs from this trend by proposing a fully unsupervised harmonic analysis method. Additionally, these corpora are processed in batch mode, meaning entire files are analyzed offline. This architectural constraint prevents live synchronization with performance or cognitive data streams.

The system proposed in this thesis addresses this problem. It eliminates the dependency on costly expert annotations by generating Roman numeral analysis in real time, synchronized with additional data streams such as EEG and facial expression. This integration enables dynamic, moment-by-moment analysis of harmonic events, aligning them with behavioral or physiological phenomena—an essential step toward therapeutic and cognition-aware applications.

In summary, the root cause of current limitations is the field's dependency on static, expert-annotated corpora-oriented machine learning offline systems and the lack of multimodal datasets. As a result, even sophisticated systems fall short of real-time interpretation. This prevents the integration of symbolic harmonic analysis with live cognitive or affective tracking, which is vital for understanding musical memory, expectation, and therapy mechanisms. The proposed system offers a potential solution by bridging these gaps and enabling cross-domain, synchronized musical analysis for interdisciplinary research—particularly in music and memory studies.

1.5 Essential Problem as the Tip of the Iceberg

Now, I will focus on selected studies in neurophysiological and behavioral research, especially those investigating short-term and long-term memory in connection with music performance within interdisciplinary contexts. These studies (Lichtenstein et al., 2024 [18]; Derks-Dijkman, 2024 [8]; Benhamou et al., 2021 [1]; Goldman et al., 2021 [12]; Byron et al., 2025 [4]; Telesco et al., 2021 [25]; Goldman, 2016 [13]; Kubit, 2024 [17]; Weiss & Peretz, 2024 [28]) represented in (Tugral, 2025 [26]) (except Goldman, 2016) mentioned at the beginning of this section, highlight the rarity of structural music analysis being systematically integrated into music and memory, or music and Alzheimer's research. This gap limits the depth and reproducibility of insights about how musical structures interact with cognitive processes like memory encoding, retrieval, and deterioration in neurodegenerative conditions.

1.5.1 Neurophysiological Work

To contextualize the aims of this thesis within existing interdisciplinary research, I first turn to neurophysiological work that uses music as an experimental framework. Studies by Goldman et al. (2021 [12]), Benhamou et al. (2021 [1]), Lichtensztejn et al. (2024 [18]), and Kubit et al. (2025 [17]) employ music as a structured paradigm for probing cognitive mechanisms such as expectation, memory, and prediction error. These works share the assumption that music's syntactic and probabilistic structure makes it an effective tool for examining complex brain processes. Across studies, this structural manipulation is paired with neurophysiological, behavioral, or linguistic measures, including EEG and behavioral responses (Goldman et al., 2021; Lichtensztejn et al., 2024), pupillometry and MRI (Benhamou et al., 2021), or natural language diary analysis (Kubit et al., 2025).

Goldman et al. (2021 [12]) investigate how listeners' brains respond to unexpected chords in popular music, focusing on the difficulty of distinguishing cognitive syntactic surprisal from basic acoustic surprisal in prior ERP research. Their goal is to determine whether sensory or cognitive computational models better explain EEG responses to harmonic expectations. While some ERP components aligned with both model types, others—especially the ERAN—did not, suggesting limits to generalizing syntax-related ERP effects. Importantly, the harmonic structure was analyzed offline, and EEG was not synchronized with moment-to-moment harmonic predictions during listening.

Benhamou et al. (2021 [1]) extend this line of inquiry by examining how frontotemporal dementia and Alzheimer's disease affect the processing of musical surprises. Using familiar melodies containing unexpected notes, they measure detection accuracy, pupillary responses, and structural brain correlates. Their results indicate that frontotemporal dementia disrupts both behavioral and autonomic responses to surprising events, while Alzheimer's patients resemble healthy controls. Here again, musical structure is modeled offline using information-theoretic estimates, without real-time integration with physiological data.

Lichtensztejn et al. (2024 [18]) investigate whether intensive singing-based training improves memory for new songs and general cognition in adults with Alzheimer's disease. EEG (N400) and ADAS-cog assessments are collected before and after training phases. Although musically informed, the protocol does not include chord recognition or Roman numeral analysis; instead, melodic violations are used to probe semantic memory. Structural processing is performed in advance, and EEG is not analyzed in real time during musical engagement.

Kubit et al. (2025 [17]) explore the feasibility of pairing individualized music listening with gamma-band light stimulation for individuals with dementia. Daily diary entries are analyzed using natural language processing, providing insight into engagement and affect. Despite using personalized musical material, the study does not incorporate detailed harmonic analysis or real-time alignment between musical structure and neural or behavioral responses.

Taken together, these four studies highlight a shared methodological pattern: musical materials—whether chord sequences, melodic deviants, or newly composed songs—are structurally analyzed or prepared offline before experiments begin. None implement dynamic, real-time integration of musical structure with neural or behavioral data streams. Synchronization occurs only through discrete event markers (e.g., ERPs or stimulus onsets), not through continuous, online mapping to harmonic parameters. This gap underscores the need for systems capable of real-time processing and multimodal alignment, a direction that motivates the methodological contributions of the present work.

1.5.2 Behavioral Studies

Now, I will focus on two behavioral studies which used structural music analysis in music and memory research. Both Byron et al. (2025 [4]) and Derks-Dijkman et al. (2024 [8]) explore how specific musical structures shape cognitive processes like attention and memory by isolating elements and testing their effects.

Byron et al. (2025 [4]) investigate how specific musical features in pop songs—such as topline vocals, choruses, and compound hooks—influence listeners' perceptions of memorability and salience. Addressing a gap in empirical studies on what constitutes a musical "hook," the authors asked participants to first listen to full pop songs and then evaluate twenty isolated excerpts per song for how much they stood out and were easy to remember. Results showed that excerpts taken from the topline vocal layer, from chorus sections, or containing compound hook structures were consistently rated as more memorable and salient. This supports the view that musical hooks function through mechanisms of attentional capture and memory encoding. Importantly, the structural elements under investigation were analyzed offline, using a combination of track separation algorithms and musicologist-led annotations. No real-time behavioral tracking or moment-to-moment synchrony with listener responses was implemented. As such, the study highlights static correlates of musical salience rather than dynamic perceptual or cognitive alignment over time.

Derks-Dijkman et al. (2024 [8]) investigates how musical mnemonics affect working memory in cognitively unimpaired older adults and individuals with amnestic mild cognitive impairment (aMCI), which is a clinical condition characterized by noticeable memory problems that are greater than expected for a person's age, but not severe enough to significantly interfere with daily life or meet criteria for dementia (like Alzheimer's disease). The problem addressed is that, while musical mnemonics are known to help episodic memory, their effects on working memory in aging and aMCI had not been tested. The goal was to assess whether rhythm, pitch, or melody could improve working memory performance in these groups. Using a forward digit span task[3] in spoken, rhythmic, pitch-based, and melodic conditions, the researchers found that rhythm enhanced performance, whereas pitch and melody hindered it, with musical expertise boosting the positive effects. They conclude that rhythm-based mnemonics may offer a simple, non-pharmacological strategy to support working memory in aging and cognitive impairment. In this study, musical material was structurally pre-recorded as unfamiliar pitch and rhythm sequences; no chord recognition or Roman numeral analysis was performed, and all musical structures were processed offline without real-time integration or synchronization with behavioral data.

In conclusion, these two studies also share the feature of using pre-analyzed, offline musical material rather than dynamic, real-time processing. In Byron et al. (2025 [4]), the pop songs are structurally prepared in advance through track separation and expert annotation to identify topline vocals, choruses, and hooks. Similarly, Derks-Dijkman et al. (2024 [8]) use pre-recorded sequences of pitch, rhythm, and melody as controlled mnemonic stimuli. Neither experiment processes or synchronizes these musical structures dynamically with participants' behavioral or neural responses during listening, which limits tracking of cognitive effects online.

1.5.3 Studies on Music Performers' Reactions

Finally, I turn to studies examining the reactions of music performers. While two of these are drawn from the data analysis in Tugral (2025 [26]), I also include Goldman (2016 [13]) in the present discussion, as it directly relates to the current work's focus on the cognition of musical performance. Goldman's study is not included in Tugral (2025 [26]) because

[3] A forward digit span task is a neuropsychological test in which participants are asked to repeat increasingly long sequences of spoken digits in the same order, assessing short-term memory and attention. The test continues until the participant fails two sequences of the same length, and their highest correctly repeated sequence determines the digit span score.

that analysis is limited to research published between 2020–2025. The three selected studies (Goldman, 2016 [13]; Telesco et al., 2021 [25]; Weiss & Peretz, 2024 [28]) underscore the importance of musician experience and training in shaping how musical memory and improvisation are understood and developed. Each connects musical structure to underlying cognitive and memory processes, demonstrating that structural features influence what musicians perceive, learn, and recall.

Goldman (2016 [13]) argues that although improvisation research spans many fields, its scientific study has struggled due to vague concepts like novelty, freedom, and spontaneity that are hard to define or test consistently. To address this, the paper proposes an alternative framework that defines improvisation as a ”way of knowing,” focusing on how different training and experience shape musicians’ perception, cognition, and performance behaviors. The method involves comparing improvisers and non-improvisers using neuroscientific and psychological experiments, especially looking at perception–action coupling. While the paper itself is theoretical, it reviews studies suggesting that musical training creates measurable differences in the brain and behavior, supporting this approach. Goldman concludes that reframing improvisation as a way of knowing can make it possible to design clearer, falsifiable experiments and connect scientific work more productively with music theory and critical studies. In this framework, musical structures like chords are analyzed theoretically or in planned experiments but not processed or synchronized dynamically in real time, meaning that music, EEG, and behavioral data are not integrated for moment-to-moment tracking—confirming that current methods still lack real-time dynamic analysis during actual performance.

Telesco et al. (2021 [25]) may be seen as complementary to Goldman (2016 [13]). The study investigates whether a retrieval practice learning strategy—known to boost long-term retention in verbal learning—can also improve memorization of piano melodies. The problem addressed is that musicians often struggle to memorize music securely, and traditional methods like restudy may not be the most effective. The goal was to experimentally test whether blocked or alternating retrieval practice helps students remember short piano pieces better than restudying alone. The method involved two controlled lab experiments comparing study and retrieval practice schedules with short piano melodies, followed by memory tests after short (10-minute) and longer (2-day) delays. Results showed a small but consistent trend: retrieval practice conditions slightly improved memory accuracy over study, especially with longer retention intervals, although differences were not statistically significant due to small sample sizes. The authors conclude that retrieval practice shows promise for music memorization but recommend larger, more robust studies to confirm its effect. The study uses simple, clear tonal melodies with predefined harmonic structure but does not perform chord recognition or Roman numeral analysis; structural aspects like cadences were composed and presented offline, and musical performance and memory retrieval were not synchronized dynamically with behavioral or neural data in real time.

In a more specific domain, Weiss et al. (2024 [28]) investigate whether the well-known advantage for recognizing vocal over instrumental melodies also extends to recall. The problem addressed is that while vocal melodies are more recognizable, it is unclear whether this benefit applies to the more demanding task of memorizing and reproducing melodies. The study's goal is to test if trained violinists remember and recall sung melodies better than violin-played ones, and whether this advantage depends on tonal structure. The method involved violinists learning four matched melodies (two tonal, two nontonal) and immediately recalling them in the same timbre (voice or violin), then recalling them again after a short delay. Results show a vocal advantage for melodic contour in immediate recall but a violin advantage for both contour and interval accuracy in delayed recall, with tonality improving recall overall. The study concludes that singing provides a short-lived contour-based memory boost, but instrumental expertise dominates long-term recall accuracy.

In terms of benchmark purposes with the system proposed in this work, Goldman (2016 [13]), Telesco et al. (2021 [25]), and Weiss & Peretz (2024 [28]) all employ pre-defined structural musical material that is analyzed offline. None of these studies implement dynamic, moment-to-moment synchronization of musical structures with behavioral or neural data. In Appendix 1, I present a mock-up experiment using the current system developed in this thesis to collect and analyze performance data from five improvisers of varying expertise. This experiment illustrates how musicians' spontaneous behavior—when analyzed in conjunction with dynamically processed, real-time multimodal data—can offer new insights not only for Goldman's later work on ERP responses to naturalistic music (Goldman, 2021 [12]) but also for his earlier conceptualization of ”perception-action coupling” (Goldman, 2016 [13]). Following the introduction and documentation of the proposed technological tool, which includes both hardware and software components (the latter designed specifically in this thesis), I will revisit these studies in the Future Directions section to discuss how this type of technology could enhance experimental methodologies across interdisciplinary fields.

Up to this point, I have demonstrated that Problem namely, the limited engagement of core music-theoretical studies with research on music and memory—has contributed to a broader absence of technological solutions capable of bridging theoretical insight and cognitive application. Regarding another problem, current music tokenization systems typically depend on expert-annotated datasets and operate exclusively in offline environments. These constraints significantly hinder the creation of interdisciplinary, multimodal corpora. As a result, neurophysiological and behavioral research—especially that which seeks real-time, biologically and ecologically valid experimental paradigms—remains limited by the absence of integrated symbolic-music processing tools that can align with cognitive or therapeutic data streams.

Bibliography

[1] Elia Benhamou et al. “Decoding expectation and surprise in dementia: the paradigm of music”. In: *Brain Communications* 3.3 (2021), fcab173. DOI: `10.1093/braincomms/fcab173`. URL: `https://doi.org/10.1093/braincomms/fcab173` (cit. on pp. 6, 15–17).

[2] Yuri Broze and Daniel Shanahan. “Diachronic Changes in Jazz Harmony: A Cognitive Perspective”. In: *Music Perception: An Interdisciplinary Journal* 31.1 (2013), pp. 32–45. DOI: `10.1525/mp.2013.31.1.32`. URL: `https://doi.org/10.1525/mp.2013.31.1.32` (cit. on pp. 12, 14).

[3] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. “An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis”. In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. 2011, pp. 633–638 (cit. on pp. 12, 13).

[4] Timothy P. Byron, Carl T. Rushworth, and Maxwell J. Stewart. “Popular Music Excerpts Are Rated As More Memorable And Salient If They Involve Vocals, Compound Hooks, And Choruses”. In: *Music Perception* 42.3 (2025), pp. 197–206. DOI: `10.1525/mp.2024.2322897`. URL: `https://doi.org/10.1525/mp.2024.2322897` (cit. on pp. 6, 16, 18, 19).

[5] Tsung-Ping Chen and Li Su. “Attend to Chords: Improving Harmonic Analysis of Symbolic Music Using Transformer-Based Models”. In: *Transactions of the International Society for Music Information Retrieval* 4.1 (2021), pp. 1–13. DOI: `10.5334/tismir.65`. URL: `https://doi.org/10.5334/tismir.65` (cit. on pp. 4, 8, 11).

[6] Michael Scott Cuthbert and Christopher Ariza. “music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data”. In: *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*. 2010, pp. 637–642 (cit. on pp. 12, 14).

[7] Trevor Declercq and David Temperley. “A generalized model of tonal attraction”. In: *Music Perception* 28.4 (2011), pp. 293–305. DOI: `10.1525/mp.2011.28.4.293`. URL: `https://doi.org/10.1525/mp.2011.28.4.293` (cit. on pp. 12, 13).

[8] Marije W. Derks-Dijkman et al. “Effects of Musical Mnemonics on Working Memory Performance in Cognitively Unimpaired Older Adults and Persons with Amnestic Mild Cognitive Impairment”. In: *Journal of Neuropsychology* 18.2 (2024), pp. 286–299. DOI: `10.1111/jnp.12342`. URL: `https://doi.org/10.1111/jnp.12342` (cit. on pp. 15, 18, 19).

[9] Chris Donahue, John Thickstun, and Percy Liang. “Melody Transcription via Generative Pre-Training”. In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*. Bengaluru, India, 2022. URL: `https://chrisdonahue.com/sheetsage` (cit. on pp. 4, 9, 11).

[10] Vsevolod Eremenko et al. “Audio-Aligned Jazz Harmony Dataset for Automatic Chord Transcription and Corpus-Based Research”. In: *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*. Paris, France, 2018, pp. 483–490 (cit. on pp. 12, 14).

[11] Lawrence Fyfe and Elaine Chew. “Traces.js: A Javascript library for presenting music, physiology, and other time-series on the web”. In: *Proceedings of the Web Audio Conference (WAC-2025)*. Camera-ready version. Paris, France, Nov. 2025 (cit. on p. 3).

[12] Andrew Goldman et al. "Reassessing Syntax-Related ERP Components Using Popular Music Chord Sequences". In: *Music Perception: An Interdisciplinary Journal* 39.2 (2021), pp. 118–144. DOI: 10.1525/MP.2021.39.2.118. URL: https://doi.org/10.1525/MP.2021.39.2.118 (cit. on pp. 6, 15, 16, 22).

[13] Andrew J. Goldman. "Improvisation as a Way of Knowing". In: *Music Theory Online* 22.4 (2016). DOI: 10.30535/mto.22.4.1. URL: https://doi.org/10.30535/mto.22.4.1 (cit. on pp. 6, 16, 19–22).

[14] Mark Gotham et al. "When in Rome: A Meta-corpus of Functional Harmony". In: *Transactions of the International Society for Music Information Retrieval* 6.1 (2023), pp. 150–166. DOI: 10.5334/tismir.165. URL: https://doi.org/10.5334/tismir.165 (cit. on pp. 9, 11–13).

[15] Mark Granroth-Wilding and Mark Steedman. "A Robust Parser-Interpreter for Jazz Chord Sequences". In: *Journal of New Music Research* 43.4 (2014), pp. 355–374. DOI: 10.1080/09298215.2014.910532. URL: https://doi.org/10.1080/09298215.2014.910532 (cit. on pp. 12, 13).

[16] Emmanouil Karystinaios and Gerhard Widmer. "Roman Numeral Analysis with Graph Neural Networks: Onset-Wise Predictions from Note-Wise Features". In: *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*. Milan, Italy, 2023. URL: https://github.com/manoskary/chordgnn (cit. on pp. 4, 9, 11).

[17] Benjamin M. Kubit et al. "Diary Analysis of an RCT: Natural Language Analyses of Gamma-Music-Based Intervention". In: *Music and Medicine* 17.1 (2025), pp. 10–16. DOI: 10.47513/mmd.v17i1.1066. URL: https://doi.org/10.47513/mmd.v17i1.1066 (cit. on pp. 6, 16, 17).

[18] Marcela Lichtensztejn et al. "Memory for Music (M4M) Protocol for an International Randomized Controlled Trial: Effects of Individual Intensive Musical Train-

ing Based on Singing in Non-Musicians with Alzheimer's Disease". In: *medRxiv* (Sept. 2024). Preprint. DOI: 10.1101/2024.09.25.24313991. URL: https://doi.org/10.1101/2024.09.25.24313991 (cit. on pp. 6, 15–17).

[19] Gianluca Micchi, Mark Gotham, and Mathieu Giraud. "Not All Roads Lead to Rome: Pitch Representation and Model Architecture for Automatic Harmonic Analysis". In: *Transactions of the International Society for Music Information Retrieval* 3.1 (2020), pp. 42–54. DOI: 10.5334/tismir.45. URL: https://doi.org/10.5334/tismir.45 (cit. on pp. 4, 8, 11).

[20] Néstor Nápoles López. "Automatic Roman Numeral Analysis in Symbolic Music Representations". PhD Dissertation. McGill University, 2022 (cit. on pp. 4, 9, 11).

[21] Néstor Nápoles López, Mark Gotham, and Ichiro Fujinaga. "AugmentedNet: A Roman Numeral Analysis Network with Synthetic Training". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*. Conference dates: November 7–12. Online, Nov. 2021. DOI: 10.5281/zenodo.5624533. URL: https://doi.org/10.5281/zenodo.5624533 (cit. on pp. 4, 8, 11).

[22] Markus Neuwirth et al. "The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets". In: *Frontiers in Digital Humanities* 5 (2018), Article 16. DOI: 10.3389/fdigh.2018.00016. URL: https://doi.org/10.3389/fdigh.2018.00016 (cit. on pp. 12, 13).

[23] Martin Pfleiderer et al., eds. *Inside the Jazzomat: New Perspectives for Jazz Research*. Schott Campus, 2017 (cit. on pp. 12, 14).

[24] Federico Simonetta et al. "Symbolic Music Similarity through a Graph-Based Representation". In: *Audio Mostly 2018: Sound in Immersion and Emotion (AM'18)*. Conference dates: September 12–14. Wrexham, United Kingdom: ACM, Sept. 2018.

DOI: 10.1145/3243274.3243301. URL: https://doi.org/10.1145/3243274.3243301 (cit. on pp. 12, 13).

[25] Paula Telesco et al. “Does Retrieval Practice Enhance Memorization of Piano Melodies?” In: *College Music Symposium* 61.2 (2021), pp. 1–23. URL: https://www.jstor.org/stable/10.2307/48645695 (cit. on pp. 6, 16, 20, 21).

[26] Oguzhan Tugral. *Data Analysis on ”Music and Memory” Research in Music and Health Science Contexts*. https://oguzhantugral.com/research/musicTheory/dataAnalysisMusicHealth.html. Accessed: 2025-07-24. 2025 (cit. on pp. 7, 10, 16, 19).

[27] Yui Uehara. “Unsupervised Learning of Harmonic Analysis Based on Neural HSMM with Chord Quality Templates”. In: *Proceedings of the International Conference on New Music Concepts (ICNMC 2024)*. arXiv preprint. 2024. URL: https://arxiv.org/abs/2403.04135 (cit. on pp. 9, 11).

[28] Michael W. Weiss and Isabelle Peretz. “The Vocal Advantage in Memory for Melodies Is Based on Contour: Evidence from Recall in String Players”. In: *Music Perception* 41.4 (2024), pp. 275–287. DOI: 10.1525/mp.2024.41.4.275. URL: https://doi.org/10.1525/mp.2024.41.4.275 (cit. on pp. 6, 16, 20, 21).

[29] Christopher W. White. *The AI Music Problem: Why Machine Learning Conflicts with Musical Creativity*. New York: Routledge, 2025. DOI: 10.4324/9781003587415. URL: https://doi.org/10.4324/9781003587415 (cit. on p. 12).