Doctoral dissertation

Acoustic and perceptual aspects of sound sources in the formation of ensemble sound

Jithin babu Pozhamkandath Thilakan

Accepted by Detmold University of Music, Germany, in fulfillment of the requirements for the degree of Doctor of Engineering Sciences.

Date of defense: 30 September 2024

Doctoral committee:

Supervisor :	Prof. Dr.-Ing. Malte Kob	Erich Thienhaus Institute, Detmold University of Music, Germany
Reviewers :	Prof. Dr. Stefan Weinzierl	Audio Communication Group, Technische Universität Berlin, Germany
	Prof. Dr.-Ing. Malte Kob	Erich Thienhaus Institute, Detmold University of Music, Germany

Erich Thienhaus Institute, Detmold University of Music, Neustadt 22, 32756, Detmold, Germany

Abstract

The acoustic and auditory properties of individual musical instruments have been extensively studied over recent decades, however, their sonic interplay within a musical ensemble remains under-explored. Given their considerable importance across various fields, aspects of ensemble sound, such as blending of instruments, perceptual relevance of directivity of instruments, and the role of room acoustics, demand comprehensive evaluations and an interdisciplinary approach. This study aims to improve the perceptually motivated acoustic representation of instruments in joint performance in both real and virtual acoustic domains, by exploring different stages of these aspects in musically realistic contexts.

An explorative listening test with live string ensemble performance suggested that the characteristics of the acoustic environment considerably influence the blending of violins playing in unison. Combining methods of Machine Learning and Music Information Retrieval, a computational modelling approach is proposed to classify sound samples from an ensemble recording according to perceived blending. Proving this classification to be effective for monophonically rendered sound samples of two violins from in-situ environments, without requiring the individual source recordings marks a first step towards comprehensive blending modelling. Furthermore, the applicability of close-microphone recordings for auralization of a perceptually convincing ensemble sound was successfully demonstrated.

Advancing previous research in directivity perception, it could be demonstrated that the room acoustics have a greater impact on the orientation perception of sources than their directivity. By involving instruments with distinct radiation directivities in a variety of acoustic environments, the major acoustical parameters influencing the orientation perception have been explored. Examining musical instruments with their inherent dynamic directivity against loudspeakers in in-situ conditions showed that their distinction becomes obscured under specific acoustic conditions. These findings led to a pilot study on the perceptual relevance of high-order directivity modelling of individual sources forming an ensemble. Results indicate, that even with an increasing number of sources, their detailed directivity characteristics remain pivotal for auralizing ensemble performance.

The role of room acoustics in shaping the overall blending is shown to be dependent on the source-level blending. A computational model for predicting overall perceived blending in musical performance using source-level blending ratings and room acoustical parameters was suggested and validated. Analysis of its feature importance revealed that the room acoustic contribution to the overall blending impression is nearly

as significant as the blending between instruments at the source level. By emphasizing and detailing relations between musical blending, directivity perception, and auralization aspects, this thesis contributes to the advancement of ensemble sound research and offers insights pertinent to music performance and perception research, virtual acoustics, and related fields.

Keywords: Musical ensembles, Instrument blending, Sound source Directivity, Room acoustics, Auralization.

Zusammenfassung

Die akustischen und auditiven Eigenschaften einzelner Musikinstrumente wurden in den letzten Jahrzehnten ausführlich untersucht - ihr klangliches Zusammenspiel in einem Ensemble ist jedoch vergleichsweise wenig erforscht. Aspekte des Ensembleklangs wie die Klangverschmelzung, der Beitrag der Abstrahlcharakteristik für die Klangformung und der Einfluss der Raumakustik sind für verschiedene Fachgebiete von Bedeutung und erfordern daher eine ganzheitliche Betrachtung und einen interdisziplinären Ansatz. Ziel dieser Arbeit ist es daher, die perzeptiv relevanten akustischen Darstellungen gemeinsam klingender Instrumente für sowohl reale als auch virtuelle akustische Umgebungen zu verbessern. Dazu werden verschiedene Abstufungen dieser Aspekte in musikalisch realistischen Kontexten untersucht.

Ein explorativer Hörtest anhand von live Einspielungen eines Streicherensembles ergab, dass die raumakustischen Eigenschaften die Klangverschmelzung unisono spielender Violinen erheblich beeinflussen. Durch die Kombination von Methoden des maschinellen Lernens und des Music Information Retrieval wird ein Modellierungsansatz vorgestellt der es erlaubt, Klangbeispiele aus einer Ensembleaufnahme nach dem Grad der wahrgenommenen Klangverschmelzung zu klassifizieren. Diese Klassifikationsmethode kommt dabei ohne nahmikrofonierte, quellgetrennte Signale aus und wurde vielmehr anhand von monophonen Raumklangaufnahmen von zwei Violinen unter realen Aufführungsbedingungen validiert. Diese Methodik stellt einen erfolgversprechenden ersten Schritt hin zu einer ganzheitlichen Modellierung von Klangverschmelzung dar. Darüber hinaus wird gezeigt, dass Nahmikrofonaufnahmen geeignet sind um einen perzeptiv überzeugenden Ensembleklang zu auralisieren.

In Weiterentwicklung früherer Forschungen zur Richtwirkungswahrnehmung konnte gezeigt werden, dass die Raumakustik einen größeren Einfluss auf die Wahrnehmung der Orientierung einer Quelle hat als deren Richtwirkung. Anhand von Instrumenten mit unterschiedlichen Hauptabstrahlrichtungen in verschiedenen akustischen Umgebungen wurden die akustischen Parameter identifiziert welche die Orientierungswahrnehmung hauptsächlich bedingen. Für Musikinstrumente als Quellen mit inhärent dynamischer Richtwirkung zeigte sich im in-situ Vergleich zu Lautsprechern, dass bestimmte raumakustische Bedingungen die Unterscheidbarkeit erschweren. Diese Erkenntnis inspirierte eine Pilotstudie zur Detailtreue der Richtcharakteristikmodellierung von Quellen innerhalb eines Ensembles in Bezug auf die Wahrnehmung des Gesamtklangs. Diese zeigt, dass selbst bei einer zunehmenden Anzahl von Quellen deren spezifische Richtcharakteristik weiterhin von entscheidender Bedeutung für die Auralisierung von Ensembledarbietungen ist. Die Rolle der Raumakustik bei der Gestaltung von Ensembleklang hängt dabei von der Klangverschmelzung auf Quellenebene ab. Ein Modell zur Vorhersage der insgesamt wahrgenommenen Verschmelzung einer Musikdarbietung wird vorgeschlagen und validiert, das auf akustischen Parametern basiert und mit subjektiven Einschätzungen

der Klangverschmelzung auf Quellenebene trainiert wird. Eine Merkmalsanalyse ergab dabei, dass die Raumakustik fast genauso wichtig ist für den Gesamteindruck von Ensembleklang wie die Klangverschmelzung der einzelnen Instrumente auf Quellenebene.

Durch die Herausarbeitung und Verdeutlichung der Zusammenhänge zwischen musikalischer Verschmelzung, Richtwirkungswahrnehmung und Auralisation trägt diese Arbeit zur Weiterentwicklung der Ensembleklangforschung bei und bietet Erkenntnisse, die für die Musikaufführungs- und Musikwahrnehmungsforschung, die virtuelle Akustik und verwandte Bereiche relevant sind.

Schlüsselwörter: Musikensembles, Instrumentenmischung, Richtwirkung der Schallquelle, Raumakustik, Auralisierung.

Acknowledgement

First and foremost, I am deeply grateful to Dr. Malte Kob for granting me the freedom, support, and opportunities to explore various aspects of my research journey. His patience and encouragement allowed a physics graduate with minimal knowledge in acoustics to evolve and grow in this field. Thanks for taking a chance on me. I am sincerely thankful to Dr. Stefan Weinzierl for serving as the external examiner of my thesis. His insightful and constructive feedback on various aspects of my research have been invaluable in refining and strengthening my work. This research was conducted as part of the Marie Skłodowska-Curie Action-funded VRACE (Virtual Reality Audio for Cyber Environments) project (grant agreement number 812719) at the Erich Thienhaus Institute, Detmold University of Music, Germany, between 2019 and 2023. Additionally, several studies presented in this thesis were developed in collaboration with the ACTOR (Analysis, Creation, and Teaching of ORchestration) project. I extend my sincere thanks to the coordinators of both projects for their financial support and for providing the resources that made this research possible.

My sincere gratitude to Dr. Balamurali B.T. for his unconditional support during difficult times, his guidance, compassionate encouragement, and, most importantly, his integral involvement in advancing this thesis. I am deeply grateful to Dr. Timo Grothe for his generosity in sharing knowledge, offering support, and always being available for discussions, all of which had a significant influence on the progress of my research. I am sincerely thankful to Dr. Chen Jer Ming for introducing me to the field of acoustics, sparking a genuine interest in doing research, and for his generous contribution to the advancement of many of the research works presented in this thesis. I extend my thanks to Dr. Eckard Mommertz for the opportunity to collaborate with Müller-BBM GmbH, and also for his kind support in many research works. I also thank Dr. Aristotelis Hadjakos, the Chair of Doctoral committee, for his kind guidance throughout the submission and evaluation process of my PhD thesis. I am grateful to Otavio and Andrea, VRACE collaborators turned good friends, for their kind support in the numerous studies involved in this work. I thank Prof. Martha De Francisco, Ying-Ying, and other researchers of the ACTOR project for helping me in various projects, and for the exchange of knowledge and support. I am thankful to Dr. Sebastia Amengual Gari, and Dr. David Ackermann, for their supportive role in sharing knowledge and offering guidance that has contributed significantly to the growth of this research.

Many thanks to the Tonmeister colleagues at ETI, whose golden ears were always available to assist me during conducting listening tests. Thanks to Malte Heinz for being there for all my needs in ETI, and making my life in ETI much nicer. I am grateful to Walter, Theresa, and Quique for being integral parts of my journey at ETI, and for learning and growing together. I am grateful to Santhannan, Vasanthy aunty, and Robinson uncle for their unconditional love, and for making Detmold feel like a second

home. Sincere thanks to Anil and Aleena for being the constant critics, supporters, and motivators in my ups and downs of this journey. Lastly, I want to express my deepest gratitude to Amma, Ammu, and Anju for being the three pillars of strength in my life. Without your support, this would not have been possible. To Thilakettan, with love, for being the guiding light in my life; this thesis is dedicated to you.

Acronyms

ASA	Auditory Scene Analysis
ASW	Apparent Source Width
BRIR	Binaural Room Impulse Response
DRR	Direct-to-Reverberant Ratio
GA	Geometrical Acoustics
HOA	Higher-Order Ambisonics
HRTF	Head-Related Transfer Function
IACC	Interaural Cross Correlation
ILD	Interaural Level Difference
ITD	Interaural Time Difference
IQR	Inter-quartile range
JND	Just Noticeable Difference
LDA	Linear Discriminant Analysis
LEV	Listener Envelopment
LOOCV	Leave-One-Out Cross Validation
LTI	Linear Time-Invariant
MFCC	Mel-Frequency Cepstral Coefficient
MIR	Music Information Retrieval
ML	Machine Learning
MUSHRA	MUlti Stimulus test with Hidden Reference and Anchor
PCA	Principal Component Analysis
PDF	Probability Distribution Function
RIR	Room Impulse Response
SC	Spectral Centroid
SH	Spherical Harmonics
SRIR	Spatial Room Impulse Response
t-SNE	t-Stochastic Neighbourhood Embedding
VR	Virtual Reality
WFS	Wave Field Synthesis
XR	Extended Reality

Contents

Chapter 1

Introduction

1.1 Motivation

Ensemble sound epitomizes the harmonious convergence of a group of voices or musical instruments that emerge from interconnectedness and collective expressions, transcending boundaries and resonating deeply within the human experience. As articulated by Daniel Barenboim, the musical ensemble functions as a dynamic organism, reliant on symbiotic collaboration, effective communication, and creative synergy. Attributes such as orchestration techniques, timbral characteristics of individual instruments, joint performance strategies, spatial arrangement of sources, room acoustics, and more, interact in complex ways to shape the formation and perception of ensemble sound, highlighting the intricate relationship involved between music, acoustics, and human perception. Ranging from the percussion ensemble, *Ilanjithara Melam* from Kerala, India to the *Vienna Philharmonic orchestra*, these characteristics of the ensemble sound exhibit cross-cultural consistency regardless of their size or genre. Although significant research has been carried out on the acoustic and perceptual attributes of individual musical instruments, the impact of these attributes in musical ensembles remains mostly under-explored. This gap is primarily due to the multitude of factors influencing the formation of ensemble sound, necessitating an interdisciplinary approach.

During a live musical ensemble performance, the listeners typically do not hear individual instruments as it is, but an immersive perceptual fusion of musical instruments that blend together to result in an 'auditory chimera', a sound with rich musical timbre, which is mostly different from the timbre of the constituent instruments. Rather than perceiving each instrument distinctly, achieving a unified and harmonious perception of instruments is a fundamental sonic objective in joint musical performances.

Therefore, auditory blending, a psychoacoustic phenomenon referring to the perceptual fusion of sound sources, stands as an integral perceptual attribute of joint musical performances and ensemble sound. The blending of musical instruments shaped by the collective effort of musicians, and the impact of room acoustics on it, are critical aspects in both music performance and perception domains. Yet, these aspects remain largely unexplored in realistic conditions, mainly due to the complex, multi-level, and multi-dimensional characteristics involved in blending.

The directivity characteristics of musical instruments, which shape the spatial distribution of energy radiated from the instrument, constitute another perceptually relevant acoustic attribute of the sound sources within a musical ensemble. Based on the spatial distribution of the constituent instruments including their position and orientation, the directivity of sound sources is observed to alter the perceived ensemble sound in spectral, temporal, and spatial domains. Attributes associated with directivity, such as source orientation, are pivotal in the arrangement of instruments in music performance, and instrument recording techniques. Moreover, modeling the directivity of sound sources with high spectral and spatial resolution holds significance in accurate room acoustic simulations and sound field reconstruction applications. While the directivity characteristics of individual instruments have been extensively investigated over the past decades, their perceptual significance in room acoustic environments has been relatively underexplored. This aspect becomes even more significant when it comes to musical ensembles consisting of multiple sources.

The Virtual orchestra is considered to be the next-generation tool of orchestration that opens up a wide range of possibilities such as customizable immersive listening experiences for entertainment purposes, experimentations on composition and orchestration, educational applications, facilitation of interactive collaborations and performances, and much more. Reconstructing a perceptually convincing spatial sound field impression of an orchestra or ensemble involves various aspects, including capturing source signals from each constituent instrument, modeling the directivity of sound sources, simulation of room acoustic environment, and sound field reproduction, etc. To ensure a musically and perceptually authentic acoustic representation of sound sources in a virtual orchestra, the initial step is to capture authentic source signals that retain the intrinsic and natural attributes of joint performance. Moreover, the assessment of the perceptual requirement of the 'detailings' in directivity modeling of individual sound sources in room acoustic environments is highly relevant for a perceptually plausible creation of ensemble sound by minimizing computational efforts. Furthermore, the assessment of blending between sound sources should be incorporated as an essential quality attribute of the virtual ensemble sound.

This thesis aims to investigate the perceptually relevant acoustic attributes of sound sources in joint performance within a musically realistic setting. This includes the evaluation of blending between instruments, perceptual relevance of directivity of instru-

ments, and capturing individual instruments during joint musical performance from in-situ conditions. Moreover, it also seeks to explore the role of room acoustic attributes in these aspects. By delving into these aspects, this thesis attempts to advance the acoustic and perceptual aspects of joint music performance research, and thereby contributing to various fields including virtual reality acoustics, music performance and perception research, music recording techniques, and beyond.

1.2 Background

1.2.1 Ensemble sound formation: an overview

When musicians perform together in a joint musical performance, the sound radiated from the instrument reaches the listeners as direct sound as well as room acoustic reflections. Such joint performances by the musicians are significantly influenced by multimodal attributes including aural, visual, and tactile feedback, and thereby shape the apparent ensemble sound generated. Figure 1.1 illustrates a flow diagram depicting the major aspects and multimodal attributes involved in a joint musical performance.

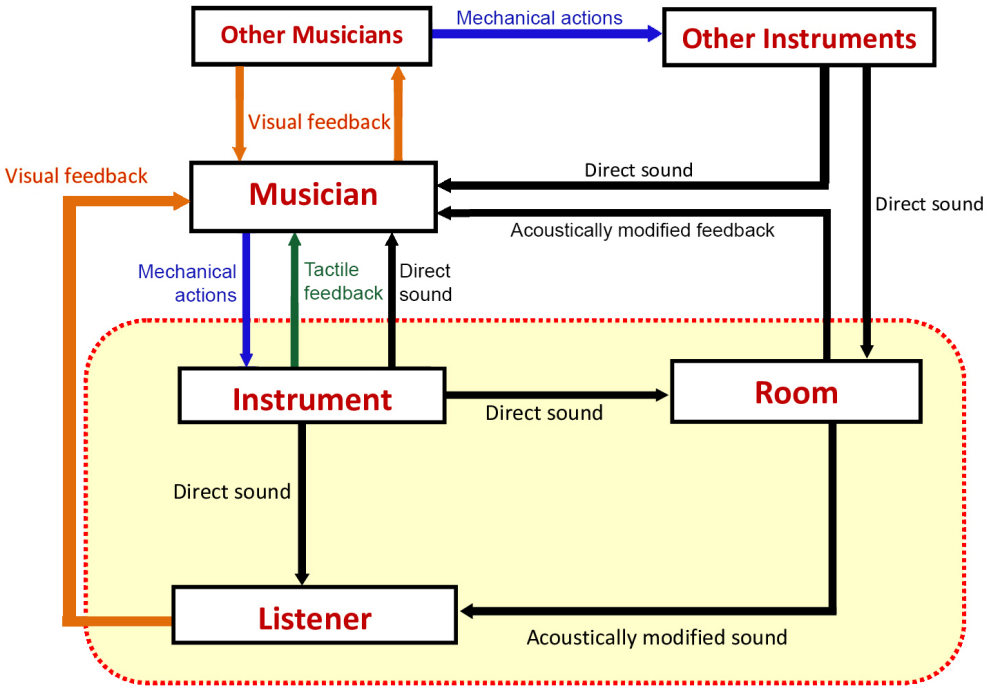


Figure 1.1: Flow diagram of major aspects and multimodal attributes involved in a joint musical performance.

The first important aspect in the ensemble sound performance involves the distribution, i.e., the position and orientation, of musical instruments. The quality of the ensemble performance is substantially impacted by performance-related attributes such as the spacing and orientation of the sources and the corresponding self-to-others ratio of auditory feedback [1; 2]. This is accompanied by cross-performer interaction using visual feedback as well as auditory feedback [3; 4], and joint action strategies such as leader-follower roles among the players [5]. The presence of a conductor can be an important aspect in this stage. By optimizing these performance-related factors and interactions, performers try to attain temporal synchronicity and pitch similarity in producing sound and also modify the instrument timbre while performing [2; 3; 4; 5; 6]. Once the sound is produced, it radiates from the instrument according to the directivity characteristics of the instruments and reaches both the listener and the performer as direct sound. Subsequently, a multitude of room acoustic reflections, that encompass strong early reflections and late reverberations, reaches the listeners and performers. These room acoustic reflections are observed to shape the spectral and spatial attributes of the perceived instrument sound; the reflections act as a filter in the frequency domain due to the absorption properties of the walls, moreover, the spatial impression including the perceived source width and spatial envelopment are altered according to the strength and directionality of the room acoustic reflections. These perceptually important attributes of room acoustics have been extensively analyzed in the past decades, particularly in the context of musical performance perception, by utilizing a wide range of acoustic parameters and verbal descriptors [7; 8; 9]. Additionally, these modified room acoustic reflections, serving as room acoustics feedback, are another major factor that has been shown to influence the timbre, tempo, dynamics, vibrato, etc., of the performance, and thereby actively control the performance strategies [10; 11; 12; 13].

Achieving an immersive perceptual fusion of musical instruments, and thereby yielding a blended auditory impression, is often the primary objective in ensemble sound. Therefore, the auditory blending of sound sources, defined as the perceptual fusion of two or more concurrent sounds where the constituent sound sources are no longer individually distinguishable [6; 14], is an integral aspect of the ensemble sound. Although the multimodal attributes and their complex interactions are important in the joint musical performance, by including the four stages of musical blending evolution, the formation and development of a 'blended ensemble sound' can be summarized in a simplified way as provided in Figure 1.2.

The process starts with the composer's formulation of the arrangement and orchestration of the musical composition. The composer's conception of the desired level of musical blending by choosing suitable instruments and musical elements including pitch range, dynamics, tempo, and articulation is involved in this stage. The seating arrangement of musical instruments in the performance space can be pre-decided by

the composer or conductor to meet specific musical or acoustical needs (as discussed in [15]), or by following conventional practices (e.g. American and German seating arrangement in orchestra). The understanding of the composer's intention by the conductor, subsequently by the musician, and its execution as a joint performance is the second stage. The blending development at this stage is referred to as 'source-level blending', where joint performance strategies and room acoustic feedback play a significant role. The next part involves the directivity-related attributes of musical instruments; depending on how the instruments are positioned and oriented on stage, the directivity characteristics of the instruments decide the spatial distribution of sound energy radiated from the ensemble, leading to changes in the perceived sound of the ensemble performance [16]. Transformation of the ensemble sound by room acoustic environment is the next part in the ensemble sound evolution, where the room acoustic reflections modify the perceived timbre, clarity, and spatial impression of the ensemble sound. As a result, the degree of musical blending impression also gets altered by the room acoustic reflections, at this stage. The final stage of ensemble sound formation encompasses the perception and interpretation of ensemble sound by the listeners. The realization of the blending occurs at this stage which varies based on the skills and background of the listener [17; 18].

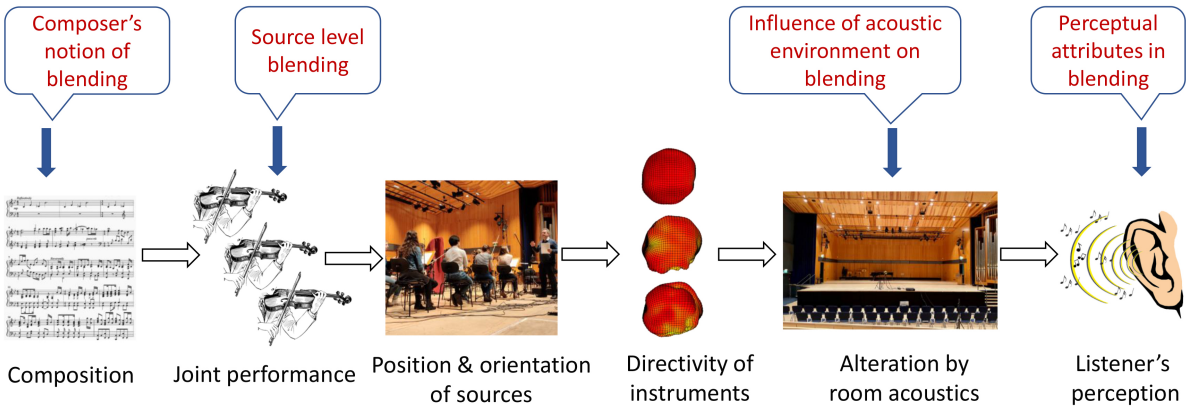


Figure 1.2: Schematic diagram of the formation and evolution of a blended ensemble sound

The formation of an ensemble sound in the virtual acoustic domain is fundamentally carried out through a process called 'Auralization'. Analogous to 'visualization', it is the process of (re)creating audible sound files of an acoustic scene, enabling people to listen to acoustics of a particular real/virtual space, by utilizing measured or simulated or synthesized data. Auralization involves several steps including sound generation and capture, modeling sound source, simulation of room acoustic environments, and sound field reproduction. In addition to its applications in Virtual Reality (VR) and

Extended Reality (XR) [19], the auralization technique has been widely used in room acoustic planning [20], environmental soundscape evaluation [21], architectural and archaeological acoustics [22; 23], and so on. An intrinsic and plausible representation of sound sources in the auralization of an ensemble performance involves two major steps: capturing the sound signals of constituent sources in an ensemble and modeling the directivities of the sources in simulations. Considering the significance of both visual and auditory feedback from musicians and from the room acoustics as discussed above, it is important to capture the source signals of individual instruments in the ensemble by preserving the musically intrinsic and perceptually authentic qualities of their joint performance. Moreover, by accounting for the influence of room acoustic reflections, estimating the perceptual relevance of complex and inherent directivity patterns of instruments in realistic room acoustic environments could contribute to optimizing the directivity filters of sound sources in auralization for a perceptually plausible rendering of an ensemble sound.

The details on the different stages and aspects of musical blending, directivity perception, and auralization are described in the following sections.

1.2.2 Blending of sound sources

Attaining a harmonious blend of sound sources is a fundamental sonic objective in collaborative musical performances. Therefore, the phenomenon of auditory blending between sound sources is established as a crucial aspect in music perception-related domains, spanning music composition and orchestration[6; 24], techniques for orchestrating recordings[25; 26], room acoustic adaptations [27; 28], development of joint performance strategies [5], and the evaluation of orchestral soundscapes in both real-world and virtual reality settings. The auditory blending phenomenon is fundamentally linked to the principles of fusion and segregation of concurrent auditory stream perception described in Auditory Scene Analysis (ASA) [29] (detailed in the following section). The degree of the perceived blending between sound sources can be assessed in two ways: either by using a rating scale to judge the blending impression or by evaluating the identifiability of constituent sound sources in the concurrent sound [6; 14; 24]. The four important stages involved in the evolution of blending and the major attributes involved in each stage are detailed below.

Blending as composer's notion:

Blend and contrast are two major musical aspects composers use skillfully to achieve different sonic goals required in musical compositions and thereby enhance the musical experience. Therefore the evolution of blending commences with the composer's conceptualization of the desired blending in a musical piece. Composers utilize the

concept of blending between instruments to obtain certain timbral outcomes by producing an augmented, emergent, or softened timbre from individual timbres of involved instruments [15]. Additionally, they use blending to attain specific sonic goals, such as enhancing expressiveness for emotional depth, ensuring balance between instruments, achieving coherence, and so on. It is achieved through careful decisions regarding the selection of musical instruments, and choices of the composition elements such as pitch range, dynamics, tempo, articulation, etc.

Musical blending is observed to be higher in unison performances, where two or more players perform in the same pitch or octaves, compared to non-unison arrangements [14; 24; 30]. The timbre characteristics of the instruments chosen for joint performance significantly influence the blending. Compositions involving lower pitch registers result in relatively darker timbral characteristics of the instruments, thereby improving the blending impression [14]. Moreover, the dynamics marking of the musical piece, representing the variation in loudness between the musical notes, also appears to influence the blending impression, where softer dynamics leading to a darker timbre are observed to enhance blending [15]. Furthermore, musical articulation-related aspects such as the excitation of the instrument by bowing or plucking the string, and the presence of temporal or spectral modulations such as vibrato are also observed to influence the blending [14; 30; 31]. The symbolic information of musical attributes extracted from musical scores, such as the onset synchronicity, pitch harmonicity, and parallelism in pitch and dynamics, are shown to provide cues on modeling the orchestral blend from musical scores [32]. While the selection of attributes influencing blending is important, the key part to achieving the intended blending lies in the musicians' flawless execution of the composer's vision. This crucial aspect of blending occurs during the joint performance stage, leading to the formation of blending at the source level.

Blending at source level

Source-level blending between instruments in a joint performance is a multifaceted process influenced by both composition-related attributes and musical performance-related parameters, which span spectral, temporal, and loudness domains. While musicians try to attain the composer's or conductor's vision of blending, their performance strategies are controlled by inter-musician coordination and room acoustic feedback. In such conditions, musicians trying to achieve a blended output will try to optimize attributes such as the timbre of individual instruments, temporal synchronicity, pitch similarity, coherence in dynamics, and so on [2; 3; 4; 5; 6]. These mentioned musically oriented attributes that influence the source level blending can be evaluated by analyzing corresponding acoustic parameters extracted from audio signals across spectral, temporal, and loudness domains.

The perception of blending between two sound sources can be assessed using acoustic parameters representing their spectral characteristics such as the composite spectral centroid, spectral envelope, the prominence and frequency relationship of formants, etc. A lower spectral centroid value, representing a darker timbre of the instrument sounds, correlates with an improvement in blending impression [14; 31]. Moreover, similarity in spectral envelope characteristics, coinciding formant locations, etc., also contribute to the improvement of blending impression [6; 24; 14; 33]. Since the difference in fundamental frequencies is observed to influence auditory stream segregation of concurrent sounds [34; 35], increasing pitch separation between the instruments in joint performance results in reduced blending [15]. When it comes to temporal domain, the onset synchronization of musical notes are very important to achieve a blended impression [14]. The attack time of onsets, influenced by musical articulation such as bowed or plucked excitation, also plays a significant role in blending. A slower attack is observed to result in better blending compared to an impulsive attack [31; 30]. Additionally, a high correlation in loudness between the concurrent sounds from instruments was also observed to influence the blending positively [14].

Attempts to statistically predict the source-level blending impression between different instrument combinations were previously investigated using linear correlation and regression of blend rating with individual acoustic parameters [14]. This method was able to account for 51% of the variance in blending ratings of unison performance using composite centroid, attack contrast, and loudness correlation [14]. In some later investigations, Multiple Linear Regression (MLR) was used to predict the blending perception in accordance with the variation in spectral characteristics such as the multi-parametric variance of the formants [33]. Due to high collinearity between the variables involved, a Partial Least Square Regression (PLSR) based model was proposed as an extension to the earlier studies to predict the blending rating on a diverse data set of audio samples that included different instrument combinations with unison and non-unison intervals, various pitch range and distinct excitation mechanisms [30]. These aforementioned studies utilized sound samples with sustained tones by limiting musical features such as loudness, dynamics, duration, vibrato, location and relative strength of formants, and so on. However, these mentioned musical features are mutually and stochastically entangled in realistic joint musical performance recordings. As a consequence, although the mentioned studies on isolated instrument tones give insights into the potential parameters and their influence on the blending in musical contexts, evaluation of the perception of blending on musically realistic 'ecological' sound samples remains unexplored.

Role of acoustics environment on blending

Considering the room acoustic environment as a Linear Time-Invariant (LTI) system, the way the system transforms the input signal (i.e., sound radiated from the sound source) to the output (i.e., sound received by the listener) can be analyzed using Room Impulse Response, which serves as the transfer function of the system. The Room Impulse Response (RIR) illustrates how a room responds to a Dirac impulse generated from a source by showing its transfer to the receiver as direct sound followed by a series of impulses as room reflections with decaying amplitude with time. The time domain representation of RIR demonstrates the strong early reflections, and late diffuse reverberation caused by the room, as well as offers cues on the absorptive and diffusive nature of the room. Based on the perceptual aspects, the RIR can be classified into three regions: the direct sound part (i.e., 0 - 5 milliseconds), the early reflections (5 - 50 or 80 milliseconds), and late reverberation (80 milliseconds - end of RIR). The direct sound from the instrument is crucial in sound source localization and distance estimation. While the early reflections contribute to the perception of clarity and source width impression, the late reverberation influences the perception of spaciousness and envelopment [36]. Starting from Sabine's reverberation time formula proposed in the 19th century, numerous room acoustic parameters developed over the last century addressing different objective and subjective features of room acoustic environments have been widely utilized in a standardized manner to characterize acoustic environments [37; 38]. A detailed description of these parameters including their equations and estimation procedures is described in the Appendix A. These parameters derived from RIRs, or a combination of them, have been shown to capture specific subjective sensations and attributes associated with room acoustics perception, thereby offering a comprehensive overview of the perceptual characteristics of the acoustic environment [39; 40; 41; 42].

The blending of sound sources has been considered to be an important subjective attribute of room acoustics and concert hall acoustics research [7; 43]. However, the influence of different room acoustic attributes that shape the perception of blending has not been thoroughly analyzed yet. Although no specific studies address the direct relationship between the room acoustic attributes and the simultaneous grouping principles of ASA, insights from studies on room acoustic perception offer valuable clues about room acoustic attributes from spectral, temporal, loudness, and spatial domains that could impact ASA and thereby alter the blending. Reverberation holds major importance as it significantly alters the perceived timbre of instrument sounds [12], influences the modulation depth of signals which in turn affects intelligibility [44], and weakens the listeners' ability to discern variations in fundamental frequencies and the spatial localization of sound sources [45], thereby presumably affecting the overall blending perception. Although not directly addressing blending perception, previous literature aligns with these observations, indicating that the reverberation enhances

the 'melting' of individual sound to form a closed overall sound [16], with some studies proposing the spatially enveloping late reverberation blends music naturally [46].

While small timing deviations in onsets/offsets in the order of 30-50 ms may remain as perceptually synchronous, the room reflections further improve temporal synchrony between concurrent sounds by smoothing transient envelopes [47]. Additionally, room acoustics display responsiveness to musical dynamics of stimuli by altering the perceived impressions of dynamic levels, apparent source width, and envelopment [48]. The spatial distribution of sound sources, including the positioning and orientation of constituent instruments within the ensemble (e.g. different orchestral seating arrangements), has been reported to influence the ensemble sound and the blending perception [16]. Accurate localization of constituent sources in the ensemble by the listener may adversely affect the simultaneous grouping principles of sources in ASA and thereby impact the blending. In that context, the room acoustic reflections have been demonstrated to impact the perception of sound source localization, with reflections from specific directions (e.g., reflection from the floor, ceiling, sidewalls, etc.) either enhancing or worsening the localization perception [49].

As an initial attempt to investigate the relationship between room acoustic attributes and instrument blending, recent research on the perception of blending in concert halls, utilizing binaural auralization of a string quartet, observed a significant correlation between blending ratings and parameters such as the reverberation, treble ratio, spatial envelopment, and sound strength [27]. However, it is unclear if room acoustic reflections always enhance blending perception, moreover, it is possible that the impact of the acoustic environment on the overall perception of blending can be different for samples with different degrees of source-level blending. Therefore, a detailed investigation is required to assess the individual contribution of source-level blend and room acoustics in the overall perception of the blend, as well as to understand the major room acoustic attributes that influence blending.

Auditory perception of musical blending

The perception of blending of sound sources is the result of the simultaneous grouping of concurrent auditory streams, an auditory phenomenon described in Auditory Scene Analysis (ASA) [29]. ASA refers to the process of extracting the auditory information involved in a complex mix of sound arriving at the listener's ears into distinct perceptually meaningful auditory objects by utilizing spectral, temporal, and spatial cues. This psychoacoustic phenomenon is the key concept involved in the identification, localization, and differentiation between various sound sources present in an 'auditory scene'. A well-familiar example of ASA in regular life is the cocktail party effect [50]. The grouping or segregating of spectrally and temporally overlapped auditory information from sound sources into separate mental representations known as auditory streams is performed by ASA [29]. The grouping can be done in two conditions; simultaneous grouping in the case of concurrent auditory streams, and sequential grouping in the

case of temporally evolving auditory streams. While simultaneous grouping leading to the fusion of sounds holds more importance in the blending of concurrent streams, the sequential grouping leading to the association between temporally evolving sounds is also relevant in musical contexts such as the judgment of rhythm and melody.

The grouping of auditory streams in ASA is fundamentally connected to the Gestalt principles in psychology in audio perception, notably the principle of 'common fate' [14; 51]. This principle suggests that the sounds that undergo similar kinds of variations, such as synchronous onsets, matching frequency or amplitude modulation, and parallel loudness variation, are grouped together and perceived as a part of a single auditory object [51]. Moreover, the auditory system also seeks other relevant cues such as timbral proximity, pitch similarity, closeness in spatial locations, etc., between the sounds involved, for the perceptual grouping or segregation of auditory streams (more details on the Gestalt principles of perceptual grouping can be found at [18; 52]). As described in the previous section, the room acoustic characteristics have been shown to impact these attributes. Therefore, the acoustic characteristics of the performance spaces are expected to play a significant role in the perception of blending.

While listening to the musical performance, listeners extract auditory information from the ensemble sound based on personal experience, preference, ability, and other factors. Previous research indicates that musicians are shown to have sensitivity in selectively attending to and analyzing the complex spectral and temporal features of sounds, as compared to non-musicians [17; 18]. Therefore, major perceptual evaluations carried out in this thesis work, including studies on blending perception and other aspects of ensemble sounds, are exclusively carried out among the trained participants including tonmeister students and musicians. By utilizing such a population of trained people, concordant and more reliable results are expected in the perceptual evaluations.

1.2.3 Directivity of musical instruments

The directivity characteristics of musical instruments, depicting the spatial distribution of energy radiated for each frequency, play a pivotal role in shaping the perceived sound field of instrument performance. By affecting the loudness, timbral characteristics, and spatial impression of the perceived instrument sound, the directivity characteristics of instruments holds a significant role across many fields such as music performance, instrument recording techniques, room acoustic simulation, and sound field reconstruction [16; 53; 54; 55; 56]. The directivity characteristics of instruments are frequency-dependent; in general, from an omnidirectional behavior at low frequencies, they transition to highly complex directional characteristics at high frequencies based on the properties of the instrument [16]. Additionally, the directivity is also shown to be influenced by the notes being played [57]; difference in playing style or fingering for performing a particular note would result in different directivity characteristics.

This thesis primarily deals with five different musical instruments with different directivity characteristics including trumpet, trombone, transverse flute, saxophone, and violin, and the directivity properties of those instruments are described here briefly. The directivity of the trumpet and trombone, the instruments from the brass family, is mainly influenced by the shape and size of their bell and bore [16; 58]. For trumpets, the bell of the instrument acts as the main radiation source, displaying an omnidirectional behavior up to around 500 Hz. As the frequency increases beyond this point, the instrument radiates mainly along the axis of the bell, showing rotational symmetry relative to the bell axis [16]. As the frequency increases, the side lobes decrease drastically in amplitude compared to the main radiating lobe along the bell axis, and thus the instrument becomes highly directive in the high-frequency part of the spectrum. Trombones exhibit directivity characteristics similar to trumpets but with a shift to lower frequencies due to the differences in geometry, particularly their larger flare size [16; 54].

The overall radiation characteristics of the flute can be represented as a dipole behavior by describing the energy emitted from the blowing hole and the first open tone hole (the far open end when all tone holes are closed while producing the lowest note) [16]. The dipole sources radiate nearly equal energy in and out of phase, depending on the order of harmonics. Unlike an omnidirectional behavior at low frequencies, this leads to both constructive and destructive interference, which results in strong directional features. When multiple-tone holes are open, they contribute individually to the overall radiation characteristics, particularly for the mid and high-frequency range (for high notes and partials of low notes) which makes the directivity pattern more complex [58]. In the case of very high frequencies, the far open end of the flute serves as the main radiating source [54]. The saxophone displays radiation properties akin to the woodwind instruments, with radiation from the bell opening complemented by radiation from the finger holes [59]. Consequently, unlike brass instruments which have consistent strong and weak radiation regions, saxophones exhibit complex directional characteristics by producing interference lobes. While individual tone holes contribute to radiation in the low-frequency range, for very high frequencies, particularly above the cut-off frequency specific to the tone hole lattice, radiation primarily emanates from the open bell [58].

Violins exhibit a rather complex directivity pattern, which is known to show rapid variations across frequencies, and whose behavior cannot be easily predicted except in the lowest frequency range. Unlike brass or woodwind instruments, violins lack a defined shape for directing the sound energy, which results in their intricate directional characteristics. The instrument's vibrating plates are primarily responsible for its directivity, with different points on the plates vibrating at varying amplitudes and phases. Additionally, the f-hole contributes to the radiation characteristics, particularly in the low-frequency range [16]. While the violin displays omnidirectional character-

istics up to approximately 600 Hz, it produces complex directional patterns for higher frequencies which are primarily radiated from the instrument's top plate. [16; 54].

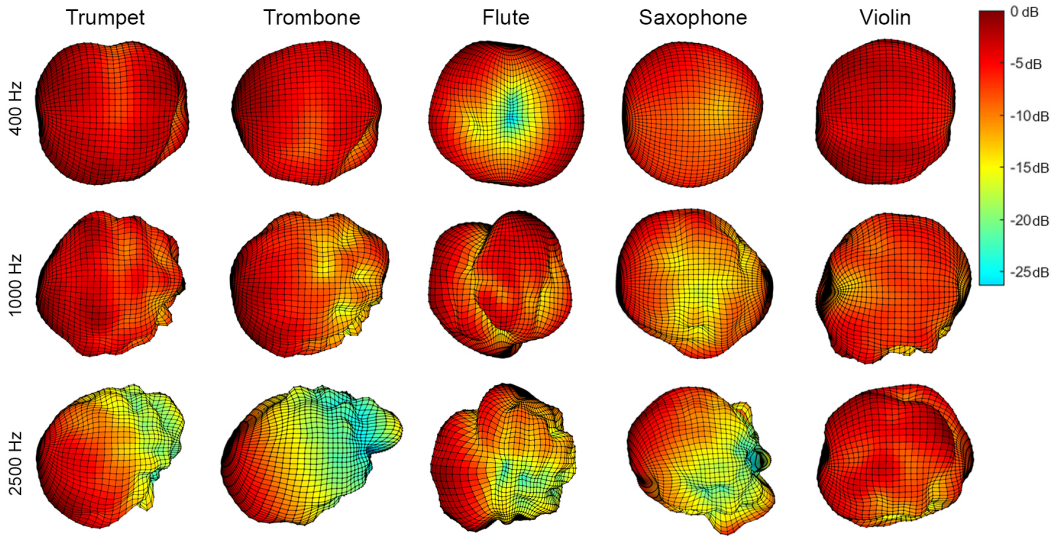


Figure 1.3: Directivities of musical instruments involved for low (400 Hz), mid (1000 Hz) high (2500 Hz) 1/3rd octave frequency bands, after [60; 61; 62; 63; 64]

Radiation patterns of musical instruments change rapidly in the temporal domain for dynamic musical samples. Furthermore, a small frequency shift within certain characteristic frequency ranges can produce significant changes in its directional patterns for specific instruments, such as violin [65]. Nevertheless, to provide an overview into the overall trend of directional characteristics, the directivity patterns of the five instruments involved in this thesis are depicted as 3D balloon plots in Figure 1.3 by averaging for 1/3rd octave bands centered at 400 Hz, 1000 Hz, and 2500 Hz, respectively (these plots are generated using the high-resolution directivity data published by spatial audio library of Brigham Young University directivity [60; 61; 62; 63; 64]). The complex directivity patterns generated across various frequencies within the specific frequency band are averaged and normalized to 0 dB as the maximum, and the plot is constrained to a side view of the 3D representation with instruments oriented towards the left side. While most instruments, with the exception of the flute, appear to exhibit omnidirectional behavior at lower frequencies, as mentioned above, they tend to display more complex radiation patterns at higher frequency bands. The highly directional characteristics of the trumpet and trombone having a strong beam along its axis, the lobes arising from the interference of multiple radiating points in saxophone and flute, and the complex directivity shapes of violins can be observed from these plots. This underscores the diversity in directional characteristics among the five sound sources examined in this study on source orientation perception.

Directivity perception in room acoustic environments

Owing to the crucial importance of directivity, extensive research has been carried out on modeling and measuring the directivity properties of individual instruments with a high degree of accuracy [16; 66; 67]. Moreover, the geometrical and physical construction aspects of the instruments that influence the directivity patterns have been extensively analyzed for a wide range of instruments [58]. Despite these extensive efforts, the perceptual aspects of directivity attributes of musical instruments in realistic conditions remain relatively less explored.

The significance of directivity analyzed using room acoustic simulations indicated that the difference between a specific and an averaged directivity filter of the sound source in a simulated acoustic environment is reflected in the resultant room acoustic parameter values [53]. Moreover, these directivity differences were noticed to be perceptually perceivable, particularly as differences in loudness and clarity impression. A study conducted on the perception of musical instruments modelled with omnidirectional, realistically directional, and extremely directional directivities using room acoustic simulations showed a significant difference between omnidirectional and extremely directional directivities in terms of the estimated room acoustic parameters and the perceptual impressions [68]. However, the differences between omnidirectional and realistically directional directivities were observed to be negligible, which could be due to the limited frequency-band directivity data utilized in the study. Recent studies have proposed that the listeners can distinguish the tone-dependent directivities from averaged directivity, especially in echoic conditions [69]. The variation in directivity caused by the movement of sound sources during the performance, which is an integral aspect in realistic performance conditions, is also found to be perceptually significant [57], thereby asertain the importance of directivity in sound source representation.

When it comes to real-world applications, room acousticians often use electroacoustic sources for the playback of musical instrument recordings to know the 'sounding of the room'. Advancing from this, loudspeaker orchestra, in which a wide range of instruments were represented as a combination of different electroacoustic sources in a sophisticated manner, was used for the perceptual evaluation of concert halls and acoustic measurements [70]. However, in these mentioned cases, the natural/realistic impression and the perceptual similarity of these electro-acoustic substitutions to the real instruments are not well-explored. As mentioned above, the dynamic directivity of instruments can get complex and drastically changing for specific frequency ranges or specific notes. Therefore, these simplified approximations of directivity require investigation to assess their perceptual relevance, particularly in the context of a dynamic musical performances in in-situ acoustic condition. Such explorations in in-situ conditions could provide insights relevant for instrument recording techniques, etc., and also contribute to optimizing the modeling of sound source directivity according to

the limits of human auditory perception that is relevant for virtual acoustics-related applications.

Another important aspect of directivity relevant in real and virtual acoustic environments is the perception of sound source orientation. The orientation of sound sources is noted to have a key role in musical performance. For instance, the different seating arrangements of instruments on stage with different orientations are used to achieve certain acoustic goals [16]. The factors influencing source orientation perception are detailed in the coming section.

Source orientation perception

The perception of the source orientation is one of the important perceptual attributes that is significantly influenced by the directivity characteristics of the sound source [71]. Depending on the sound source directivity and room acoustic properties, different source orientations around its acoustic center would create differences in the energy and spectral content of both direct sound and room acoustic reflections, especially first and second early reflections, in a specific listener location [71]. This effect is most notable in high frequencies due to the complex directivity characteristics arising from the instruments. These variations could lead to alterations in the loudness, timbre, intelligibility, and spatial impression of the perceived sound, which have significant implications in fields such as music performance and perception, communication acoustics, and virtual acoustic applications.

While the auditory perception of sound source localization and distance perceptions has been extensively analyzed in the past decades [72; 73; 74], most studies on source localization and distance perception have typically focused on the speaker facing the receiver condition and did not explore the role of other 'facing angles' on their perception. Despite its significant importance, the perception of the source orientation and the factors influencing it in in-situ conditions have remained largely unexplored. While the majority of research conducted on this topic was focused on orientation perception and its influential factors in anechoic or semi-anechoic conditions [75; 76], only a limited number of studies considered echoic environments to examine the influence of room acoustic reflections on orientation perception [77; 78; 79; 71].

The accuracy of orientation prediction is observed to be higher when the source is oriented toward the receiver [80; 76; 79], and this trend appears to hold true in both echoic and anechoic conditions. The acoustic attributes influencing the orientation perception in anechoic conditions are observed to be from different domains for lateral and medial orientations; lateral orientations depend on spatial cues, whereas medial orientations rely on monaural cues due to the absence of spatial cues. The Interaural Level Difference (ILD), the difference in Sound Pressure Level between the two ears, is observed to be a key binaural cue for lateral (i.e., left and right) direction judgment for lateral orientations in anechoic conditions [75; 76]. Since ILDs are absent in medial

(front, back) orientations [75], monaural cues such as overall level difference, access to high-frequency sound, and spectral tilt variation in high-frequency regions are observed to provide cues for medial orientation judgments [76; 79; 81].

When it comes to source orientation perception in room acoustic environments, certain studies propose that orientation perception is generally more challenging in echoic environments compared to anechoic conditions [77]. However, there are contradictory findings suggesting that listeners also perform well in room acoustic environments, with certain specific directions having high prediction accuracy [78; 79]. Room acoustic reflections have been demonstrated to impact the perception of sound source localization, with reflections from specific directions (e.g., reflection from the floor, ceiling, etc.) either enhancing or impairing it [49]. Likewise, certain room acoustic reflections are expected to impact the perception of source orientation [82]. A recent study conducted on the ability to perceive the orientation of a human speaker in simplified simulated room acoustic environments suggests the importance of strong early reflections in improving orientation perception [71]. This study also demonstrates that the presence of first-order reflections carrying directivity information strongly supports the source orientation perception while the higher-order reflections with directivity information might not be necessary [71]. This aligns well with the observations from previous studies where the presence of first-order reflections such as strong side-wall and back-wall reflections resulted in easy identifications in those specific orientations [78; 79]. ILDs generated from strong early reflections are observed to be a significant parameter in predicting orientation in the left-right directions in room acoustic environments, while less pronounced ILDs are also observed to hinder left-right orientation prediction [71]. Moreover, the prediction accuracies are also observed to decrease with a decrease in Direct-to-Reverberant Ratio (DRR) and change in spectral coloration from low-pass characteristics of source directivity in room acoustic environments [79; 71].

Studies on source orientation conducted so far have been primarily concentrated in the area of communication acoustics. Consequently, they were constrained by the use of human speakers [77; 81; 76; 82; 71] or loudspeakers [75; 79] as the sound sources, which possess relatively simplified directivity patterns, for producing voice signals or broadband noise. While the directivity of the sound source is observed to be a factor influencing orientation perception results [79], it still demands a comprehensive investigation, especially for musical instruments. Furthermore, the role of room acoustic attributes in orientation perception in in-situ conditions needs to be analyzed, by incorporating real room acoustic environments with diverse acoustic characteristics rather than simplified simulated environments utilized in earlier studies [71].

1.2.4 Auralization of ensemble sound

"Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space", defines M. Kleiner, who coined the term Auralization [83]. This process can be performed using simulated, measured, or synthesized numerical data [84].

As previously noted, the process of auralization involves different stages. It starts with capturing the source signal, which can be live for real-time auralization, or recorded. The second stage involves convolving it with the Room Impulse Response (RIR), which encapsulates the transfer function of sound from the instrument to the receiver. The RIRs can be of two kind; Spatial Room Impulse Responses (SRIRs), which capture the spatial sound field using an array of microphones (e.g. ambisonics microphone), and Binaural Room Impulse Responses (BRIRs), which capture the spatial sound field using a binaural head. These SRIRs or BRIRs represent the transfer function of an impulsive sound generated by the source to the receiver, which can be measured from in-situ conditions, simulated from computer models, or synthesized from existing data. Modeling the source and receiver characteristics, such as source directivity and Head Related Transfer Function (HRTF), is important at this stage. The in-situ measurement of these impulse responses is conducted by exciting the room acoustic environment with a Dirac impulse using an electroacoustic source and capturing the resultant sound field at the receiver location using a microphone array or a binaural head. While it can capture complex geometrical and acoustical properties of the room, the electroacoustic source exciting the room with a directivity that is close to the real instrument is the crucial part here. While simulated environments allow modeling the sound source with a directivity close to the real instruments, accurately capturing the physical and acoustic aspects of room acoustics is a challenge in acoustic simulations. The final stage of auralization involves the reproduction of sound using loudspeaker arrays for the spatial sound field, or using headphones for the binaural sound field. While spatial sound field reproduction advances in accurately reproducing sound fields with an increasing number of channels, it is often restricted to laboratory conditions due to the complexity of the hardware setup. On the other hand, the binaural sound field reproduction using headphones offers advantages for real-world applications. The binaural sound field reproduction can be advanced further by implementing real-time rendering to allow head movements, and customization of HRTFs for better accuracy.

Figure 1.4 represents the flow diagram of the different stages involved in the auralization of an ensemble sound in virtual acoustic environments. Additionally, when it comes to real-time auralization of ensemble performance, incorporating interactions between musicians and feedback from the simulated acoustic environment is essential for achieving real-time performance and perception of the virtual orchestra. The details of each individual aspect involved here are described in the coming sections.

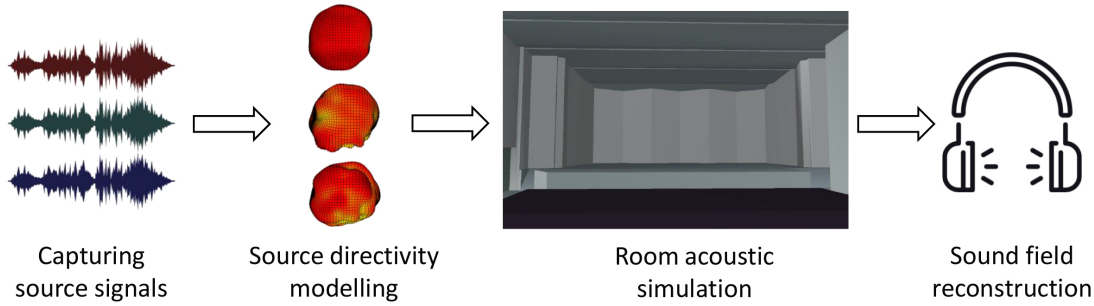


Figure 1.4: Flow diagram of the factors involved in the auralization of a musical ensemble in virtual acoustic environment.

Capturing of source signals

When it comes to capturing acoustically 'clean' source signals for the auralization of individual sources, recording a musical instrument in an anechoic environment with/without directivity is a trivial and direct solution. However, in the case of auralization of joint musical performance, accurate capturing of the individual source signal is one of the most important but less explored aspects. Three notable ways followed to capture the signals of individual sound sources in a joint musical performance are; (1) recording each instrument in the orchestra individually in an anechoic chamber by providing external visual/aural cues[85; 86; 87], (2) keeping the whole orchestra instruments in the anechoic chamber and recording them simultaneously by close-miking techniques [88; 89], and (3) recording individual sound sources using close-miking techniques during a joint musical performance in a regular musical performance space [19].

According to previous studies, factors such as the visual and audio feedback from the musicians[1; 2], joint performance strategies[5; 4], and the room acoustic feedback to the musicians[10; 13] are shown to have a significant impact on the individual performances of instruments in ensemble and thereby influence the resulting ensemble sound. Although the research carried out using the first two aforementioned recording techniques had secondary means to compensate for the performance attributes, their recorded outputs might not be an intrinsic and natural representation of the orchestral/ensemble sound. However, the third method includes the natural and intrisic attributes of joint performance and acoustic feedback in their recordings. Moreover, it is particularly advantageous when it comes to real-time auralization of an ensemble performance in in-situ conditions. The spectral colouration that can occur due to the spatial positioning and frequency response of the microphones used can influence the plausibility and timbre of the auralized output signal, but it's an unavoidable error that is present in all three discussed methods to different extends. While the close-mic

recordings are expected to have a high direct sound contribution compared to the reverberation, the influence of cross-talk between sound sources due to spatial proximity, noise from the musician and instrument, and the strong room acoustic feedback in reverberant environments could degrade their recording quality. Therefore, it is required to analyze the quality of clip-on microphone recordings in delivering a perceptually convincing sound field of an ensemble by reconstructing an orchestral performance.

Modelling of directivity of sound source

Figure 1.5: A high resolution directivity measurement setup of a trumpet with mannequin, employing a 3D turn table, and offering a 5° spatial resolution resulting 2,522 unique measurement points (from [67]).

Given that physically accurate sound field reconstruction necessitates detailed directivity information of sound sources, substantial efforts have been invested toward capturing the high spatial and spectral resolution directivity of musical instruments and human voice for virtual acoustic applications [66; 67]. Figure 1.5 presents high resolution directivity measurement setup of a trumpet with mannequin, by utilizing a 3D turn-table, to achieve 5° spatial resolution directivity data which results in 2,522 individual measurement points (presented at [67]). Despite the considerable efforts involved in these high-resolution directivity measurements, the perceptual significance of the high spatial and spectral resolution of the directivity of sound sources remains unclear. While representing instruments with tone-dependent directivity filters with high spectral and spatial resolution can result in accurate reproduction of sound fields, it is constrained in practical conditions due to the heavy computational efforts. Therefore, to have a balance between the perceptual relevance and practical applicability,

the 1/3rd octave band frequency-averaged directivity data is being commonly used in room acoustic simulations. Regarding spatial resolution, in certain GA-based simulations such as Odeon, the spatial resolution of directivity is possible until 5° angular resolution. However, it is not clear whether representation of instrument directivities with such a spectral and spatial resolution of could produce a perceptually plausible auralization of a sound source. Therefore, it is essential to evaluate the perceptual thresholds of spectral and spatial resolutions of directivity data after which no perceivable changes in the auralization can be detected. This would enable computationally efficient modeling of sound sources for a perceptually plausible sound field synthesis in comparison to the ones with high-resolution directivity data.

Recent research have explored the perceptual threshold of spatial resolution of directivity data by incorporating controlled variation in the spatial resolution of directivity shapes through truncation of Spherical harmonics (SH) orders [90; 91]. This was primarily conducted for voice directivity using voice samples and broadband noise, where a perceptual threshold of SH order between 3-4, and 8.4 are detected. Future studies should extend on wide range of musical instruments with complex directivity characteristics, and also the role of factors such as room acoustic attributes should also be explored. Such perceptual threshold related studies are particularly important when it comes to the auralization of ensemble performances in virtual reality applications. As multiple sound sources need to be rendered simultaneously in real time, knowing the perceptual threshold would significantly impact the computational efforts.

Room acoustic simulation

The origin of computer-based room acoustic simulation was over six decades ago [92], but it became widely utilized in general practice in the 1990s with the advancement of computer technology [93]. Modern room acoustic simulations offer various features such as visualization of sound propagation in 3D space, estimation of RIRs for particular source-receiver positions, assessment of room acoustic parameters, and so on. This helps to modify the room geometries and materials applied to the boundary surfaces to achieve certain acoustic characteristics tailored to specific sonic goals. By facilitating customization and personalization of room acoustic environment design through saving cost and time, the room acoustic simulation has proven to be a key tool in the design and construction of rooms according to regulatory standards and auditory perception-oriented requirements. Consequently, it has been utilized in various areas such as room acoustic planning [20], architectural and archaeological acoustics [22; 23; 94], music performance research [11], Virtual Reality (VR) and Extended Reality (XR) [19], and so on.

The room acoustic simulations can be classified into two categories based on their approaches, namely wave-based simulations and Geometrical Acoustics (GA) based simulations. Wave-based room acoustic simulations model the sound propagation

as a wave phenomenon and solve it using wave equations. They take into account the complicated wave-related attributes such as diffraction and interference, that are especially important in low frequencies. Finite-Element Method (FEM), Boundary-Element Method (BEM), and Finite-difference Time-Domain (FDTD) are some of the common techniques utilized in wave-based room acoustic simulations (more details can be found at [84]). While this simulation method can accurately model the complex room geometries and get more accurate results across the different frequency bands, it is computationally demanding and relatively difficult to implement. On the other hand, Geometrical Acoustics (GA) based simulations neglect the wave properties of sound and treat it as rays. These simulations follow the principles of optics to model the propagation, reflection, and absorption of rays. As a result, the GA-based simulations are valid in mid to high-frequency ranges and are suitable for conditions where the wavelength of sound is much smaller than the surface and overall dimension of the room, which is true in most cases [93; 84].

Whereas classical GA-based simulations can not account for the wave phenomena, which result in a higher error rate in low-frequency bands, the state-of-the-art GA simulations have incorporated methods to introduce wave phenomena like diffraction to some extent [93; 95; 96], thereby reducing the error in the low-frequency range. Although having limitations in the accurate reproduction of real environments [97], they have been shown to provide a perceptually plausible recreation of room acoustic environments. Moreover, despite having limitations in handling complex low-frequency wave phenomena, the GA-based modeling technique offers advantages for controlled auditory experiments, including creating physically invalid imaginary rooms, estimating numerous room acoustic parameters accurately that can be laborious in real-life situations, and enabling flexibility for controlled adjustments of geometrical and acoustic attributes without background noise and distortion with fast and computationally efficient performance. As a result, they are widely utilized in regular practice as well as in numerous auditory perception-related investigations for simulating acoustic environments [98; 68; 99; 100].

Two major methods involved in GA-based simulations are the Ray-Tracing method and the Image-Source method. In ray tracing, the sound is modeled as a set of rays from the source, with each ray carrying a certain energy [101]. These rays interact with the boundaries (walls and other surfaces), lose energy according to the wall properties, and finally get collected at a particular receiver location using a surface or a volume detector. It is a stochastic simulation technique based on Monte Carlo methods, and therefore the results will have certain fluctuations based on the number of rays emitted. On the other hand, originating from concepts in electrostatics, the image source method is a deterministic model that ideally works with specular reflections. In this method, a ray from the source hitted and reflected from a wall can be considered as originating from an 'image source', a mirror image of the actual source [102]. This image source solu-

tion can be recursively applied for each reflection of a ray until it reaches the receiver, which is modeled as a point detector in this method. While the image source method accurately estimates sound reflections, particularly specular reflections, it struggles to include complex geometries and diffuse reflections. Conversely, ray tracing can accommodate complex geometries and scattering reflections, but an accurate estimation of a particular reflection requires a high number of rays that results in higher computational effort. Combining their advantages and disadvantages, a hybrid model as a combination of these two methods, the image source method for the accurate estimation of early reflections and the ray tracing method for the estimation of diffuse reverberation for complex geometries, has been widely utilized for an improved and optimized modeling in GA-based simulations [84].

Considering the advantages of recreating room acoustic environments and introducing controlled variations in the acoustics of simulated environments, GA-based hybrid simulations are utilized in certain investigations involved in this thesis for the creation and controlled variation of virtual room acoustic environments. Two major simulation software utilized in this thesis are ODEON version 17, a commercially used room acoustic simulation software [96], and RAVEN (Room Acoustics for Virtual Environments), a simulation environment developed for academic purposes [95; 103]. Both ODEON and RAVEN incorporate a hybrid approach that combines the image source method for early reflections and the ray tracing method for late reverberations. Moreover, both of them have incorporated diffraction effects to an extent to achieve physically valid results. Furthermore, by including features such as frequency-dependent absorption and scattering properties of the boundaries, as well as unique directivity characteristics of both the source and receiver, they represent state-of-the-art platforms in their field.

Sound field synthesis

Integrating the techniques of signal processing, electro-acoustics, and psychoacoustics, the Sound field synthesis aims to create or reproduce a spatial sound field through audio playback systems. In contrast to the basic sound field production techniques possessing spatial cues such as stereo panning and commercial surround sound systems that are utilized for aesthetic or entertainment purposes such as movies or music production, sound field synthesis in auralization aims to achieve a physics-based physically more accurate creation of sound fields. This is vital for preserving a precise directionality and timbre of direct sound and early reflections, that are highly relevant for auditory tasks such as localization and characterization of sound sources and auditory scenes. As a result, the spatial sound field synthesis of auralization achieves a relatively higher degree of realism that is essential for VR-related tasks. The spatial sound field synthesis can be divided into two major classes that are loudspeaker-based reproduction and headphone-based reproduction. Some of the well-known tech-

niques in loudspeaker-based spatial sound field synthesis include Wave field synthesis (WFS), Vector-based amplitude panning (VBAP), and Ambisonics, while headphone-based spatial sound field synthesis generally refers to binaural reproduction.

VBAP utilizes amplitude panning across a set of spatially distributed loudspeakers to place a virtual sound source at a desired position. This is achieved by estimating a combination of loudspeakers, typically triplets of loudspeakers, that approximates the source direction and applying the source signal to these speakers with appropriate amplitudes [104]. The Ambisonics technique encodes the sound field as a linear combination of Spherical Harmonics in a specific Ambisonics format, typically in B-format. The synthesis of the sound field is carried out by decoding the signal according to the given loudspeaker layout, during which the amplitude and phase of individual loudspeakers are controlled to reconstruct the encoded sound field [105]. Higher-Order Ambisonics (HOA) utilizes the same principle to recreate the sound field with more detailing in directionality and higher spatial resolution resulting in a higher degree of realism. However, this method requires a higher number of microphones and loudspeakers to capture and reconstruct the sound field. Based on the Huygens-Fresnel principle, the WFS technique synthesizes the spatial sound field by reconstructing the wavefront radiated from the sound source placed at a virtual source position [106]. This is carried out by precisely controlling the magnitude and phase of an array of closely spaced loudspeakers. Although this method can theoretically recreate complex wavefronts, the sound field synthesis using WFS is mostly constrained to the horizontal plane due to practical difficulties.

Binaural sound field reproduction can be achieved either by convolving the source signals with the BRIRs or by performing binaural recording of a sound source in the acoustic environment and reproducing it using headphones. While the multichannel loudspeaker-based sound field reproduction aims to create a spatial sound field impression for a particular listening point (in VBAP and ambisonics) or a particular listening area (WFS), the headphone-based binaural reproduction only considers providing the appropriate signals at the two channels for the two ears. Given its straightforwardness and minimal hardware requirement, it becomes more important in real-world conditions such as emerging VR and XR applications along with head-mounted displays. The binaural sound field capture using the microphones placed at the ear canal of the dummy head is intended to capture interaural cues such ILD, and Interaural Time Difference (ITD), etc. This enables binaural reproduction to deliver these cues directly to the listener's ears, which helps in perceiving a spatial auditory scene [74]. The spectral content of sound arriving from certain directions is filtered by the pinna (outer ear) structure, while the anthropometry of the head, shoulders, and torso, are also observed to contribute to the filtering of the sound signals. Together, these spectral and spatial cues facilitate auditory perception-related tasks such as source localization in both horizontal and vertical planes, as well as distance estimation.

The way outer ear alters the sound signals of different frequencies from various spatial orientations before they reach the inner ear is referred as the Head-Related Transfer Function (HRTF). Since the physical geometries of head and ear can vary between person to person depending on their shape of pinna, size and geometry of head and shoulders, etc., each person is expected to have unique HRTF, resulting in individual ways of perceiving the auditory scenes. Therefore, using a typical HRTF of a dummy binaural head may not necessarily match every individual, which would result in issues such as lack of externalization, and front-back confusions [107; 108; 109]. A solution to this is to use personalized HRTFs for binaural rendering. Apart from conventional methods to capture HRTFs in anechoic environments, predicting the HRTFs from the image data of the outer ear is currently developing research area [110; 111], which is expected to take the binaural rendering forward. While head rotation is an inherent attribute in loudspeaker based sound field reproduction, by incorporating a head tracker and a real-time binaural rendering technique, head rotation and head movements can be possible in this technique as well.

1.3 Scope of the thesis

While previous studies addressed blending from two distinct directions – one as a music-perception problem at the instrument level without an acoustic environment, and the other as a subjective attribute in the perceptual evaluation of acoustic environments – a unified approach that integrates these aspects has not yet been introduced. Previous investigations on blending between instruments had limitations in utilizing musically realistic audio stimuli, rather they were restricted to musical notes or chords. Moreover, the distinct contributions made by source-level blending and the room acoustic environment, in the overall blending, as well as the major room acoustics attributes involved in it have not been thoroughly investigated yet.

When it comes to instrument directivity, while numerous instruments are observed to show rapid variations in directivity patterns across frequencies that are dynamically varying with musical signals, their perceptual relevance in in-situ performance environments remains underexplored. Furthermore, while localization of sound sources is well explored, the investigation of directivity-related attributes such as source orientation perception is limited to sources with simplified directivity patterns by overlooking the influence of room acoustics in it.

Exploring the quality requirements of source signals for ensemble performances and evaluating the perceptual relevance of directivity representation of sources in the simulation are important aspects related to the sound source representation in the auralization of joint performances. Although this thesis does not directly focus on the advancement of virtual reality acoustics, investigations of the aforementioned aspects are expected to contribute to the advancement of a perceptually plausible recreation of an authentic and intrinsic musical ensemble sound.

Since most of the aforementioned problems are complex and multidimensional, analyzing all the involved aspects in each problem exceeds the scope of a thesis. Nevertheless, as an initial step towards evaluating the ensemble sound in a realistic performance context, certain aspects of these individual problems are investigated in several stages. The contents involved in the thesis can be grouped into three modules.

1. Investigations on musical performance-based representation of sound sources: this includes an initial exploration of ensemble sound and musical blending using live ensemble performance, assessment of source level blending between instruments in joint performance for musically realistic sound samples, and evaluation of the quality of close-mic recordings of instrument in joint performance for auralization of ensemble sound.

2. Exploration of directivity perception-related aspects: this addresses the sound source orientation perception for diverse musical instruments and also analyses perceptual differences caused by differences in source directivity characteristics, by utilizing the performance of individual musical instruments in diverse room acoustic conditions. Based on the results, the perceptual relevance of spatial resolution of directivity filter in simulations is analyzed for ensemble performance.

3. Analysis of the role of room acoustics on ensemble sound; this final module analyzes the role of room acoustic environments in shaping musical blending by estimating the major attributes involved and thereby presenting its role in ensemble sound.

These aspects are investigated in this thesis by utilizing in-situ recordings of individual and joint musical performances of different instruments in diverse acoustic environments. Additionally, auralization techniques including GA-based room acoustic simulation and binaural rendering are also utilized for experimenting in controlled room acoustic environments. Combining the acoustic and perceptual attributes of sound sources in the ensemble sound formation and auralization, Figure 1.6 presents a schematic diagram depicting the major topics covered in the thesis and highlighting the structure of the research problems addressed, thereby providing an overall perspective of the thesis framework. Collectively, these studies are expected to integrate and thereby improve the understanding of perception-based acoustic representation of musical instruments in joint performances.

1.4 Structure of the thesis

Chapter 2 details the performance and recording of a string ensemble in various acoustic settings, accompanied by an in-situ live listening test. This listening test explores the overall blending in a broader perspective and offers initial insights about the definition of 'ensemble sound'. The recordings presented in this chapter serve as material for some of the following investigations involved in the thesis.

Chapter 3 introduces a computational modeling approach to classify musically realistic sound samples according to their source-level blending impression. Combining the methods of Machine Learning (ML) and Music Information Retrieval (MIR), this modeling approach incorporates 'ecological' score-independent sound samples without requiring access to individual source recordings, thereby contributing to the holistic modeling of source-level blending.

Chapter 4 investigates the perceptual quality of utilizing clip-on microphone recordings for the auralization of an ensemble performance with varying numbers of instruments. This is accomplished by auralizing a joint performance using in-situ measurements and room acoustic simulations and comparing them with a binaural recording of an actual musical performance.

Chapter 5 investigates the perception of sound source orientation by analyzing the role of source directivity and room acoustic attributes in it. Moreover the study explores the potential acoustic parameters influencing orientation perception in in-situ conditions.

Chapter 6 examines the perception of dynamic directivity of variety of musical instruments in in-situ conditions by comparing the binaural recordings of real instrument performances against those generated by two electroacoustic sources (omnidirectional source and studio monitor).

Chapter 7 explores the perceptual relevance of high-spatial resolution directivity data of instruments in auralization of ensemble performance with different number of sources. This is carried out for sources with distinct directivity characteristics (violin and trumpet), in echoic and anechoic conditions.

Chapter 8 presents a statistical modeling approach to evaluate the perceived overall blending between instruments in joint performance by evaluating the contribution of source-level blending and its alteration brought by the room acoustics. The chapter underscores the intricate relationship between room acoustic attributes with the different degrees of source-level blending, and also demonstrates the major factors influencing the overall blending.

Finally, **Chapter 9** presents the overall findings and conclusions derived from these investigations, and also discusses the potential future work directions.

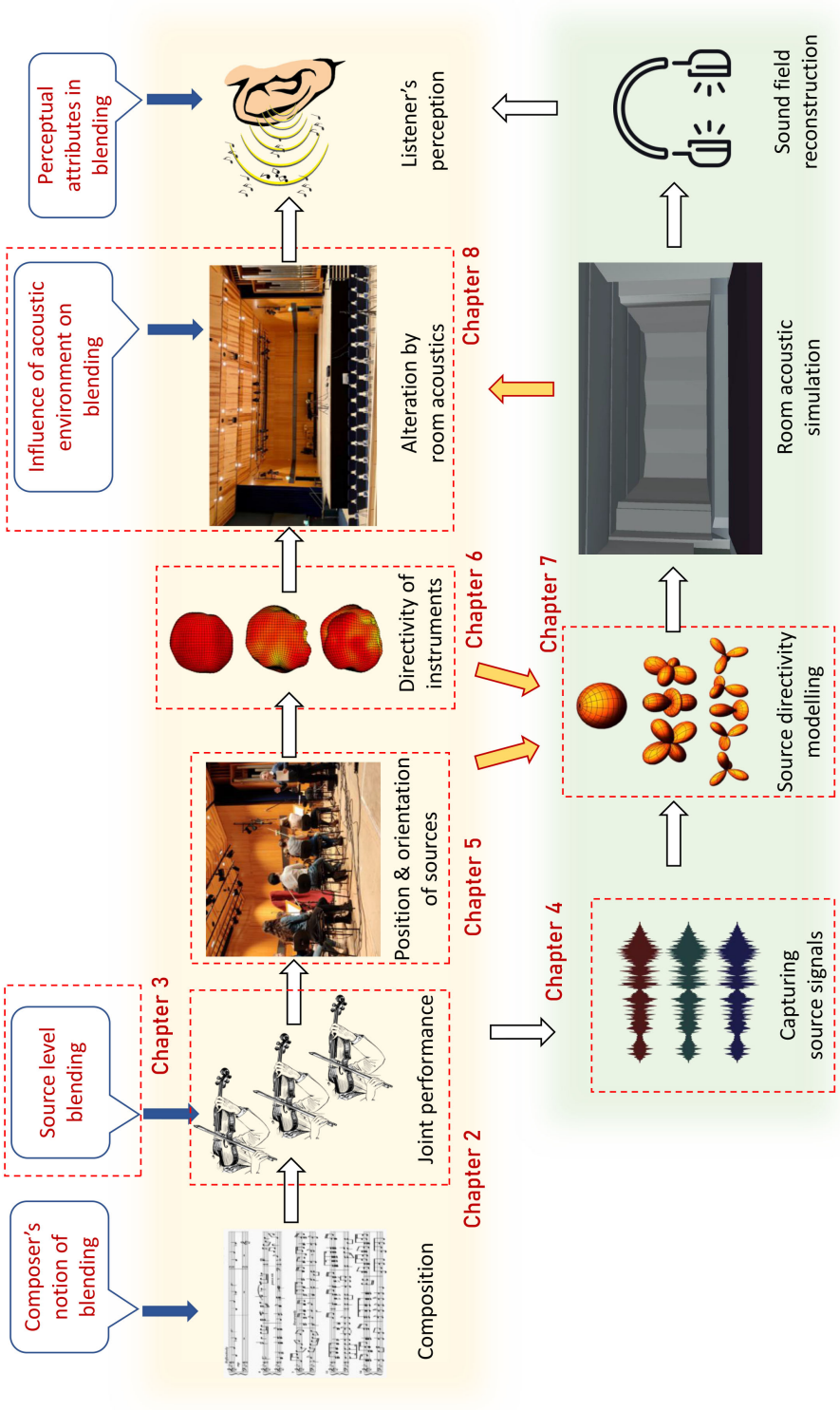


Figure 1.6: A schematic diagram depicting the major topics covered in the thesis.

Chapter 2

Exploring Ensemble sound: ensemble recording and live listening test

"How many instruments are necessary to generate an ensemble sound?" is a fundamental question that is crucial for the perceptually relevant acoustic representation of sound sources within the ensemble. Although this is influenced by the musical context, genre of music, and the expected sonic characteristics, apart from the loudness-related aspects, the minimum number of instruments needed to achieve the 'richness' in the ensemble sound is still subject to research. The impression of blending, another major attribute of ensemble sound, can be analyzed from macroscopic and microscopic perspectives. The macroscopic perspective considers the overall impact of blending by a musical performance on the listener, while the microscopic perspective focuses on the detailed analysis of minute-level variations in specific regions of sound samples. Utilizing a live performance of a violin ensemble, this chapter presents a pilot study that explores the blending perception from a macroscopic perspective. This was carried out by assessing the ability of listeners to predict the number of sources in a musically realistic ensemble performance in variable acoustic conditions as well as evaluating the ensemble sound impression. Furthermore, this chapter also outlines the methods used for simultaneous recording of the ensemble performance to gather materials for studies related to ensemble sound and blending perception presented in this thesis. A part of the content presented in this chapter is reproduced from the following research article with the permission of the Deutsche Gesellschaft für Akustik e.V:

J. Thilakan, Malte Kob, "Evaluation of subjective impression of instrument blending in a string ensemble", Fortschritte der Akustik- DAGA, Vienna, (2021).

2.1 Materials and methods

A string ensemble consisting of 9 violins was performed at Detmold concert house as a part of a live listening test as well as an ensemble recording process. This ensemble performance was conducted as the third ODESSA (Orchestral Distribution Effects in Sound, Space, and Acoustics) project [112]. Previous studies have shown that factors such as the acoustic characteristics of the performance space, the number of constituent sound sources and their distribution and orientation on stage, etc., significantly impact the resultant sound field of an ensemble, thereby affecting the perception of blending [16; 26; 27]. Therefore, the variations in the room acoustic characteristics of the concert house, the number of violins in the ensemble, and the spatial distribution of both the musicians and the listeners are diversified in this investigation to gain a preliminary understanding of how these factors influence the ensemble sound and blending perception in a live musical performance context. Additionally, the performances of the ensemble were simultaneously recorded using various methods such as individual clip-on microphones, stereo pair microphones, binaural heads, and so on. The procedure of the listening test and the methods of ensemble performance recording in this investigation are explained in the following sections.

2.1.1 Performance of listening test

The objective of the listening test was to investigate the blending perception from a macroscopic perspective within a musically realistic real-world setting while also exploring the perception of 'ensemble sound'. Ensemble sound remains an underexplored topic that may not be directly related to blending perception and can be researched as an independent aspect of ensemble perception, as performed in earlier studies [26]. While blending is likely a contributing factor to ensemble sound, it can be hypothesized that a performance may convey a convincing ensemble sound impression even when the degree of blending is not high. As previously discussed in section 1.2.2, apart from assessing the blending on a rating scale, the identifiability of the constituent sound sources involved in the concurrent sound of joint performance can serve as an indicator of perceived blending impression, where high identifiability of constituent sources corresponds to a poorer blending impression. Therefore, this test evaluates both the identifiability of individual sound sources in an ensemble performance in diverse acoustic conditions and the impression of 'ensemble sound', using live ensemble performances with different numbers of violins in diverse acoustic environments.

A group of 16 participants (7 female, 9 male), with 14 of them having musical backgrounds, participated in the listening test. Unlike the upcoming perceptual evaluations presented in this thesis, not all of the test participants in this pilot study were musically and technically ear-trained. The overall goal of the experiment and the procedure of

the listening test were explained to them at the beginning of the test to make the test participants aware of the objective of this investigation. The musicians in the violin ensemble were students of HfM Detmold. In the beginning of the test, the musicians were asked to perform two musical pieces utilized in the earlier ODESSA recordings (Symphony No. 6 in B minor, I. Adagio – Allegro non troppo by Tchaikovsky, and Sonata No. 12 in A flat major, II. Scherzo by Beethoven) sequentially in unison with a group of violins starting from 1 to 9 with an increment of one violin. As a result, musicians got familiar with the piece and adapted to the joint performance with others whereas the test participants were able to develop an idea of the overall sounding impression with the increase in the number of violins. Although the acoustic conditions and number of instruments changed during the test, the musicians were requested to perform with the highest possible impression of blending in all the conditions.

The formal test started following the training session, and it was performed in two parts: in the first part, the listeners were advised to sit in the prescribed seats in the predefined locations of the Concert House, labelled as A, B, C, D, and E as shown in Figure 2.1, and in the second part, the listeners were free to choose seats according to their individual preferences. Out of the 16 participants, each of the groups A, B, and C had four participants, and D and E were combined into one consisting of four participants collectively. Listener locations A and B were within the critical distance from the source, where there is a strong impression of direct sound and the sound pressure decreases with distance. In contrast, locations C, D, and E were outside the critical distance, where the reverberation of the room dominates, and the sound pressure level remains nearly uniform.

In each part of the listening test, four different acoustic variations were presented. This included changing the seating arrangement on the stage and altering the acoustic characteristics of the concert house using the room acoustic enhancement system installed in the concert house. Firstly, two instrument seating arrangements in the natural room acoustic condition of the Detmold concert house (RT_{60} = 1.6 seconds) were included in which violins are mainly radiated toward the listeners (analogous to the German way of string section arrangement; denoted as S1 in Figure 2.1), or toward the rear wall of the stage (similar to the American way of string section arrangement; denoted as S2 in Figure 2.1). Secondly, apart from the natural acoustics of the concert house, two artificial reverberation conditions with reverberation times of 2.3 seconds and 3.2 seconds with S1 seating arrangement were also included, by using a room acoustic enhancement system installed in the concert house.

In the first part of the test, for a specific acoustic condition, a music conductor on the stage arbitrarily decided the number of violins to play together in the ensemble for each take from one of six scenarios: 1, 2, 3, 4, 6, and 9 violins. Accordingly, the string ensemble with the chosen number of violins performed the two musical pieces which lasted around 60 seconds. For each round of performance, the conductor gave

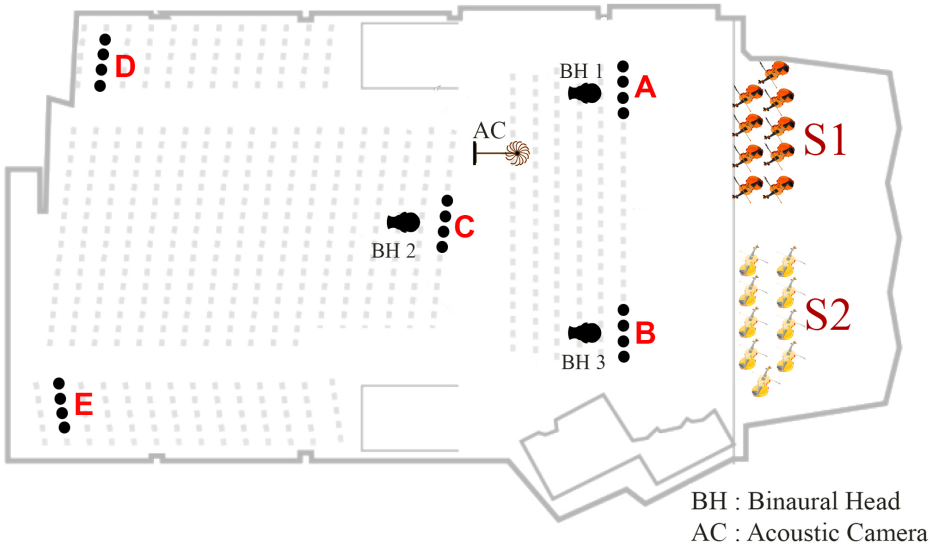


Figure 2.1: The position and orientation of sound sources, binaural heads, and listening test participants in Detmold Concert House.

a cue to everyone, and the listeners were asked to close their eyes and listen to the samples. Once the performance of the two pieces was over, the listeners were asked to predict the number of violins involved in the ensemble performed and also to indicate whether it sounded like an ensemble or not (yes or no). As per definition, the increase in the confusion in identifying the correct number of instruments in the ensemble corresponds to a higher degree of blending. After the performances in the four acoustic conditions with six different combinations of violins in the first part of the test, the listeners were allowed to choose their favorite seats in the concert house and repeat the same process in the second part. The test altogether took 2 hours to complete with a small break in between parts 1 and 2.

2.1.2 String ensemble recording setup

The performance of the string ensemble was captured using stereo pair microphones, an acoustic camera, three binaural heads, and clip-on microphones attached to the individual instruments. The stereo pair recordings were intended for artistic recording purposes, while the acoustic camera was used to analyze the radiation characteristics of the ensemble as a whole as well as spot significant room acoustic reflections. The close microphone recordings of the individual instruments as well as the binaural head recording of the ensemble performance were mainly utilized in this thesis work. Therefore, the methods and equipment utilized in these two types of recordings are given below.

close microphone recordings: Considering the influence of factors such as coordinated action, joint strategies among musicians, and room acoustic feedback on the resultant ensemble sound, the finest way to obtain the authentic source signals of instruments in joint performances is to record the sources individually in in-situ conditions. This is specifically important in phenomena like musical blending where joint performance strategies play a major role. Therefore, in this study, the individual violins in the string ensemble were recorded using 'DPA 4099 Core Violin' clip-on microphones attached to the body of the instrument. The DPA mics were positioned close to the violin bridge in order to better capture individual source signals from the joint performance as shown in Figure 2.2. These microphones have a frequency response of 20 Hz – 20 kHz with an effective frequency range of 80 Hz–15 kHz ($\pm$ 2 dB) at 20 cm distance. The musicians in the ensemble were seated with a separation of roughly 0.8 to 1 meter, and equal gain was applied for all the DPA microphone tracks in the sound card while recording. Due to the super-cardioid directivity characteristic of the DPA microphones and their placement close to the instrument, the DPA recordings are expected to minimize cross-talk from other instruments and room acoustic reflections [113]. Since these close microphone recordings can be assumed to be authentic and intrinsic representatives of realistic musical performances possessing minimal ambient noise and microphone cross-talk, they were utilized in this investigation to obtain sound samples of joint performances in the upcoming investigations.

Figure 2.2: Position and orientation of DPA clip-on microphone on violin.

Binaural recordings: Three binaural heads were utilized in this study to capture the sound field of the ensemble performances. The location and orientation of the individual and binaural head (denoted as BH) in the concert house are presented in Figure 2.1. A portable BHS II headphone unit SQobold–4 data acquisition system from head acoustics [114] was used to capture the binaural signal at the location of BH1. The Head Acoustics HSU III.2 binaural head, equipped with a head-shoulder unit and

ICP measurement microphones [115], was placed at the location of BH2 in the far field. Additionally, Neumann's KU-100 binaural head [116], a commonly used dummy head in the audio recording domain, was placed at the location of BH3 close to the violins.

2.2 Results and discussion

Overall prediction accuracy for different number of instruments:

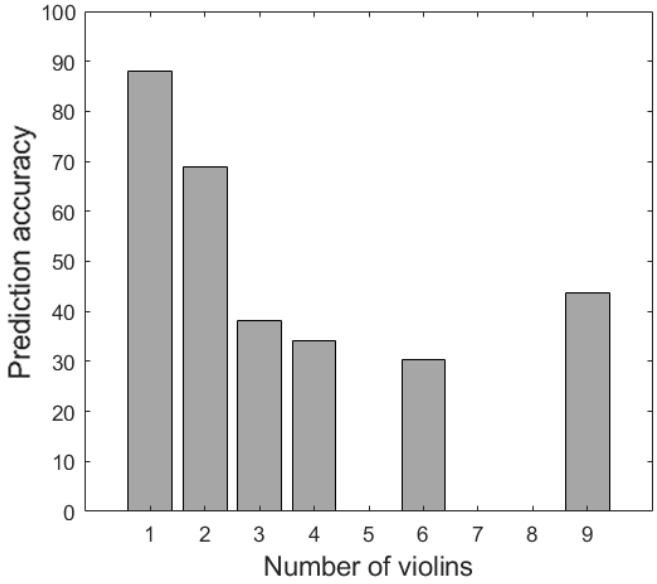


Figure 2.3: The variation of prediction accuracy of the number of violins played in the ensemble.

The overall predictability of the number of instruments involved in musical ensemble performance, averaged across all acoustic variations in the test is presented in Figure 2.3. The results show that the listeners' ability to predict the correct number of violins diminishes with an increase in the number of violins played. A significant drop in the prediction accuracy is observed after 2 violins, while only a little change in accuracy is noted from three to six violins. Interestingly, a slight improvement in accuracy is observed for the nine-violin condition compared to the three, four, and six-violin scenarios. This observation could be attributed to the listeners' prior knowledge of the maximum number of violins involved in the test. Factors such as an increased loudness impression compared to all the conditions could be one of the potential cues that influenced a higher accuracy in nine violins, but further analysis is required to identify the cues listeners utilized to differentiate the nine violin performance from other conditions.

To have a more detailed view of the predictability of violins, Figure 2.4 presents the distribution of the predicted number of violins against the actual number of violins, averaged across all acoustic variations. Consistent with the high prediction accuracy observed in Figure 2.3, the distribution of the predicted number of violins converges closely to the actual number of instruments for the one and two violin conditions, with only a few outlier points. From four violins onwards, a high variation in the prediction of the number of violins is observed among listeners. Particularly, a trend of overestimation in the perceived number of violins is observed for four violins, with an interquartile range (IQR) spanning from 4 to 6 and the upper whisker extending to 9. Conversely, the distribution for the nine violins exhibits an IQR ranging from 6 to 9, with the lower whisker extending to 2. Given that the listeners were already influenced by their prior knowledge of the maximum number of violins involved in the test, the prediction of values above 9 would not be expected.

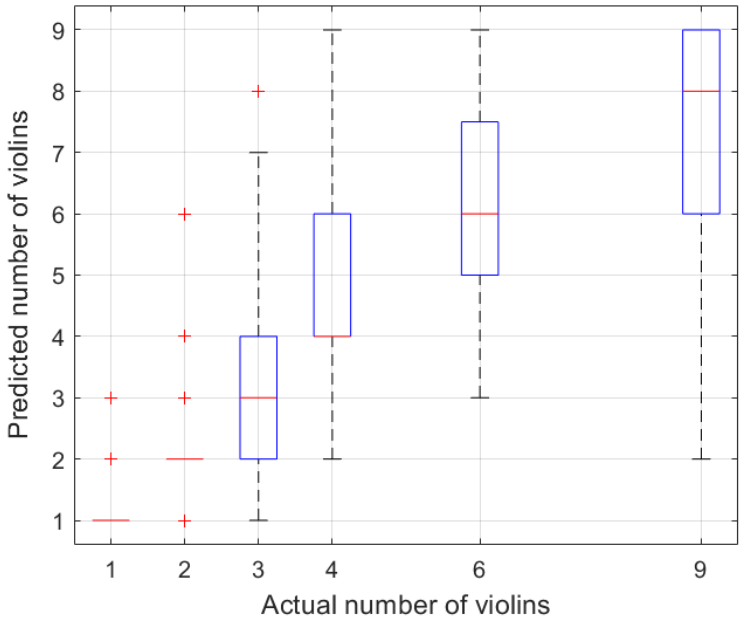


Figure 2.4: The distribution of the predicted number of violins corresponding to the actual number of violins played.

These observations suggest that the identifiability of the number of sound sources involved in a joint performance decreases with an increase in the number of constituent sources. The inability to distinguish the constituent sources in a concurrent sound is directly related to the impression of blending. Therefore, it is reasonable to anticipate an increased blending impression for ensemble performance with an increase in the number of sources. This observation aligns with the findings from [26], which

analyzed the identifiability of the number of instruments in an ensemble performance using different recording techniques. However, based on the individual prediction accuracies and the distribution of the predicted number of violins, a possibility could be hypothesized that the inability to predict the number of sources may not vary beyond a specific number of violins, and reach a saturated level. In that case, there may not necessarily be a significant enhancement in the contribution of additional sources to the blending impressions when increasing the number of sources beyond this threshold value. This phenomenon may be influenced by the acoustic environment-related attributes, as well as the characteristics of the sound stimuli, that need to be explored further.

The role of room acoustic attributes:

To better understand the role of room acoustic attributes in the prediction accuracy of the number of sources involved in ensemble performance, this section individually examines the overall variation in prediction accuracies across different room acoustic conditions and seating locations utilized in the test. Figure 2.5(a) illustrates the variation of the actual and predicted number of violins across four room acoustic variations involved in the test. Among the four conditions, the natural acoustic condition with S2 source distribution seems to possess relatively higher prediction accuracies for different numbers of violins. In contrast, the condition featuring a higher reverberation of 3.2 s with S1 seating arrangement exhibits the relatively weaker prediction accuracy. The decreased accuracy could be attributed to the increased reverberation time of the acoustic environment, which is previously shown to influence blending [27], and/or the seating arrangement of the instruments. It is observed that the participants overestimated the number of violins until six violins in the reverberant condition with 3.2 s artificial reverberation. However, they could not discern the difference between six and nine violins in this reverberant acoustic environment. This observation aligns with the previously-mentioned threshold hypothesis, suggesting that after a certain number of sources, no major changes in identification accuracy as well as the blending perception can be observed while increasing the instruments in an ensemble.

Figure 2.5(b) illustrates the overall variation in the prediction of the number of violins across different seating locations in the Concert House during the first part of the listening test. Out of the four predefined listener locations, the listeners in the far location (D and E) are observed to have relatively lower prediction accuracies compared to other locations. However, the trend is minimal, and not strong enough to draw conclusions. Unlike the earlier comparison on the room acoustic environment variation, where each sample point had 16 independent ratings from 16 individual listeners, the 16 ratings for each sample in this comparison are from 4 listeners rated across the 4 room acoustic variations. Therefore, the chances of bias and errors due to the involved listeners' skill and ability can be high for each seating location.

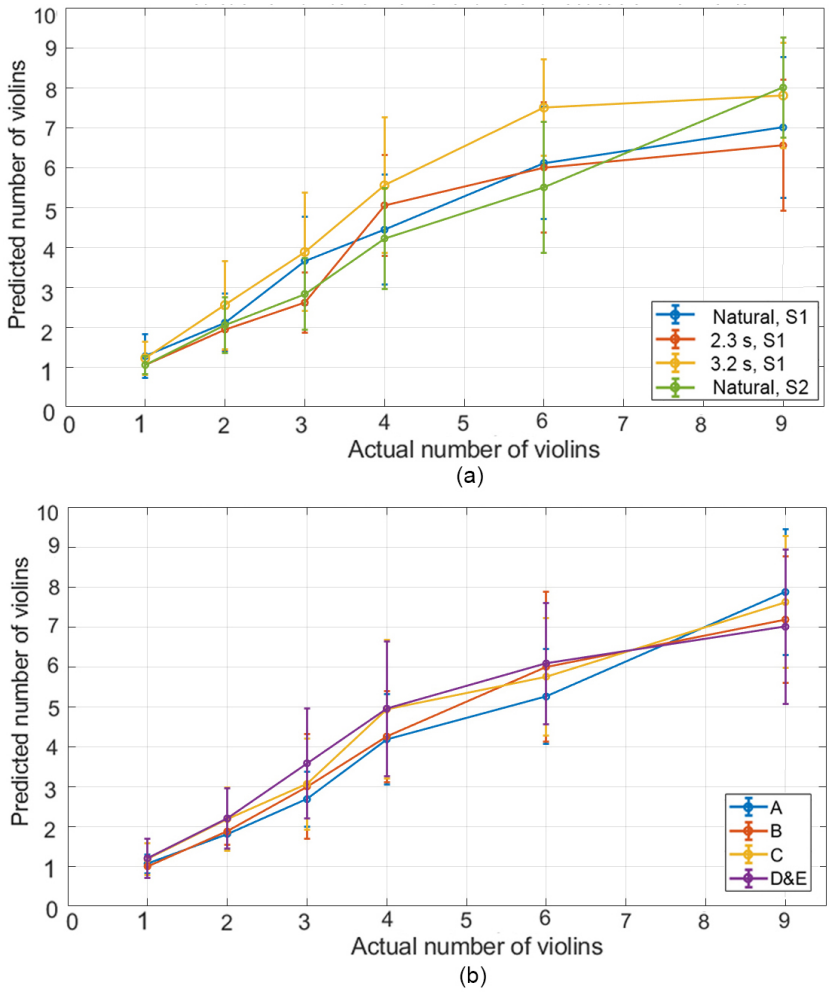


Figure 2.5: Variation in the prediction of number of violins with (a) different acoustic environments, (b) different seating locations.

As a broader approach, the responses of 12 participants who performed the test in both the direct sound field (i.e., within the critical distance from the source) and the diffuse sound field (i.e., outside the critical distance from the source) during the two parts of the listening test (the first part with listeners in the prescribed locations and the second part with listeners at their preferred locations) under natural acoustic conditions were taken for further analysis. The variation in prediction accuracies between the direct and diffuse sound fields is illustrated in Figure 2.6 for different numbers of violins. In general, the percentage of correctness is observed to drop with an increase in the number of violins. A significant difference in the percentage of correct predictions is noted between the direct and diffuse fields, particularly for the case of one violin where it drops from 100% correctness in the near-field to around 60% in the diffuse field. The

direct sound field generally holds a higher prediction accuracy in all conditions, except for the case of six violins. For the direct sound field, the prediction accuracy dropped from 100% to 20% with an increase in the number of sources from one to six violins, but the accuracy was improved for nine violins. As previously mentioned, the increase in sound pressure level along with more access to direct sound than the room reflections could be a potential reason for judging the loudest condition having 9 violins. In contrast, for the diffuse sound field, the prediction accuracy dropped only from 60% to 30% with an increase in the number of violins from one to four. The increase in accuracy for nine violins is marginal in this condition, moreover, the four, six, and nine violins have comparable prediction accuracy values. This confirms the saturation effect in the prediction accuracy after reaching a threshold number of sources, where no significant change in prediction accuracy is observed with an increase in the constituent sources in the ensemble. The significant differences observed between the direct and diffuse sound fields in terms of the prediction accuracies for different numbers of sources and also the saturation trend, validate that the characteristics of room acoustic environment play a significant role in identifying the constituent instruments in an ensemble sound, and thereby influence the impression of blending.

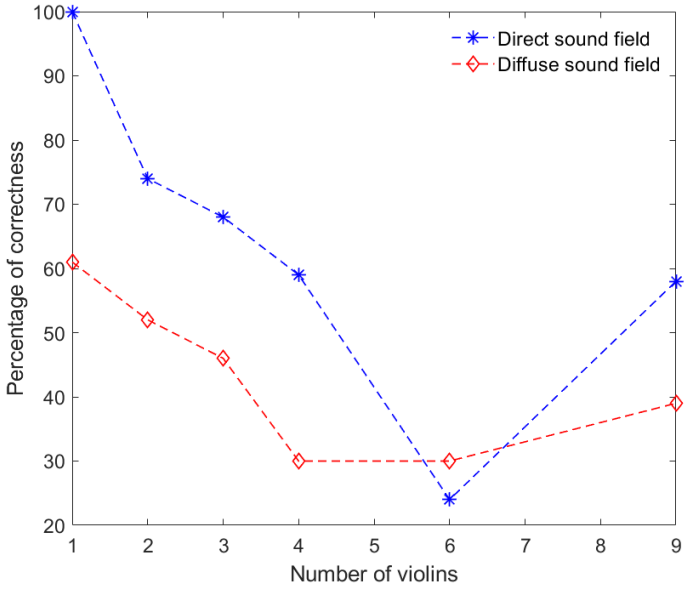


Figure 2.6: Variation of the prediction of number of violins in direct and diffuse sound fields.

Impression of Ensemble sound:

The percentage of agreement to the question 'whether the performance sounds like an ensemble or not' is presented in Figure 2.7 by averaging across the acoustic variations involved in the test. As expected, the percentage of agreement was zero in the case of one violin. While the percentage of agreement increases with the increase in the number of instruments from one instrument, it is observed to reach a plateau level at 4 violins after which no major improvement is found. This suggests an overall trend that the joint musical performances tend to sound like an ensemble from around 4 violins onwards. Although the percentage of agreement slightly increased for three violins in reverberant conditions, this trend remained almost consistent across the different acoustic variations utilized in the test, thereby aligning with the previous findings [26].

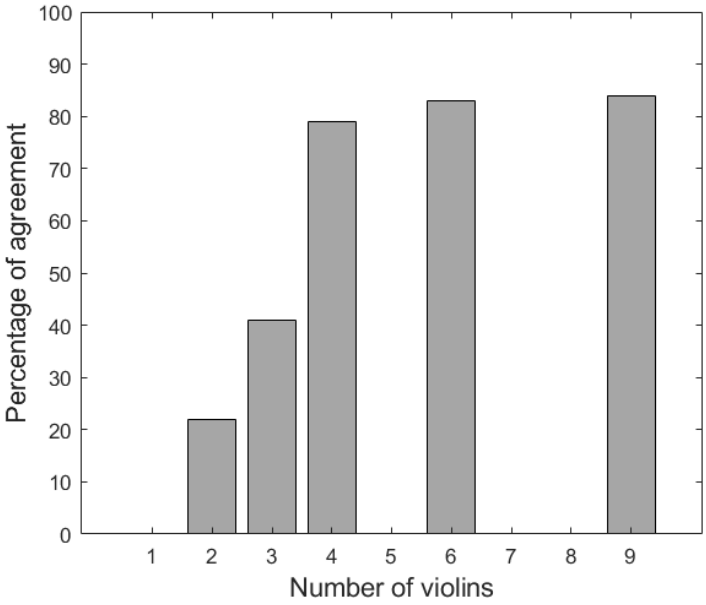


Figure 2.7: Agreement to the ensemble sound impression for different number of violins.

2.3 Summary

This pilot study investigated the blending of sound sources and the characterization of ensemble sound using a live listening test conducted with musical ensemble performance by analyzing the influence of room acoustic attributes in it. Additionally, extensive recordings of the ensemble performance were collected using diverse recording methods, which serve as material for the ensemble sound-related investigations of this

thesis. These recordings will soon be open for scientific purposes in a public repository, and they can be utilized to explore ensemble sound research by analyzing the role of musical and room acoustic aspects.

The findings from the listening test suggest that the listeners' ability to predict the number of constituent instruments involved in a joint performance of an ensemble tends to diminish with the increase in the number of instruments involved. Since the decreased identifiability of the constituent sound sources contributes to a better sensation of blending, the blending can be expected to improve with an increase in the number of sound sources. The acoustic variations utilized in the test by changing the characteristics of the concert house using an artificial reverberation system and changing the location of listeners appear to influence this phenomenon. Particularly reverberant acoustic environment and a listener's location in the diffuse sound field are noted to have a lowered ability to predict the constituent sources, thereby potentially enhancing the blending impression. In specific cases, a trend is noticed that the ability to predict the number of violins diminishes after a specific number of violins in the ensemble and reaches a saturated level. This trend is observed to be influenced by the characteristics of the acoustic environment, with reverberant and diffuse room acoustic conditions favoring this effect. In such a condition, there may not necessarily be a significant enhancement in the contribution of additional sources to the blending impressions when increasing the number of sources beyond a threshold value.

The listeners' agreement on the impression of ensemble sound is observed to increase with the number of sources in the joint performance to an extent, beyond which no major change is observed. In the variations involved in the test, the condition with four violins seems to show a high percentage of agreement of the ensemble sound impression, which is closely comparable with the six and nine violin conditions. Additionally, the four violin condition also appears to be a transition point at which a high variance was observed in the predicted number of instruments that is comparable to conditions with six and nine violins.

Since the listener was aware that the maximum number of violins involved in the ensemble was nine, it seemed to have caused a bias that restricted the interpretation of the results to an extent. Therefore, the trend of saturation effect observed in the prediction accuracy of violins needs to be evaluated with an improved test procedure or by including more instruments. Furthermore, based on the previous research findings, it can be stated that the way the musicians perform the same musical piece will not be necessarily the same across the different room acoustic variations involved in the test, due to the difference in room acoustic feedback. Therefore, it is acknowledged that the source-level blending may also have changed across these room acoustic conditions. While the influence of room acoustic variation on the musician's performance to achieve blending is an important and broad topic in the context of the musical ensemble, it is beyond the scope of this thesis.

Since the musicians in the string ensemble performance were asked to blend well during the performance, they reported at the end of the test that they relied significantly on the cues from the conductor as well as from the first violinist for synchronous performance to achieve a blended impression. This pilot investigation only focussed on the macroscopic perception of blending by evaluating the overall impression of blending from a joint performance lasting a few minutes. Considering the feedback from the room and co-performers, the degree of blending at the source level achieved from their joint strategies might have an effect on these results. In other words, a poor source-level blending achieved by the musician's performance at some trials can have a deteriorating effect on the final ratings, even if the room acoustic environment enhances it to some extent. Therefore, it requires a detailed analysis with a microscopic perspective to examine the blending at the source level and its alteration by the room acoustic environment separately, and thereby to have a better understanding of the evolution of blending. Based on these insights gained from this study, the evaluation of blending at the source level and the room acoustic level are analyzed in detail in the upcoming chapters of this thesis.

Chapter 3

Source level blending: development of a classification model

Assessing the auditory perception of source-level blending between sound sources is a crucial research topic within the fields of music perception and performance evaluation, but remains poorly explored due to its complex and multidimensional nature. Previous studies were able to estimate the source-level blending in musically constrained sound samples such as notes or chords, but comprehensive modeling of blending perception that involves musically realistic samples was beyond their scope. Combining the methods of Music Information Retrieval (MIR) and Machine Learning (ML), this chapter presents an investigation that attempts to classify sound samples from real musical scenarios having different musical excerpts according to their overall source-level blending impression. Rather than demanding access to acoustically clean individual source signals as done in earlier studies which poorly represent realistic musical listening situations, this investigation is performed in more realistic musical settings by utilizing in-situ recordings of monophonic, musically realistic, and score-independent sound samples of two violins from unison performances. This offers a first step toward the comprehensive modeling of the overall source-level blending impression. The content of this chapter is reproduced from the following research article with the permission of the Acta-Acoustica:

J. Thilakan, B.T. Balamurali, J.M. Chen, Malte Kob, "Classification of the perceptual impression of source-level blending between violins in a joint performance," Acta Acustica 7, 62 (2023), https://doi.org/10.1051/aacus/2023050,

3.1 Materials and methods

The overall block diagram of the investigated classification modelling is depicted in Figure 3.1. Monophonic sound samples of two violins extracted from live ensemble performances were perceptually evaluated in terms of the overall blending impression by a group of expert listeners, and subsequently labelled into 'blended' and 'non-blended' classes. The preparation of sound samples, execution of the perceptual test, and labelling are described in section 3.1.1. In contrast to the conventionally established parameters that explained the blending in previous research which were only accessible from the individual source channels (detailed in section 1.2.2), this study utilizes the Mel Frequency Cepstral Coefficients (MFCCs) extracted from the monophonic sound samples as input features for the classification model. The process of extraction of MFCC and its deployment in the study are described in section 3.1.2. Three commonly used feature transformation methods – Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-distributed Stochastic Neighbour Embedding (t-SNE) are employed here to project the high-dimensional features into a lower dimension by retaining the important information and avoiding redundancy.

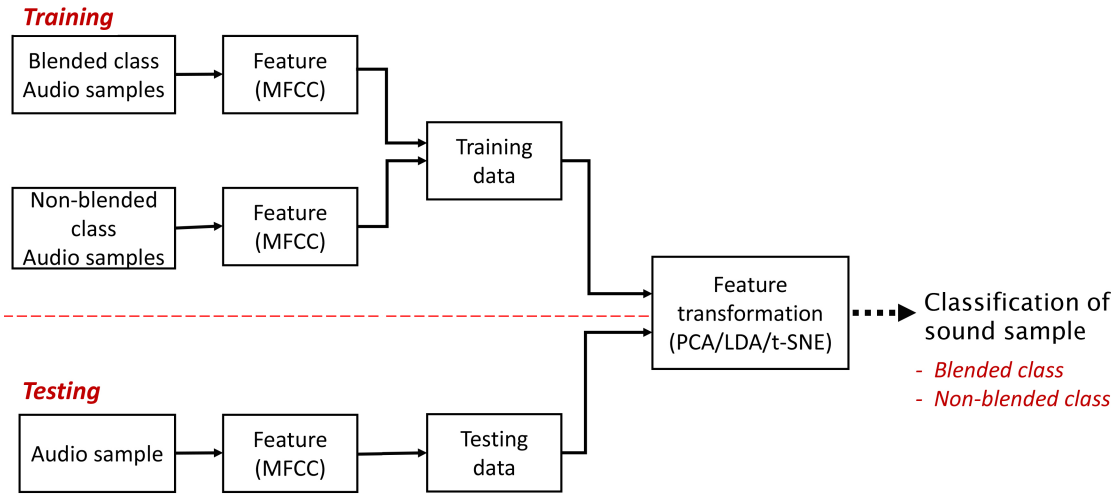


Figure 3.1: Block diagram of the proposed classification model.

Audio samples were grouped into training and test sets for this blending classification modelling process. The training data, which includes the pre-defined blended and non-blended classes, and test data are then transformed using the chosen feature transformation methods to a low-dimensional feature space, and the distance between clusters is used as a parameter for the blending classification of the test sample. The feature transformation techniques and modelling are explained in section 3.1.2.

3.1.1 Preparation of sound samples

The clip-on microphone recording of the performance of a string ensemble consisting of 9 violins presented in section 2.1.2 is utilized in this investigation for the preparation of sound samples. Due to the super-cardioid directivity characteristic of the DPA clip-on microphones used in the recordings, the instrument recordings are expected to have minimal cross-talk from other instruments and room reflections [113]. These close-miking recordings are assumed to be authentic and intrinsic representatives of realistic musical performances possessing minimal ambient noise and microphone cross-talk, and hence they were utilized in this study to obtain sound samples of joint performances.

From these recordings of the nine violins, 50 sound samples consisting of two violin signals were extracted for evaluation. These samples had a duration of approximately 3 to 5 seconds each, and they included different musical fragments. The two violins in each sample were randomly chosen from the 9 violin tracks, and hence the influence of the coordination effect due to spatial proximity can be negligible in the selected samples. The basis of the selection of these samples was that these samples should not provide salient cues for distinguishing the two constituent violins such as pitch difference, onset timing asynchrony, etc., and significant noise level from bow & musician. Nevertheless, these 50 samples are expected to differ in terms of blending impressions due to the unavoidable differences in the musical attributes described in section 1.2.2.

The samples were extracted and post-processed in Reaper digital audio workstation; breathing and violin bow noises were minimized using a high pass filter with a cutoff frequency of 200 Hz without attenuating the low notes in the musical pieces. Furthermore, fade-in and fade-out filters of less than 0.3 s duration were added at the beginning and end of the signals. Subsequently, the two channels from the violins were rendered by downmixing to a mono-aural sound sample with equal gains on each track at 44.1kHz/16-bit depth, and these samples were used for the perceptual evaluation of source-level blending.

Perceptual labelling of sound samples:
Fourteen musically ear-trained participants including Tonmeister students and professional musicians (4 female, 10 male, mean value of age 28.7 $\pm$ 7.2) participated in the listening test. The participants had prior experience in critical listening, and previous studies have shown their sensitivity over non-musicians in selectively attending to and analyzing the complex spectral and temporal features of sounds [17; 18]. So, such a population is expected to better conceive the notion of blending and provide the blending ratings with concordance.

The objective of the listening test and the test procedure were described to the participants at the beginning of the test. As discussed in Section 1.2.2, blending can be assessed using a rating scale or by evaluating the identifiability of constituent sound

sources in a joint performance. Building on previous studies on statistical modeling of source-level blending, this study used a 0–10 rating scale to assess perceived blending, where lower values indicate poor blending and higher values indicate strong blending. The working definition of blending was stated to the participants as "the perceived fusion of violin sounds where the constituent instruments are indistinguishable". Since the standard examples of the possible extrema of the blending impression between violins are not known, it was not possible to provide the reference samples for the training phase. This might have limited the listeners in conceiving the possible variation in the blending impression in the chosen set of samples and forming their inner-scale of blending rating. Nevertheless, to prime the listener with the sound of the instrument in the DPA microphone recordings, familiarization audio examples consisting of sound samples from the recording were provided at the beginning of the test.

Five audio files, each with 10 sound samples, were generated for the listening test. The sound samples were randomized in the audio files in order to reduce the memory retention effects in ratings. Each sound sample was played three times and the listeners were asked to rate the blending impression of the particular sample on a scale of 0 to 10 in a test response form. The participants performed the test using studio-grade headphones of their choice in acoustically treated quiet environments. After each audio file, participants had the option to take a short break and resume the test which helped them to reduce mental fatigue. Including short breaks between each set of audio samples, participants took an average of 30 minutes to complete the test.

Consistency and reliability of listening test rating: When looking at the listening test responses, some of the samples chosen for the study had a high variation in the sample rating, indicating a high inter-participant disparity in the perceived blending impression among the trained participants. This could be due to the different levels of attention given to the musical aspects (e.g., pitch, timbre, onsets, etc.) by the participants while judging the sample. To tackle this problem, and to use sound samples with a considerably consistent rating, a threshold value of standard deviation of 2 was chosen in this study. Accordingly, sound samples with a standard deviation value smaller than 2 were chosen for the classification evaluation which resulted in a final set of 31 sound samples from 50. The assessment of inter-rater reliability was performed to further evaluate the degree of agreement between the listeners using Cronbach's alpha [117]. The Cronbach's alpha value for the selected 31 sound samples was obtained to be 0.93, indicating high reliability among the test participants on the rating of blending.

The probability distribution of the blending ratings of the short-listed 31 sound samples is shown in Figure 3.2, which demonstrates a bi-modal distribution having two maxima around 5.25 and 7.75 and a minimum around 6.5. Based on the bi-modal distribution of blending perception ratings, two classes of samples were established - the "blended class" that includes samples with a mean rating $>$ 6.5/10, and the "non-blended class" that includes samples with a mean rating $<$ 6.5/10. Resulting in, the

selected sample set having 13 samples from the blended class and 18 samples from the non-blended class[1].

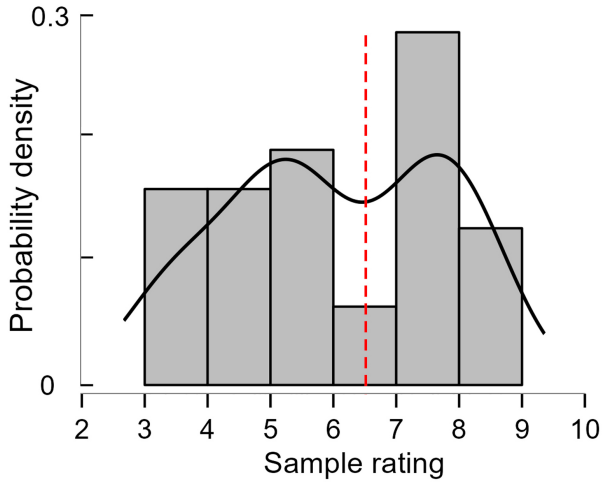


Figure 3.2: Probability distribution of the blending ratings of 31 sound samples (the thick black line shows the probability distribution function; the dashed red line indicates the minimum arising between the two maxima in the distribution function.

3.1.2 Classification modelling

Feature extraction

The Mel Frequency Cepstral Coefficients based subjectively on a nearly logarithmic sense of human auditory pitch perception [119] have been shown to successfully represent perceptually-related characteristics of signals [120]. Therefore, they have been used widely in speech signal analysis including speech recognition [120; 121], speaker identification [122], and verification [123]. Other studies further demonstrate their relevance in music modeling [124], musical instrument recognition [125], and voice and musical emotion detection [126; 127], thereby showing performance superior to conventional audio features in MIR applications.

The process of extraction of MFCC features begins with converting the audio signals into frames using a moving time window and performing Discrete Fourier Transformation (DFT) on each frame to get the power spectrum. A filter-bank derived from the Mel scale is then applied to the power spectrum to obtain the Mel-scale power spectrum. The logarithm of the amplitude spectrum is then taken. Finally, a Discrete Cosine Transform (DCT) of the log filter-bank energies generates the Mel Frequency Cepstral Coefficients.

[1]The selected sound samples are available at [118]

Silent regions at the start and end of the audio samples were removed, and the first 14 MFCCs [128] were extracted for every 100 ms of the audio signal with an overlapping length of 50 ms using a Hamming window. Along with the raw MFCC features, standardized (Z-score normalized) MFCC features were also computed for further analysis due to their applicability in similar related studies [129].

Feature transformation methods

The fundamental purpose of feature transformation using techniques adapted from dimensionality reduction is to project the higher-dimensional data into a lower dimensional space yet retaining most of the relevant information and removing the redundant or correlated information as well as the undesired noise. It helps in decreasing the complexity of high-dimensional features and also supports in low-dimensional visualization of the features. They have also been shown to improve the performance of the statistical modeling method or machine learning algorithms [130; 131]. Depending on the size, quality, and characteristics of input data (i.e., the feature set), different types of feature transformation algorithms can be used for dimensionality reduction. They can be classified into three main groups — linear vs. non-linear, supervised vs. unsupervised, and random-projection vs. manifold-based [130]. The three feature transformation techniques used in this investigation are detailed below.

Principal Component Analysis (PCA): PCA is a widely used unsupervised and linear dimensionality reduction method. It linearly projects higher-dimensional data into Principal Components (PCs) while maximally preserving input data variance [132]. The principal components are mutually orthogonal and they represent directions of the data that explain a maximal amount of variance.

Estimation of PCs starts by calculating the covariance matrix of the n-dimensional (n=14 here) input data (X). Next, the Eigenvalue decomposition is done on the covariance matrix to estimate the Eigenvalues and Eigenvectors. A transformation matrix ($W_{\mathrm{n\,x}\,k}$) made up of top k Eigenvectors is used to project X onto a lower-dimensional feature space. The PC transformation minimizes redundancy, noise, and feature collinearity. Non-linear feature extraction techniques outperform the PCA on artificial tasks and can deal with complicated data structures, but studies suggest that they do not outperform PCA in natural data sets [133]. Furthermore, PCA was shown to enhance modeling accuracy and efficiency by transforming MFCCs [121; 125; 134].

Linear Discriminant Analysis (LDA): LDA is a supervised and linear feature transformation technique that uses Fisher's criterion – maximizing inter-class variance while minimizing intra-class variation – resulting in minimal overlap of features corresponding to different classes (maximum class separation) in new dimensional transformed space [135].

To derive low-dimensional features, two scattering matrices are estimated for the predefined classes – (1) within the class scattering matrix (S_{wc}), and (2) between the class scattering matrix (S_{bc}). The Eigenvalue decomposition is done on the matrix $S_{wc}^{-1}.S_{bc}$ to derive the Eigenvalues and Eigenvectors. Eigenvectors corresponding to the highest Eigenvalues in the new feature space maximize class separation in the transformed space. Similar to the PCA, a transformation matrix W with the top k Eigenvectors is constructed, which transforms the input data X onto a lower dimension. Unlike PCA, the features in the low dimensional basis are not necessarily orthogonal in LDA [135; 136].

t-Stochastic Neighbourhood Embedding (t-SNE): t-SNE is an unsupervised, non-linear feature transformation that can capture most of the necessary local structure information from a high-dimensional feature space while simultaneously revealing information about the global distribution of the data [137]. The dimensionality reduction of t-SNE is performed similarly to the Stochastic Neighbourhood Embedding (SNE) in which the distance measures between data points in high dimensions are converted into conditional probabilities. A symmetrized cost function based on 'student t-distribution' is implemented in t-SNE which improves the optimization and crowding problems of SNE [137]. The non-linearly extracted probability values refer to the similarity between all the pairs of data points. Afterward, this process is performed for all the pairs of data points in lower dimensional feature space (which is normally 2 or 3 for visualization purposes), and the probability values are extracted. Finally, the embedding of the high dimensional data to a lower dimensional feature space is performed by minimizing the difference between probabilities using the optimization of Kullback–Leibler divergence [137; 130]. Euclidean distance was used in this method for the similarity estimation between data points.

Test-train split up and classification criteria

Due to the limited sample size available in our study, implementation of advanced machine learning modelling algorithms like Deep Neural Networks has limitations. Comparison of similarity using distance measures between clusters is a conventional method in statistical modelling [138; 139], and a similar technique is used in this investigation. The first phase of the modelling process started with randomly dividing the samples into training and testing data sets; this investigation included 23 training and 8 test samples, respectively. The training data included 10 samples from the blended class and 13 samples from the non-blended class, and the test data included 3 samples from the blended class and 5 from the non-blended class. The transformation of the training data (including pre-defined blended and non-blended classes) and test data to a low-dimensional feature space is performed using the proposed feature transformation methods.

The centroid of the data distribution–the Euclidean coordinate which corresponds to the arithmetic mean of data points across the dimensionality-reduced feature space–is estimated for the blended class, non-blended class, and test data. The Euclidean distance between the centroids of these blended, non-blended classes and the testing audio sample in the low-dimensional feature space was used as the metric for the classification of blending impression. In our classification criteria, if the Euclidean distance between the centroid of the blended class and test data is less than the distance between the centroid of the non-blended class and test data, then the test sample is classified as 'blended', and vice versa. For each test sample, the predicted class from the model was compared with its perceptually labelled class. Accordingly, the performance of each feature transformation technique is estimated.

The results of this evaluation could be biased due to the chosen samples in the training and testing data sets and the limited sample size. To overcome this issue of bias and the possible randomness in the result, the accuracy of the best-performing feature transformation models when evaluated using a distinct training and test set is further validated using Leave-One-Out Cross-Validation (LOOCV). LOOCV involves training the model with all of the data except for one data point, for which a prediction is made [140]. A total of 31 unique models must be trained using 31 data samples; while this is a computationally expensive strategy, it ensures an accurate and unbiased measure of model performance.

3.2 Results

3.2.1 Statistical Analysis of Transformed Features

The hypothesis that the distribution of two classes (blended and non-blended) in transformed features have equal mean values (null hypothesis) or not was tested using the Mann-Whitney U test [141]. Transformed features corresponding to all 31 samples (13 blended and 18 non-blended) were considered for this analysis. In PCA and LDA, the transformation that preserves maximum data variance (95% in this investigation) was considered. This has resulted in four transformed features for raw MFCC and nine transformed features for standardized MFCC for PCA transformation, and one transformed feature for both raw and standardized MFCCs for LDA transformation. For t-SNE, the default transformed dimension of three was used for both raw and standardized MFCCs.

PCA feature analysis: The Mann-Whitney U test was performed on the PCA- transformed raw and standardized MFCC features, and the results are shown in Table 3.1. Additionally, the box plot and probability density function of the two classes corresponding to each feature shown in Figures 3.3 and 3.4 describe the distribution of PCA-transformed features. Since the distributions of PC7 to PC9 are very similar to those

of PC6, they were omitted from Figure 3.4.

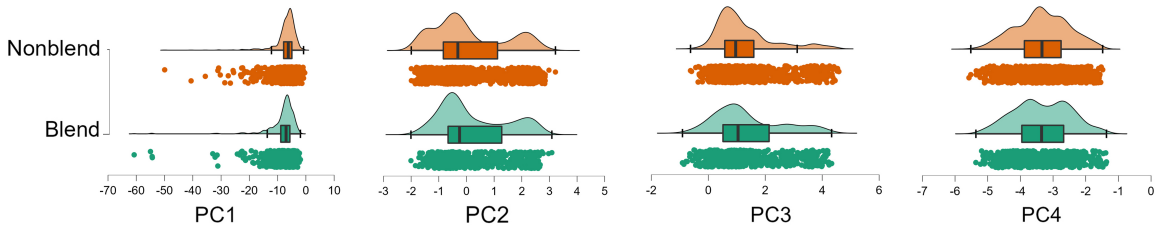


Figure 3.3: Distribution of PCA-transformed raw MFCC.

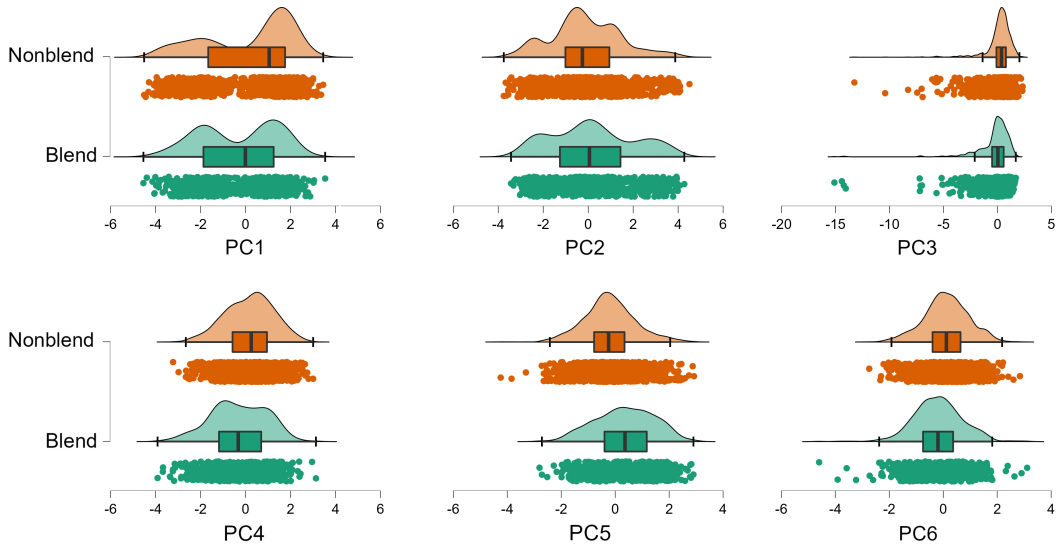


Figure 3.4: Distribution of PCA-transformed standardized MFCC.

For Mann-Whitney U test results comparing two transformed feature sets (see Table 3.1), all p-values (with the exception of PC4) are less than 0.05, rejecting the null hypothesis that there is no difference between the mean values of PC1, PC2, and PC3 corresponding to the PCA transformed MFCC features of blended and non-blended samples. The exception (p-value $>$ 0.05) for the standardized MFCC was PC7.

LDA feature analysis: Mann-Whitney U test result of LDA transformed raw and standardized MFCCs is shown in Table 3.2 and their distributions are depicted in Figure 3.5 (a) and (b), respectively. The p-value of the transformed features corresponding to blended and non-blended samples are less than 0.05, implying that the null hypothesis of equal means is rejected once again. The distribution plots of raw and standardized MFCCs clearly demonstrate the differences between the two classes of data and validate the Mann-Whitney U test findings.

Input feature	Feature	p-value
Raw MFCC	PC1	<0.001
	PC2	0.002
	PC3	0.138
	PC4	0.863
Standardized MFCC	PC 1,2,3,4,5,6,8,9	<0.001
	PC7	0.421

Table 3.1: Mann-Whitney U test summary of the PCA features.

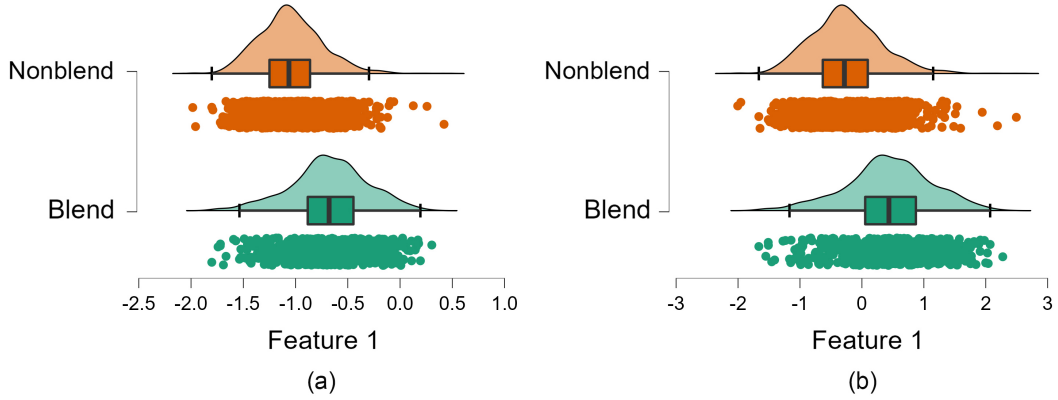


Figure 3.5: Distribution of LDA-transformed (a) raw MFCC, (b) standardized MFCC.

Input feature	Feature	p-value
Raw MFCC	Feature 1	<0.001
Standardized MFCC	Feature 2	<0.001

Table 3.2: Mann-Whitney U test result summary of the LDA features.

t-SNE feature analysis: The Mann-Whitney U test findings for t-SNE transformed raw and standardized MFCCs are shown in Table 3.3, and their respective distributions are illustrated in Figures 3.6 and 3.7. The p-value is significant for the second and third t-SNE transformed raw MFCC features, however, the first t-SNE transformed feature is significant for standardized MFCC. Furthermore, the distribution of t-SNE differs from that of PCA and LDA, where notable bimodal characteristics can be detected in the former.

3.2.2 Cluster visualization of PCA, LDA, and t-SNE

Figure 3.8 shows the cluster distribution of transformed raw MFCC features for blended and non-blended samples. This helps in visualizing the transformation in lower dimen-

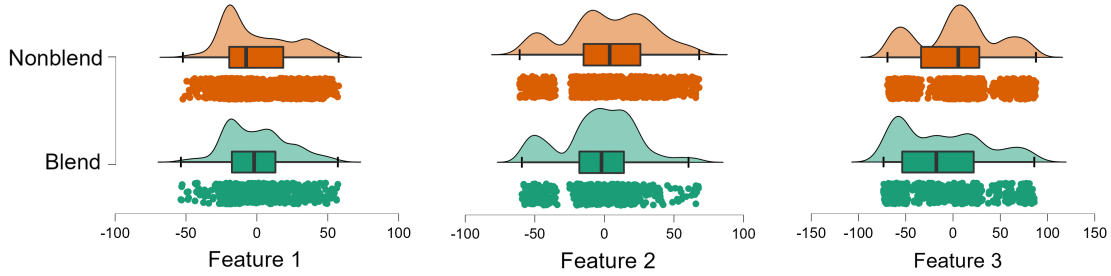


Figure 3.6: Distribution of t-SNE transformed raw MFCC.

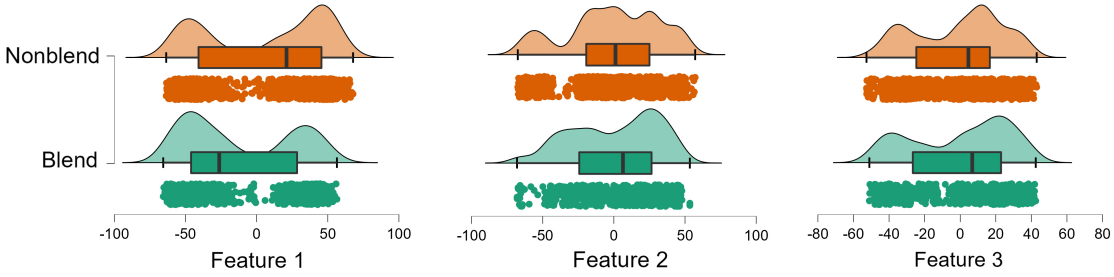


Figure 3.7: Distribution of t-SNE transformed standardized MFCC.

Input feature	Feature	p-value
Raw MFCC	Feature 1	0.135
	Feature 2	<0.001
	Feature 3	<0.001
Standardized MFCC	Feature 1	<0.001
	Feature 2	0.507
	Feature 3	0.107

Table 3.3: Mann-Whitney U test result summary of the t-SNE features.

sional space, and the first three transformed features of MFCC were compared across the three transformation techniques. The blended audio features transformed from the training set are represented by red dots, while the non-blended features are represented by green dots. The centroids of blended and non-blended training data distributions are highlighted using red and green spheres. Furthermore, the centroids of the transformed blended audio samples from the test data are shown as red triangles, while that of the non-blended samples are shown as green triangles (See Figure 3.8 (a) and (b)). Because the t-SNE transformation matrix is dependent on the test data, the resulting centroid of a non-blended sample (selected as an example) is shown in Figure 3.8 (c).

Since the overall blending rating is considered in this investigation rather than the time-varying blending parameter, the two classes in training data may overlap,

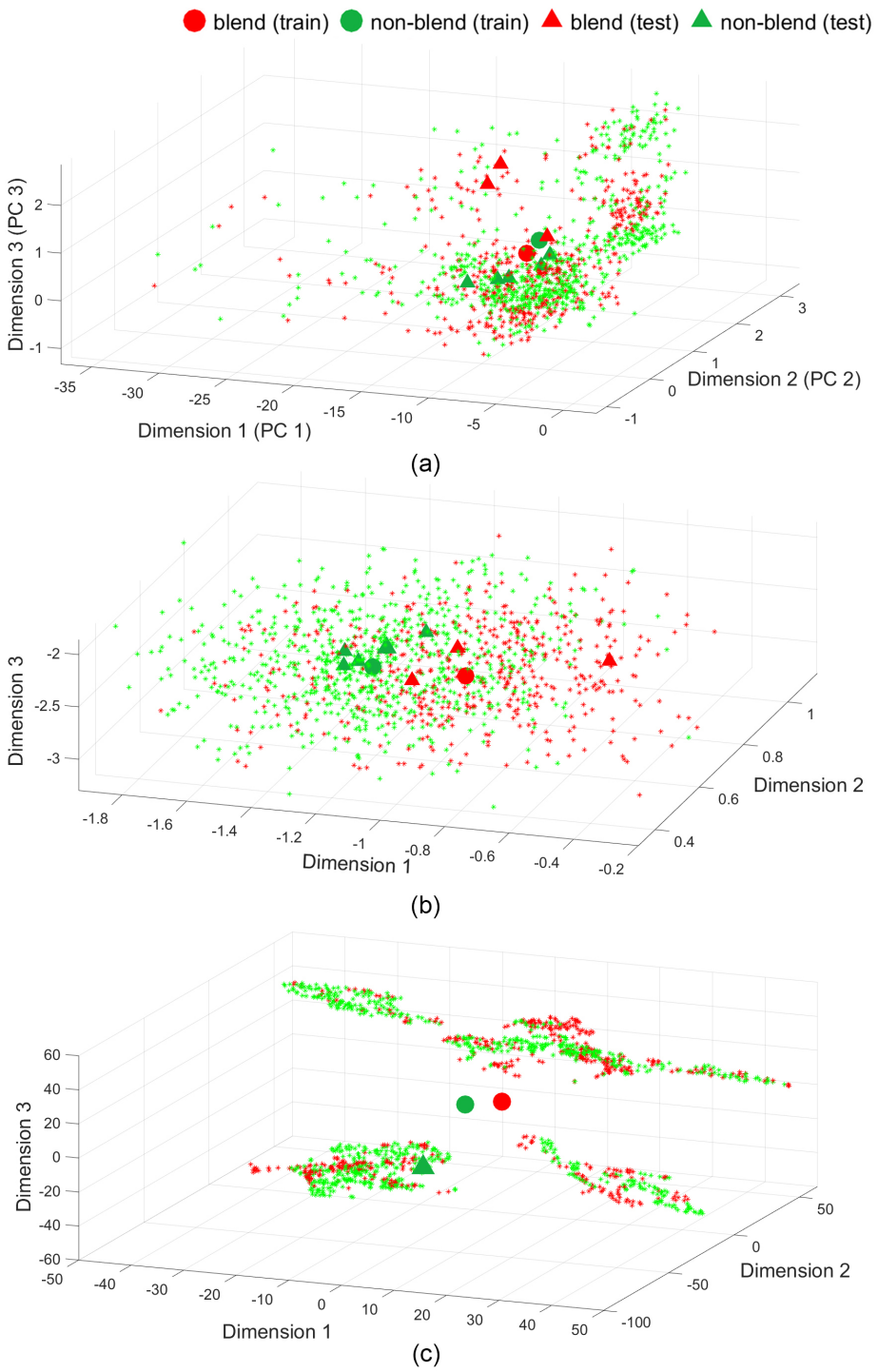


Figure 3.8: Cluster distribution of transformed raw MFCC features for blended and non-blended samples using (a) PCA, (b) LDA, (c) t-SNE. Spheres indicate centroids of blend (red) non-blend (green) training data, while triangles indicate centroids of blend (red) and non-blend (green) test data.

which means that the non-blended sound sample may contain many data points with blended characteristics and vice versa. Looking closely at Figures 3.8 (a), (b), and (c), the class overlap in PCA, t-SNE, and LDA is visible, though slightly less overlap in the latter. This less overlap could be attributed to the supervised nature of the LDA transformation, which could eventually aid in class identification.

3.2.3 Classification Model Result

With Separate Train-Test Samples

Table 3.4 shows the performance of various models for blended-non-blended classification when the model is trained using a fixed training sample set (23 samples) and tested against the remaining eight samples. The table clearly shows that the raw MFCC consistently outperforms the standardized MFCC, which is surprising and contradictory to many previous results on audio classification [142; 143].

Among the six classification models, the transformation of raw MFCC using LDA and subsequent similarity estimation produced the highest accuracy (87.5%). The LDA-supervised transformation, which results in less overlap between the two classes (see feature distribution in Figure 3.5), could be responsible for this superior performance. PCA and t-SNE transformations of raw MFCCs were relatively worse (75%) as compared to LDA. All the remaining models have resulted in the same accuracy (62.5%). Figure 3.9 shows the resulting confusion matrices for these models. When using raw MFCC for transformation, LDA misclassified one of the blended signals as non-blended (see Figure 3.9 c). PCA and t-SNE transformation has an additional non-blended misclassification. The remaining models perform poorly and consistently misclassify the blended signals as non-blended.

Transformation method	Input data	Accuracy
PCA	Raw MFCC	6/8 (75%)
PCA	Standardized MFCC	5/8 (62.5%)
LDA	Raw MFCC	7/8 (87.5%)
LDA	Standardized MFCC	5/8 (62.5%)
t-SNE	Raw MFCC	6/8 (75%)
t-SNE	Standardized MFCC	5/8 (62.5%)

Table 3.4: Performance of PCA, LDA and t-SNE transformation models trained and validated using separate train and test samples.

To confirm that the technique selected for feature transformation and classification is free of bias and to assess the likelihood of possible overestimation of accuracy with selected samples from the testing set, a leave-one-out cross-validation technique was

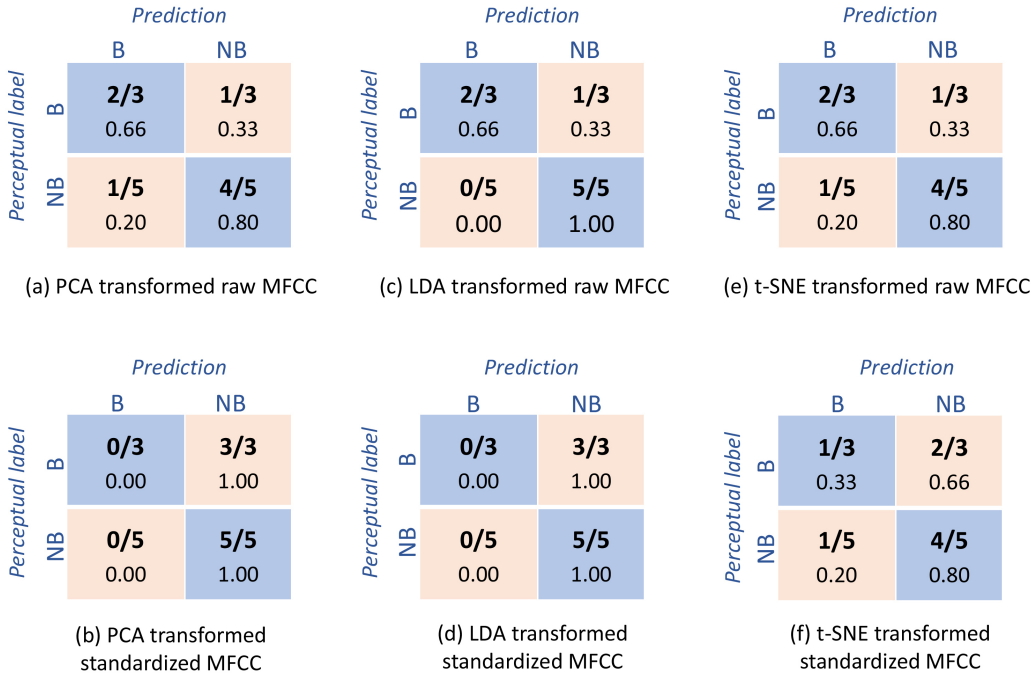


Figure 3.9: Confusion matrices depicting the correct and misclassification rates of the six transformation models trained and validated using separate train and test samples, (number of test samples n=8; B and NB represent Blended and Non-Blended classes).

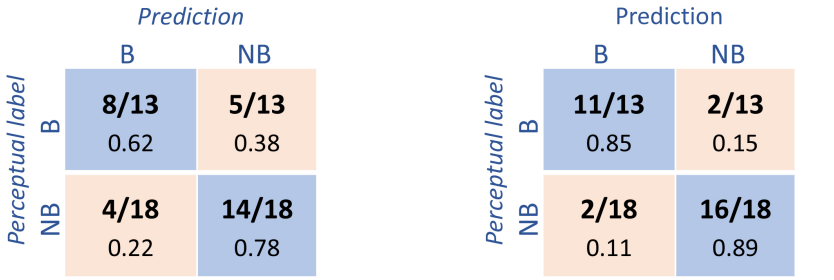
finally carried out in this investigation. Although the t-SNE transformation of MFCC exhibits a comparable result to PCA, it can't be considered to be a generalized solution since the t-SNE transformation is test data-dependent [137]. Further, t-SNE was employed in this study for the completeness of the dimensionality reduction method due to its ability to visualise more sophisticated higher-dimensional clustering of data using a non-linear approach. Hence, the investigation using LOOCV is limited to the top-performing models from this analysis, i.e., PCA and LDA transformation of raw MFCC. The results of LOOCV are discussed in the following section.

Cross validation of the model

In LOOCV, the model is trained using all of the samples except for one sample, for which a prediction is then made. Table 3.5 shows the performance of PCA, and LDA transformation models validated using LOOCV, and Figure 3.10 shows the Confusion matrices depicting correct and misclassification rates in LOOCV (models were trained and validated with 31 separate iterations).

LOOCV result is comparable to the earlier result (i.e., from a distinct train-test split). A prediction accuracy of 87.1% for MFCC features transformed using LDA was

achieved with 27 correct classifications out of 31. There are two misclassifications for both the blended and the non-blended classes. The PCA-transformed MFCC performed marginally worse with 22 correct classifications out of 31. In PCA transformed features, the blended classes misclassified as non-blended are higher than the LDA equivalent (5 out of 13 in PCA compared to 2 out of 13 in LDA). So overall following the result based on the given set of samples, this investigation demonstrates that the MFCC features transformed using Linear Discriminant Analysis are a suitable method for classifying musical score-independent, monophonically rendered dynamic musical signals in terms of their overall perception of the source-level blending impression.



(a) PCA transformed raw MFCC

(c) LDA transformed raw MFCC

Figure 3.10: Confusion matrices depicting correct and misclassification rates in LOOCV (models were trained and validated with 31 separate iterations; B and NB represent Blended and Non-Blended classes).

Transformation method	Input data	Accuracy
PCA	Raw MFCC	22/31 (70.9%)
LDA	Raw MFCC	27/31 (87.1%)

Table 3.5: Performance of PCA and LDA transformation models validated using LOOCV.

3.3 Discussion

This study used a computational approach to classify monophonic recordings of unison performances by two violins into blended or non-blended classes based on their overall impression of source-level blending, while the two latter classes were perceptually validated through a listening experiment with expert listeners. The classification

accuracy reached up to 87%, indicating a promising method that considered realistic sound samples with different musical content without accessing individual source recordings. Our study shows that a computational approach can model certain aspects of complex psychoacoustic phenomena such as musical blending, and thus introduces a new tactic to address a stubborn and difficult auditory puzzle. In this regard, this study differs from earlier studies on understanding and assessing blending in that it directly incorporated "ecologically" realistic sound samples containing musical excerpts representing diverse musical contents, and was not restricted only to musical notes or chords [14; 33; 30]. Given that this ecologically representative approach has never been attempted before, it increases the novelty and impact of this study. Moreover, this investigation advances the state-of-the-art by incorporating methods from the disciplines of Music Information Retrieval (MIR) and Machine Learning in this music perception-related research problem.

Unfortunately, a direct comparison of this study with earlier investigations that treated blending as a continuous variable is not possible since this study performed the binary classification of sound samples into blended and non-blended classes. Further, while rating sound samples that contained musical excerpts, only a single-valued rating that represents the 'overall' impression of blending was given for each sample, which was not finely resolved in time. Hence, it is acknowledged that this subjective rating may not always be useful in explaining the temporally continuous variation of musical attributes in other contexts (e.g. a short performance mistake that occurred for 100 ms in a highly blended sample may have a stronger effect on its final overall impression of the blend for the whole sample). Nevertheless, we observed the bi-modal probability distribution of sample ratings (see Figure 3.2), therefore the naive classification of the samples into two categories is not unreasonable.

The notion and perception of musical blending can differ across individuals from a heterogeneous background, thus this feasibility study was performed 'only' among musically ear-trained critical listeners (who are expected to be sensitive and perceptive to audio cues and hence offer good convergence in terms of the musical agreement), and thereby expected to reduce variance and inconsistencies in perceptual labeling. Additionally, in contrast to previous studies which employed instrument recordings from sophisticated recording conditions [33; 30], this study utilized in-situ recordings of an ensemble performance from a concert hall which was carried out by intentionally not restricting the natural environment of the musicians and the auditory and visual feedback from the performers and acoustic space, and thus represents an "ecological" musical and listening context. Therefore, unlike the necessity of having 'acoustically clean' signals, this study made use of authentic and natural representative sound samples of joint musical performance having natural ambient noise and negligible microphone crosstalk – this increased representativeness of natural musical performance thus broadens the relevance when studying other in-situ joint musical performances.

Although MFCCs are widely used in Music Information Retrieval (MIR) to describe spectral and timbre information [144], it is noted that the coefficients alone offer limited insight into music perception [145] and by extension, limited utility to explain blending impression in terms of temporal, spectral, and energy-based musical attributes. In this approach, we did not incorporate certain prominent feature parameters from time and frequency domains that were explored in earlier blending studies [14; 33; 30] such as pitch, spectral centroid, attack contrast, and loudness correlation, etc. The reason for not using them in this approach is that our study focuses on monophonically rendered audio samples, while those parameters were specifically designed for individual source channels. Since this work is limited to MFCC features, alternative MIR features would be studied and included in future studies for modelling the blending perception.

The bias or chance predictions arising from the proposed classification model have also been checked by performing a two-fold evaluation (test-train split up, and cross-validation). Having developed the blending classification model, we now have a tool that allows future work to focus on the estimation of blending impressions of dynamic music samples using large and diverse datasets (that includes samples having different numbers and combinations of instruments, or samples manipulated to explore audio features like pitch, spectral centroid, onset time, loudness, etc.,) and also further extending the modelling to machine learning models such as decision trees, Support Vector Machines (SVM), Neural Networks, etc. The final goal would be to finally comprehensively assess the blending of sound sources as a time-varying parameter by also incorporating room acoustic contribution.

3.4 Summary

This investigation demonstrates the feasibility of classifying musically realistic sound samples based on their overall blending impression perceived between two musical instruments in a unison performance. Musical score-independent monophonically rendered sound samples extracted from in-situ recordings of an ensemble performance were used for the perceptual evaluation of blending impressions and the corresponding classification modelling. The results show that the Linear Discriminant Analysis paired with the Euclidean distance measure performed on the raw MFCC features extracted from the sound samples is an effective method of classifying these sound samples into blended and non-blended classes. The model was tested and verified using a separate train-test data set, and leave-one-out cross-validation which showed an accuracy of 87.5%, and 87.1% respectively. This outperforms the other models tested in this study which were developed using the PCA and t-SNE transformations of raw and standardized MFCC features. Unlike the previous research on the estimation of source-level blending impression which employed musically constrained sound sam-

ples (such as notes or chords) of instruments from sophisticated recording conditions, this study surpasses earlier limitations by implementing the classification of 'ecological' sound samples of joint performances even without accessing the individual source recordings.

This investigation serves as a proof of concept for the capability of feature transformations to categorize the perception of multivariate psychoacoustic blending phenomenon, despite the fact that the perceptual rating scale is subjective and may vary depending on the listeners' backgrounds and the characteristics of the audio samples. The proposed method could be further expanded using a larger sample size and applied in various domains such as joint musical performance training, and real and virtual orchestral sound evaluation because it is independent of the musical content of the signal and does not require access to the individual source signals. Given that the method is applicable for sound signals from joint performance recording with minimal microphone cross-talk and background noise, this expands its utility in in-situ applications.

Chapter 4

Quality assessment of auralization of ensemble sound

An accurate representation of individual sound sources is required to achieve a perceptually convincing auralization of joint musical performances. Therefore, it is essential to capture clean signals of each musical instrument in the joint performance with minimized microphone cross-talk and room acoustic feedback for this purpose. Recording instruments in anechoic environments is a widely used method, but it typically lacks the natural and intrinsic characteristics of a joint performance due to limited room acoustic and inter-musician feedback. An alternative to this is to use close miking techniques to capture individual sound sources in a joint performance. Although these recordings have the potential to capture authentic attributes of joint performance, the challenge here is to improve the quality of recordings by minimizing the microphone cross-talk and room acoustic contribution in reverberant environments. This study investigates the perceptual quality of auralization of an ensemble using clip-on microphone recordings in comparison to a binaural recording of musical performance. The content of this chapter is reproduced from the following research article with the permission of the Acoustical Society of America:

J. Thilakan, O. C. Gomez, E. Mommertz, M. Kob, "Pilot study on the perceptual quality of close-mic recordings in auralization of a string ensemble", Proceedings of Meetings on Acoustics, vol. 49, Acoustical Society of America, (2023), https://doi.org/10.1121/2.0001686.

4.1 Materials and Methods

4.1.1 Preparation of sound samples

The close-microphone recordings of an in-situ performance of a string ensemble with a different number of violins ranging from 1, 2, 3, 4, 6, and 9 was utilized in this investigation (from chapter 2). The ensemble performance and its recording was conducted at Detmold Concert House (the details of the performance and the recording techniques are presented in the section 2.1), and the ensemble recordings with natural acoustic conditions of the concert house (T_{30}= 1.6 s) were utilized in this investigation.

DPA 4099 clip-on microphones were used to capture individual source signals by attaching them to the instrument and placing them close to the bridge of it. These microphones have a frequency response of 20 Hz – 20 kHz with an effective frequency range of 80 Hz–15 kHz ($\pm$ 2 dB) at 20 cm distance, and possess a super-cardioid directivity [113]. Due to this directivity characteristic, the contribution of the room acoustic reflections was minor in the recordings. However, in a pilot listening test, the recordings are reported to have a non-trivial amount of cross-talk between the sound sources. Head Acoustics HSU III.2 binaural head having head-shoulder unit and ICP measurement microphones (±2 dB for 3.5 Hz – 20 kHz) [115], placed in the far-field of the concert house was used to capture the sound field of the musical performance, and thereby serve as the reference. The seating arrangement of the instrument and the position of the head for this specific investigation is as shown Figure 4.1.

The musicians in the ensemble were separated by roughly 80 cm to 1 meter of seating distance, and equal gain was applied for all the DPA microphone tracks in the sound card. Due to the relatively adjacent position of the DPA microphones, the major issues that deteriorated the quality of the clip-on mic recordings were the breathing noise of the musician, noise from bowing & tapping on the fingerboard, etc. To reduce the noisy components due to breathing, etc., a smooth high-pass filter centered at 200 Hz was applied to the DPA recordings using REAPER digital audio workstation. For later direct comparisons with auralized signals, the same filter was applied to the binaural recordings to avoid spectral dissimilarity. Two different musical fragments with fixed melodies having a length of 5-6 seconds were extracted for different numbers of violins from the DPA microphone signals and their corresponding Binaural recordings to serve as two independent observations. The musical fragment with the lowest noise level in the recording among the multiple repetitions of the performance in each condition was chosen for the extraction.

Using good quality close mic recordings of individual sources, the sound field of the joint musical performance was reconstructed by convolving them with the Binaural Room Impulse Responses (BRIRs) of each source estimated from two independent methods; (1) in-situ measurements conducted at the concert house using a loudspeaker,

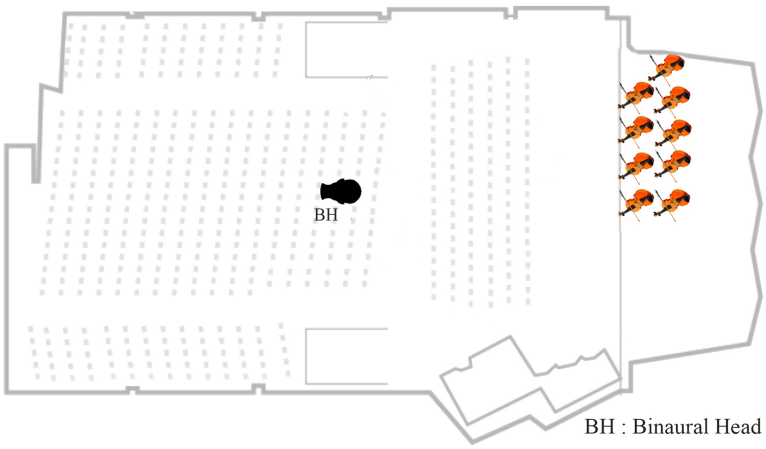


Figure 4.1: Position and orientation of violins and the binaural head (denoted as BH in the figure) in the concert house.

(2) a room acoustic simulation of the concert house. Employing a loudspeaker (typically a studio monitor with flat frequency response) to measure BRIRs has limitations due to its difference in the directivity with the musical instrument which would lead to the difference in the spatial distribution of the radiated energy. However, similar methods have been followed in earlier studies to have an approximate representation of sound sources like violin using a simplified loudspeaker setup in concert halls [70]. Implementation of GA-based room acoustic simulation, which employs a lower computational effort in comparison with a wave-based acoustic simulation is another well-used alternative to re-create acoustic environments. The GA-based room acoustic simulations have limitations in dealing with complex wave phenomena and are observed to have perceptual differences from actual rooms, but they also have benefits like the ability to estimate Spatial Room Impulse Responses (SRIRs) for sources with specific directivities, ease of changing the acoustic properties of materials, absence of background noise and distortion, etc [146]. The process of estimation of BRIRs using these two independent methods for auralization is described in the following section.

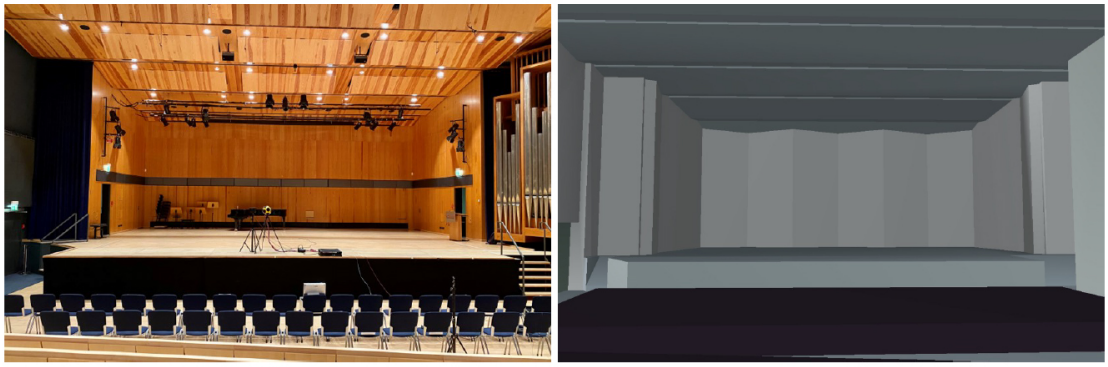
Measurement of BRIR from the concert house:

A set of acoustic transfer function measurements was carried out in the concert house using Neumann KH120 A, a commonly used studio monitor speaker, and the HSU III.2 binaural head to obtain the individual BRIRs for each source-to-receiver position. KH120A loudspeaker consists of a 5.25" woofer and 1" tweeter with a frequency response of $\pm$ 3 dB for 52 Hz–21 kHz, and it exhibited a directional characteristic similar to a trumpet[147]. The loudspeaker was held at a height of 1 metre, and oriented towards the front side of the musician's orientations. An exponential sine sweep signal

(f_s = 44100, number of samples = 65536) averaged with three repetitions was used to measure the transfer function of each channel of binaural recording. The BRIRs obtained for each individual sound source were convolved with the DPA recordings to produce auralized sound samples. The recorded samples were extracted from a musical composition without a reverberation tail using a smooth fade-out filter. Therefore, following the same procedure, the reverberation tail of the convolved samples was also removed using a fade-out filter for direct comparisons.

Estimation of BRIR from the simulation of concert house:

This study utilized ODEON version 17, a commercially used GA-based room acoustic simulation software, to generate virtual room acoustic environments. Since ODEON incorporates state-of-the-art room acoustic simulation techniques (detailed in section 1.2.4) and yields reliable results, it has been employed in many auditory perception-related investigations for simulating acoustic environments [68; 99; 100]. A geometrical room acoustic simulation model (GA model) of the Detmold concert house was created and developed in ODEON version 17, and the scattering and absorption coefficients of the materials used in the concert house were chosen within their typical range of physically valid values in the simulation. The picture of the Detmold concert house from the listeners' area and its corresponding view in the GA model in ODEON are given in Figure 4.2.



(a) (b)

Figure 4.2: (a) Picture of Detmold Concert House from the listeners' position on the left, (b) Corresponding view in the GA model in ODEON.

To start with the optimization of the GA model of the concert house, the impulse response of the hall for an omnidirectional loudspeaker on stage and an omnidirectional microphone at the location of the binaural head (as shown in Figure 4.2a) was measured, and important acoustic parameters, such as Early Decay Time (EDT), Reverberation time (T_{30}), and Clarity index (C_{80}) were estimated (described in Appendix

A) from the measured impulse response. Subsequently, the room acoustic parameters from the ODEON simulation of the concert hall were obtained for the same source-receiver positions and compared with the measured values. Similar to procedures followed in earlier studies on the reconstruction of acoustic environments [23; 19], the GA model was further optimized by modifying the absorption and scattering coefficients of the materials within the physically reasonable range in order to achieve similarity in measured and simulated IRs by bringing EDT, T_{30}, and C_{80} parameters closer to the measured values. The room acoustic parameters estimated from the RIR measurement and optimized simulation for different frequency bands are provided in Table 4.1.

Parameter	RIR source	250 Hz	500 Hz	1000 Hz	2000 Hz
EDT (s)	Measurement	1.24	1.37	1.63	1.49
	Simulation	1.23	1.35	1.65	1.44
T_{30} (s)	Measurement	1.57	1.59	1.65	1.56
	Simulation	1.29	1.36	1.62	1.39
C_{80} (s)	Measurement	2.32	0.34	-1.24	2.54
	Simulation	1.80	0.5	-1.5	-0.3

Table 4.1: Room acoustic parameters estimated from measured and simulated RIRs for different frequency bands.

The Just Noticeable Difference (JND) values of these parameters show the least difference in the parameter value that a human ear can detect. According to the ISO standards with frequency averaging across 500-1000 Hz octave bands [37], the Early Decay Time (EDT), which better reflects the perception of reverberation, estimated from the simulation (1.50 s) is in good agreement with the measurement (1.50 s $\pm$ 0.07 s JND). Similarly, the C80 value of the simulation (-0.90 dB) is within the JND of measurement (-1.00 dB $\pm$ 1 dB JND). The T30 value estimated from the simulation (1.49 s) is marginally different from the measurement (1.62 s $\pm$ 0.08 s JND) with an 8% deviation compared to the 5% JND level. This optimized GA model having a considerable degree of similarity with the measured room acoustic parameters was finalized and taken further for the auralization of the ensemble.

Sound sources were placed in the ODEON model according to the location of musicians during the performance of recordings, and an inbuilt directivity of violin which is averaged over 1/3rd octave bands with spatial resolution of 5° was applied on each source. The SRIRs from each sound source to the pre-defined location of the binaural head was obtained as third order (16 channel) ambisonics file. To auralize sound samples with different number of sources, the individual DPA recordings of each source were convolved with its corresponding SRIR obtained from ODEON using the MCFX convolver plug-in in the REAPER digital audio workstation. The gain of each source was adjusted according to the attenuation factors obtained from ODEON. The HRTF of

the HMS II binaural head from Head Acoustics which has the same geometrical shape of HSU III.2 that is designed in conformity to ANSI S3.36 was used for transforming the 3D audio file into the binaural format, and it was performed using SPARTA AmbiBIN plugin [148]. After rendering the samples, the reverberation tail at the end of the sample was cropped out using a fade-out filter as followed in the earlier cases.

4.1.2 Perceptual evaluation of sound samples

A group of thirteen ear-trained expert listeners that included tonmeister students and musicians participated in the listening test. All of them had a background in musical and technical ear training and they were very accustomed to the sound of the string ensemble. With a background in music and experience in critical listening, they are expected to be more capable and sensitive in selectively scrutinizing and evaluating the complex spectral and temporal features of sounds than non-musicians [17; 18]. The objective of the study and the characteristics of sound samples were not disclosed to them at the beginning of the test to prevent pre-judgmental biases. A familiarization audio file of the binaural recording of the string ensemble performance was provided at the start of the test to make the listeners familiar with the binaural head recordings. Then, a dedicated Graphical User Interface designed using MATLAB's app designer as shown in Figure 4.3 was used to perform the listening test.

The application consisted of 24 trials with a pair of sound samples in each (6 source combinations $\times$ 2 musical stimuli $\times$ 2 synthesized methods), and the order of pairs of sound samples was randomized according to the number of violins and the musical stimulus to minimize the direct comparison due to retaining memory of listening. Each trial consisted of two types of sound samples: one binaurally **recorded sample** (abbreviated in this study as **"Rec"**), and the corresponding synthesized counterpart which is auralized using either a direct convolution with measured BRIRs (referred to as **convolved sample** and abbreviated as **"Conv"**) or the BRIRs obtained from GA simulation (referred to as **simulated sample** and abbreviated as **"Sim"**). These two kinds of samples (recorded and synthesized) were randomly assigned to samples A and B in the GUI of the application.

The participants were asked to rate two aspects of sound samples in each trial: firstly, to rate the naturalness (realism) of each sound sample on a scale of 0 to 10, and secondly to rate the degree of similarity between the two samples on a scale of 0 to 10. A higher value rating on the scale referred to a very natural/highly similar impression. The listeners had an opportunity to perform the test remotely, and the listeners were permitted to play back the samples multiple times as they desired. The test was performed using Beyerdynamic DT 770 Pro headphones in an acoustically treated room, and it took an average of 15 minutes to complete the test. Finally, a short discussion with each participant was conducted to review the perceived characteristics of the sound samples.

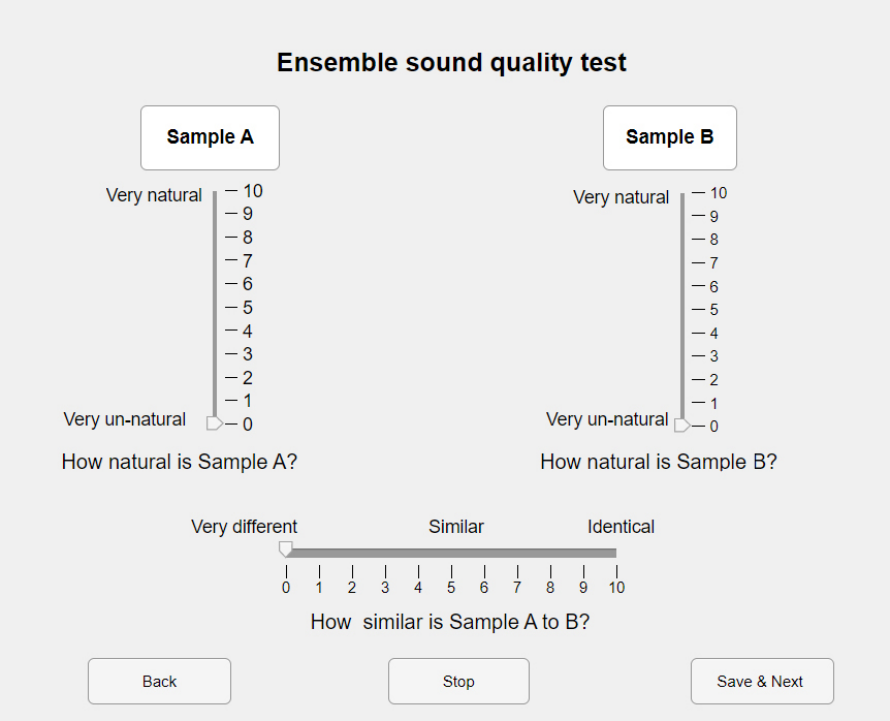


Figure 4.3: Graphical user interface of the listening test application.

4.2 Results and Discussion

4.2.1 Assessment of naturalness of sound samples

The distribution of naturalness ratings of the recorded and convolved sound samples for different numbers of violins is provided in Figure 4.4. Since the assessments of samples having two different musical stimuli were regarded as individual observations in the test, each box plot in Figure 4.4 includes 26 observations (13 participants $\times$ 2 stimuli). The results show a trend that the binaural recordings are not always rated to be highly natural in comparison to the convolved samples, and their naturalness ratings decreases as the number of violins increases as reflected by a reducing median value and a shift in the interquartile range (IQR) toward lower ratings. Moreover, in most conditions, recorded and convolved samples exhibit a relatively comparable distribution in naturalness ratings, as indicated by the close median values and overlapping IQRs. Mann-Whitney U test [141] was conducted to compare the statistical differences in naturalness distributions between real recordings and convolved samples for different numbers of violins. As discussed in chapter 3, this non-parametric alternative to Student's t-test was chosen due to violations of normality in the distributions, as confirmed by the Shapiro-Wilk test [149] The test results show that while the recorded

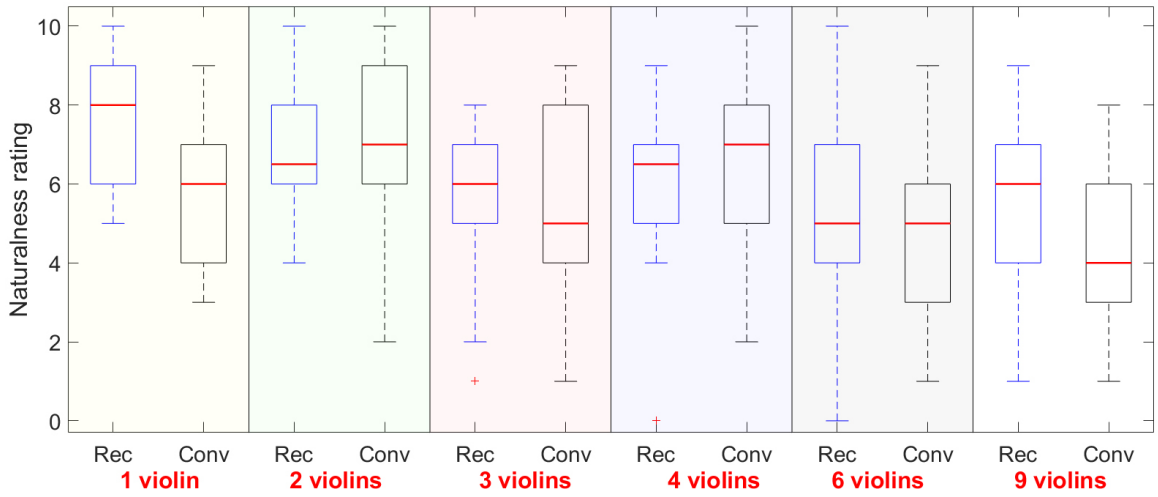


Figure 4.4: Distribution of naturalness ratings between recorded and convolved audio samples.

and convolved samples exhibit a statistically significant difference in their distributions for 1-violin condition ($p<0.001$), no statistical evidence was found to conclude that the pairs of distributions differ significantly in the other five conditions (2, 3, 4, 6, and 9 violins) at the 5% significance level. While this result does not confirm that the pairs of distributions are similar, the lack of strong statistical evidence for a difference supports the argument that close microphone recordings can be capable enough to deliver a perceptually convincing and natural auralization of joint musical performance in comparison with a real binaural recording.

Figure 4.5 shows the distribution of naturalness ratings for the recorded and simulated sound samples for different numbers of violins (26 observations in each box plot). Corresponding to the preceding observation, the naturalness of the recorded sound samples is observed to decline with an increase in the number of violins as reflected by a reducing median value and a shift in the interquartile range (IQR) toward lower ratings. This indicates the deficiency of the incorporated binaural recording and reproduction techniques in the re-synthesis of a relatively complex sound field with perceptually convincing authenticity and realism.

When comparing the recorded and simulated sound samples, the distribution of naturalness ratings of simulated samples is shown to be significantly different from that of the recorded samples across the 6 conditions, as shown in Figure 4.5. The distinction between recorded and simulated samples is observed to diminish with an increase in the number of violins. However, The Mann-Whitney U test conducted on the pairs of distributions of naturalness ratings for recorded and simulated samples suggests a statistically significant difference between the pairs of distributions in all six conditions at the 5% significance level. This reveals the limitation of simu-

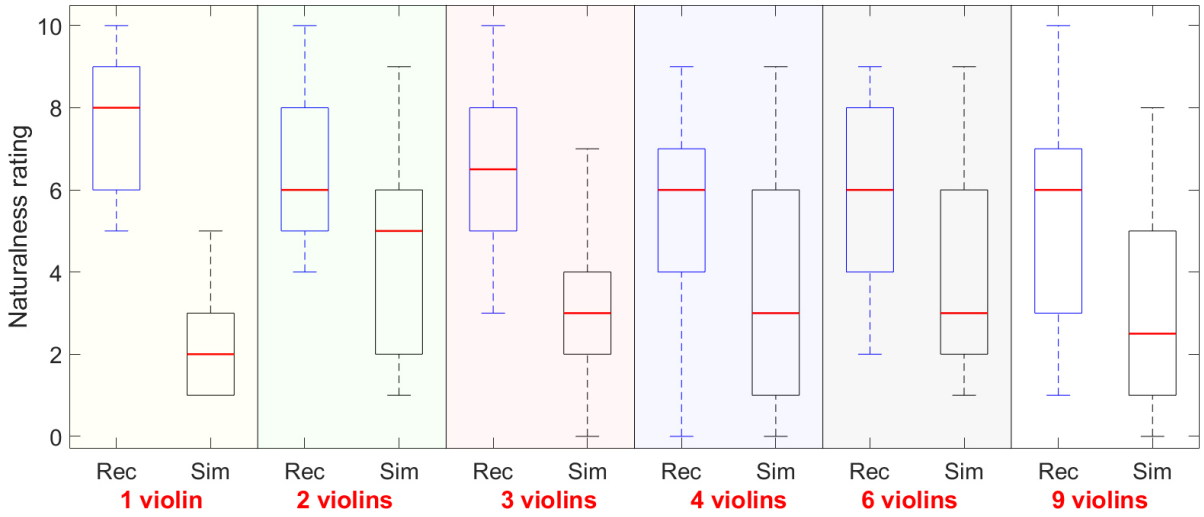


Figure 4.5: Distribution of naturalness ratings between recorded and simulated audio samples.

lated samples in conveying a perceptuall natural impression in comparison to the real recordings. The dominance of high-frequency components in the simulation outputs is reported in earlier studies on the reconstruction of sound fields using GA room simulations, along with a proposal for implementation of a filter to adjust for it [19]. This could be a potential reason for the degraded naturalness impression of the auralized sound samples. Additionally, test participants reported that some of the sound samples of joint performance were significantly distinguishable due to timbre difference, low spatial envelopment, and narrow apparent source width, which could account for the trend of lower naturalness impressions with simulated samples.

	1 violin	**2 violins**	**3 violins**	**4 violins**	**6 violins**	**9 violins**
Convolved	0.056	0.242	0.559	0.537	0.655	0.662
Simulated	0.007	-0.036	-0.029	0.126	0.294	0.376

Table 4.2: Lin's concordance correlation between the naturalness rating of recorded and synthesized (convolved and simulated) samples.

Lin's concordance correlation coefficient [150] was employed to assess how well the naturalness ratings of the recorded and auralized samples conform relatively to each other. It is a statistical method to assess the reproducibility of measurement or the inter-rater reliability by calculating how closely the two measurements of the same variable lie on the 45-degree line through the origin. The concordance correlation between the naturalness ratings of the recorded samples (reference) and auralized samples (replica) was estimated, and the correlation coefficients are provided in Table 4.2.

The correlation coefficient values show a systematic trend in which, the naturalness ratings of the auralized samples tend to be closer to the recorded samples with an increase in the number of violins, and this holds for both convolved and simulated samples. Similar to the observations from Figures 4.4 and 4.5, in comparison to the simulated samples, the naturalness ratings of convolved sound samples have a relatively higher correlation to that of recorded samples. Moreover, the observation remains valid for all the cases with different numbers of violins involved. This indicates the limitations of geometry-based room acoustic simulations in re-synthesis of the ensemble performance.

4.2.2 Similarity between recorded and auralized samples

The similarity of convolved and simulated samples with the recorded samples was measured using a rating scale, and the distribution of similarity ratings is presented in Figure 4.6 (26 observations in each box plot). Analogous to the perceived impression of naturalness, the convolved samples are assessed to be considerably more similar to the recorded samples than the simulated ones, as indicated by higher median values non-overlapping IQRs. The Mann-Whitney U test conducted on the pairs of distributions of similarity ratings suggests that there exist a statistically significant difference between the pairs of distributions for all the 6 conditions at a 5% significance level. This indicates a significantly higher simiarity impression of convolved samples to recorded ones as compared to simulated ones. With an increase in the number of violins, the similarity ratings of the convolved samples appear to improve, evidenced by the increasing median value and upward-shifting trend in IQR. However, no significant trend is observed for the simulated samples across the 6 conditions.

Based on the distribution of similarity ratings shown in Figure 4.6, the convolved samples are not perceived as extremely similar to or identical to the recorded ones in any of the six cases. The difference in timbre colouration of instruments which can occur due to close-micing of the clip-on microphone recordings could be a contributing factor for that. In addition, the difference in the directivity of the sound sources (violin, and the studio monitor) would be favoring reason for lowering the similarity rating of convolved samples. The difference in spectral coloration and variation in spatial impression such as source width and envelopment, as reported by the listeners, could be the potential reason for the poor performance of the samples created using GA-simulation. However, the shortcomings of GA-based simulations in reconstructing the sound field of an ensemble performance need further research. This can involve fine-tuning the GA model by matching the strength and spatial distribution of major early reflections and incorporating perceptual experiments with an extended vocabulary of perception-oriented verbal attributes, such as the Spatial Audio Quality Inventory (SAQI) [151].

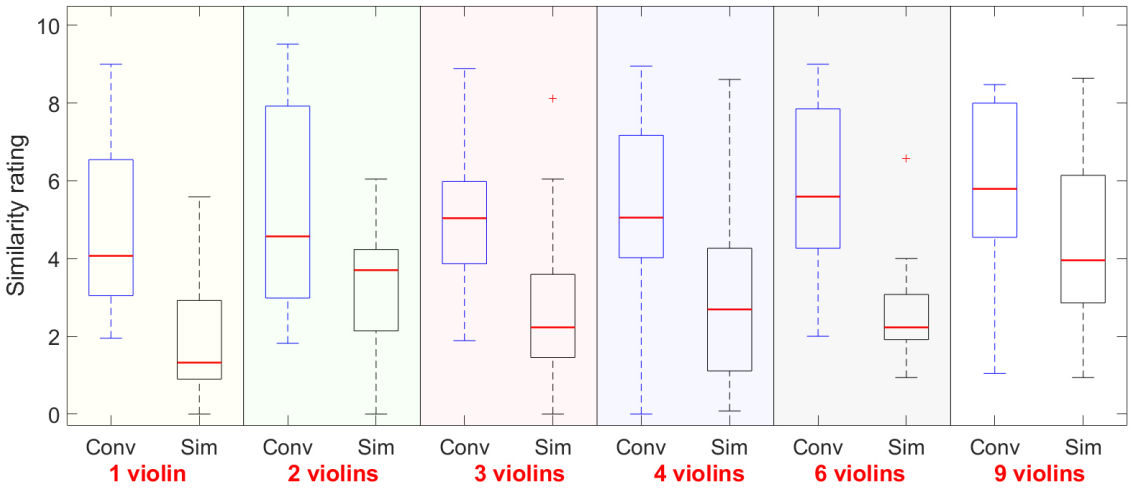


Figure 4.6: Distribution of similarity ratings of convolved and simulated samples with the recorded samples.

4.3 Summary

The perceptual quality of clip-on microphone recordings in auralization of an ensemble was investigated in this study by evaluating the perceptual attributes of auralized sound samples in comparison with the binaural recordings. Sound samples from binaural and clip-on mic recordings of a string ensemble with different number of violins were collected, and its auralization using the RIRs from in-situ measurements and GA simulation of the concert house were generated. Afterwards, the perceptual attributes including impression of naturalness and perception of similarity between binaurally recorded and auralized sound samples were evaluated.

The findings reveal that the binaural recordings prepared are not always rated to be highly natural, especially with an increase in the number of violins. Nevertheless, the clip-on microphone signals auralized using in-situ BRIR measurements have a similar distribution of naturalness impression to that of binaural recordings whereas the auralization using RIRs from GA simulation showed a poorer naturalness impression. This demonstrates the applicability of clip-on microphone recordings for auralization-related applications. In both auralization methods, the naturalness was shown to improve with an increase in the number of violins in the ensemble and thereby tending to mask the deficiencies in the clip-on mic recordings. Furthermore, analogous to the preceding findings, samples generated using measured BRIRs were consistently rated to be more similar to the actual recording than to the simulated BRIRs. This highlights the deficiencies of geometry-based room acoustic simulations in re-synthesis of complex acoustic sound fields.

The results derived from the study are only based on the evaluation of representative sound samples with the minimal background noise level. In realistic cases, the artefacts due to noise from the musician or the instrument should be accounted for in the auralization. These results propose further investigations on the degree of perceived realism/ authenticity in different binaural recording and reproduction methods for relatively complex acoustic environments, and the quality requirements needed for the GA simulations for the reconstruction of sound fields.

Chapter 5

Sound source orientation perception in in-situ conditions

While source localization has been a key topic in auditory perception research and speech communication, the perception of source orientation is seldom researched. Although orientation perception holds great importance in VR,XR domains, factors influencing it such as source directivity require detailed evaluations. This study aims to explore the relevance of sound source directivity in orientation perception by utilizing five distinct musical instruments that demonstrate a broad spectrum of directivity characteristics, in ecological conditions. Furthermore, the significance of room acoustic environments in orientation prediction is further examined by incorporating three acoustic environments in the study, spanning from a recording studio to music performance halls, each characterized by its own unique acoustic properties. The in-situ performance of musical instruments for four source orientations with 90-degree spacing – front, back, left, and right – are recorded using a binaural head positioned in the far field of the three acoustic environments, and predictability of sound source orientation of these samples are analyzed for each instrument and room variants. Additionally, the potential features that influence the orientation perception including binaural, and monaural parameters are evaluated and presented. By incorporating an 'ecological' performance condition for the source orientation perception, this study is expected to offer insights relevant to music performance, sound recording techniques, and virtual reality applications. The content of this chapter is reproduced from the following research article:

J. Thilakan, B.T. Balamurali, W. Buchholtzer, J.M. Chen, Malte Kob, "Source orientation perception; exploring the role of directivity of sound sources in diverse acoustic environments," (manuscript under preparation).

5.1 Materials and methods

5.1.1 Characterization of acoustic environments

To understand the influence of room acoustic attributes on the perception of source orientation, three different music performance spaces with distinct acoustic characteristics were chosen for this study which comprise the Recording studio-1 of Erich Thienhaus Institute (abbreviated as 'RS' henceforth), Sommertheater Detmold (abbreviated as 'ST'), and Brahmssaal of HfM Detmold (abbreviated as 'BS'). The Recording studio (volume of ca. 110 m^3) is an acoustically treated room specifically intended for music recording purposes. The Sommertheater (volume of ca. 2930 m^3 with a seating capacity of 320) serves as a spacious venue utilized for a variety of events, ranging from stage plays and music performances to presentations and symposiums. In contrast, the Brahmssaal (volume of ca. 775 m^3 with a seating capacity of 110) is a dedicated room tailored for musical performances such as solo performances, chamber music, and small ensemble presentations. The room acoustic parameters comprising the reverberation time (T_{30}), Early Decay Time (EDT), and Clarity index (C_{80}) were estimated from these performance spaces in accordance with the ISO standards [37] (detailed in Appendix A) using omnidirectional sound source and NTi M2010 measurement microphone, and their averaged values for 500-1000 Hz octave bands are presented in Table 5.1. The extracted parameters indicate that the perceived sound field in the recording studio would exhibit lower reverberation (evident from EDT and T_{30} values) and enhanced clarity (evident from C_{80} value). Conversely, the Brahmssaal is expected to have a more reverberant acoustic environment with a diminished clarity perception. Meanwhile, the Sommertheater, characterized by a moderately reverberant sound field, occupies an intermediate position in between.

Parameter	Room Acoustic environment		
	Recording studio	Sommertheater	Brahmssaal
EDT (s)	0.16	0.94	1.39
T_{30} (s)	0.20	1.02	1.23
C_{80} (dB)	23.77	3.44	1.20

Table 5.1: Room acoustic parameters assessed from the three acoustic environments (averaged for 500-1000 Hz octave bands).

5.1.2 Preparation of sound samples

Five musical instruments including trumpet, trombone, transverse flute, saxophone, and violin were chosen as the sound sources for the orientation perception investigation, and the directivity characteristics are observed to be different across these instruments 1.2.3. While trumpet and trombone have relatively similar directivities,

they were included as two independent observations in the objective investigation conducted in this chapter (detailed in the coming section).

Given that directivities are frequency-dependent, a dedicated musical fragment that covers the entire pitch range of each instrument was composed for each one to excite all the possible directivity patterns (these compositions from [152] are presented at Appendix B). The position and orientation of sound sources and receivers in the three-room acoustic environments are illustrated in Figure 5.1 (this top-view schematic is only made for visual comparison, and the rooms are not on the same scale). To minimize undesirable acoustic effects within the room, both the sources and receivers were deliberately positioned off-centered from the symmetrical axis of the room. The sound source was positioned on the stage in both the Sommertheater (ST) and Brahmssaal (BS), while a typical recording position was selected in the Recording studio (RS). The four potential orientations of the sound sources denoted as 'F', 'B', 'L', and 'R', represent the orientation of 'the musician with the musical instrument'. Neumann KU-100 Binaural head was utilized in this study to capture the sound field as binaural recording. The binaural head was consistently positioned in the far field (i.e., beyond the critical distance limit) in all three rooms, thus minimizing the influence of the acoustic parallax effect from the direct sound [71]. Furthermore, the binaural head was consistently aligned along the frontal source orientation.

In each source orientation condition across the three rooms, musicians were instructed to perform the designated musical fragment with maximum consistency in tempo, articulation, intonation, and dynamics, to make the samples as similar as possible. A metronome was also provided to the musicians to support this throughout the performance. After performing the musical piece for a particular orientation, the musicians were instructed to revolve by 90 degrees while maintaining the acoustic center of the instrument fixed at the specified source location, until they performed across the four orientations. With this approach, the distance between the acoustic center of the instrument and the receiver is expected to remain constant during the rotation, and as a result, potential differences arising from the shift of acoustic centre of the instrument in overall sound level or source localization are avoided. While instruments like the trumpet and trombone have relatively obvious acoustic centers due to a single radiating source opening, instruments such as the flute, saxophone, and violin exhibit complex radiation characteristics, leading to frequency-dependent acoustic centers. Nevertheless, the primary radiation points, such as the bell opening of the trumpet, trombone, and saxophone, the embouchure hole of the flute, and the bridge of the violin, were considered as the acoustic centers of the respective instruments in this study. Moreover, the receiver was positioned in the far field to mitigate the auditory parallax effect from sound sources, ensuring that Interaural Level Differences (ILD) primarily arise from room acoustic reflections. The shadowing effect of musicians in the rotation can be assumed to be negligible as it has shown a marginal deviation in the directivity

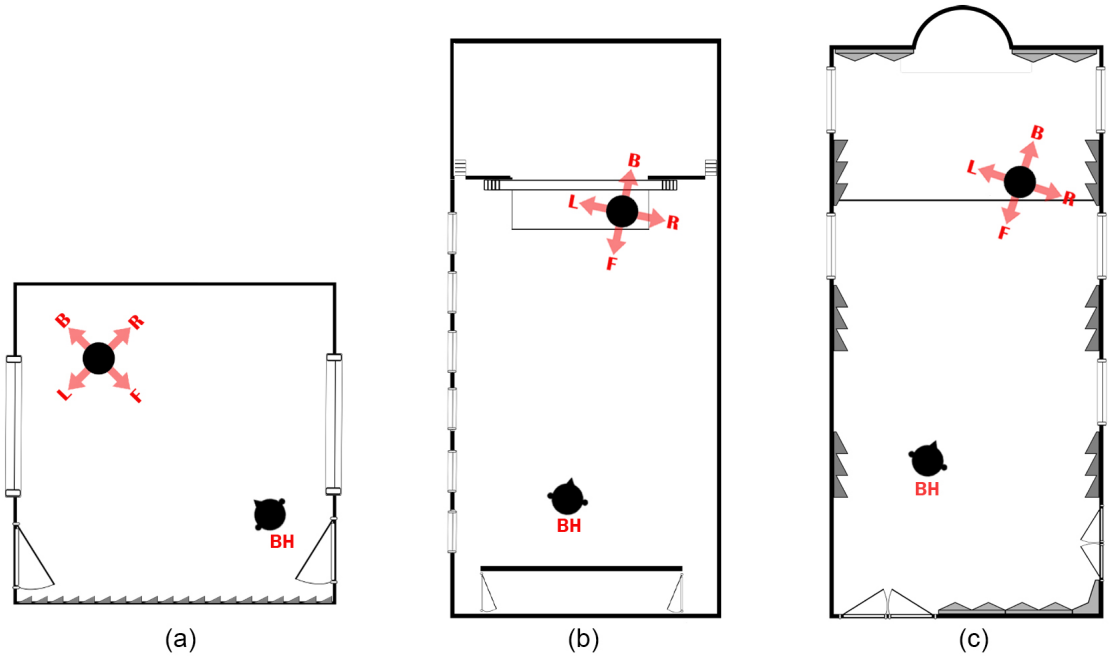


Figure 5.1: Schematic diagram of three acoustic environments denoting the position and orientations of the sound source and binaural head (BH) in the three acoustic environments; (a) Recording studio, (b) Sommertheater, (c) Brahmmsaal (the schematic made for visual comparison are not in the same scale).

measurements [67]. The sound samples were extracted from these binaural recordings using the REAPER Digital Audio Workstation. Given that the primary focus of the study is on the relationship between the room acoustics and the dynamically varying directivity patterns associated with a musical signal, the reverberation tail at the end of each sample was eliminated by applying a fade-out filter.

5.1.3 Perceptual evaluation of sound samples

A group of 15 participants, consisting of Tonmeister students and experienced musicians, participated in the listening test. All the participants had undergone musical ear training, possessed prior experience in critical listening assessment, and had a minimum of 12 years of musical experience. Research indicates that trained musicians typically exhibit ability and sensitivity in selectively analyzing and assessing the intricate spectral and temporal characteristics of sounds compared to non-musicians [17; 18]. Therefore, it was expected that the test participants would provide concordant test responses. The goal of the test was to predict the source orientation in each sound sample by identifying one of four possible directions (front, back, left, and right), solely based

on auditory cues. To mitigate potential confusion arising from the varied ways of holding the involved instruments by the musician while performing (e.g., the trumpet is held straight by the musician, while the violin is held at an angle), participants were instructed to predict the 'orientation of the musician holding the instrument' relative to the listener. A listening test application was created using MATLAB app designer with a dedicated Graphical User Interface (GUI) (see Figure 5.2) to perform the listening test. This application allowed participants to listen to each sound sample multiple times and make their orientation predictions accordingly.

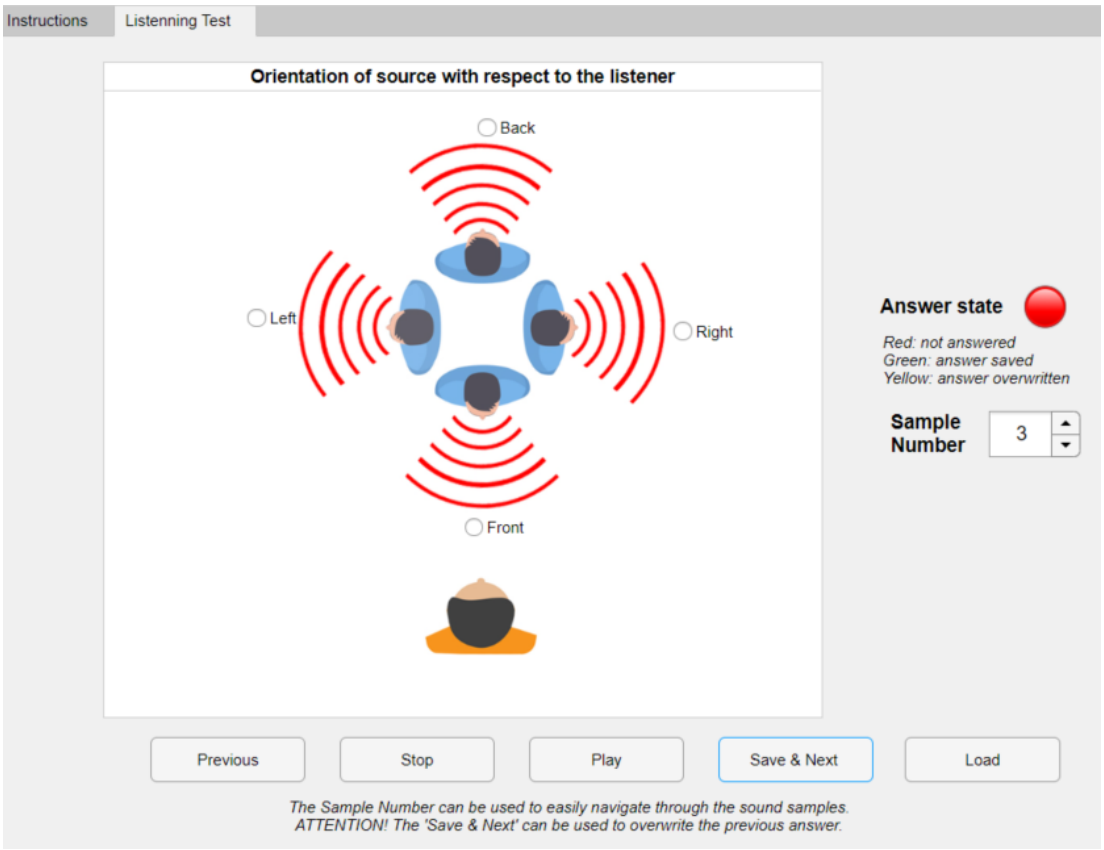


Figure 5.2: The Graphical User Interface (GUI) of the listening test application.

A training audio file containing the binaural audio samples of the instruments involved in the study was provided prior to the commencement of the listening test, with the aim of familiarizing listeners with the binaural head recording. The participants were permitted to adjust the gain during this training phase, but they were asked to maintain a fixed gain afterward for the performance of the listening test. Following the completion of the training phase, participants were introduced to the test GUI (see Figure 5.2), and instructed to proceed with the listening test on orientation prediction of

60 audio samples (5 instruments $\times$ 3 rooms $\times$ 4 orientations). The sound samples in the test were presented in a randomized order to avoid direct comparison between samples and mitigate the influence of memory retention effect on the ratings. Additionally, the randomization was unique for each participant to prevent any potential sequential effects in the sample ratings. The test was conducted in an acoustically treated studio room, utilizing Beyerdynamic DT 770 Pro closed-back studio headphones and an RME Babyface Pro sound card for playback of the binaural audio files from the computer. On average, participants took approximately 30 minutes to complete the listening test. Following the test, a brief discussion session was held to explore the potential subjective cues they found useful for orientation perception.

5.1.4 Extraction of acoustic features

Binaural Room Impulse Responses (BRIRs) were measured for each of the four orientations in the three rooms, using the Neumann KH120A studio monitor speaker as the source and the same Neumann KU100 binaural head as the receiver. The Neumann KH120A studio monitor loudspeaker is shown to possess a directivity profile that is very close to that of a trumpet [147]. Additionally, the trombone exhibits a similar directivity profile to the trumpet, albeit with a relatively lower pitch range. Consequently, the way the loudspeaker excites different room acoustic reflections for each orientation is expected to be similar to the way these instruments do. As a result, the interaural parameters derived from these BRIRs can serve as representatives of trumpet and trombone samples for each orientation condition, due to the similar directivity characteristics between the loudspeaker and these instruments. However, these parameters may not necessarily correspond with other instruments due to their different directivity characteristics compared to the loudspeaker.

Along with the established Interaural Level Difference (ILD) parameter in source orientation, an extended set of interaural parameters including Interaural Time Difference (ITD) and Inter Aural Cross Correlation (IACC) that are known for source localization and spatial perception [153; 154], were extracted from the BRIRs. ILD denotes the difference in sound level (intensity) perceived between ears, whereas ITD represents the difference in the time taken for sound to reach each ear; both parameters are considered pivotal in sound source localization within the horizontal plane [74]. IACC refers to the correlation between the signals received at the two ears, signifying the spatial impression of the Apparent Source Width (ASW) of the sound source. A high IACC value indicates a focused or centralized source with a narrow apparent source width, providing localization cues, while a wide and diffuse source typically yields lower IACC values [154; 155]. To analyze the significance of direct sound and early reflections versus late reverberation on source orientation perception, these interaural parameters were extracted specifically from the direct+early (0 - 80 ms) and late (80 ms - end) segments of the impulse responses of each source orientations within the three rooms

using the ITA-toolbox[156]. Following the proposed standards [37; 18; 153], a single value for each parameter representing their overall sensation was obtained by averaging across specific frequency bands, and utilized for each specific case: the ITD value was averaged across 250 - 1000 Hz octave bands, the ILD value was averaged across 500 - 4000 Hz octave bands, and the IACC value was averaged across 250 - 4000 Hz octave bands (the frquency band centered at 125 Hz was avoided, as it was removed from the audio samples due to background noise). While ILD and ITD may possess positive or negative values corresponding to dominant left or right ears, their absolute values were considered to evaluate their relationship with orientation prediction accuracy in subsequent sections.

Apart from the interaural parameters, spectral and temporal parameters such as Spectral Centroid (SC), and temporal energy ratios including Direct-to-Reverberant Ratio (DRR) and Clarity parameter (C_{80}) were computed from the individual channels of the binaural impulse responses. The spectral centroid parameter, denoting the 'center of mass' of the spectrum, is calculated as the weighted mean of the frequencies present in the signal, and it is observed to better represent the brightness of the instrument timbre [14].

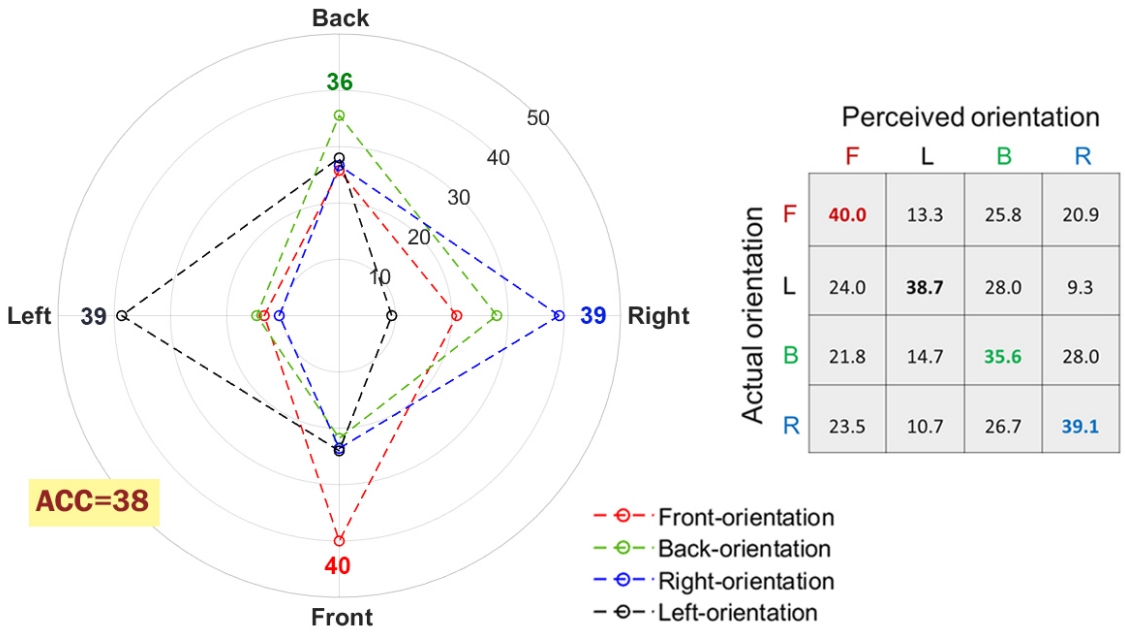
The influence of spectral centroid values from IRs in orientation perception was analyzed for two different scenarios here. Firstly, to explore the change in the 'spectral-tilt' in the direct sound component from the rotation of the sound source (referred in [76]), the centroid values were estimated from the direct sound (0-5 ms) of the BRIR. Secondly, to investigate the influence of spectral coloration differences in perceived instrument sound introduced by the various room acoustic conditions on orientation perception, centroid values were estimated from the 0-RT_{60} portion of the BRIR, which covers a 60 dB decay from direct sound by excluding the noise floor. For the first case, a single time window of 5 ms was utilized, while for the second case, a moving time window of 20 ms with an overlap length of 10 ms was employed. In both instances, a single centroid value is calculated as the mean of centroid values from the two channels of the BRIRs.

The DRR quantifies the strength of direct sound relative to room reflections, calculated as the logarithmic ratio between the energy of direct sound (0-5 ms) and the energy of room reflections (5 ms-end). Since it is on a decibel scale, the log is performed on the mean value of the ratios from the two channels. On the other hand, the Clarity parameter (C_{80}) assesses the strength of direct sound and early reflections (0-80 ms) compared to late reverberation (80 ms - end). While these energy ratio parameters are typically extracted from monaural room impulse responses and presented as an averaged value across multiple measurements [37], in this study we have extracted them from the BRIRs as individual observations to better align them with the perceptual impressions in each condition.

5.2 Results

5.2.1 Perceived source orientation in different conditions

The prediction rates of four actual (physical) source orientations in the four perceived conditions are presented as a radial plot as well as a confusion matrix in Figure 5.3 by averaging across all instruments in all room acoustic conditions. Rows in the confusion matrix show the percentage of responses perceived across the four orientations for each actual orientation, and they are depicted in the polar plot using different colors for improved visualization. The diagonal elements of the confusion matrix represent the prediction accuracy for each physical orientation, and they are highlighted in the appropriate directions of the polar plot. The incorrect responses of a particular orientation, i.e., mistakenly perceived orientations, are indicated by the other elements in each row of the confusion matrix. The accuracy (abbreviated as 'ACC') of overall orientation prediction averaged across all instrument, room, and orientation combinations is determined to be 38.3%, and it is also depicted in the plot.



Perceived orientation

Actual orientation	F	L	B	R
F	40.0	13.3	25.8	20.9
L	24.0	38.7	28.0	9.3
B	21.8	14.7	35.6	28.0
R	23.5	10.7	26.7	39.1

Figure 5.3: Overall prediction rates perceived in four orientations (in percentage scale) averaged across instruments and rooms, with prediction accuracy (ACC) of 38%.

Influence of orientation direction in prediction

The four source orientations are observed to have prediction accuracy within a range of 36-40% when averaged across all instruments and room combinations (see Figure 5.3). Although previous research suggests that the sound source oriented toward the listener, i.e., the front orientation, is relatively easier to predict than in other cases [80], such a trend is not evident in this study. This could possibly be due to the broader variations in the distinct directivity and acoustic characteristics of the musical instruments and performance spaces incorporated. The result also suggests that the misclassification of laterally oriented samples in the opposite directions, i.e., perceiving a left-oriented sample in the right direction and vice versa, was minimal in the overall perception of source orientation.

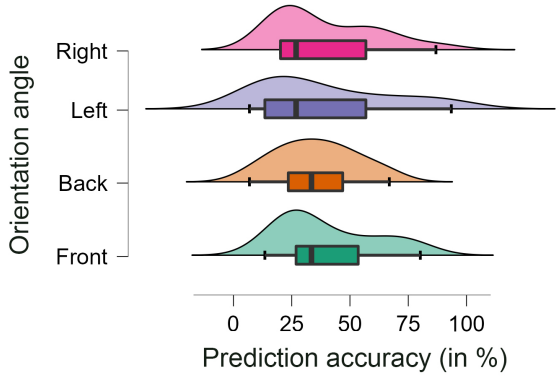


Figure 5.4: Distribution of prediction accuracies for four orientation angles (in %).

For a comprehensive understanding of the prediction accuracy variation of each source orientation condition, the distribution of individual prediction accuracies of the four orientations from the 15 conditions (5 instruments$\times$3 rooms) is depicted in Figure 5.4 using a box plot and the corresponding Probability Distribution Function (PDF). Except for the back orientation, the remaining three orientations appear to exhibit a skewed distribution, characterized by a tail in the PDF toward high prediction accuracies. Since the normality assumption is violated in the distribution of ratings for certain orientations (validated using the Shapiro-Wilk test [149]), the Kruskal-Wallis test is conducted to examine the statistical difference among the four groups under consideration. The Kruskal-Wallis test, the non-parametric alternative of one-way Analysis of Variance (ANOVA), evaluates the statistical difference between two or more groups by comparing their mean ranks [157]. The result of the Kruskal-Wallis test conducted among the prediction accuracies of four angles showed that there is no statistical difference between the mean ranks of the four groups examined ($\chi^2(3)=0.36$, $p=0.95$). These results suggest that the prediction accuracy for the front orientation is not significantly higher compared to the accuracies in the other three orientations.

As an extended evaluation, the Mann-Whitney U test [141] was performed to analyze whether there exists any statistical difference between the distribution of prediction accuracies of lateral (left and right) and medial (front and back) samples. As mentioned in chapter 3, the Mann-Whitney U test is a non-parametric version of Student's t-test that is utilized due to the non-normal distribution of the two classes [141]. The results showed no statistical difference between the prediction accuracy values of lateral and medial samples (p=0.77).

Role of instrument directivity in prediction

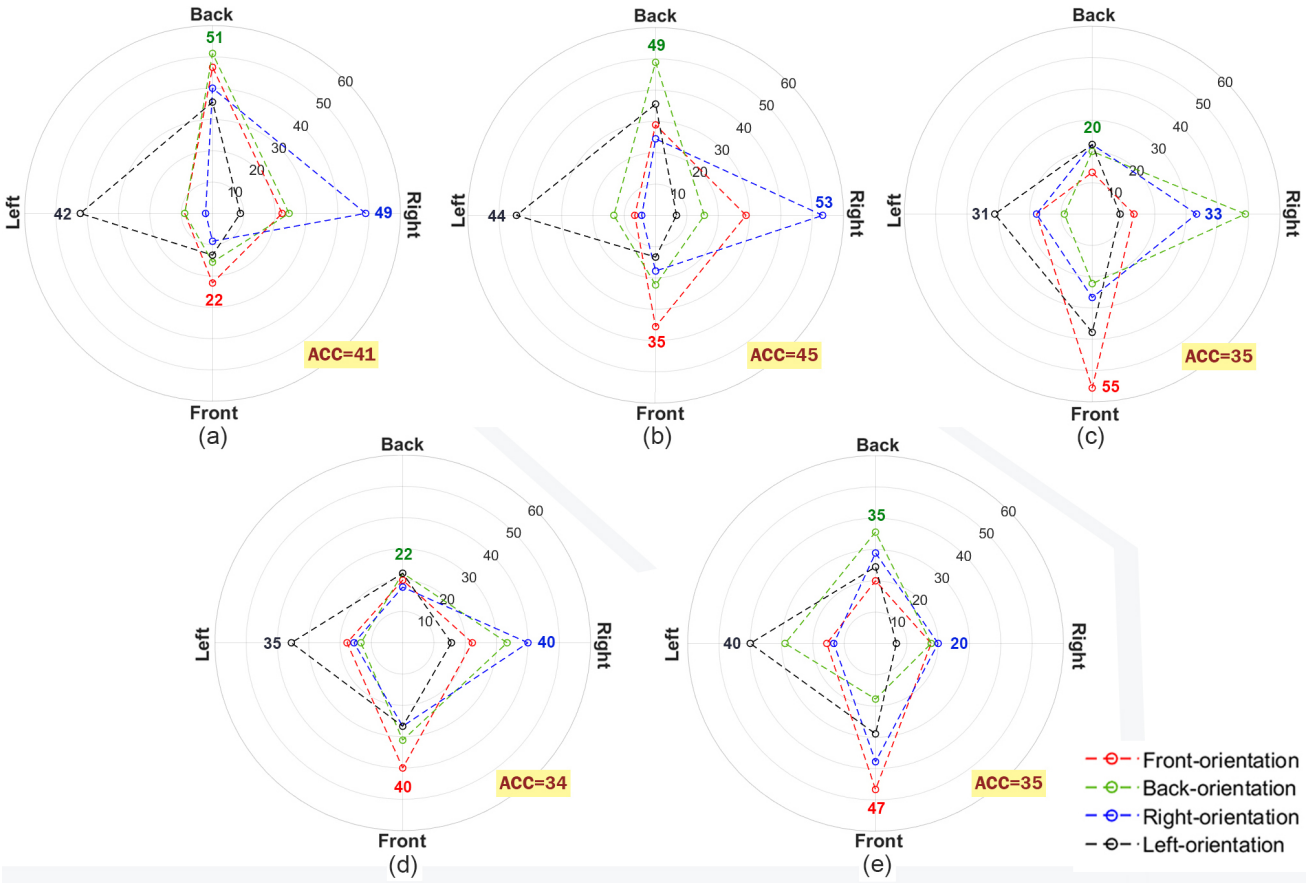


Figure 5.5: Variation of prediction rates in each orientation for five instruments; (a) trumpet, (b) trombone, (c) saxophone, (d) flute, (e) violin.

The variation in prediction rates of source orientation for the five musical instruments involved is presented in Figure 5.5, by averaging across the three different acoustic environments. Similar to the previous case, the perceived identification rates of each physical orientation in the four directions are illustrated, with the prediction ac-

curacy of each specific orientation highlighted in the figure. The overall prediction accuracy (highlighted in the figure as 'ACC') ranges from 34 to 45% among the instruments involved. Highly directional instruments such as the trumpet and trombone exhibit relatively higher accuracy (41%, 45% respectively) compared to instruments like the saxophone, flute, and violin (35%, 34%, and 35% respectively), which possess more complex directivity. However, upon examining the detailed responses of each instrument in the four orientations, it becomes evident that all the instruments have diverse prediction accuracies ranging from 20% to 55%, with some specific orientations having strong or weak predictability. The trumpet and trombone are shown to have relatively well-perceived across all orientations except the front one. Although they exhibit the lowest rate of lateral misclassification (perceiving left orientation as right, and vice versa), when it comes to the medial directions, the front orientation of the trumpet with the lowest prediction accuracy is shown to be perceived more to the back direction. Unlike these highly directional instruments exhibiting the lowest prediction accuracy towards the front direction (front accuracy of 22%, 35% respectively), instruments like the saxophone and violin demonstrate relatively stronger accuracy to the frontal direction compared to other orientations (front accuracy of 55%, 47% respectively). Given that the performers tried to play the musical stimuli with the maximum possible consistency, the spectral content radiated from the source is anticipated to be consistent across the four orientations. Therefore, these disparities in prediction accuracies of instruments in specific directions could potentially stem from the radiation characteristics of individual instruments that cause variations in the strength and coloration of direct sound, as well as different room reflections.

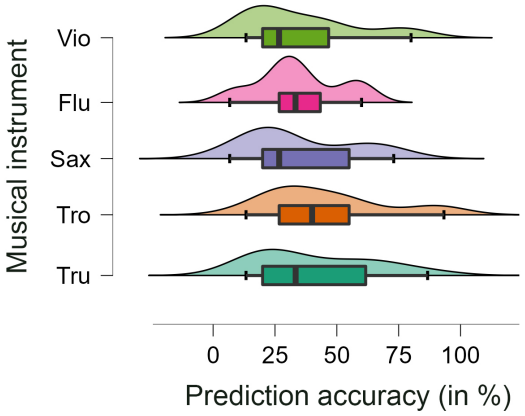


Figure 5.6: Distribution of prediction accuracies (in %) for five instruments; Trumpet (Tru), Trombone (Tro), Saxophone (Sax), Flute (Flu), and Violin (Vio).

The distribution of individual prediction accuracy values of the five instruments in 12 conditions (3 rooms×4 orientations) is presented in Figure 5.6 using box plot and

probability density function. Although the interquartile range and whiskers of the box plots largely overlap between the instruments, the trombone exhibits a relatively higher accuracy distribution, whereas the flute demonstrates a narrower distribution in relatively low accuracy values. The Kruskal-Wallis test, conducted due to the non-normal distribution of prediction accuracy ratings, indicated no statistical difference between the five instruments involved ($\chi^2(4)=2.24$, $p=0.69$). Therefore, although individual instruments exhibit relatively high and low prediction accuracies only in specific directions, likely due to their directivity features, no significant difference was observed among the instruments when comparing their prediction accuracies across all orientations in the different room acoustic environments. Consequently, no specific instrument with particular directivity characteristics appears to outperform others in this orientation perception study.

Role of room acoustics in orientation prediction

The variation of prediction rates of each source orientation across the three acoustic environments is illustrated in Figure 5.7. The recording studio, characterized by low reverberation with strong direct sound and early reflections (as indicated by EDT, T_{30}, and C_{80} values in Table 5.1), demonstrates an overall prediction accuracy of 48%. In contrast, the Brahmssaal, featuring a high reverberance with weaker clarity perception, achieves an accuracy of 28% which is just above the chance level of 25%. With a prediction accuracy of 38%, the Sommertheater environment falls between the other two environments in terms of accuracy, and this ordering seems to be consistent with the room acoustic parameters values as well.

Every orientation in the recording studio except the back one exhibits high prediction accuracy compared to the conditions in other rooms. However, the accuracy of the back orientation was 28%, slightly above the chance level, and it is observed to be more frequently misclassified in the front direction. This could be attributed to the presence of strong reflection from the back wall or the presence of high-frequency information in the perceived sound, however, advanced studies are required to validate this observation. On the other hand, misclassification among the lateral samples (perceiving left orientation as right, and vice versa) reaches its minimum in the recording studio environment ($<$5%). Within the Sommertheater, each orientation demonstrates moderate prediction accuracy, spanning from 31% to 40%. In contrast to the two mentioned acoustic environments, three orientations (front, left, right) in Brahmssaal are observed to possess prediction accuracies close to the minimum chance level of 25%, while only the back orientation in Brahmssaal demonstrates relatively improved prediction accuracy of 40%. Moreover, those three orientations (front, left, right) are noted to be predominantly misclassified towards the back direction with a prediction rate ranging between 30–40%, exceeding their true prediction accuracies. This tendency of perception of back orientation might possibly be due to the strength and spectral

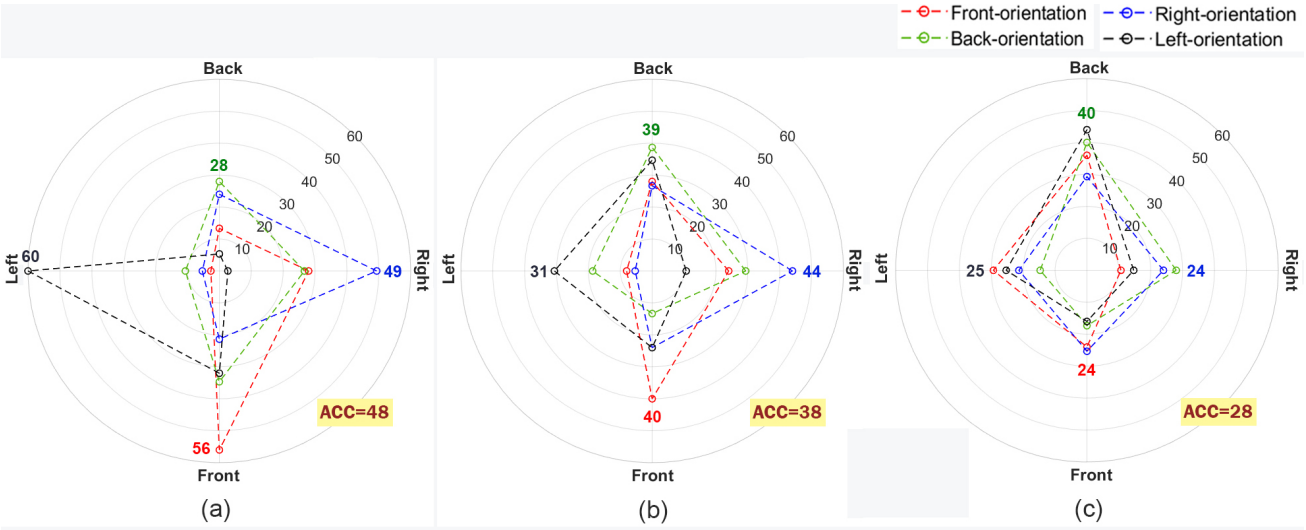


Figure 5.7: Variation of prediction rates in each orientation for the three acoustic environments; (a) Recording studio, (b) Sommertheater, and (c) Brahmssaal.

coloration of late reverberation of the room which needs further investigation. Based on the systematic trend of decreasing prediction accuracy with decreasing C_{80} and increasing EDT and T_{30} parameters, it can be hypothesized that high reverberance and low clarity sensations may negatively impact the perception of source orientation.

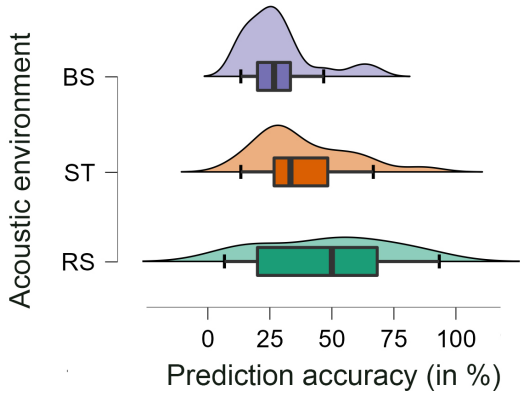


Figure 5.8: Distribution of prediction accuracies (in %) for three rooms: Brahmssaal (BS), Sommertheater (ST), and Recording studio (RS).

Figure 5.8 illustrates the distribution of prediction accuracies of 20 different conditions in the three acoustic environments involved (5 instruments×4 orientations). While the recording studio shows a median centered around 50% with a wide interquartile range (IQR), the median value is observed to be 27% in the Brahmssaal, accompanied by a narrower distribution with an IQR ranging between 20% and 32%.

The Kruskal-Wallis test conducted between the three groups of data showed that it failed to reject the null hypothesis that there is no difference in the mean ranks of the three groups ($\chi^2(2)=6.83$, $p=0.03$). This suggests the distinction of prediction accuracies across the three room acoustic environments involved in the test. Dunn's post hoc test, conducted to evaluate which groups differed significantly from one another, indicated a significant difference between the Brahmssaal and recording studio ($p=0.03$). However, the Sommetheater did not show a significant difference with the Brahmssaal ($p=0.29$) or the Recording studio ($p=0.90$).

5.2.2 Exploring acoustic parameters in orientation perception

Based on insights derived from previous research, this section explores the impact of potential interaural and monaural parameters discussed in earlier studies on source orientation and source localization perception, aiming to assess their influence on source orientation perception in musically realistic in-situ conditions. As previously mentioned in section 5.1.4, the loudspeaker employed for measuring the impulse responses is expected to have a similar spatial distribution of energy to that of trumpet and trombone in the room acoustic environments due to their similar directivity characteristics. Therefore, the binaural and monaural parameters extracted from the impulse responses for specific orientations are expected to correspond with the orientation perception of the trumpet and trombone in those directions. Following this hypothesis, the variation of prediction accuracies of the two instruments against the acoustic parameters extracted from the BRIRs is illustrated here, and the correlation between the prediction accuracies and the parameters is estimated and presented. The prediction accuracies of the trumpet and trombone are regarded as two independent observations here for the analysis against each parameter extracted from a particular impulse response.

Interaural parameters derived from BRIRs

Along with the established ILD parameter in source orientation, the relationship of prediction accuracy with an extended set of interaural parameters, encompassing ITD and IACC is analyzed here. Since features influencing lateral (i.e., left, right orientation) and medial (i.e., front and back orientation) orientations are observed to be different in the previous studies, lateral and medial samples are analysed separately for each parameter.

Figure 5.9 illustrates the variation between the prediction accuracies against the ILD parameter values for early reflections (0-80 ms) and late reverberation (80 ms-end). Acoustic conditions that correspond to each of the obtained parameter values are highlighted on the top x-axis with their room abbreviation (RS, ST, and BS) followed by the orientation in subscript (F, L, B, R for front, left, back, right orientations). Since left and right channels would have positive/negative values, the absolute of ILD values averaged over 1000-4000 Hz octave bands is presented here for each orientation condition.

Moreover, the Spearman's rank correlation coefficient (abbreviated here as ρ), a non-parametric correlation that asseses the monotonic relationship between variables[158], is estimated between the extracted parameters and their corresponding prediction accuracy values from different groups of samples (all samples, lateral samples only, and medial samples only), and presented in Table 5.2.

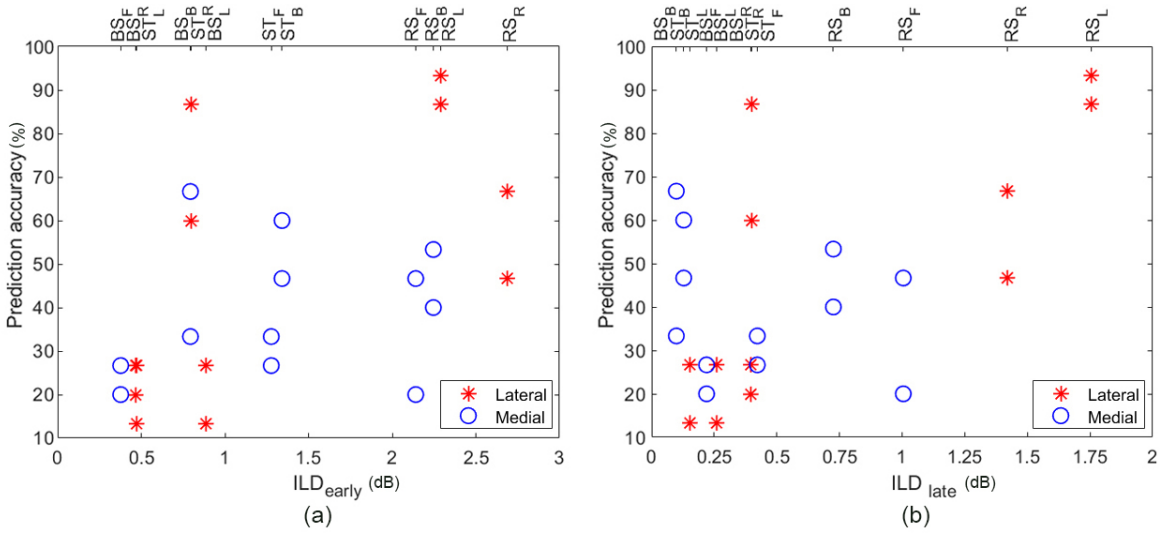


Figure 5.9: Variation of true positive values with Interaural Level Difference (ILD) estimated for (a) early reflections, (b) late reverberation.

Considering the variation of ILD_{early} across all samples, a statistically significant positive trend can be observed between ILD_{early} and prediction accuracies (ρ=0.56, p<0.05), while no clear trend is observed between the ILD_{late} and overall samples. When examining the samples of lateral and medial orientations separately, prediction accuracies of lateral samples are observed to have a strong tendency to increase with the increase in ILD values which holds true for both ILD_{early} (Spearman's rank correlation coefficient ρ=0.54, p<0.05) and ILD_{late} (ρ=0.85, p<0.01). Since the binaural receiver is placed in the far field of the room, the ILD caused by direct sound (0-5 ms) can be neglected from these observations due to the distance between source and receiver, and the ILD_{early} can be considered specifically caused by the early reflections from the room acoustic environments (5-80 ms). This is in accordance with previous studies on orientation perception [71]. Interestingly, the ILDs from late reflections also seem to have a direct positive relationship with the orientation perception in lateral directions, which needs further exploration. Although source orientation in medial directions (i.e., front and back) does not produce ILDs between the ears in reflection-free environments, ILDs are observed to occur in real rooms due to the presence of room acoustic reflections which may positively or negatively impact the orientation perception. However, unlike the lateral samples, no significant trend is observed between

the prediction accuracies of samples from medial orientations and their corresponding ILD values.

The Recording Studio featuring dry acoustic characteristics seems to possess strong early reflections that produce ILDs between the two channels, which could be the reason in higher overall lateral prediction accuracy (mean prediction accuracy of 90% & 56.7% respectively for right and left orientation, in the case of trumpet and trombone). On the other hand, the Brahmssaal, characterized by a strong diffuse field, lacks strong lateral reflections that produce ILDs and thereby result in the least overall lateral prediction accuracy (mean prediction accuracy of 20% & 23% respectively for right and left orientation). Looking at the individual observations, specific cases like the rightward orientation in Sommertheater (ST_R) having ILD values within the JND range of 0.5-0.8 dB is observed to possess relatively high prediction accuracy (mean prediction accuracy of 73.3% for trumpet and trombone; 44% for all instruments), which opens up the possibility of advanced features involved in the lateral orientation perception.

Type of parameters	**Parameter**	**Overall samples**	**Lateral samples (L,R)**	**Medial samples (F, B)**
Interaural parameters from BRIRs	ILD_{early}	**0.56****	**0.54***	0.34
	ILD_{late}	0.35	**0.85****	-0.29
	ITD_{early}	-0.26	-0.44	0.19
	ITD_{late}	-0.46	-0.52	-0.58
	$IACC_{early}$	**0.48***	**0.85***	-0.01
	$IACC_{late}$	0.10	0.36	0.02
Spectral parameters from BRIRs	SC	**0.67****	**0.83****	0.23
	SC_{Dir}	-0.24	0.34	**-0.60***
Temporal parameters from BRIRs	DRR	0.03	**0.68***	**-0.58***
	C_{80}	0.24	**0.70****	-0.29

Table 5.2: Spearmann correlation coefficients estimated between the prediction accuracies of each acoustic environment against its corresponding parameter explored in the study.

Figure 5.10 illustrates the variation between the prediction accuracies against the ITD parameter values for early reflections and late reverberation. Unlike ILD, no statistically relevant trend is observed between the prediction accuracies of both lateral and medial samples and its corresponding ITD_{early} and ITD_{late} values (see Table 5.2). As previous studies suggest, the ITD parameter is observed to be important for localization aspects for low to mid-frequency range [74; 153]. Within this frequency range, the instruments mostly show more omnidirectional characteristics, with slight variations in the directivity shapes. This could be a potential reason for ITD having no direct role in orientation prediction.

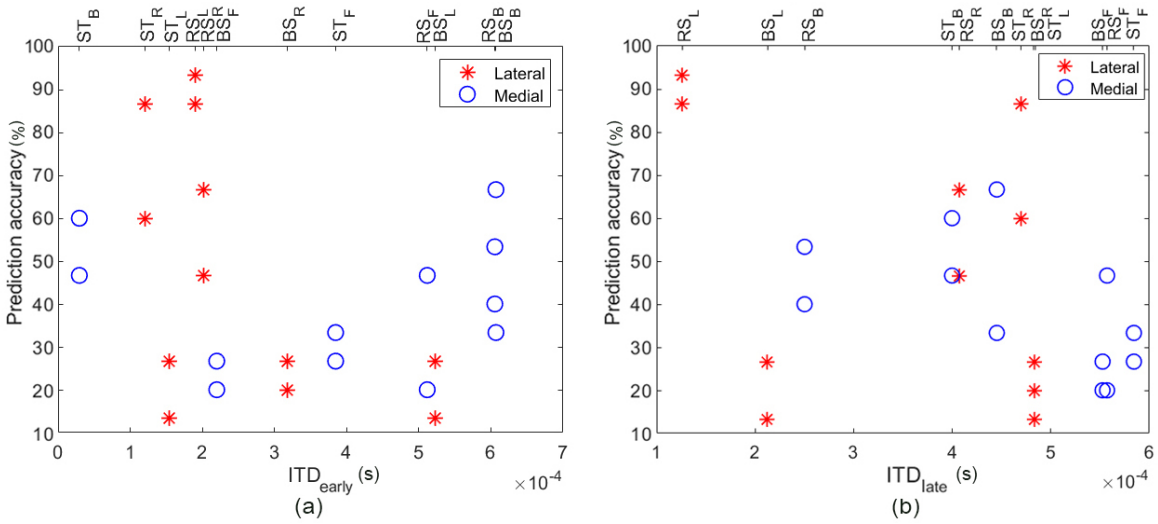


Figure 5.10: Variation of true positive values with Interaural Time Difference (ITD) estimated for (a) early reflections, (b) late reverberation.

The variation of prediction accuracies of acoustic conditions against the IACC values is presented in Figure 5.11 for the early and late parts of the impulse response. A trend of increase in the overall prediction accuracy values with an increase in the $IACC_{early}$values is observed (ρ=0.48, p<0.05), and it is highly pronounced for the lateral samples (ρ=0.85, p<0.05). However, unlike ILDs, no such trend is observed for the late part, $IACC_{late}$. In real performance spaces, depending on the acoustic properties of rooms, the room acoustic reflections decorrelate the signals received in the two ears which results in a decreased IACC value with a broader sensation of Apparent Source Width (ASW). A higher IACC value, representing an increased similarity between signals received in the two ears, corresponds to the perception of a narrow ASW with focused and localized sound source perception. Therefore, a focused and localizable source perception which improves with an increase in IACC may have an influence on the orientation perception of lateral conditions, which shall be explored further. However, no such trend is observed between prediction accuracies of medial samples against their perceived source width. Although ILD and IACC refer to two different aspects of sound perception, and they are estimated differently from the impulse responses, it should be noted that their values across 12 acoustic conditions are observed to show a high linear correlation (ρ=0.97 between ILD_{early} and $IACC_{early}$, and ρ=0.88 between ILD_{late} and $IACC_{late}$).

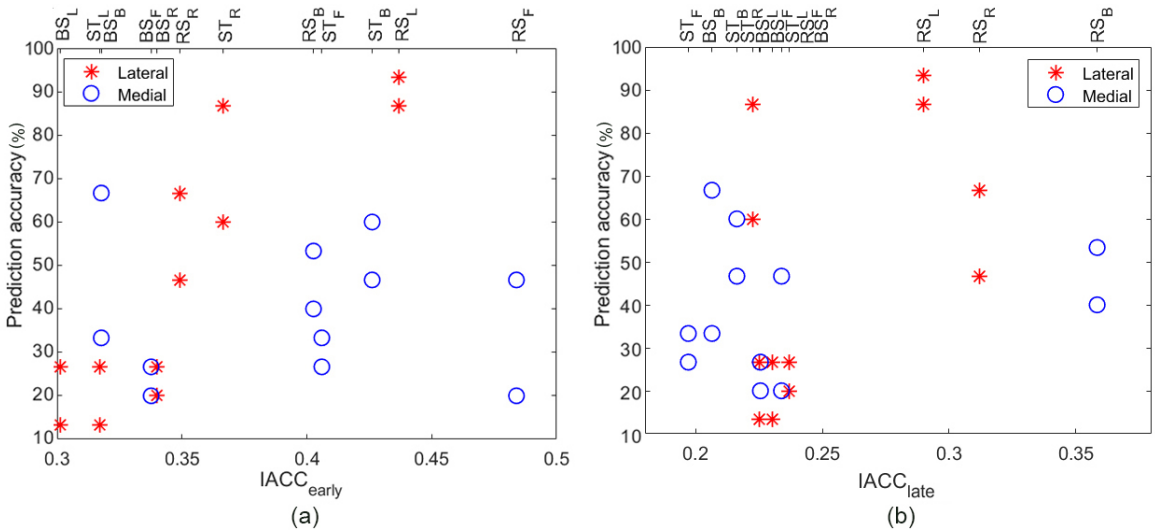


Figure 5.11: Variation of true positive values with Interaural Cross Correlation (IACC) estimated for (a) early reflections, (b) late reverberation.

Other spectral and temporal parameters from BRIRs

To understand the influence of spectral coloration brought by the room acoustic environment and the timbre difference in direct sound introduced by the rotation of the source on the orientation perception, the variation of spectral centroid values for both the overall BRIR and the direct sound part of the BRIR against the prediction accuracies of trumpet and trombone was analyzed, and the results are presented in Figure 5.12 (a) and (b) respectively. For the overall spectral centroid parameter from BRIR, a positive relationship is observed between the centroid value and the prediction accuracy for the overall samples (ρ=0.67, p<0.001), and this association is stronger for lateral orientations (ρ=0.83, p<0.001). When examining the overall spectral centroids of room impulse responses, which depict the transformation of sound by the acoustic environment from the source to the receiver, a high spectral centroid value is expected to enhance the energy in high frequencies, leading to a brighter timbre of the perceived instrument sound. As the high-frequency sound component of the instrument carries directivity information, it can be hypothesized that the room acoustic environments with high centroid values, featuring reflections having high-frequency content, are expected to support to the orientation perception. Conversely, environments with low spectral centroid are expected to carry a lower degree of directional information due to the attenuated high-frequency components, leading to a deteriorated prediction accuracy. Consistent with this hypothesis, in the case of Brahmssaal featuring low centroid values across all directions, all data points exhibit accuracies below 35% except for the trumpet in the back orientation as an outlier at 66%, while the other two environments featuring high centroid values showing relatively better accuracies than Brahmssaal.

When comparing the centroid values of the direct sound part, no significant trend is observed across the overall samples and samples from lateral orientations. However, when comparing the centroid values of samples from medial orientations, the data points are observed to be clustered into two distinct groups where, alongside the centroid values, these two groups do not overlap in prediction accuracy values, except for an outlier point. Consequently, a statistically significant negative correlation is observed between the centroid values of direct sound and their prediction accuracies of medial samples (ρ=-0.60, p<0.001), whereas no such trend is observed for the lateral orientation samples. The reduced centroid values of back orientation in comparison to the front orientation, referring to the spectral tilt mentioned in [76], is expected due to the highly directional beam-like radiation characteristics exhibited by both the loudspeaker and the musical instruments. Following that, the decline in centroid values observed in the Figure 5.12, could serve as a potential attribute for distinguishing back from the medial samples.

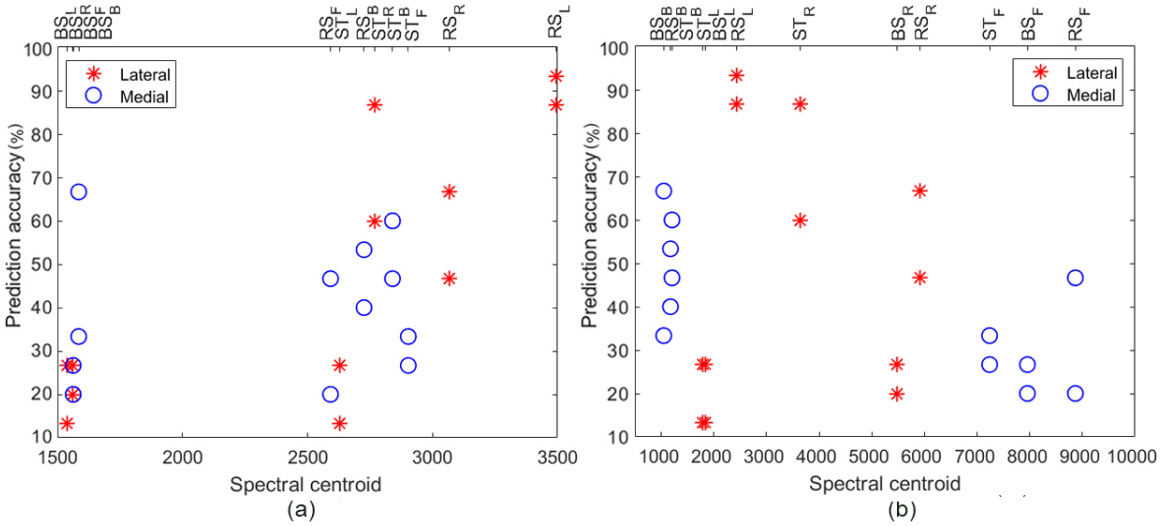


Figure 5.12: Variation of prediction accuracies with spectral centroid of BRIR estimated for (a) the overall BRIR (b) only for direct sound.

Figure 5.13 shows the variation of prediction accuracies against the two temporal energy ratio parameters from BRIRs, the DRR and the C_{80} parameter. While there is no overall trend visible between the DRR and prediction accuracy values of all samples, a statistically significant negative relationship is observed between the DRR values and the prediction accuracies of medial samples (ρ=-0.58 p<0.05). Although the front orientations possess higher DRR than the back orientations, only the front orientation in the recording studio seems to have a positive DRR value in the given 12 conditions. This demonstrates the presence of a stronger direct sound component than the room reflections in this specific condition. A lower DRR value for the non-facing back angle,

with approximately a 13 dB difference from the facing front angle, may have served as a potential cue for higher prediction accuracy in distinguishing back orientation from front in echoic environments, consistent with observations from previous studies [79].

The front orientations are observed to have lower accuracy than the back orientation, and it is valid for both trumpet and trombone samples in the three acoustic environments. While the negative DRR values representing weak direct sound can be a potential reason that may hinder the prediction accuracy of front orientation in ST and BS, the front orientation in RS with a strong direct sound component still possesses weak prediction accuracy. This can be attributed to the influence of other potential room acoustic reflection-related attributes that deteriorate the orientation perception in front direction, making this orientation perception a multifaceted problem. While the variation in DRR is relatively limited, a trend of enhancing prediction accuracy through improved DRR can be inferred for lateral samples (ρ=0.68, p<0.05). The presence of both strong early reflections, known to support lateral orientation prediction, and diffuse reverberation, which is expected to oppose orientation prediction, will result in a lowering of DRR values. Thus, interpreting the influence of DRR alone on lateral orientation prediction can be challenging, especially within a limited range of variation observed in lateral samples.

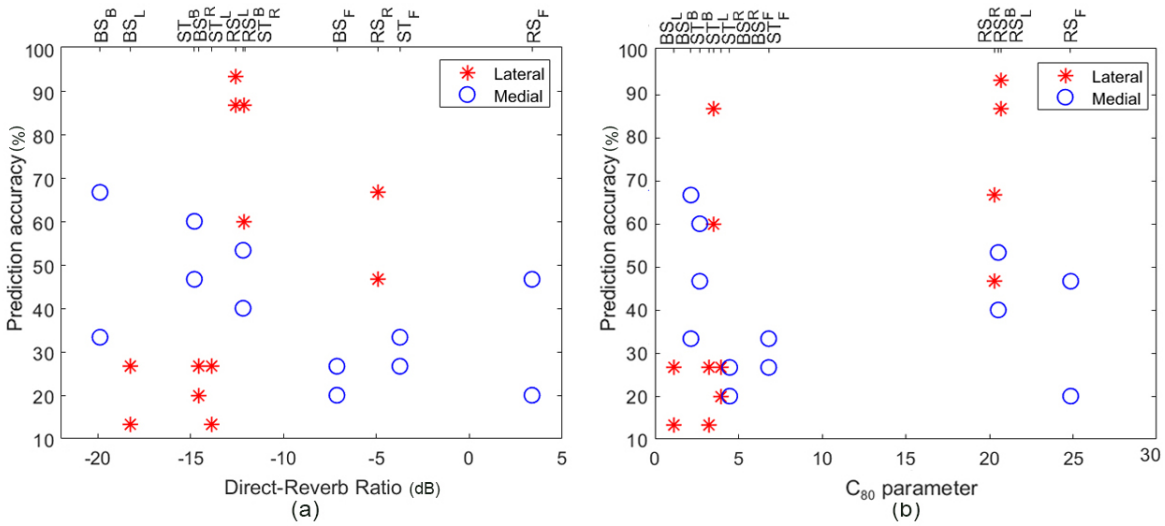


Figure 5.13: Variation of prediction accuracies with (a) Direct-to-Reverb Ratio (b) C_{80} parameter.

When analyzing the C_{80} parameter, RS featuring a high clarity impression seems to have a higher prediction accuracy. On the other hand, the BS and ST, featuring a low clarity perception, seem to have relatively low prediction accuracy. Observing the trend of medial and lateral samples, while no specific trend is observed for medial orientations, a statistically significant positive relationship is observed between the clar-

ity parameter and prediction accuracy of lateral samples (ρ=0.70, p<0.01). Previous observations suggesting that environments characterized by strong early reflections support orientation perception [71] are observed to be valid here for samples of lateral orientation, while no such trend is observed for medial orientation in the given conditions. The presence of strong early reflections, particularly from lateral directions, resulting in a higher C_{80} value would have supported the orientation perception of lateral samples by providing binaural cues. This could be a potential reason for such a trend. However, room acoustic reflections from other directions apart from lateral ones (such as strong back-wall reflections) might have also contributed to an increased C_{80} value, however, the role of such reflections in these trends needs further exploration.

5.3 Discussion

While previous studies have primarily focused on source orientation perception using loudspeakers and human speakers by often overlooking the significance of room acoustic attributes, this study examined the role of directivity of the sound source in orientation perception for the first time in diverse room acoustic conditions. This was accomplished by conducting in-situ recordings of five musical instruments having varied directivity properties in three acoustic environments characterized by distinct perceptual properties and assessing their outcomes in orientation perception in a static binaural listening condition. In this investigation, the perceptual evaluation of audio samples was exclusively conducted with tonmeister students and musicians who have received technical and musical ear training. This ensures reliable responses based on their experience in critical listening and familiarity with the instruments involved.

The influence of already established and other potential parameters on orientation prediction was investigated independently for lateral and medial samples by extracting them from binaural room impulse responses (BRIRs) recorded using a loudspeaker for each orientation. These parameters were then compared against the prediction accuracies of the trumpet and trombone, which exhibit comparable directivity characteristics to that of the loudspeaker. When looking at the overall trend of variation of parameters extracted from BRIRs against prediction accuracies of all samples, parameters including ILD_{early}, $IACC_{early}$, and Spectral Centroid seem to show a positive relationship with the overall prediction accuracy values. While this positive relationship strengthens for lateral orientations, it remains insignificant for medial ones, suggesting that the acoustic cues used to judge lateral and medial orientations differ.

When it comes to parameters influencing to lateral judgment, the ILD parameter, previously identified as essential for orientation prediction in lateral directions [75; 76; 71], is confirmed to be a significant factor in lateral orientation perception in in-situ conditions involved in this study. While the ILD resulting from early reflections was emphasized as a pivotal cue, ILDs from late reflections also appear to exhibit

a positive correlation with lateral orientation perception in this study, which needs to be validated further. Furthermore, the IACC parameter extracted from the early portion of the impulse response, reflecting apparent source width perception, exhibits a positive correlation with lateral orientation prediction accuracy; higher IACC values, reflecting a focused and localized source perception, seem to lead to enhanced prediction accuracy. However, it should be noted that despite the two parameters representing distinct aspects of sound perception and being calculated differently, they demonstrate a high correlation among the 12 environments examined in the study. The ITD parameter, known for its significance in source localization, nonetheless, exhibits no significant trend with prediction accuracies, which could be due to its minimal influence in high-frequency regions above approximately 1000 Hz.

The spectral brightness of overall impulse responses assessed using the spectral centroid seems to have a positive correlation with the lateral sound samples. Access to high-frequency spectral content, including both direct sound and room reflections, is expected to convey variations in source directionality, potentially explaining the positive relationship between spectral centroid and orientation perception, as noted in previous studies [81]. However, this relationship does not hold for medial samples. As the spectral centroid values are observed to decrease systematically with increasing reverberation time among the three rooms, this observation warrants further investigation. Considering the temporal energy distribution in the impulse responses, analysis employing the C_{80} parameter reveals that strong early reflections relative to the reverberation are particularly significant for lateral conditions. This finding is in line with the observations from the ILD analysis and remains consistent with previous results [71]. While early reflections from lateral directions are expected to generate high ILD cues, a higher C_{80} value can also be due to reflections from other possible directions as well. Given that the directional attributes of early reflections have been demonstrated to affect source localization properties [49], it is plausible that the order and directional properties of early reflections could also impact orientation perception. Moreover, the ILD variation within the conditions explored in this study is limited to 2-3 dB, which may or may not be significant enough compared to its JND value of 05-0.8 dB [159; 160]. Therefore, the strength and directionality of early reflections causing ILD in orientation perception shall be explored in future research.

Regarding the medial orientations, while front orientations are observed to be relatively easy to predict in previous studies [80; 79], a lower prediction accuracy was observed for the front than the back orientation in this study, and it is consistent across the three rooms and two instruments involved in the objective analysis. Factors such as high ILD values for the front orientations in specific cases could be a potential reason for this. Although this study focused on analyzing the individual variations of parameters, the observations from medial samples suggest that the decision of orientation prediction may be influenced by multiple attributes involved. This points to the

multidimensionality of the factors involved in the orientation perception judgment. Although no specific trend is evident between the centroid values of front and back orientations from BRIRs, when specifically analyzing the direct sound excluding reflections (0-5ms of the BRIR), a spectral-tilt in the high-frequency region is observed between the front and back orientations which results in a decrease in spectral centroid values from front-to-back orientations. Given that this spectral-tilt is recognized as a cue for medial orientation identification in anechoic conditions [76], it is plausible that it also played a role in perceiving the front and back orientations in the in-situ conditions examined in this study. The DRR, which is found to be a potential cue for distinguishing between facing and non-facing angles in medial orientations, exhibits a relatively higher value for front-facing angles compared to the back ones with a difference of around 13 dB. The low DRR values, which represent dominating room reflections over the direct sound, could also be a supportive cues listeners utilized in identifying the back orientation in the medial samples.

Considering the overall response in Brahmssaal, all orientations except back orientation had an accuracy of just the chance level, and they seemed to be dragged towards the back orientation by exhibiting a higher prediction rate in that direction (see Figure 5.7). The darker timbre of the perceived sound due to suppressed high-frequency information, and a low DRR value, accompanied by a diffuse field that lacks ILD and IACC cues for lateral samples, could be the reasons behind this phenomenon. However, when specifically analysing the spectrally bright instrument like the trumpet, these factors appear to have an opposing effect in Brahmssaal which results in easy identification of back orientation with an accuracy of 66%, making it an outlier point in the objective analysis. Similarly, while the accuracy values for the other three angles are high, the back orientation is noted for its poor accuracy in the overall orientation perception of the recording studio. Factors such as high spectral centroid and improved clarity of the acoustic environment in the presence of rear wall reflection may also have influenced this reduced accuracy for back orientation. Hence, while this study primarily delved into individual features affecting orientation perception, it underscores the multifaceted nature of orientation perception and its associated parameters. Consequently, it necessitates future research to investigate the inter-relationship and contribution of involved parameters in orientation perception.

A univariate analysis using Kruskal-Wallis test was conducted in this study to independently assess the influence of facing angle, instrument directivity, and room acoustics on sound source orientation perception. Since these mentioned three aspects can be considered to be independent aspects of variations with no direct interactions, concerns about multicollinearity can be neglected. Moreover, the data used for the Kruskal-Wallis test is perfectly balanced; each level of one factor is equally represented across all levels of the other factors (e.g., when comparing three room acoustic conditions, each room contains samples from five instruments across four

orientation angles, etc.). This balanced data structure in the univariate analysis ensures that confounding effects between factors are unlikely, strengthening the validity of independent factor assessments. While this univariate analysis provides an initial understanding of each attribute's contribution to orientation perception, future studies shall employ mixed-effects models to further investigate their combined influence on orientation prediction and assess the individual contributions of these factors.

Previous studies have indicated that the position of the sound source relative to the receiver's head affects orientation prediction, with the position of sound source in front of the head demonstrating the highest accuracy [79]. As an initial step towards a comprehensive evaluation of orientation perception of diverse sound sources in in-situ conditions, this investigation is limited to the front source position. Studies have demonstrated that dynamic cues from the movement of sound sources, such as the rotation of the source, can enhance accuracy [80]. Furthermore, although head rotation movements have been demonstrated to be crucial for enhancing source localization tasks [161], they were restricted in this study to gain insights into analyzing and interpreting individual parameters based on room acoustic reflections. Therefore, it necessitates further exploration of the impact of attributes relevant to real-world conditions such as the location and movement of sound sources and receivers on orientation perception. While earlier studies typically employed either noise or speech signals, this research utilized distinct compositions tailored to each instrument which covers its whole pitch range, and recorded their performances in realistic musical environments. Nevertheless, it's worth noting that the attributes of the musical content, such as dynamics and tempo, may have impacted the predictive outcomes. Given that binaural reproduction has constraints, such as front-back confusion, future investigations could integrate spatial audio capture and reproduction methods to enhance the exploration of orientation perception.

5.4 Summary

This investigation delved into analyzing how the directivity of the sound source and acoustic attributes of performance space influence the perception of source orientation in real-world conditions. This was performed by utilizing static binaural recordings of the performance of five different instruments with diverse acoustic characteristics in three performance spaces featuring contrasting acoustic characteristics. Contrary to the previous findings, none of the four orientations analyzed in this study (front, back, left, right) demonstrate statistically higher prediction accuracy among the various conditions considered in the study. Although the brass instruments (trumpet and trombone) are observed to have relatively higher overall prediction accuracies of source orientation, when comparing the individual prediction accuracies across different conditions involved, no statistically significant differences were observed between the five instruments involved in the study. Conversely, a statistically signifi-

cant difference was observed for the distribution of prediction accuracies across the three acoustic environments for various conditions involved. A systematic trend is noted across the three analyzed acoustic environments, wherein prediction accuracies increase with a decrease in reverberation, improved clarity impression, and better access to high-frequency information (assessed using spectral centroid). Based on these observations, the room acoustic variations incorporated in the study appear to have a greater influence on source-orientation prediction than the directivity attributes of the sound sources. Therefore, within the constrained scope of the experiment, it can be hypothesized that the distinct directivity characteristics, which result in varied spatial energy distributions from the instrument, are being obscured in orientation perception by the influence of the room acoustic reflections encountered in real-world conditions.

The role of already established and other potential parameters influencing orientation perception in in-situ performance context was systematically analyzed by extracting them from BRIRs of each orientation and comparing them afterward against the corresponding prediction accuracies of the trumpet and trombone, both having a directivity close to the loudspeaker used for BRIR measurement. While some parameters demonstrate significant influence on the prediction accuracies of lateral orientations (left, right), specific parameters previously speculated to affect medial orientation (front, back) in previous controlled experiments also provide cues for medial orientation in this in-situ investigation. The ILD is shown to be an important factor for the orientation perception of lateral samples involved in this study, which is consistent with previous findings. The increase in the C_{80} parameter, reflecting stronger early reflection compared to late reverberation, was also found to improve the orientation perception, further supporting the observation regarding ILD_{early}. Additionally, an increase in $IACC_{early}$, indicating a focused and localized source perception, seems to positively enhance orientation perception.

For medial orientations, consistent with previous observations, the DRR is observed to provide cues for distinguishing between back and front orientations. Additionally, the spectral centroid value assessed from the direct sound, which reflects the spectral tilt of the direct sound between front and back orientations, appears to provide cues for distinguishing front and back orientations. However, certain observations point towards the multidimensionality of the orientation perception judgment. Factors such as ILD cues arising from medial orientations due to specific room acoustic reflections in in-situ conditions, access to high frequencies from rear room acoustic reflections in the back orientation, etc., appear to challenge the orientation predictability of samples from medial orientations. Therefore, it necessitates future research to investigate the orientation perception as a multi-faceted problem by analysing the interrelationship and contribution of involved parameters.

While this investigation has limitations due to incorporating steady source and static listening conditions, it offers the first step in analyzing the important factors

that influence orientation perception in in-situ conditions. This investigation shall be further extended by incorporating dynamic sources and receivers that resemble real-world conditions and analyzing the contribution of the individual parameters discussed in this investigation in the orientation perception judgment. Given that this investigation integrates the ecological performance of diverse musical instruments in realistic performance spaces for the first time in orientation perception, it offers valuable insights into music instrument arrangement, musical recording techniques, auralization, communication acoustics, and related areas.

Chapter 6

Directivity perception in room acoustic environments

Room acousticians often use electro-acoustic sources, which possess simplified directivity characteristics, for the playback of musical instrument recordings to know the 'sounding of the room'. In a sophisticated manner, loudspeaker orchestras, in which the instruments were represented as a combination of different loudspeakers, were used for the perceptual evaluation of concert halls and acoustic measurements. However, in these cases, the natural/realistic impression and the perceptual similarity of these electro-acoustic substitutions to the real instruments are not well-explored yet. By analysing the similarity and naturalness perceived in the sound samples recorded from the real instruments and electro-acoustic counterparts in different acoustic environments, this chapter aims to investigate the perceptual quality of sound fields produced by electroacoustic sources and thereby understand the perceptual relevance of the dynamic directivity of the musical instruments in in-situ conditions. Morevoer, a potential statistical modeling approach is also introduced for assessing the perceived similarity between real instrument and electroacoustic sources by incorporating the in-situ binaural recordings. A part of the content presented in this chapter is reproduced from the following research article with the permission of the Deutsche Gesellschaft für Akustik e.V:

J. Thilakan, W. Buchholtzer, M. Kob, "Evaluation of subjective impression of instrument blending in a string ensemble", Fortschritte der Akustik- DAGA, Vienna, (2021).

6.1 Materials and methods

6.1.1 Collection of sound samples

This investigation utilizes the in-situ binaural recordings of five different musical instruments (trumpet, trombone, violin, saxophone, and flute) having unique directivity features, from various room acoustic conditions, as presented in Chapter 5. Almost all these instruments show omnidirectional behavior roughly up to 500 Hz, and distinctive complex directional shapes above this frequency range [16]. Whereas the instruments like trumpet and trombone show less complex but highly directive radiation patterns for high frequencies, instruments like the flute and saxophone show relatively complex directivity patterns for high frequencies due to the multi-pole source radiation behavior. The same trend is seen in the violin which exhibits highly complex radiation characteristics due to its vibrating body (The specific directivity characteristics of each instrument are explained in detail in Section 1.2.3).

In addition to these musical instruments, two electro-acoustic sources, Neumann KH 120 A and Outline Globe Source Radiator, with entirely distinct directivity characteristics were chosen as the electro-acoustic counterparts of these instruments. Neumann KH 120 A (abbreviated as 'KH120' in this study), a commonly used studio monitor speaker, consists of a 5.25" woofer and 1" tweeter with a frequency response of 52 Hz–21 kHz ($\pm$ 3 dB), and it exhibited a directional characteristic similar to a trumpet, especially for mid and high frequencies [147]. The Outline Globe Source Radiator (abbreviated as 'GSR') consists of 12 individually driven loudspeakers with a frequency range of 90 Hz–12.5 kHz, and it is intended to be used for room and building acoustic measurements. The GSR exhibits omni directional characteristics until approximately 2 kHz [162]. GSR gives more complex directivity shapes above this frequency range, but still, it satisfies the ISO 3382 required for an omnidirectional source [37].

As detailed in Chapter 5, the performance of the instruments was carried out in three room acoustic environments characterized by unique room acoustic properties, including studio-1 of Erich Thienhaus Institute (with a volume of 110 m^3), Brahmssaal of HfM Detmold (775 m^3), and Sommertheater Detmold (2930 m^3). The room acoustic parameters of these environments measured with ISO 3382-1 standards [37] are provided in Table 5.1. In a pilot test on the perceptual impression across the four orientations, the right orientation with respect to the listeners was observed to be perceptually closer to the back orientation. Consequently, to minimize the number of samples for comparison, the front, back, and left orientations with respect to the listeners, which are observed to have perceptually unique sounding impressions across the three rooms, were chosen for this analysis, while the right orientation was dropped out from the further analysis (these source orientations in the three rooms are illustrated in Figure 5.1). The bell opening in the trumpet, trombone, and saxophone, the

embouchure hole of the flute, and the f-hole of the violin were considered to be the main radiating acoustic centers. These acoustic centers were maintained at a specified source location for the entire recording session. Since directivity patterns of musical instruments are a function of frequency and it is observed to have dramatic changes over the performing range of an instrument [16], as mentioned in Chapter 5, the instruments performed dedicated compositions (provided in Appendix B) which covered the whole pitch range of the instrument, with an expectation to excite most of the possible variations in the directivity patterns.

A DPA 4099 clip-on microphone attached to the musical instruments (positioned towards the bell opening of the brass instruments & saxophone, near to embouchure hole in flute, and close to the f-hole on the violin) was used to record the instrument signals with minimum room acoustic contribution. Although the DPA microphone recordings had exhibited a minimal room acoustic contribution, noises from the breathing of musicians, scratching of the bow, etc, were present in it, consequently also in the playback recordings as well. In addition, in certain environments, the binaural recording had background ventilation noise as well. Therefore, a smooth high-pass filter centered around 200 Hz was applied to all samples to reduce these noisy components. These processed signals were used for the playback through the electro-acoustic sources. During the playback of the signals, the KH120 was also kept at the specific orientation corresponding to the orientation of the instrument in the particular samples whereas no rotation was performed in the case of GSR. Neumann KU-100 binaural head [116] placed at the far-field of the room (as illustrated in Figure 5.1) recorded the resultant sound field of the real & electro-acoustic sources.

Since the goal of this study is regarding the overall perception of the 'dynamic directivity' of instruments during the performance, but not the static directivity pattern of instrument tones, the reverberation tail in the samples is cropped, and smooth fade in and fade out filters were added to the samples. Finally, the loudness level of each group containing one real and two electro-acoustic samples (one particular instrument in a specific room at a definite orientation, and its corresponding two electro-acoustic counterparts) was manually equalized in REAPER. Moreover, the high-pass filter at 200 Hz was also applied to the real instrument performance recordings, for the direct comparison with the resynthesized recordings.

Description of the listening test

A dedicated application under the framework of MATLAB's App Designer platform was created for the performance of the listening test; the Graphical User Interface of the application is given in Figure 6.1. 13 participants (4 female, 9 male) consisting of Tonmeisters, sound engineers, and professional musicians participated in the listening test. All of the listeners had undergone musical and ear training, and are experienced in critical listening. Moreover, they were either trained or familiar with the instruments

chosen in this study. Due to this reason, no special training/familiarization session was conducted prior to the listening test to evaluate the sound of real instruments.

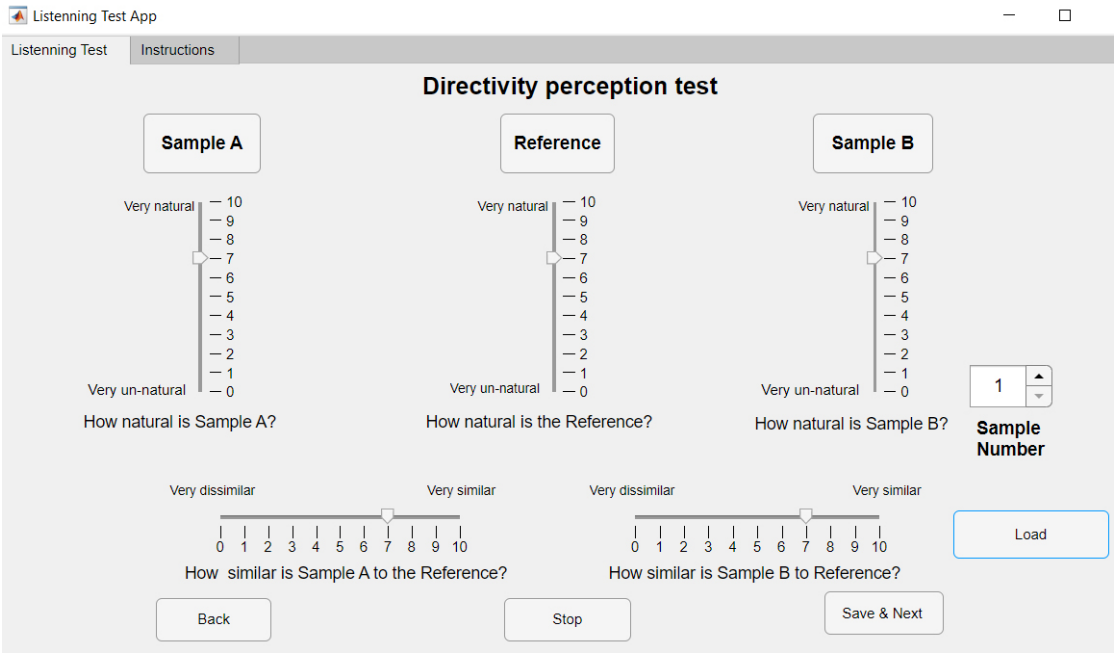


Figure 6.1: User Interface of the Listening test application

The application consisted of 45 trials (5 instruments x 3 rooms x 3 angles of orientations), and each trial included a group of three audio files which are named 'Sample A', 'Reference', and 'Sample B' as shown in Figure 6.1. The recording of the real instrument was always assigned at the 'Reference' button, whereas the samples from KH120 and GSR corresponding to the given reference were randomly assigned to Sample A and Sample B. In order to reduce the direct comparison due to memory of listening, the order of the 45 audio groups was randomized in the test for each participants.

The listeners were not aware of the characteristics of the samples i.e, they didn't know that two of the three samples were playback through loudspeakers. But they were informed that there could be different orientations of the instruments possible (instruments might not always pointing towards the listeners) between different audio groups. For each trial, the listeners were asked two aspects; (1) to independently rate the naturalness (realism) of the instrument for each of the three audio samples on a scale of 0 to 10 (a high value corresponds to a high natural impression), (2) to rate the similarity of the sounding impression between Sample A & Reference, and Sample B & Reference on a scale of 0 to 10 (high value corresponds to high similarity in sounding). The listeners had the choice to perform the test remotely, listen to the samples many times, and also to pause and resume the test at any moment they needed. Since the au-

dio files were binaural, headphones were used to perform the test, and approximately 30 to 45 minutes were taken to complete the listening test.

6.1.2 Similarity estimation modeling procedure

Similar to the classification modeling procedure discussed in Chapter 3, the modeling attempt proposed in this chapter to assess the perceptual similarity between the binaural samples of real instrument and electroacoustic counterpart utilized Mel Frequency Cepstral Coefficients (MFCCs) as the input feature. This is because of its wide usage in different areas such as musical instrument recognition [125], speech recognition [120; 121], speaker identification [122], and so on. Additionally the Principal Component Analysis (PCA) [132] was utilized as the feature transformation method to project the higher dimensional MFCC data into a lower dimensional space by retaining the important information. The process of extraction of MFCC features, and the method of PCA transformations are described in section 3.1.2.

The perceptual test on similarity analysed 90 pairs of sound samples, comprising 45 samples of instrument performances (5 instrument $\times$ 3 rooms $\times$ 3 orientations) reproduced via two electroacoustic counterparts. For each of the 90 pairs of samples, the silent regions at the start and end of the audio samples were removed, and the first 14 MFCCs [128] were extracted from each channel of the binaural recording for every 100 ms of the audio signal with an overlapping length of 50 ms using a Hamming window. Subsequently, for each channel of the binaural audio file, MFCC feature matrices from the real instrument and electroacoustic source were concatenated, and the PCA was performed on the concatenated matrix. After the PCA transform, the transfromed MFCC features were dissassociated for the two sources, and the first three principal components of the transformed features were utilized afterwards for similarity estimation.

As followed in Chapter 3, the centroids of the data distribution (i.e., the Euclidean coordinate which corresponds to the arithmetic mean of data points across the dimensionality-reduced feature space) was estimated for each channel of the binaural files of the real instrument and the electroacoustic source using the first three principal components. The Euclidean distance between the centroids of the two sound samples was estimated for each channel, and their mean value was used as a metric for the similarity estimation model. In this perceived similarity estimation modeling, if the Euclidean distance between the pair of sound samples is relatively lower, it is hypothesized that the two samples possess a high similarity, and vice-versa. The mean value of the similarity rating averaged across 13 listener ratings was used as the other input variable in this study. The process of estimation of Euclidean distance on the PCA transformed lower-dimensional MFCC feature space is repeated for the 90 pairs of samples, and their corresponding Euclidean distances were estimated and compared against the corresponding perceived similarity ratings.

6.2 Results and discussion

6.2.1 Naturalness and similarity perception

The variation of the naturalness rating of the real instrument (player) and two electro-acoustic sources in different room acoustic environments is shown in Figure 6.2. To have a generalized view of the room acoustics' influence on the sounding impression of sources with distinct directivity, responses from three different source orientations were considered here. As mentioned above, the front, back, and right orientations were observed to give distinct perceptual impressions due to the differences in direct sound, strong early reflections, and late reverberation. By considering the three different orientations of instruments in one specific room as unique observations, the distribution of each sound source in one particular room in Figure 6.2 consists of 39 independent observations (13 listeners x 3 trials).

The difference in the distribution of naturalness between the real instrument and electro-acoustic sources is observed to vary between the instruments and the acoustic environments. The real instruments are observed to possess a higher naturalness than the electro-acoustic counterparts. However, in specific cases, electro-acoustic sources exhibit a naturalness rating distribution similar to that of a real instrument, characterized by the same median value and a comparable interquartile range (IQR). Additionally, in certain cases, the KH120 and GSRs show a similar distribution of naturalness ratings with same median value and comparable IQR, despite differences in their directivity characteristics. Looking at the variation of naturalness ratings between rooms, a relatively lower impression of the naturalness of electro-acoustic sources was observed in Brahmsaal for instruments like flute and violin, but this trend is not significant in other instruments. In general, spectral coloration introduced by the close microphone recrodings may also have an impact on the lower naturalness impression of the synthesized samples. However, further exploration is needed to validate this.

To statistically compare the differences in the distributions of naturalness ratings among the three groups (real instrument and two electroacoustic sources, each with 39 independent observations) across the 15 different conditions involved (5 instruments $\times$ 3 rooms), the Kruskal-Wallis test [157] was employed. As previously discussed in chapter 5, this non-parametric alternative to one-way Analysis of Variance (ANOVA) was chosen here due to the violation in the normality condition, as confirmed by the Shapiro-Wilk test [149]. The results suggest that, out of the 15 conditions, 9 show statistically significant differences in rating distributions at a significance level of 5%, with six conditions exhibiting differences among all three groups and three conditions where one group differs from the other two. However, for the remaining 6 conditions, the test does not provide statistical evidence to conclude that the three distributions are different, at a significance level of 5%. These conditions include trombone and

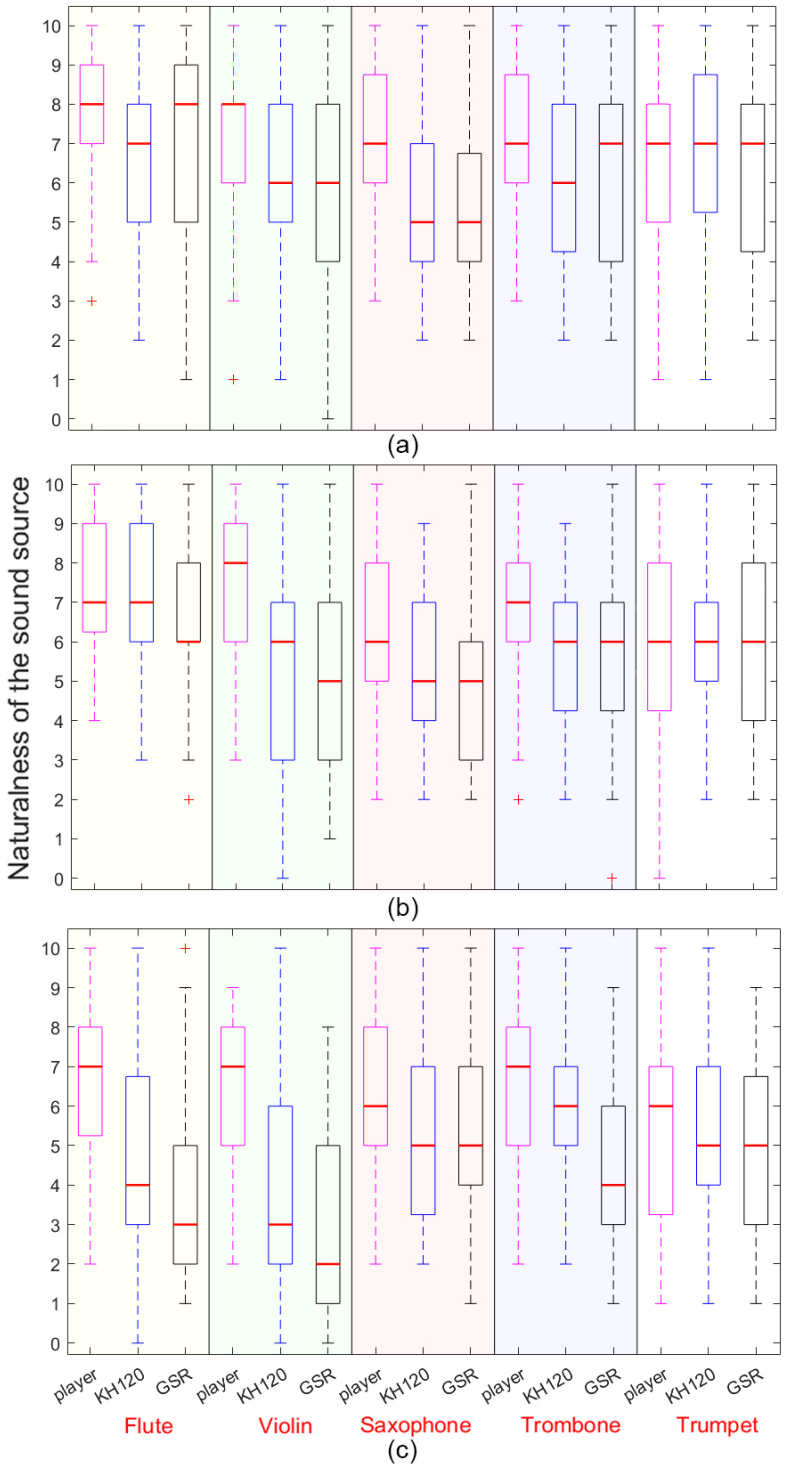


Figure 6.2: Distribution of naturalness ratings of the sound sources in different acoustic environments; (a) Recording studio, (b) Sommertheater, (c) Brahmssaal (39 observations in each condition).

trumpet in recording studio ($p = 0.349$, 0.838, respectively), flute, trombone, and trumpet in Sommertheater ($p = 0.09$, 0.072, 0.918, respectively), and trumpet in Brahmssaal ($p = 0.546$). While this does not prove that the null hypothesis (the three groups have same distribution) is true, rather this supports the argument that the observed distributions of naturalness ratings in these conditions can be comparable, given the lack of strong statistical evidence for a difference in the distributions.

Figure 6.3 shows the variation in the similarity ratings between electro-acoustic sources and real instruments in different room acoustic environments (39 observations for each instrument in a room). While the KH120 mostly exhibits relatively higher distributions, the GSR and KH120 show comparable similarity rating distributions in many conditions characterized by the same median value and overlapping IQR, across different instruments with diverse directivity characteristics. Therefore, the distinct directivity characteristics between the electroacoustic sources do not appear to play a significant role in the similarity ratings.

To analyze the statistical differences in the similarity rating distributions of the two electroacoustic sources against real instruments in each condition, the Mann-Whitney U test [141] was performed. This non-parametric alternative to Student's t-test was chosen due to violations of normality in the rating distributions (validated using Shapiro-Wilk test [149]). The Mann-Whitney U test was conducted on the pairs of distributions (each with 39 independent observations) across 15 different conditions. The results did not provide strong statistical evidence to conclude that the two groups are significantly different at the 5% significance level in any of the conditions. While this does not confirm that the pairs of distributions are similar, it supports the initial argument that the two distributions of similarity ratings can be comparable, given the lack of strong statistical evidence for a difference.

Given that the input signal, the room acoustic environment, and the loudness were kept the same, a major difference in the sounding impression between the electroacoustic sources is expected due to their distinct radiation characteristics. Yet, the perceptual evaluation reveals that the two sources are observed to have comparable distributions of similarity ratings in many conditions. As mentioned earlier, instruments like trumpet exhibits similar radiation characteristics to that of KH120 that is far different from GSR [147]. Yet, no major difference was observed in the distribution of the similarity ratings in the KH120 and GSR when compared to the real trumpet recording in the Studio and Sommertheater. This suggests that the difference in directivity of sound sources gets obscured in specific acoustic environments.

Examining variations across room acoustic environments, a slight reduction in the overall similarity ratings for instruments is observed in Brahmssaal. Additionally, while electroacoustic sources receive relatively higher ratings for instruments like the trumpet, instruments like the violin tend to have lower ratings. These observations may be influenced by the orientation of the sound source. Therefore, future studies

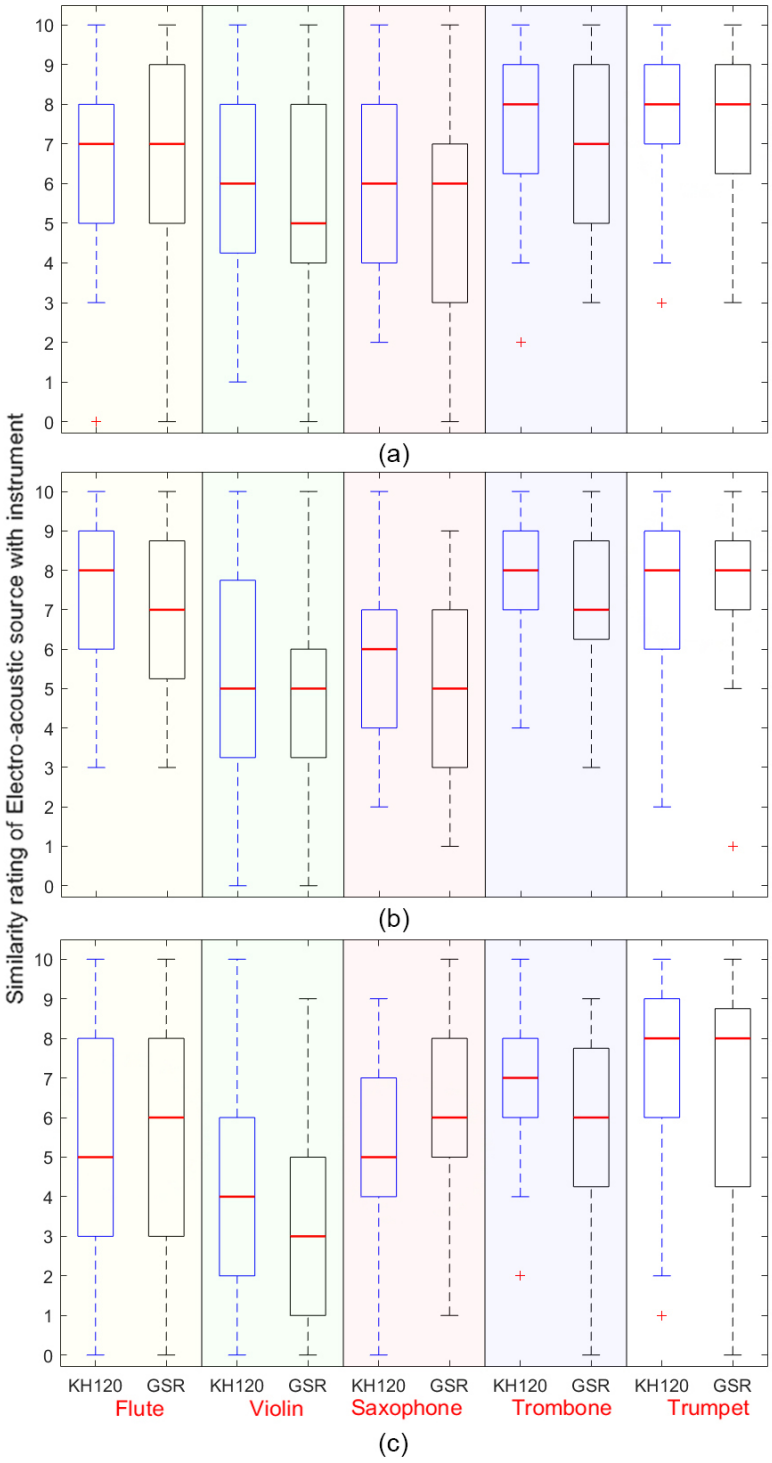


Figure 6.3: Distribution of similarity of the electro-acoustic sound sources with the real instrument at (a) Recording studio, (b) Sommertheater, (c) Brahmssaal (39 observations in each condition).

will further investigate the effects of source directivity, room acoustics, and orientation angle on the perceived similarity between instruments and their electroacoustic counterparts.

6.2.2 Similarity modeling result

The Euclidean distances between the centroids of two samples in the PCA-transformed MFCC feature space was estimated for the two channels of the binaural recording of a real and electroacoustic sources. The mean value of Euclidean distance estimated for the left and right channels was calculated for the 90 pairs of samples having a real instrument an electroacoustic counterpart. The variation of mean value of perceived similarity ratings against the mean Euclidean distance measure for these 90 samples is presented in Figure 6.4. In order to have a better perspective on the viability of this approach, the 'extreme' sound samples that possess high and low similarity ratings are highlighted using the red color.

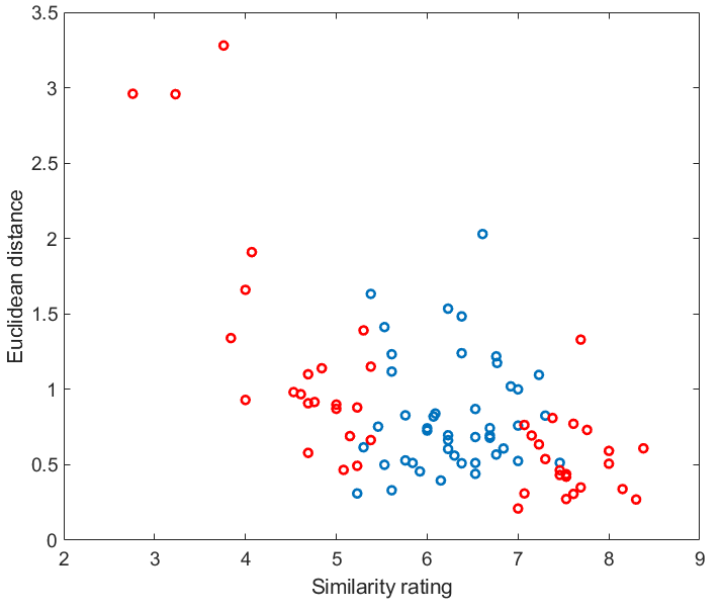


Figure 6.4: Variation of the similarity ratings of sound samples against their corresponding Euclidean distance, estimated from PCA transformed MFCC feature space for 90 pairs of sound samples involved in the study (The 25% of samples with the highest similarity ratings and the 25% with the lowest ratings are highlighted in red).

The results suggest a negative relationship between the similarity of sound samples and the derived Euclidean distances from the PCA transformed MFCC featurespace. Though some data points with similarity ratings between 5 and 7 exhibit outlier

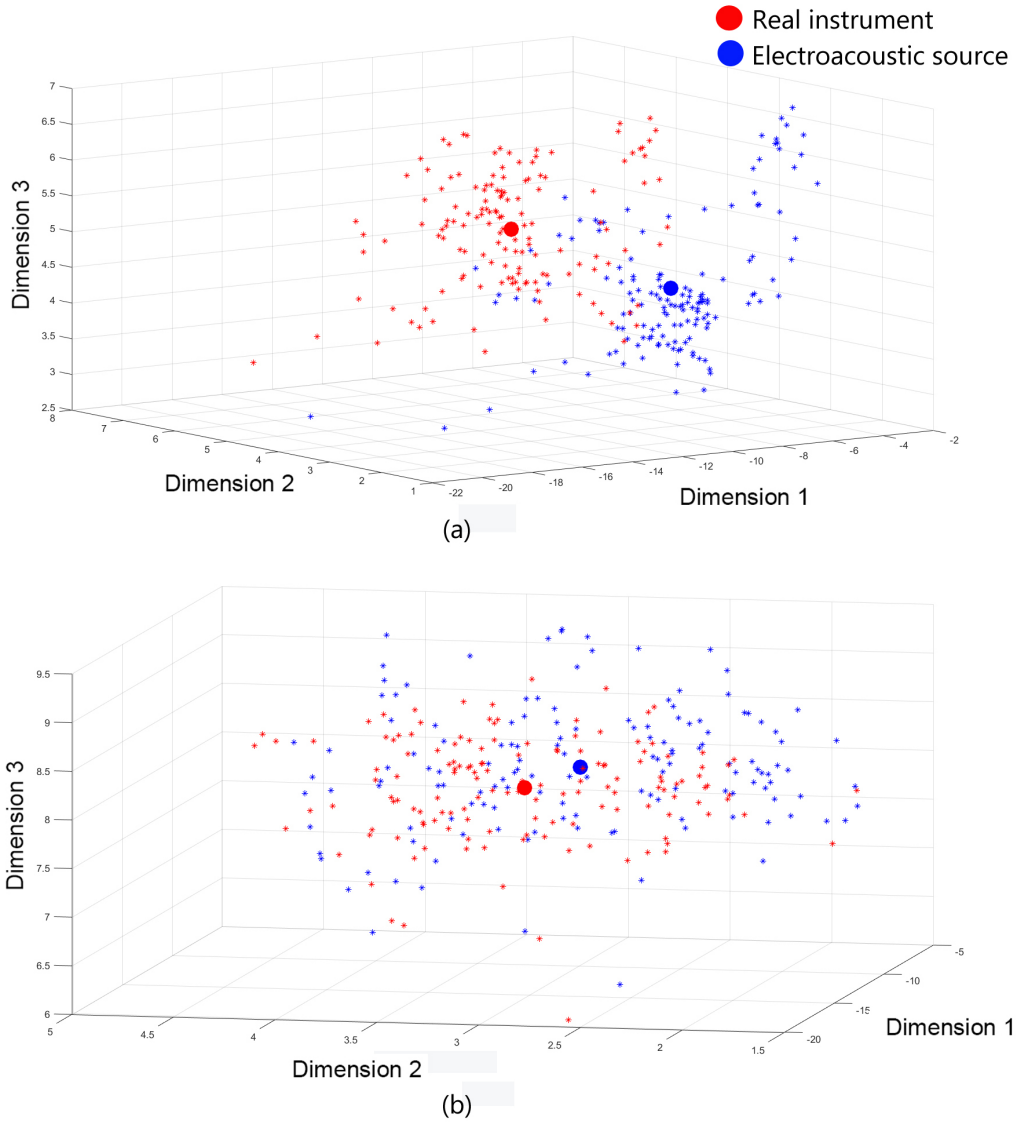


Figure 6.5: Cluster distribution of PCA transformed MFCC features for (a) a sample with low similarity rating of 3.76 and Euclidean distance of 3.27, (b) a sample with high similarity rating of 7.91 and Euclidean distance of 0.41 (red and blue represents data points of the real instrument and electroacoustic source respectively, with the spheres with specific colors denoting their centroids).

behavior, deviating from this negative relationship and potentially limiting its generalizability; in contrast, the most extreme samples (highlighted in red in Figure 6.4) consistently demonstrate a negative correlation, supporting an overall negative trend. This demonstrates the potential of a modeling approach to assess the perceptual similarity between binaural sound samples of diverse musical instruments captured from in-situ conditions having unique room acoustic characteristics.

As a visual representation of the transformed MFCC featurespace, Figure 6.5 shows the cluster distribution PCA transformed MFCC features derived from every 100 ms time window of real instrument and electroacoustic source. Two particular source pairs having very high and very low similarity ratings (7.91, 3.76 respectively) are presented here by plotting the data points of the transformed MFCC feature space using the first three principal components, and highlighting the centroids of the two sound samples (red sphere corresponds to the centroid of real instrument data, while blue sphere corresponds to the electroacoustic source data). According to the previous observations, the highly dissimilar pair of samples are shown to have separated distributions of data points with relatively distant centroids. On the other hand, for a highly similar pair of sound samples, the data points of the two samples are overlapped, and the centroids are closely spaced.

While this modeling approach is basic, and the trends are clearly observable for extreme samples, it may not be generalizable. However, this approach establishes a basis to explore further on modeling the binaural similarity between sound samples even for in-situ conditions with different acoustic conditions.

6.3 Summary

The perceptual differences in the sounding impression due to the dynamic directivity of real instruments and their electro-acoustic counterparts were analyzed in terms of naturalness and similarity ratings. Although the real musical instrument was rated to be more natural in most cases, the electro-acoustic sources also exhibited similar naturalness ratings to that of a real instrument in specific acoustic environments. Even if considering the colouration difference that can occur in close-miking, the electro-acoustic sources with distinct directivity patterns showed a similar distribution of the naturalness rating for a highly directive instrument like a trumpet.

When it comes to similarity ratings, although a rudimentary approximation of a real instrument by an electroacoustic counterpart mostly does not achieve perceptual closeness to the real instrument, certain acoustic conditions–characterized by room acoustic attributes and relative source orientation–tend to obscure the large directivity differences between the sound sources. An interesting thing observed in this study is that, while keeping the input signal, room acoustic environment and loudness to be the same, no major perceptual differences were observed between the recordings of

the two electro-acoustic sources in specific acoustic environments, despite their significantly different radiation characteristics. Even for a highly directional instrument like a trumpet, this trend remains valid. This indicates that the large differences in directivity between the real and electro-acoustic sources is somehow masked in certain acoustic environments. To validate this observation, further studies shall be performed by incorporating advanced listening tests with wide range of verbal attributes of sound, analyze variations in the (spatial) room impulse responses of different sources in temporal, spectral and spatial domain for in-situ performance conditions.

By utilizing MFCC features extracted from the sound samples and performing PCA feature transformation technique, a modeling approach to assess the perceived similarity between binaural audio samples is proposed. Despite its simple and basic nature, the initial result of the model demonstrate noticeable trend for extreme samples with very low and very high similarity impressions. Although the trend is not generalizable, the early-stage modeling idea shall be explored further by utilizing advanced methods using better distance measures and improved features.

Chapter 7

Relevance of high-resolution directivity in ensemble auralization

Given that physically accurate sound field (re)synthesis necessitates detailed directivity information of sound sources, substantial efforts have been invested toward capturing the high resolution directivity of sound sources for virtual acoustic applications. However, the perceptual significance of the high spatial and spectral resolution of the directivity of sound sources remains unclear. When it comes to practical applications, instead of the high-resolution measurements, incorporating a perceptually significant representation of a sound source directivity could facilitate the reduction of computational efforts in a perceptually plausible modeling of the source for auralization. This is particularly important when it comes to the auralization of ensemble performances in virtual reality applications, as multiple sound sources need to be rendered simultaneously in real time. As a first step towards this goal, this study explores the perceptual relevance of spatial resolution of directivity for different numbers of sound sources (from 1, 2, to 5). This is carried out by analyzing the perceptual similarity of sound samples created with various spatial resolutions of directivity. This study also employs two extreme cases for the room acoustic environment (echoic and anechoic) as well as the instrument type (trumpet with 'unidirectional' characteristics, and violin with 'multi-directional' characteristics) to explore their role in directivity perception. The content of this chapter is reproduced from the following research work:

J. Thilakan, A. C. Marruffo, L. R. Paz, D. Ackermann, T. Grothe, M. Kob. "Perceptual relevance of high-resolution directivity in the simulation of musical ensembles" (manuscript under preparation).

7.1 Materials and methods

The perceptual relevance of high-spatial-resolution of directivity patterns in the simulated acoustic environments was tested by perceptually comparing sound samples having directivities with different degrees of spatial resolutions, in the context of an ensemble performance with varying numbers of constituent sound sources (1, 2, and 5). This was done by considering a 15^{th} order Spherical Harmonics (SH) representation of directivity data as a reference and comparing it against lower resolutions created by the truncation of SH. Binaural Room Impulse Responses (BRIRs) of individual sound sources to the receiver were created by varying the directivity filters with different degrees of resolutions in both echoic and anechoic conditions using a GA-based room acoustic simulation software. Binaural audio samples of each source with varying directivity resolutions were generated by convolving these BRIRs for individual sources with anechoic recordings of individual instruments. Following this, except for the single instrument performance scenario, these binaural audio files from multiple instruments were rendered together as a stereo file to create a binaural audio sample of ensemble performance in a particular acoustic setting. A MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) test was employed to compare the similarity of a set of sound samples from a specific performance condition (featuring a specific number of sources in either echoic or anechoic environment), generated with different degrees of spatial resolution of directivity, against the reference sample having the high-spatial-resolution directivity. The preparation of sound samples, including the anechoic source signal recording, directivity processing using SH representation, and room acoustic simulation, are detailed in the coming section. Additionally, the details of the perceptual evaluation using the MUSHRA test are elaborated.

7.1.1 Preparation of audio samples

Recording of sound stimuli

Given that the study examines an ensemble performance with a maximum of five sources, five trumpets and five violins were individually recorded in the anechoic chamber of the Detmold University of Music. The DPA 4099 Core clip-on microphones were employed to capture the individual sources, by positioning them near the bell of the trumpet and the bridge of the violin. These microphones have a frequency response of 20 Hz – 20 kHz with an effective frequency range of 80 Hz–15 kHz ($\pm$ 2 dB) at 20 cm distance, and possess a super-cardioid directivity [113]. Moreover, the potential chance of the movement of the source are not expected to influence these recordings. Two musicians, specializing in trumpet and violin, were hired for recording purposes, and they were asked to perform a dedicated musical piece utilized in the previous studies (see Appendix B), which covered the entire pitch range of these instruments. Covering the

whole pitch range of the instruments is expected to excite the possible variations in the directivity filter employed in the simulations. As an effort to achieve a blended ensemble sound, the musicians were asked to perform the piece with maximum consistency across the different instruments, and they utilized a metronome to maintain the temporal synchrony in each take with different instruments. Despite lacking the intrinsic attributes of ensemble sound evolved by joint performance strategies and acoustic feedback, these anechoic recordings of individual instruments are expected to be 'clean' and noise-free, which is particularly essential for this study. A high-pass filter at 150 Hz was applied to the selected samples of individual instrument recordings to minimize the noises from breathing, etc., and these samples were subsequently utilized for auralization purposes.

Directivity representation using Spherical Harmonics

The Spherical Harmonics (SH) represent solutions of a Helmholtz equation in the spherical coordinate system, and it is commonly used to represent functions varying on a sphere. Since higher-order spherical harmonics are shown to efficiently and accurately model a spatial function using a compact representation, it has been widely utilized in the field of spatial audio and virtual reality acoustics for diverse applications such as the representation of the directivity of sound sources, modeling the HRTF of the receiver, sound field decomposition, and many more [163; 164; 165; 166]. The normalized spherical harmonic base functions Y_l^m that are mutually orthogonal, are defined as,

$$Y_l^m(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(cos(\theta))e^{im\phi} \tag{7.1}$$

where θ, ϕ represent the azimuth and elevation of the spherical coordinates, and the P_l^m represents the associated Legendre polynomial with degree l (l=0,1,2..) and order m (m=-l...+l). A spherical function, $f(\theta,\phi)$, that is assessed on the surface of a sphere can be represented as a weighted sum of spherical harmonic base functions:

$$f = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} a_l^m Y_l^m \tag{7.2}$$

where with a_l^m being the weights of the corresponding SH functions. Therefore, any complex spatial function can be represented as a linear combination of necessary higher-order spherical harmonic functions with appropriate weights. In practical situations, the representation of directivity of sound sources sampled using q number of spatially distributed measurement points, which estimate the pressure function $p = f(\theta_q, \phi_q)$, can be represented with an N^{th} order SH functions using a Least-Squares method [163]:

$$f(\theta_q, \phi_q) = \sum_{l=0}^{N} \sum_{m=-l}^{l} a_l^m Y_l^m(\theta_q, \phi_q) \tag{7.3}$$

where the maximum value of N is chosen such that $(N+1)^2 \leq q$. This can be represented in a matrix form:

$$f = Ya \tag{7.4}$$

where

$$f = \begin{bmatrix} f(\theta_1, \phi_1) \\ f(\theta_2, \phi_2) \\ \vdots \\ f(\theta_q, \phi_q) \end{bmatrix} \tag{7.5}$$

with Y of dimension $q \times (N+1)^2$ having q number of equations with $(N+1)^2$ unknown variables

$$\mathbf{Y} = \begin{bmatrix} Y_0^0(\theta_1, \phi_1)\, Y_1^{-1}(\theta_1, \phi_1) \ldots Y_N^N(\theta_1, \phi_1) \\ Y_0^0(\theta_2, \phi_2)\, Y_1^{-1}(\theta_2, \phi_2) \ldots Y_N^N(\theta_2, \phi_2) \\ \vdots \\ Y_0^0(\theta_q, \phi_q)\, Y_1^{-1}(\theta_q, \phi_q) \ldots Y_N^N(\theta_q, \phi_q) \end{bmatrix} \tag{7.6}$$

and

$$a = \begin{bmatrix} a_0^0 \\ a_1^{-1} \\ \vdots \\ a_N^N \end{bmatrix} \tag{7.7}$$

A solution to this from Least-square method with $(N+1)^2 < q$, an over-determined condition [163; 166], is given to be

$$a = Y^{\dagger} f \tag{7.8}$$

where $Y^{\dagger}$ is the pseudo-inverse. The solution when $(N+1)^2 = q$ [166] is given to be ,

$$a = Y^{-1} f \tag{7.9}$$

While an accurate reproduction of a high resolution directivity can be achieved with an adequate high SH order, truncation of the SH order (i.e., reducing the N value) reduces the complexity by employing a smaller number of SH basis functions and their associated weights.

Processing of directivity data

Two instruments involved in the study, trumpet and violin, exhibit contrasting directivity characteristics. The bell of the trumpet serves as the main radiation point, exhibiting omnidirectional characteristics up to 500 Hz, and then radiating mainly along the axis of the bell, showing rotational symmetry relative to the bell axis [16; 58]. Unlike the trumpet, the violin lacks a defined shape for directing the sound energy, leading to intricate directional characteristics. The vibrating plates, with different points on the plates vibrating at varying amplitudes and phases, and the f-hole of the instrument mainly contribute to the directivity of violins [16; 54]. While the violin exhibits omnidirectional characteristics up to approximately 600 Hz, it produces complex directional patterns for higher frequencies which are primarily radiated from the instrument's top plate [16] (further details on the directivity attributes of these instruments are discussed in section 1.2.3).

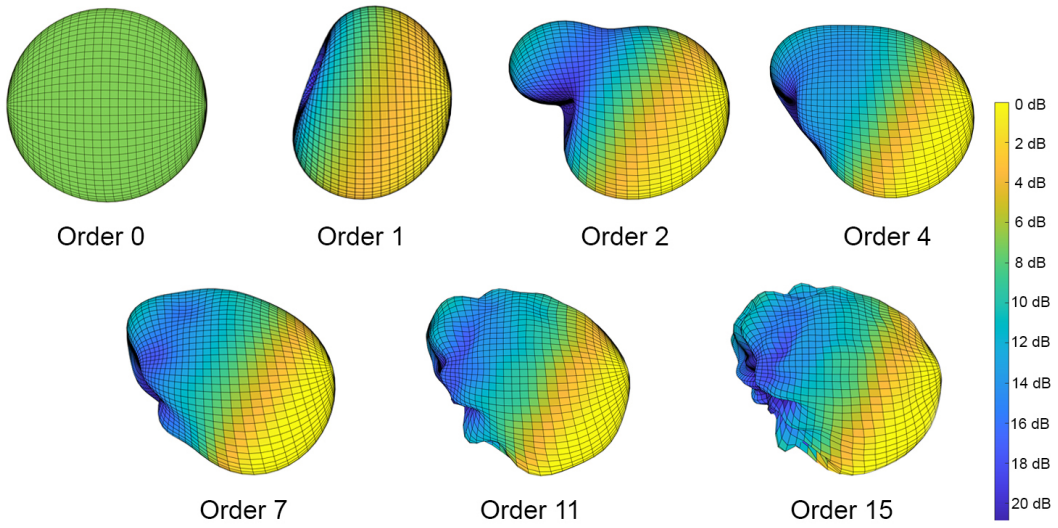


Figure 7.1: Visualization of directivity patterns of trumpet created with truncation at different SH orders (based on the data from [63]).

The high spatial resolution directivity data of these instruments were obtained from the database published by the Spatial Audio Laboratory of Brigham Young University [63; 64]. Although these directivity data were captured with a 5° angular resolution, the data published in the repository are derived from a 15^{th} order SH expansion of these measured data averaged for $1/3^{rd}$ octave bands. Previous studies on directivity perceptions with various kinds of sound sources under different acoustic conditions haven't reported perceptual differences between samples from a 10^{th} order SH and higher-orders [90; 91]. Therefore, the 15^{th} order SH representation is considered to be adequate for serving as a high-resolution reference. Although the Chebyshev quadra-

ture method was employed for the SH expansion of the published data, previous studies suggest that both the Chebyshev quadrature and the least-square methods exhibit the same error level up to around the SH order of 35 [163]. Therefore, the Least square method was utilized in this study for the SH transformation and truncation processes.

The directivity data utilized in the investigation were normalized to 0 dB as the maximum across all frequency bands. Consequently, because of the omnidirectional characteristics, the overall energy radiated in low-frequency bands was relatively higher than in the more directionally focussed high-frequency bands, and it is particularly valid for highly directional instruments like trumpets. This could potentially introduce an unnatural spectral coloration in the auralized output. To compensate for this, a method known as 'Diffuse equalization' [167] was employed to normalize the overall radiated energy across frequency bands. These diffuse normalized directivity data of the two instruments were afterward utilized for the spherical harmonics transformation. As the directivity data utilized were constrained to the 1/3rd octave bands from 160-3150 Hz for trumpet and 200-2500 for violin, the directivity of the highest available octave band in the data was used for the frequency bands above this range, in the simulations.

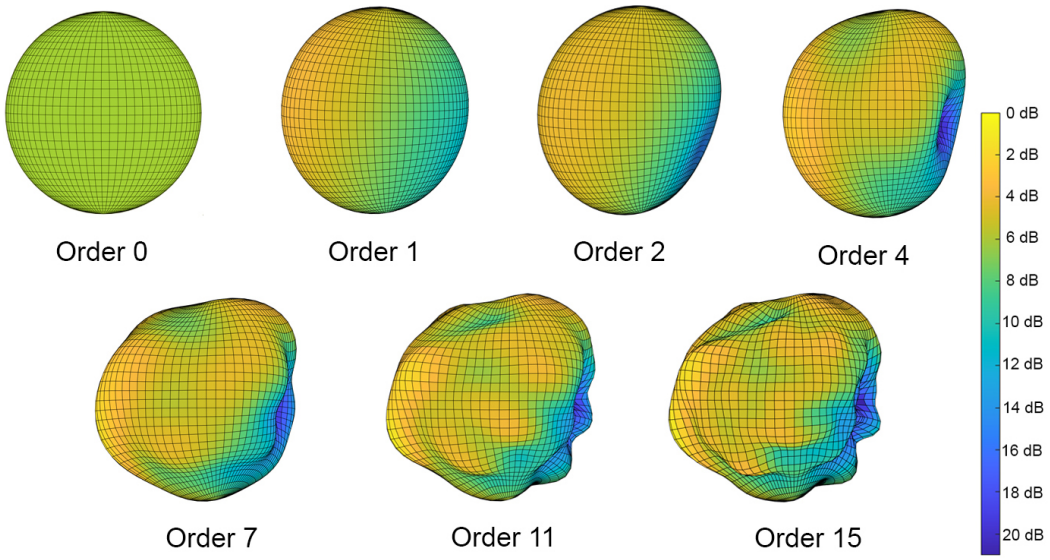


Figure 7.2: Visualization of directivity patterns of violin created with truncation at different SH orders (based on the data from [64]).

The spherical harmonic transform, utilizing the least-square method, was applied to the diffuse equalized directivity 1/3rd octave band directivity data to generate the 15th order high-resolution reference for the investigation. Following this, the coefficients of the 15th order directivity were truncated to obtain lower-order SH shapes that represent lower-resolution directivity data. For this investigation, directivity filters with

SH orders 0 (omnidirectional), 1, 2, 4, 7, and 11 were generated by truncation, and later rendered into OpenDAFF format [168] files for utilization in RAVEN room acoustic simulation software. The directivity patterns of the chosen SH orders for trumpet and violin are presented in Figure 7.1 and Figure 7.2 for visual comparison.

Room acoustic simulation

This study utilized RAVEN (Room Acoustics for Virtual ENvironments), a room acoustic simulation platform developed for academic purposes [95; 103], for the estimation of BRIRs from the individual sources to the receiver with varying source directivity filters. RAVEN is a GA-based hybrid room acoustic simulation platform that integrates the image source method for early reflections and the ray tracing method for late reverberations, and it also incorporates frequency-dependent absorption and scattering properties of the boundaries (more details on the GA-based simulation is given in section 1.2.4).

A simplified model of a chamber music hall (previously utilized in [57]), with a volume of 1953 m^3 and reverberation time of 1.02 s, was chosen as the room acoustic environment for the ensemble sound auralization. The 3D model of the room acoustic environment is presented in Figure 7.3. Rather than reproducing a specific acoustic environment accurately, the objective of this investigation was to analyze the role of room acoustic reflections within a simplified acoustic environment. Accordingly, the utilized room acoustic model had a relatively simplified geometry with only three boundary materials, each having different absorption parameters for the walls, stage area, and audience area. As reported in [57], the major room acoustic parameter values (described in the Appendix A) estimated according to the ISO 3382-1 standards [37] are T_{20} = 1.02 s, C_{80} = 4.55 dB, EDT = 1.06 s, G = 12.40 dB, J_{LF} = 0.24, respectively.

The five sound sources on the stage were placed with a separation of 1 meter, and the middle source (S1) was positioned 1 meter away from the symmetric plane of the hall to avoid unwanted symmetric effects. The receiver was placed at a distance of twice the critical distance from the sources to have a better impression of the diffuse field. Moreover, instead of placing it right in front of S1, it was 1.5 away from the middle plane to have binaural cues from the sources. Additionally, the height of the sources and the receiver were kept at 1.7 meters. The HRTF of Fabian HATO [169] was used to model the receiver, and the different directivity filters with varying spatial resolutions were utilized to model the sound sources. For each directivity filter of a particular instrument, the simulation was performed by employing a hybrid algorithm that combined a second-order image source method and 100,000 rays for ray tracing, to capture BRIRs from the individual sources to the receiver. The simulations with the described conditions were performed, and the the BRIRs corresponding to each directivity filter applied to the five sources were captured in this echoic condition.

In addition to this, all the boundaries (including the stage and audience area) of the chamber music hall model were replaced with a material with 100% absorption, which led to an anechoic environment condition. In this way, while the room acoustic reflections were eliminated, the spatial attributes such as the source localization and distance perception remained unchanged in the two conditions, which facilitated future comparison on the role of room reflections. The simulations with the above mentioned settings were repeated in this anechoic condition, and the BRIRs corresponding to each directivity filter applied to the five sources were captured.

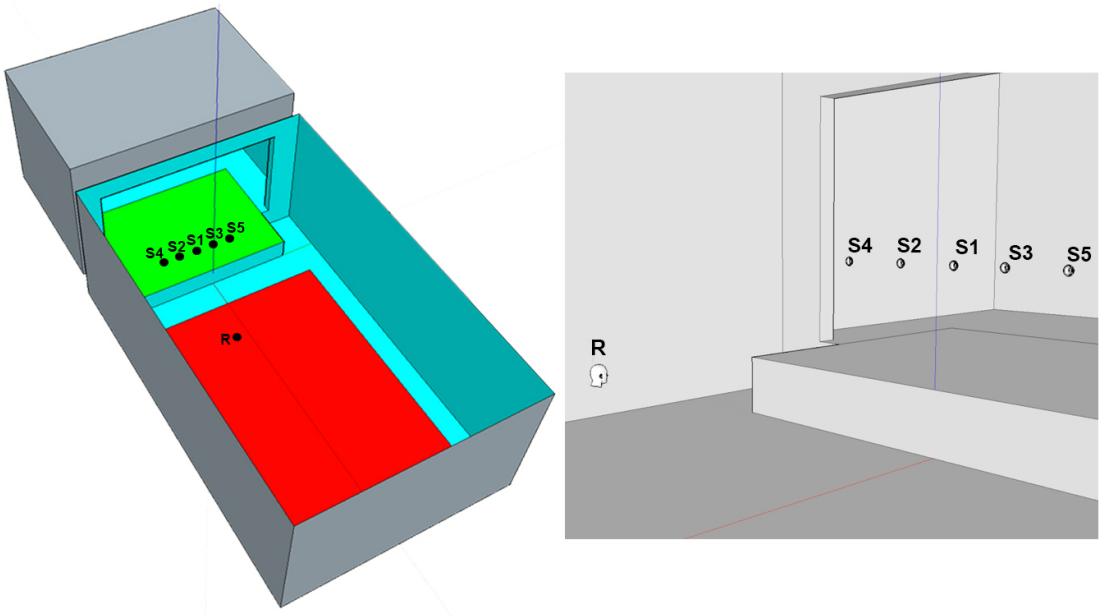


Figure 7.3: The 3D model of the room acoustic environments (the left image represents the geometry of the chamber music hall while the right image represents a zoomed view of anechoic version with all boundaries with 100% absorption, with sources and receivers denoted as 'S' and 'R').

Processing of sound samples

BRIRs corresponding to a particular source, with different directivity filters, were convolved with an anechoic recording of a specific instrument to generate auralized samples of the source with different directivity variations. Apart from solo instrument performance (with S1), for configurations involving two instruments (with S1 and S2), or five instruments (with S1 to S5), the binaural sound samples from individual corresponding were rendered together into stereo format to create the binaural audio file of the ensemble performance.

Variations in the room acoustic conditions and the type and number of instruments involved cause differences in the loudness across the 12 conditions (3 different num-

bers of sources × 2 different instruments × 2 room acoustic conditions). Along with the spectral and spatial attributes, the changes in the directivity filter within a specific condition are expected to produce differences in loudness as well. Therefore, for the perceptual comparison across these conditions, the loudness levels of reference samples (SH order 15) across the 12 conditions were normalized to the same loudness level, and the scaling factor of the reference sample was used to adjust the loudness of other lower order directivity samples in each condition. In this way, the overall loudness between the 12 test conditions was maintained to be similar, while the difference in loudness caused by the different directivity filters in each condition was retained.

7.1.2 Perceptual evaluation

The MUSHRA test was performed using the Web-MUSHRA [170], a web audio API test platform that is compliant with the ITU-R BS.1534 recommendations [171]. In contrast to other test designs, the MUSHRA test allows a simultaneous comparison of a relatively higher number of samples, which is particularly useful in the context of evaluating different kinds of audio processing methods. The perceptual evaluation involved a group of 21 participants, consisting of tonmeister students, sound recording engineers, and expert musicians. The participants had musical and ear training backgrounds and also were experienced in critical listening. Moreover, the ability of musicians over non-musicians to selectively attend and analyze complex features of sound, as observed in previous studies [17; 18], qualifies them as expert listeners for this perceptual evaluation.

The graphical user interface of the WebMUSHRA platform utilized in this study is presented in Figure 7.7. The test consisted of 12 trials, each corresponding to 12 different conditions involved. In each trial, 7 sound samples with different degrees of directivity resolution were provided which included samples created with SH resolution of order 0, 1, 2, 4, 7, 11, and the hidden reference with SH order 15. Each trial featured a separate reference sample button, and the participants were asked to rate the 7 test conditions (including the hidden reference) in terms of the similarity impression by comparing them with the given reference on a scale of 0 to 100. A highly dissimilar sample compared to the reference corresponded to a low similarity rating, and vice versa. The order of trials, and the test conditions in each trial, were randomized between the participants to mitigate errors from direct comparison and to minimize the memory retention effects on the ratings.

Following the ITU recommendations, the WebMUSHRA GUI interface allows users to choose a portion of the samples, listen to it in a loop, and seamlessly switch between test samples without interrupting the audio playback, etc. Although MUSHRA tests usually include hidden anchors, typically low-pass filtered versions of reference with cut-off frequencies at 3.5 kHz and 7 kHz, such anchors were not utilized in this test. In some of the conditions involved, the perceptual differences between some of the

samples generated with different directivity filters were very subtle. Consequently, employing a highly different anchor sample would increase the separation of ratings between the anchor and the 7 samples, while reducing the separation between the 7 test conditions, resulting in a reduced range of variation of similarity ratings among the test conditions. A pilot test conducted with the samples and anchor supported these observations. Given that the test examines the perception of details in the directivity filter, the condition of no directivity, i.e. omnidirectional source having SH order=0 could be regarded as a lower anchor.

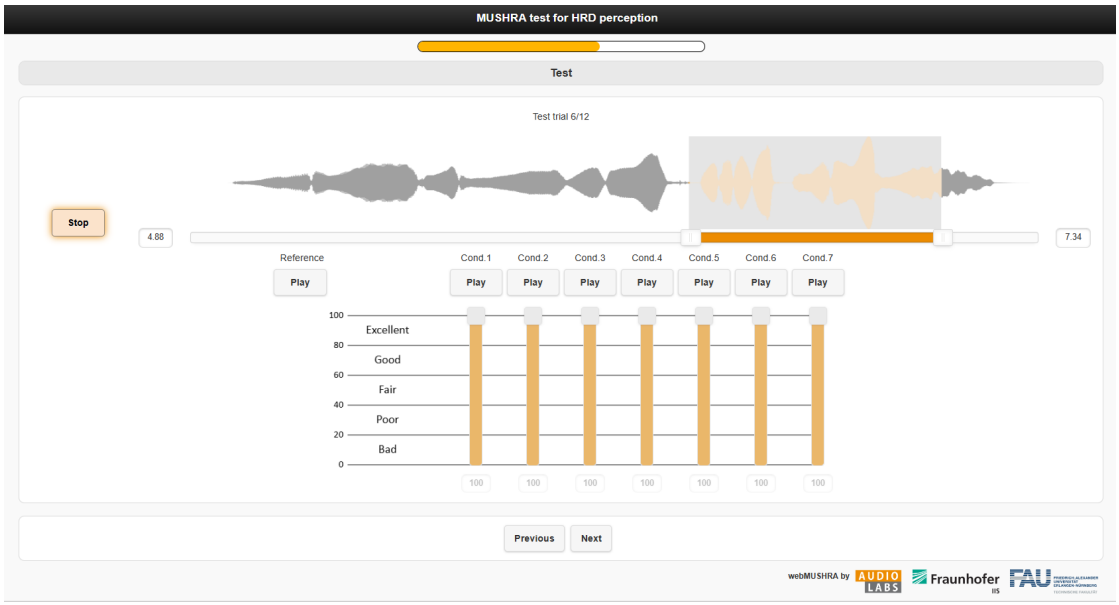


Figure 7.4: The graphical user interface of MUSHRA test.

While the Web-MUSHRA allows hosting the test online, it was decided to perform the test in a controlled environment with proper guidance to achieve more reliable results. Therefore, the test was conducted in an acoustically treated room of Erich Thienhaus Institut, and the binaural samples were presented using Beyerdynamic DT 990 pro closed-back headphones to the listeners. Both verbal and written instructions on the objective of the test and the test procedure were given to the participants at the beginning of the test. To make the participants familiar with the test design, calibrate their hearing, and minimize their initial biases, a training session consisting of 8 trials (only 1 and 5 sources $\times$ 2 different instruments $\times$ 2 room acoustic conditions) was conducted at the beginning of the test with samples of SH order 0, 2, 7, and 15 as reference. Following this training session, the actual test with 12 trials took place, which took an average of 30 minutes to complete. At the end of the test, a short questionnaire was presented to the participants to evaluate the features they utilized to judge the dissimilarity between samples. This included rating the features used in similarity

judgment—such as *timbre, spatial cues, loudness, reverberance*, and *clarity* —on a scale of 1 to 5, and also proposing any other features they might have utilized.

According to the MUSHRA guidelines, if an assessor fails to spot the hidden reference for more than 15% of conditions involved in the test (ca. 2 out of 12 conditions), by providing a rating below 90%, the assessor should be excluded from the test to get consistent and reliable results. Accordingly, 6 out of the 21 participants were excluded, and the further analysis was performed using the test responses of 15 participants.

7.2 Result and discussion

The distribution of the similarity ratings of different directivity resolution samples from 15 participants is shown to be non-normally distributed (validated with the Shapiro-Wilk test, $p < 0.05$). Therefore, a box plot, which excludes outliers and does not assume any specific distribution, is employed here for comparison between the similarity ratings.

Figure 7.5 illustrates the distribution of similarity ratings of trumpet samples of different spatial resolutions for various numbers of sources under echoic and anechoic conditions. Across all conditions, the 0^{th} order omnidirectional sample received the lowest rating and median. Similar to the previous findings, the similarity ratings of samples significantly increase from 0^{th} order with an increase in the order of SH utilized for directivity filter modeling, reaching a plateau after a threshold with no major improvement. By attaining a comparable similarity impression close to that of the reference sample, the truncated SH order at this threshold point is expected to provide a perceptually plausible representation of the high-resolution directivity reference. Observing the trend of the median curve across the conditions, it seems that the overall threshold point is likely around the 4^{th} order, after which slight improvements occur.

Considering the anechoic conditions, from the 4^{th} order onwards, all the conditions are shown to have a median value surpassing 90. Except for a slight deviation of the 11^{th} order, ratings from the 4^{th} order to the 15^{th} order reference demonstrate reasonably comparable distributions. Therefore, the 4^{th} order SH representation can be expected to be sufficient for a perceptually plausible auralization of the trumpet in such a condition. Although details on the sidelobes of the instrument are missing, the relatively simplified directivity feature of the trumpet, characterized by its main radiating lobe, is almost present in the 4^{th} order directivity shape, as illustrated in Figure 7.1. Therefore, the variation in sound from the detailed characteristics of trumpet directivity may not be necessarily discernible in a reflection-free environment. This observation holds true across the different numbers of sources (1, 2, and 5), indicating that the increase in the number of sound sources does not alter the similarity perception across different SH orders. Therefore, it can be hypothesized that the requirement of directivity resolution needed for a perceptually plausible auralization is insensitive to the increase in the number of sources involved.

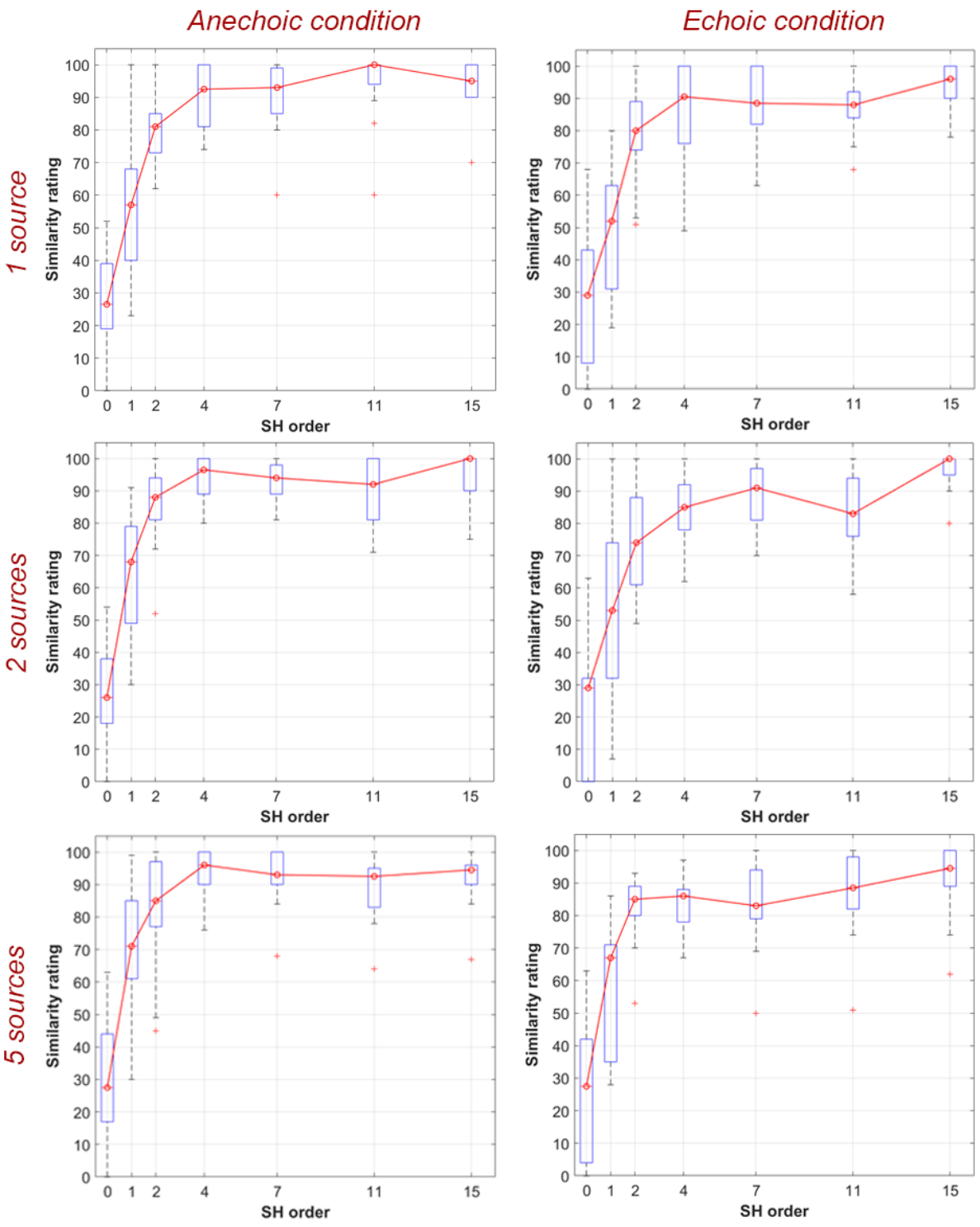


Figure 7.5: Distribution of similarity ratings of different SH order samples for trumpet: left and right columns indicate anechoic and echoic conditions, respectively, with the number of sources increasing from 1 to 2 to 5 from top to bottom.

The echoic condition exhibits a relatively lower similarity rating distribution compared to the anechoic condition, which is consistent with the observations from previous studies. Except for a few cases, the median values of the samples under echoic conditions are lower than those in anechoic conditions. Moreover, the interquartile range (IQR) and whiskers are wider in echoic conditions, indicating a relatively higher variance in the ratings. Recent studies on directivity modeling in simulated room acoustic environments indicated a need for higher order detailing of directivity for the simulation of the early-reflection part of the RIR, while an averaged directivity suffices for the diffuse reflections [71]. Therefore, as speculated in [91], the early reflections could be the potential reason for this difference, although further investigations are needed to verify this. Similar to the anechoic condition, no significant changes were observed between the distributions of similarity ratings for different numbers of sources. Consequently, the hypothesis that the number of sources doesn't alter the importance of directivity resolution is being supported in this condition as well. Identifying the perceptually plausible directivity threshold is challenging here due to the non-overlapping and relatively wider distributions compared to the reference. Based on the median curve, it can be expected between the 4^{nd} and 7^{th} order, however more sophisticated listening tests are required to confirm this threshold.

Condition	0^{th} order	1^{st} order	2^{nd} order	4^{th} order	7^{th} order	11^{th} order
1 trumpet in AC	**<0.001**	**<0.001**	**<0.001**	0.323	0.251	0.523
2 trumpets in AC	**<0.001**	**<0.001**	**0.016**	0.396	0.179	0.151
5 trumpets in AC	**<0.001**	**<0.001**	0.092	0.304	0.780	0.380
1 trumpet in CH	**<0.001**	**<0.001**	**0.003**	0.195	0.138	0.083
2 trumpets in CH	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.023**	**0.002**
5 trumpets in CH	**<0.001**	**<0.001**	**0.009**	**0.021**	0.075	0.404

Table 7.1: p-values of Mann-Whitney U test comparing similarity ratings of 15^{th} order reference with lower orders for trumpet samples in Anechoic chamber (AC) and concert hall (CH).

To examine statistically significant differences in the distribution of similarity ratings between the reference (15^{th} order) and lower order samples in each conditions, the Mann-Whitney U test [141] was conducted. This non-parametric alternative to Student's t-test was selected due to violations of normality in the distributions, as confirmed by the Shapiro-Wilk test ($p < 0.05$). Table 7.1 presents the p-values from pairwise Mann-Whitney U tests comparing lower-order samples against the 15^{th} order reference. As observed above, the lower order samples including 0^{th} and 1^{st} order samples consistently exhibit a significant difference in distribution of ratings compared to the reference. In anechoic conditions, it is failed to detect statistically significant difference in the distribution of ratings from the 4^{th} order onward, supporting the ear-

lier observation that improvements in perceptual similarity ratings become negligible beyond this order. When it comes to echoic condition (i.e., in the concert hall), while the 0^{th}, 1^{st}, and 2^{nd} order samples remain significantly different from reference for all conditions, no clear threshold can be observed here beyond which no statistically significant differences are observed, as higher-order samples also yield statistically significant p-values. This supports the requirement of relatively higher order directivity representation for room acoustic environments as compared to anechoic conditions for perceptually convincing auralization.

The distribution of similarity ratings of violin samples for different numbers of sources under echoic and anechoic conditions is presented in Figure 7.6. Although SH order 0 possesses the lowest ratings and the similarity ratings generally increase with higher orders, unlike the trumpet case, the 0^{th} and 1^{st} order violin samples exhibit improved similarity ratings compared to the reference. While the median values of similarity distributions for SH orders 0 to 4 varied from about 25% to 95% for trumpets, they converged to about 70% to 95% for violins. This suggests that the perceptual distinction between the samples is considerably smaller in the case of violins across SH orders 0 to 15. Unlike the trumpet, demonstrating a highly directional beam-like directivity characteristic at high frequencies (see Figure 7.1), the violin's more complex directivity characteristics at high frequencies might bear a relatively better resemblance to its corresponding truncated SH order 1 and 2 directivities (see Figure 7.2). This could be the potential reason for the improved similarity of the lowest orders with reference, in violins. Although identifying the perceptually plausible directivity threshold is challenging here, mostly the median values from the 4^{th} order onward remain close to or above 90% across the three different instrument conditions, suggesting the possibility of the fourth order being the tentative threshold point.

Considering the role of room acoustic reflections, the echoic condition consistently exhibits lower similarity rating distributions in comparison to the anechoic conditions, except for some outlier points. These two independent observations from the trumpet and violin samples underscore the degradation of similarity ratings by room reflections, which is consistent with [91]. Furthermore, no notable trend was observed among the three different conditions with varying numbers of violins, which remains consistent across both echoic and anechoic conditions. This supports the earlier observation that the number of sources does not alter the importance of the directivity resolution.

Table 7.2 demonstrates the p-values from pairwise Mann-Whitney U tests comparing the distributions of similarity ratings of lower-order violin samples against the reference across different conditions. Despite having relatively higher similarity ratings than the trumpet, the 0th and 1st order violin samples remain significantly different from reference in all source conditions from both echoic and anechoic environments. In anechoic conditions, except for an outlier, no statistically significant difference was

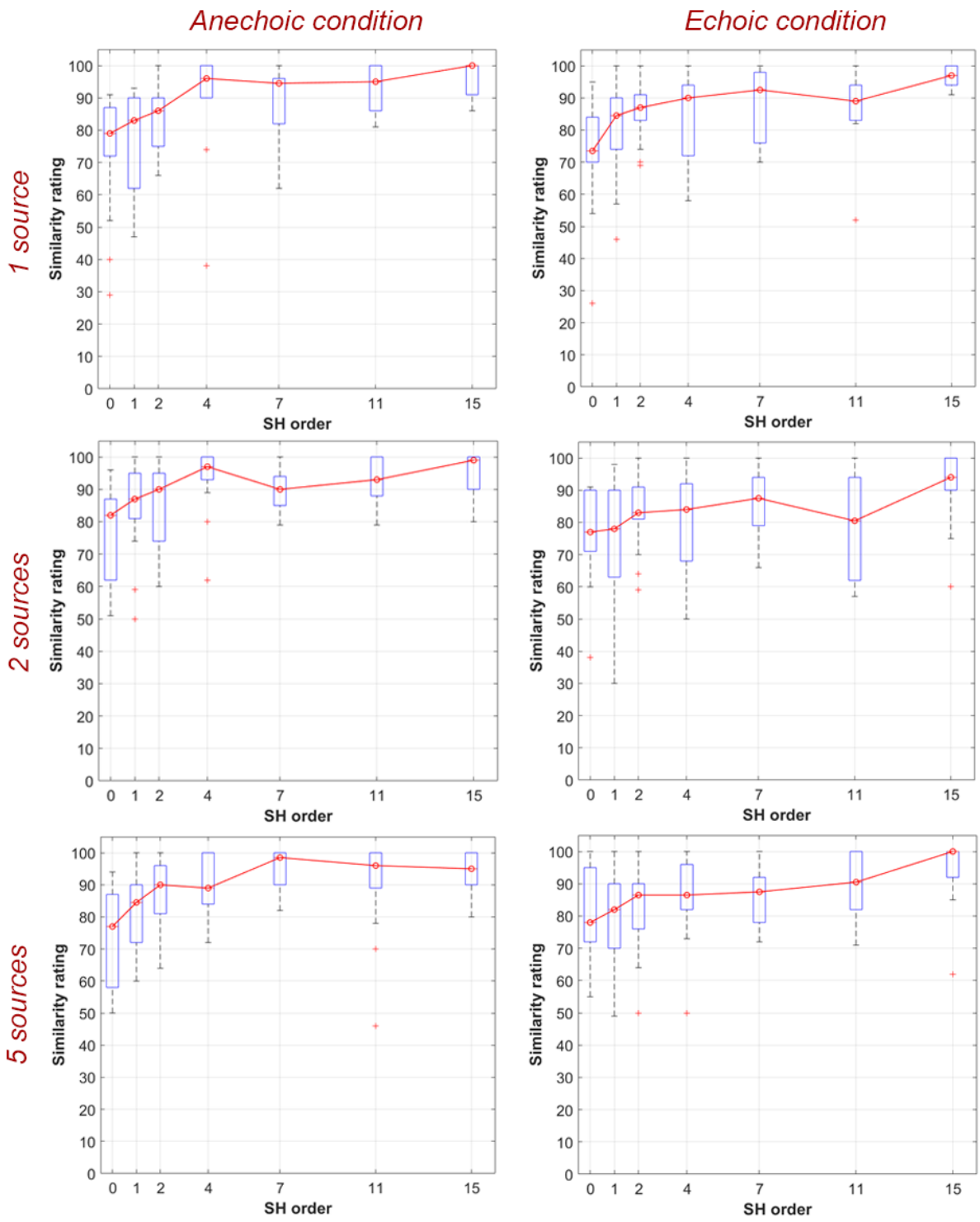


Figure 7.6: Distribution of similarity ratings of different SH order samples for violin: left and right columns indicate anechoic and echoic conditions, respectively, with the number of sources increasing from 1 to 2 to 5 from top to bottom.

Condition	0^{th} order	1^{st} order	2^{nd} order	4^{th} order	7^{th} order	11^{th} order
1 violin in AC	**<0.001**	**<0.001**	**0.003**	0.281	0.079	0.114
2 violins in AC	**<0.001**	**0.007**	0.057	0.790	**0.028**	0.255
5 violins in AC	**<0.001**	**0.005**	0.148	0.128	0.737	0.796
1 violin in CH	**<0.001**	**0.003**	**0.001**	**0.005**	0.075	**0.002**
2 violins in CH	**0.003**	**0.006**	0.089	0.084	0.074	**0.035**
5 violins in CH	**0.003**	**0.001**	**0.002**	**0.017**	**0.008**	0.062

Table 7.2: *p*-values of Mann-Whitney U test comparing similarity ratings of 15^{th} order reference with lower orders for Violin samples in Anechoic chamber (AC) and concert hall (CH).

found for samples from 4th order onward when compared to the reference, supporting the possibility of the perceptual threshold being close to the 4th order. However, in echoic conditions, 4th and higher order samples are also shown to have statistically significant difference from the reference, suggesting that a relatively higher-order source directivity representation is necessary in the presence of room acoustic environments.

The distribution of the ratings for the cues that listeners utilized to assess the dissimilarity between the given samples and the reference is depicted in Figure 7.7. The utility of each cue was rated on a scale of 1 to 5, and the results show that the loudness and the timbre emerge as the most effective cues utilized by listeners in the test.

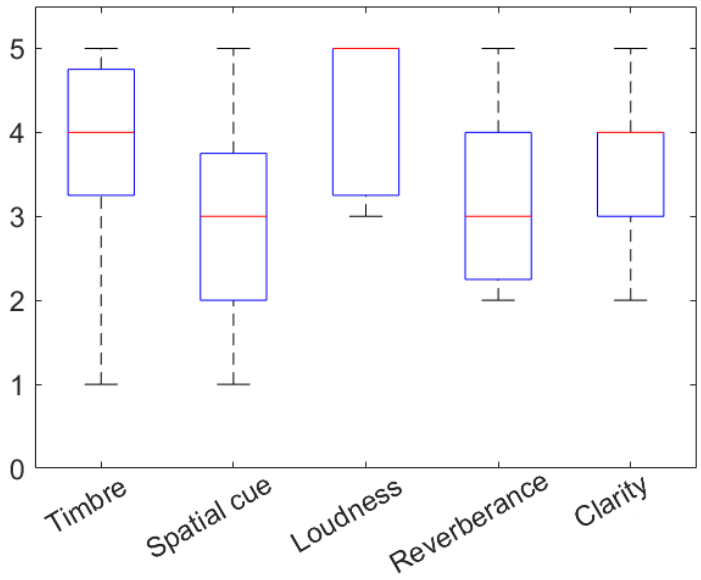


Figure 7.7: The distribution of ratings of cues utilized to assess dissimilarity between sound samples.

7.3 Summary

This study explored the perceptual significance of high spatial resolution of directivity of instruments within the context of auralizing ensemble performances. This was carried out by changing the instrument type (trumpet, and violin) and the number of instruments in the performance (1, 2, and 5) in simulated echoic and anechoic conditions. A MUSHRA test comparing the audio samples created with various degrees of detailing of the directivity of sound sources, generated by truncated Spherical Harmonics (SH) orders of high-resolution reference data of 15th order, demonstrated that a lower-resolution directivity representation could achieve perceptually plausible auralization in comparison to high-resolution directivity data.

The perceptual similarity across different SH order samples when compared to the reference is observed to be significantly influenced by the inherent directivity characteristics of the instruments. While the 0^{th} and 1^{st} order samples were significantly different from the higher order and reference samples of the trumpet, they demonstrated an improved similarity to the higher orders in violins. Furthermore, the lower-order samples in echoic conditions were shown to have relatively lower similarity to the reference as compared to the anechoic environment. This observation suggests that the presence of room acoustic reflections highlights the difference in the directivity between the sound sources, and thereby reduces the perceived similarity between them. This consistent trend, observed across the two independent instruments, is in agreement with previous research findings.

When it comes to the variations in the number of sources involved, the results suggest that the increase in the number of sources from 1 to 5 does not seem to influence the similarity perception ratings across the different SH orders. This trend was consistent for both the trumpet and violin under both echoic and anechoic conditions. These observations suggest that the requirement of relatively lower-order directivity resolution needed for a perceptually plausible auralization is insensitive to the increase in the number of sources involved. This underscores the importance of directivity in the auralization of an ensemble performance.

A truncated SH order that can deliver a perceptually plausible auralization was explored here by analyzing the threshold point where the similarity ratings saturate in comparison to the reference. In the case of the trumpet in anechoic condition, the 4^{th} order SH sample can be considered to be the threshold showing a comparable distribution to the reference, with no significant improvement in similarity ratings observed beyond this. However, in other scenarios, proposing a clear threshold point is challenging, thus necessitating advanced perceptual evaluations and statistical analysis to estimate the threshold. The threshold value could be a function of the characteristic feature of the instrument and also influenced by the room acoustic attributes. However, as stated above, the number of instruments doesn't seem to influence it. Consequently,

although the high spatial resolution is not a requisite for a perceptually plausible auralization, the minimum lower-order resolution representation of an instrument should be utilized in the ensemble auralization with multiple instruments.

The MUSHRA test carried out as an exploratory investigation facilitated the utilization of a higher number of sound samples across different conditions. Based on the insights gained from this test, other test designs, such as adaptive testing or ABX test on a limited set of samples that are perceptually very close, shall be utilized to advance further by precisely estimating the threshold point. Moreover, by advancing from the simplified room acoustic models, the test should incorporate more sophisticated room acoustic simulations that better match real-world conditions. Performing this perceptual evaluation on a wide range of room acoustic environments can provide a better understanding of the room acoustic attributes that emphasize the dissimilarity in the directivity.

Chapter 8

Role of room acoustics in blending perception

While the phenomenon of blending remains a subjective attribute with considerable importance in room acoustic research, only a limited number of studies have explored the direct relationship between room acoustic attributes and perceived blending. Having said that, the absence of source-level blending considerations and limited variations in acoustic environments further constrain these studies. Therefore, several essential questions need to be answered to have a thorough understanding of the formation and evolution of blending. Firstly, it is uncertain if room acoustic reflections always enhance blending, and the extent to which the acoustic environment contributes to samples with different degrees of source-level blending remains unresolved. Furthermore, a comprehensive investigation is required to understand the role of different room acoustic attributes in the blending perception and fine-assess the distinct contributions of source-level blending and room acoustics to the overall perceived blending. This chapter attempts to answer these questions by performing a controlled experiment in which musically realistic in-situ recorded sound samples of two violins having contrasting degrees of perceived source-level blending were auralized through different simulated room acoustic environments for the perceptual analysis of overall blending. Based on the perceptual test ratings, this study proposes a computational modeling approach to evaluate the blending of sound sources in a musically realistic performance setting through the estimation of the distinct contributions made by instrument-level blending and the room acoustic environment. The content of this chapter is reproduced from the following research article:

J. Thilakan, B.T. Balamurali, O.C. Gomez, J.M. Chen, M. Kob, "Exploring the role of room acoustic environments in the perception of musical blending," Journal of the Acoustical Society of America 157.2 (2025), pp. 738-754, https://doi.org/10.1121/10.0035563 (Licensed under a Creative Commons Attribution (CC BY 4.0) license).

8.1 Materials and methods

8.1.1 Data acquisition procedure

Simulation of room acoustic environments

This study utilized ODEON version 16, a GA-based room acoustic simulation software, to generate virtual room acoustic environments. Since ODEON incorporates state-of-the-art room acoustic simulation techniques (described in section 1.2.4) and yields reliable results, it has been employed in many auditory perception-related investigations for simulating acoustic environments [68; 99; 100].

The architectural features of performance spaces have been shown to influence the subjective sensation and judgment of acoustic environments[41]. Therefore, the acoustic environments prepared for the study were generated by diversifying architectural, and acoustical attributes such as room geometry, absorption characteristics of the rooms, and the receiver position, with the aim of eliciting distinct perceptual impressions of blending. Four rooms with rectangular geometry (shoebox shape) were selected for this study. Figure 8.1 demonstrates a schematic diagram of the geometry of four room models incorporated in this study. These rooms had an approximate volume of 500 m^3 (length $\times$ width $\times$ height as $10\times10\times5$ m), 5000 m^3 ($33\times14\times11$ m), 10,000 m^3 ($36\times20\times14$ m), and 15,000 m^3 ($36\times29\times14$ m) respectively which falls within the physically acceptable range of volume for realistic music performance spaces. These rooms will be referred to as 'R1', 'R2', 'R3', and 'R4' respectively henceforth. To make the rooms sound more realistic and convincing, a stage block of height 100 cm (highlighted as pale green regions in Figure 8.1) and an audience block of height 50 cm (highlighted as pale red regions) were incorporated into each of the rooms according to the available size.

To further diversify the acoustic environments, three different variations of each room were generated by changing the absorption coefficients of the wall surfaces from realistically feasible low, medium, and high values. These three variations will hereafter be referred to as 'wet', 'normal', and 'dry' versions, in the given order. Table 8.1 shows the absorption coefficient values implemented for the walls and floor in the three acoustic variants for different frequency bands (higher frequency bands greater than 8000 Hz have same values as 8000 Hz). Typically used absorption and scattering coefficients were applied for the audience area to achieve a realistic listening situation, and they were maintained to be the same in all acoustic variations.

Each of the simulated acoustic environments contained two sound sources (referred to as 'S1', and 'S2') and two receivers (close/near receiver referred to as 'c', and far receiver as 'f'), and their positions are depicted in Figure 8.1. The built-in directivity pattern of violins in ODEON, which is averaged over octave bands with 5°spatial resolution, was used as the directivity filter of sound sources in the simulation. Addi-

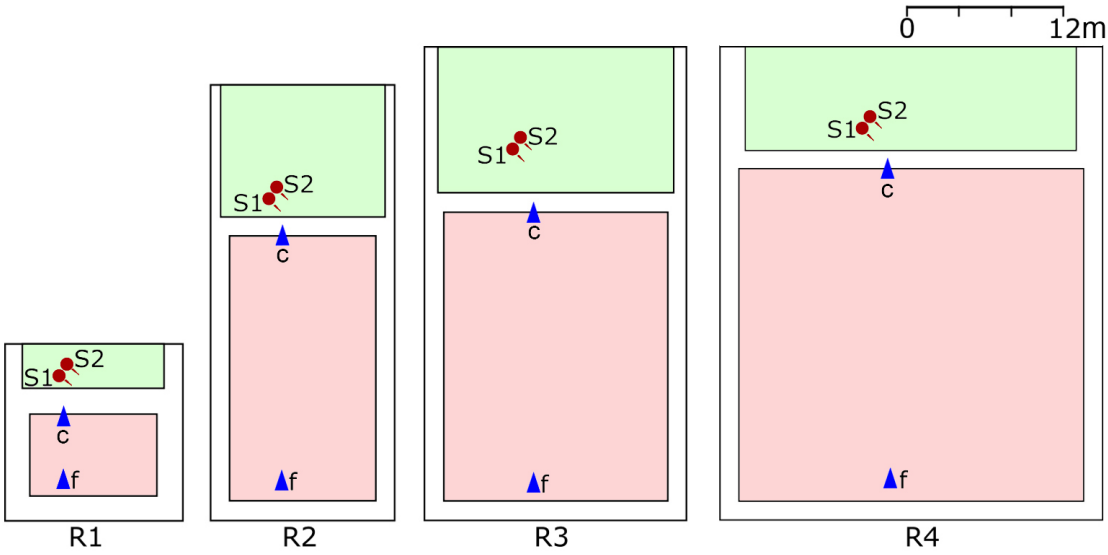


Figure 8.1: The schematic diagram of the geometry of four room models (top view) with stage block (highlighted with pale green region) and audience block (highlighted with pale red region).

Variant	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
Dry	0.18	0.31	0.36	0.40	0.42	0.42	0.43	0.43
Normal	0.18	0.18	0.16	0.14	0.13	0.12	0.11	0.10
Wet	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.05

Table 8.1: Absorption coefficient values applied to dry, normal, and wet variants across different frequency bands.

tionally, the far-field Head Related Transfer Function (HRTF) of the Neumann KU-100 binaural head[172] has been employed for the binaural receiver. Since the directivity of the sound source was not omnidirectional, but rather that of a violin, most parameters are not fully compliant with ISO 3382–1[37] standards. Nevertheless, these objective parameters are still capable of reflecting the perceptual attributes of the sound field generated by the source with distinct directivity [53; 68]. Therefore comparing different acoustic environments using these parameters is still possible.

The sound sources were positioned one meter apart and oriented slightly towards the wall of their left side. To prevent undesirable acoustic effects caused by room symmetry, both the sources and listeners were positioned slightly off-centered from the symmetrical axis of the room. The position of the two receivers was at the first and last rows of the audience area, facing in the direction of the sound sources on stage. Due to the differences in the contribution of direct sound and room reflections, these two receiver positions are expected to represent two distinct acoustic scenarios from

one another. It should be noted that the geometry of the room, coupled with the placement and orientation of the source and receivers, determine the direction and strength of direct sound, as well as the direction of the following early reflections from the room. Consequently, these attributes would be identical among all three acoustic variants of a specific receiver in a room geometry. However, altering the absorption coefficient of the walls and floor changes the strength of early reflections and the late reverberations which results in the divergence between the three acoustic variants of a room objectively and perceptually. Overall, 24 distinct acoustic environments were produced by varying the room geometry, the absorption coefficient, and the receiver position (4 rooms $\times$ 3 acoustic variants $\times$ 2 receiver positions). In addition, an anechoic room was also simulated using R1 and far receiver position by applying 100% absorption to the walls, audience area, and floor.

The following abbreviation scheme will be used in this chapter to denote each of the 25 auditory environments: The room geometry (R1, R2, R3, and R4) followed by acoustic variation ('D', 'N', 'W' for dry, normal, and wet) followed by receiver position ('c', 'f' for close and far). For example, the close receiver in the R3 room geometry with a dry acoustic variation is represented by the notation 'R3Dc'.

Collection of source stimuli

Sound samples were picked from a pool of 50 audio files presented in Chapter 3 that were perceptually assessed in terms of source-level blending by expert listeners. Unlike previous studies on source-level blending which focussed on audio samples of musical notes or chords [14; 30; 33], three musically realistic sound samples featuring two violins having a length of 3 to 5 seconds, with distinct degrees of perceived source-level blending ratings (high, moderate, and poor), were chosen to serve as the test stimuli; they were labeled as 'Stimulus A' with a source-level blending rating of 7.9 ± 1.6 out of a 10 point scale, 'Stimulus B' with a rating of 5.5 ± 2.1, and 'Stimulus C' with a rating of 3.3 ± 1.8. Unlike the previous investigation on source-level blending which utilized a monophonically rendered version of these samples, the individual instrument tracks of the two violins of the selected sound samples were used in this investigation to convolve with the corresponding impulse responses to create the audio samples. The selection of these specific samples was supported by their low standard deviation values in blending ratings, indicating internal consistency, and the absence of prominent indicators allowing easy differentiation between the two violins, such as noticeable deviations in pitch or note onset. Additionally, they were found to possess a relatively lower amount of microphone cross-talk and room reflections in a pilot listening test performed by tonmeisters on individual violin tracks. Along with these three stimuli, four more audio samples with varying source-level blending ratings were also chosen for the purpose of training prior to the listening test (detailed in the coming section).

Preparation of test samples

In each of the 25 simulated acoustic environments, the Spatial Room Impulse Responses (SRIRs) for each source-receiver combination were generated and extracted from ODEON simulations as third-order B-format ambisonics files. For each acoustic environment, the convolution of two SRIRs obtained from the two sound sources to a particular receiver position with the individual monophonic samples of two violins corresponding to each test stimulus is performed in REAPER using MCFX convolver [173]. The levels of each convolved track were further adjusted using a gain attenuation factor obtained from ODEON in order to maintain the realistic scaling of levels between source-receiver combinations in each virtual acoustic environment, thus keeping the simulation sound more authentic. Subsequently, the 3D spatial audio file was converted to a binaural audio format by utilizing the SPARTA Ambibin Plugin in REAPER [148]. This was carried out by convolving the 3D audio file with the far-field Head Related Transfer Function (HRTF) of the Neumann KU-100 binaural head [172]. Since the study focuses on how the room acoustics interact with a dynamically changing musical signal, the running room reverberance is the major point of interest here. Therefore, the reverberation tail at the end of the convolved samples is removed by trimming and applying a fade-out filter.

Listening test procedure

The perceived impression of blending can be assessed using a rating scale or by evaluating the identifiability of constituent sound sources in a joint performance, as discussed in Section 1.2.2. Following previous studies on modeling of blending perception, the blending impression of sound samples in this study was evaluated using a categorical judgment test employing a 10-point scale with values ranging from 1 to 10, where a low value corresponds to the least blended impression and a high value to the most blended impression [28]. The test was conducted using the SQALA platform [174], which incorporated verbal anchors on the rating scale using labels such as 'very poor' to 'excellent' to represent varying degrees of perceived blending. A group of 16 participants comprising Tonmeister students and experienced musicians performed the listening test. All the test participants had undergone musical ear training. Moreover, they had prior experience in critical listening assessment and had at least 12 years of musical experience. Since trained musicians tend to be more capable and sensitive in selectively scrutinizing and evaluating the complex spectral and temporal features of sounds than non-musicians [17; 18], similar to the previous investigation on source-level blending, it was expected that the test participants would have an almost unanimous understanding of blending and provide concordant test responses.

The test was carried out in an acoustically treated room (RT_{60} = 0.1 s), and the Beyerdynamic DT 770 Pro closed-back studio headphones and RME Babyface Pro sound card were used to playback the binaural audio files from the computer. The objective of

the experiment, the working definition of blending, and the procedure of the test were explained to the test participants prior to the start of the listening test. Afterwards, listeners underwent a familiarization/training phase where they were introduced to the test platform and asked to rate the blending impression of 20 trial samples which were generated using 4 sound stimuli with different degrees of source-level blending, auralized in 5 distinct acoustic environments. As the ideal examples representing extreme blending impressions (most and least blended samples) are undefined, participants may not be able to set anchor points on the inner scale developed for the blending assessment. Nevertheless, the training phase is expected to provide an initial understanding of the possible variations involved in acoustic environments and source stimuli characteristics and help them create a rating scale with a reduced central biasing tendency of the ratings. Listeners were allowed to replay the samples and adjust the loudness of the playback, but they were instructed to maintain a fixed volume level after the training phase.

The listening test with 75 samples started at the completion of the training phase. In order to avoid direct comparison and minimize the memory retention effects on the ratings, the sound samples in the test were presented in randomized order. Moreover, the randomization was distinct for each participant to prevent any potential sequential effects in the sample ratings. Listeners were instructed to take small breaks of 3 minutes after rating a set of 20 samples in order to reduce the impact of mental or listening fatigue on their ratings. At the end of 75 samples, a discussion with the test participants was conducted regarding the subjective assessment of blending, perceived aspects of the influence of room in blending from the test, etc. The listeners took around 45 minutes to 1 hour to finish the test.

The internal consistency or the reliability of the listeners' ratings was further assessed by estimating Cronbach's alpha [117]. The Cronbach's alpha value is estimated to be 0.901 which denotes a high internal consistency and reliability in the sample ratings among the test participants.

8.1.2 Data analysis procedure

Extraction of room acoustic parameters

Room acoustic perception-related studies over many decades have demonstrated that established room acoustic parameters derived from Room Impulse Responses (RIRs) or a combination of them capture specific subjective sensations related to different room acoustic attributes, thereby offering a comprehensive overview of the perceptual characteristics of the acoustic environment[39; 40; 41; 42]. The conventional room acoustic parameters that correspond to each SRIR were obtained from the ODEON simulation. Based on the subjective sensation, the room acoustic parameters chosen for the study can be grouped into the following categories as described in earlier studies[37; 175]:

- **Perceived reverberance**: Reverberation time (T_{30}), Early Decay Time (EDT), Bass Ratio (BR), Treble Ratio (TR).
- **Clarity and intelligibility measures**: Clarity (C_{80}), Definition (D_{50}), Speech Transmission Index (STI).
- **Sound strength**: Strength parameter (G) estimated for three cases; (1) strength of direct and early reflections (G_{early}), (2) strength of early reflections ($G_{5\text{-}80}$), (3) strength of late reflections (G_{late}), and Sound Pressure Level of direct sound (SPL_{direct}).
- **Spatial impression**: Early Lateral Energy Fraction (J_{LF}), Early Lateral Energy Fraction Cosine (J_{LFC}), Late lateral sound level (L_j)

The definition and the estimation procedure of these parameters from the RIRs are detailed in the Appendix A. While the reverberation time T_{30} is conventionally regarded as the primary parameter for characterizing room acoustic response, Early Decay Time (EDT) was hand-picked here due to its demonstrability to better capture the subjective sensation of reverberance [39; 40]. Since the spectral centroid of the perceived sound is shown to play a significant role in the blending perception[14], the Bass Ratio (BR) and Treble Ratio (TR) [176], which reflect spectral coloration brought by the reverberation, were selected for the investigation. While the Clarity index (C_{80}) indicates the perceived clarity of music, the Definition parameter (D_{50}) is a similar energy ratio which better represents clarity in speech [37]. Additionally, the Speech Transmission Index (STI) [38], which evaluates speech intelligibility in rooms is also assessed in this investigation.

The Strength parameter (G) indicates the ability of the acoustic environment to amplify sound energy from the source, often describing the subjective sensation of loudness[40]. While not standardized in ISO, studies suggest that analyzing early and late arriving sound energy separately can offer valuable insights into the subjective and objective characteristics of the acoustic environment[177; 175]. An increase in the early arriving results in a better clarity sensation while an increase in late arriving energy leads to a higher sensation of reverberance and envelopment. Therefore three separate parameters, G_{early} (energy in the early part of RIR, i.e., in 0 - 80 ms), $G_{5\text{-}80}$ (energy of early reflections in 5 - 80 ms excluding direct sound), and G_{late} (energy of late reflections in 80 ms - ∞), are chosen for this analysis.

Spatial perception in acoustic environments is primarily influenced by two key concepts: Apparent Source Width (ASW) and listener envelopment (LEV) [178]. Apparent source width, linked to the energy of early reflections from lateral sides, is evaluated through parameters such as Early lateral energy fraction (J_{LF}) and Early lateral energy fraction cosine (J_{LFC}). While J_{LF} is typically used to depict ASW, it varies with the square of the cosine of the angle of incident reflections. J_{LFC} compensates for this effect and

is therefore expected to better represent the subjective sensation of perceived source width. However, studies suggest that LEV may be more crucial in spatial perception aspects, as ASW can be masked by late energy [178]. Although late arriving energy from back, overhead, and front regions contribute to the sensation of envelopment, the Late lateral sound level (L_j) is shown to better represent the listener's envelopment sensation[179].

The values of room acoustic parameters for each of the 25 acoustic environments were estimated by averaging the values acquired from the RIRs estimated from two sound sources to the receiver. The parameters of ISO 3382–1 [37] reported in this study (including 3 variants of G) are the averaged values of the 500–1000 Hz (mid-frequency) bands except the spatial parameters J_{LF}, J_{LFC}, and L_j that are averaged for 125–1000 Hz frequency bands.

Although each of the discussed parameters relates to distinct objective or subjective attributes of the acoustic environments, many of them are known to exhibit a strong correlation between themselves, particularly those within the same group [37; 175]. The Pearson Correlation Coefficient (PCC) estimated between the room acoustic parameters derived from the acoustic environments chosen for this study is presented in Table 8.2. The correlation coefficient values conform with the observation of multi-collinearity by exhibiting a strong correlation between specific groups of parameters.

Prediction modeling of blending perception

In this work, a regression model is formulated utilizing the Random Forest regression method to predict the perceived degree of blending impression using source-level blending ratings and room acoustic parameter values. Random Forest (RF) modeling is one of the widely used ensemble learning approaches for regression and classification problems that work by combining multiple decision trees as base learners [180; 181; 182]. This supervised Machine Learning technique takes into account the multicollinearity and multi-dimensionality of involved predictor variables and is fast to train, robust to outliers and noise, and resilient to overfitting [181; 183].

The decision tree, the fundamental component of RF modeling, is a binary recursive partitioning method in which each node point is split into two successor nodes according to the value of a particular predictor variable at the node. The best predictor variable and its 'splitting threshold point' for partitioning at each node are estimated by identifying the variable that maximizes the decrease in its variance from parent-level to child-level nodes from all the possible combinations. The splitting process will eventually result in homogeneity and reduced impurities in two child nodes, and the process continues until a predefined criterion, such as a limit on the number of nodes or a specific minimum number of samples in each node, is met.

Variable	T_{30}	EDT	BR	TR	C_{80}	D_{50}	STI	G_{early}	$G_{5\text{-}80}$	G_{late}	SPL_{Dir}	LF	LFC
EDT	0.95**												
BR	-0.38	-0.36											
TR	-0.67**	-0.52**	-0.19										
C_{80}	-0.88**	-0.83**	0.55**	0.43*									
D_{50}	-0.78**	-0.67**	0.52**	0.44*	0.95**								
STI	-0.78**	-0.68**	0.54**	0.40	0.96**	0.98*							
G_{early}	-0.36	-0.19	-0.14	0.76**	0.33	0.56*	0.42*						
$G_{5\text{-}80}$	-0.17	-0.09	-0.38	0.73**	-0.06	0.05	-0.04	0.79**					
G_{late}	0.28	0.384	-0.65**	0.42*	-0.41*	-0.24	-0.32	0.66**	0.76**				
SPL_{Dir}	-0.28	-0.09	0.06	0.47*	0.46*	0.63**	0.58**	0.77**	0.37	0.37			
LF	-0.21	-0.24	-0.32	0.54**	-0.17	0.23	-0.28	0.35	0.71**	0.45*	-0.23		
LFC	-0.16	-0.23	-0.30	0.39	-0.23	0.33	-0.35	0.12	0.55**	0.34	-0.42*	0.94**	
L_j	0.93**	0.87**	-0.51*	-0.53**	0.93**	-0.93**	-0.95**	0.39	-0.04	0.34	-0.46*	0.07	0.13

Table 8.2: Pearson correlation coefficient estimated between the room acoustic parameters (n=24, * $p<0.05$, ** $p<0.01$).

The flow diagram of the Random Forest modelling is demonstrated in Figure 8.2. The RF model trains an ensemble of decision trees, where each tree is trained using an independent sample space that is generated using the bootstrapping method – a resampling method that creates distinct datasets by randomly sampling the original dataset iteratively with replacement [184]. Accordingly, no assumptions about the underlying distribution of the data, such as normality, are considered in the modeling process. Additionally, predictor variables from the dataset are also randomly assigned to each decision tree, reducing the correlation between trees. While testing the RF model, the features associated with the test sample are passed through the individual decision trees until the sample reaches the terminal nodes, and the prediction value of the sample is obtained by averaging the response values of training samples in the given terminal node. The final prediction of the Random Forest model for each test sample is obtained by averaging the predicted responses of each decision tree without any weighting. Although single decision trees are relatively weak in prediction accuracy and prone to overfitting, the Random Forest model is famous for its high prediction accuracy and ability to provide feature importance of involved predictor variables, especially for small sets of samples with large numbers of features.

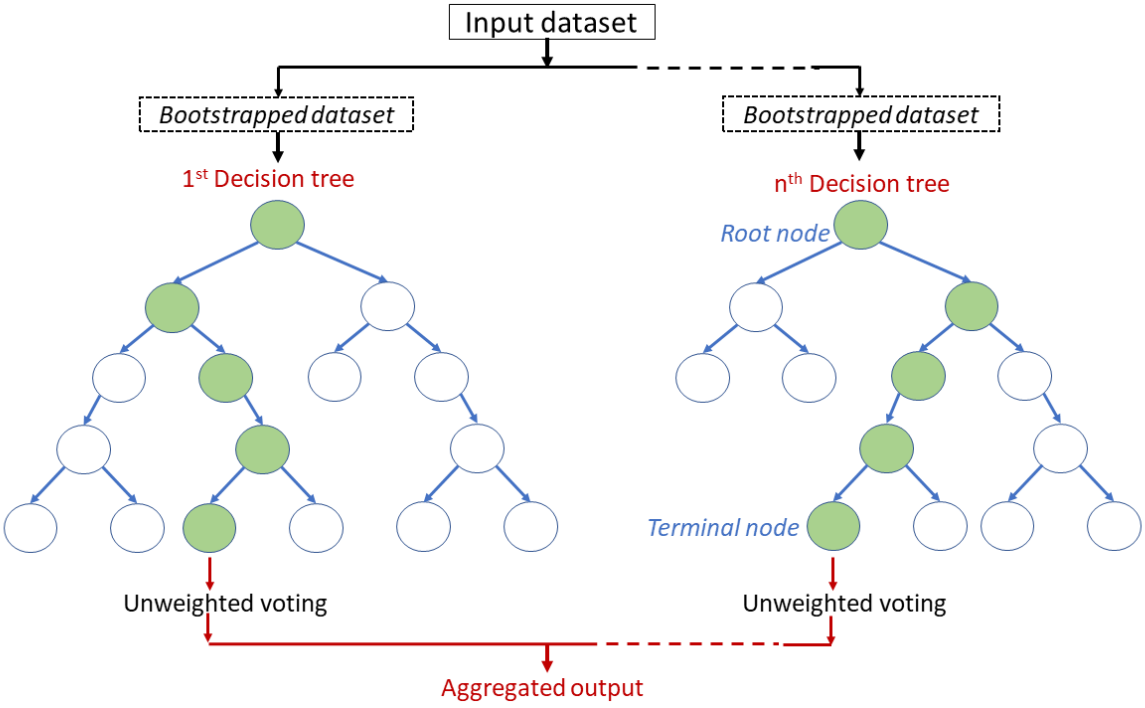


Figure 8.2: The flow diagram of Random Forest modelling.

One advantage of the Random Forests algorithm is that it can provide a measure of feature importance which indicates how much each predictor variable contributes to the performance of the model. A most typical way to estimate feature importance is by estimating the Mean Decrease in Impurity (MDI) for all nodes where the predictor is used in the decision trees. A higher MDI at a node corresponds to the ability of the predictor to split a large set of samples into two homogeneous classes by maximizing the decrease in impurity.

8.2 Results

8.2.1 Univariate exploratory analysis of musical and architectural variables:

Univariate exploratory data analysis is performed here to explore the impact of musical, architectural, and acoustic features incorporated in the test design on musical blending perception.

Influence of source-stimuli on blending ratings: The distribution of blending ratings of sound samples for different source stimuli is plotted in Figure 8.3 using horizontal box plots and Probability Distribution Functions (PDF). The Figure shows a decline in the mean value of rating values from stimulus A to C, congruent with the order of their source-level blending ratings. In contrast to the evenly spaced source-level blending ratings, stimulus B shows a relatively similar distribution of blending ratings to stimulus A with a relatively similar median value when passed through different acoustic environments, whereas stimulus C seems to stand out with relatively lower rating values. Notably, the tails of the distributions for high source-level blending (stimulus A) and low source-level blending (stimulus C) are extended to very low and high ratings respectively, underscoring the impact brought by the acoustic environments on the perceptual outcomes for samples with distinct source-level blending.

Since the distribution of the three groups of ratings does not follow the normality assumption (validated using Shapiro-Wilk test [149], $p<0.001$), the Kruskal-Wallis test, the non-parametric equivalent of one-way ANOVA based on ranks [157], was performed to analyze the statistical difference between the distribution of three groups of samples by assessing differences in the mean ranks. The results showed that it failed to reject the null hypothesis that there is no difference in the mean ranks of the groups ($\chi^2(2)=278.1$, $p< 0.001$), therefore there exists a statistical difference in the distribution of ratings of the three groups. Dunn's Post Hoc test conducted to assess which groups were substantially different from the others revealed that the three groups were statistically different from one another ($p< 0.01$ for three pairs of groups, adjusted using Bonferroni correction), underscores the unique role of the source stimulus in shaping perceived ratings.

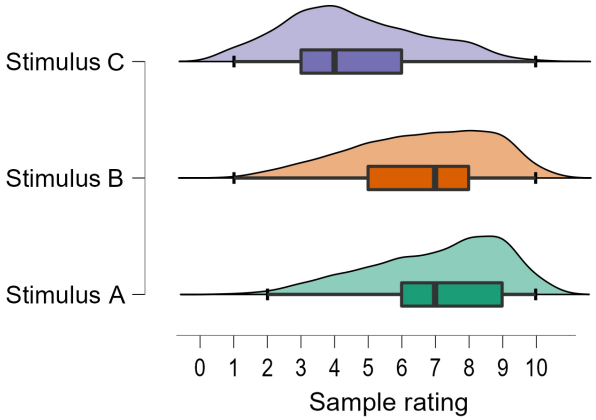


Figure 8.3: Distribution of blending ratings for three source stimuli with different degrees of source level blending.

Influence of room geometry on blending ratings The variation in the distribution of blending ratings of samples from different room geometries is depicted in Figure 8.4. The blending ratings for R1 are relatively lower, while R2 has a broader distribution of blending ratings. Conversely, R3 and R4 show relatively similar distributions of ratings encompassing higher values with the same median values and inter-quartile ranges, indicating slight improvement over R1 and R2. This observation indicates that, within the chosen set of rooms, the increase in the volume of the rooms tends to enhance the blending impression up to a certain level, beyond which no substantial improvement is observed.

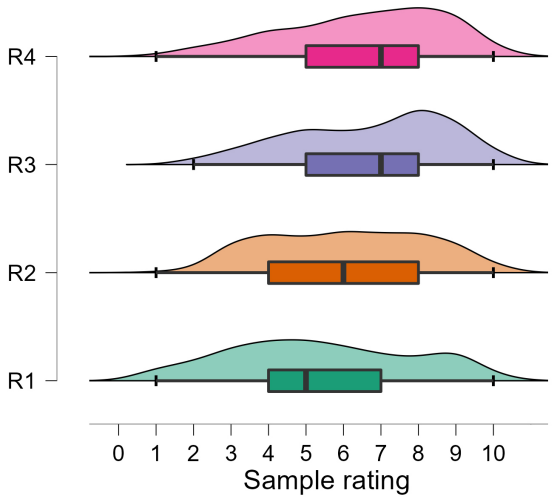


Figure 8.4: Distribution of blending ratings for the four acoustic environments having different geometries.

Since the distributions of these 4 groups do not meet the normality condition (Shapiro-Wilk test, $p<0.001$), the Kruskal-Wallis test was performed here similar to the preceding case to assess the statistical difference among the groups. The results suggest that there exists a statistical difference between the ratings of samples from four room geometries ($\chi^2(3)=56.02$, $p<0.001$). The Dunn's Post Hoc test conducted with Bonferroni correction showed that while there is no statistically significant difference between the pair R3 and R4 ($p=0.935$), the other two groups, R1 and R2, are different from each other ($p=0.009$) and they are also different from R3 ($p<0.001$, $p=0.002$) and R4 ($p<0.001$, $p=0.034$) respectively.

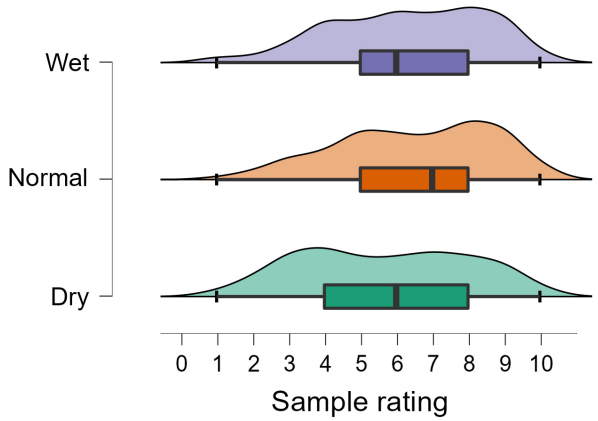


Figure 8.5: Distribution of blending ratings for three variations in the absorption coefficients utilized.

Influence of acoustic absorption variation Figure 8.5 demonstrates the distribution of rating of samples having different absorption properties. The dry acoustic environment has resulted in the relatively lowest blending ratings, while the normal acoustic environment resulted in higher ratings as seen in Figure 8.5, indicating an improved impression compared to the rest two cases. This suggests that the low and high extremes of absorption could result in very strong or very weak reflections may lead to a lowering of blending impression.

The Kruskal-Wallis test, performed due to the deviation from normality (Shapiro-Wilk test with $p<0.001$), indicated a statistical difference in ratings among samples from different acoustic variants ($\chi^2(2)=28.91$, $p<0.001$). The Post Hoc analysis revealed no statistically significant difference between Normal and Wet acoustic conditions ($p=0.933$), while the Dry condition showed a significant difference from both Normal ($p<0.001$) and Wet ($p<0.001$) conditions. These results imply that the acoustic variations introduced in the test design must have diversified the perceived blending.

Influence of listener's position The variation in the distribution of ratings of samples for the near and far positions is shown in Figure 8.6. With a maximum centered around the rating value of 8 in their probability density function, the far position is shown to have relatively higher blend ratings than the near position. The Mann-Whitney U test[141], a non-parametric version of the Student's t-test, was performed here due to deviations from normality conditions (Shapiro-Wilk test, $p<0.01$), confirmed a statistical difference between two groups ($p<0.01$). This suggests that the introduced distance variations should have contributed to altering the perceived blending impressions of the sound samples.

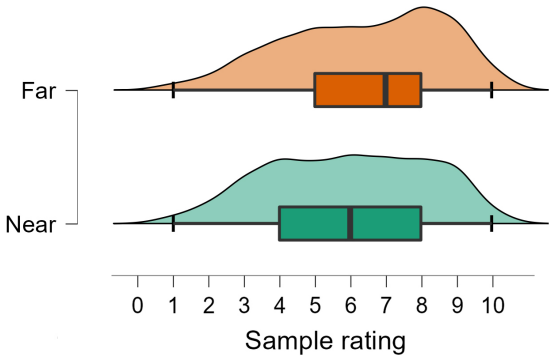


Figure 8.6: Distribution of blending ratings for the near and far listener's position.

8.2.2 Correlation between blending impression and room acoustic parameters

The correlation between room acoustic parameters estimated for each of the acoustic environments and the mean value of ratings of samples of three stimuli with different degrees of source-level blending is evaluated. A number of room acoustic parameters involved in this investigation exhibit a non-linear higher-order variation with the test ratings (see section 8.2.3). Therefore, Spearman's rank correlation coefficient[158], which is based on the monotonic relationship between variables, is used for this analysis. In contrast to the established Pearson's correlation coefficient which measures the linear relationship between the two variables involved, Spearman's correlation Coefficient, a non-parametric measure of the correlation of ranks, is observed to be more appropriate for non-normally distributed data and more robust to outliers than Pearson's correlation. Since certain acoustic parameters such as C_{80}, G_{late}, BR, TR, etc., lack a physically meaningful value in the anechoic environment, the ratings corresponding to the acoustic environment 'R1A' were avoided from this correlation analysis. The Spearman correlation coefficients derived from correlating the mean value of ratings of samples of three different stimuli and the room acoustic parameters extracted from the 24 acoustic conditions are presented in Table 8.3.

Room Acoustic Parameter	Stimulus A (7.9±1.6)	Stimulus B (5.5±2.1)	Stimulus B (3.3±1.9)
EDT	**0.48***	**0.51***	**0.73****
T_{30}	**0.44***	**0.60****	**0.78****
BR	-0.07	-0.30	**-0.50***
TR	**-0.50***	-0.32	**-0.54****
C_{80}	-0.36	**-0.73****	**-0.78****
D_{50}	-0.39	**-0.77****	**-0.78****
STI	-0.32	**-0.76****	**-0.75****
G_{early}	**-0.57****	**-0.45***	**-0.51***
$G_{5\text{-}80}$	**-0.56****	-0.07	-0.23
G_{late}	-0.25	0.07	0.11
SPL_{dir}	-0.32	**-0.58****	**-0.42***
J_{LF}	-0.34	0.13	-0.12
J_{LFC}	-0.25	0.20	-0.03
Lj	**0.44***	**0.72****	**0.83****

Table 8.3: Spearman's correlation coefficient computed for ratings of three stimuli in different acoustic environments and corresponding room acoustic parameters (n=24, *$p<0.05$, **$p<0.01$).

Many of the room acoustic parameters were shown to have a statistically significant correlation with blending ratings. Notably, specific parameters such as EDT, T_{30}, C_{80}, D_{50}, STI, and L_j reveal a systematic relationship in the correlation with the changes in the degrees of source blending of samples. EDT and T_{30} are shown to be positively correlated with the blending ratings of samples, and the magnitude of correlation increases with a decrease in the degree of source-level blending (from 0.48 to 0.73 for EDT, and 0.44 to 0.78 for T_{30}). This suggests a direct positive relationship between the reverberance of the acoustic environment and the perceived blending rating, further the magnitude of the correlation is attributed to the source-level blending characteristics. Conversely, parameters such as C_{80}, D_{50}, and STI exhibit a negative correlation with blending ratings, with the correlation being statistically insignificant for stimulus A (p-value>0.05) but highly significant for stimulus B and C. This suggests that the perceived blending of samples with a moderate or low degree of source-level blending improves with the degradation of clarity and intelligibility of the acoustic environment, although this trend is not observed in samples with high source-level blending (see correlation values of C_{80}, D_{50}, and STI in Table 8.3). Furthermore, the parameter L_j also displays a significant correlation with the three stimuli, with its magnitude of correlation increasing as the source-level blending of the sample decreases. This suggests that an enhanced sensation of spatial envelopment can positively contribute to im-

proving the blending of samples, with this effect being more pronounced for samples featuring a lower level of source-level blending.

On the other hand, parameters like G_{early} show almost equivalent statistically significant correlations ($p<0.05$) with samples of three stimuli irrespective of the variation in their degrees of source-level blending (see correlation values of G_{early} in Table 8.3). Other strength parameters like $G_{5\text{-}80}$ exhibit a negative correlation solely with sample A, while G_{late} demonstrates no statistically significant correlation with blending (see Table 8.3). Additionally, SPL_{direct} appears to have a statistically significant negative correlation with samples B and C, therefore the stronger direct sound from the sound source seems to degrade the perceived blending for samples having moderate or poor source-level blend. A trend of negative relationship is observed between the treble ratio and the blending ratings which is relatively the same for samples having high and poor source-level blend, however, their correlation is not statistically significant for stimulus B. Other than L_j, the spatial parameters J_{LF} and J_{LFC}, which represent the perceived sensation of apparent source width, do not exhibit any significant correlation with the blending ratings.

8.2.3 Variation of blending with room acoustic parameters

In this section, we visualize the relationship between the room acoustic parameters calculated for each of the 25 acoustic environments and the mean value of test ratings of three stimuli characterized by different degrees of source-level blending. One representative parameter from each of the four classes representing different subjective sensations of the acoustic environment (discussed in section 8.1.2), namely T_{30}, D_{50}, G_{early}, and L_j is presented here, all of which exhibit a significant correlation with the blending ratings.

The variation of mean values of sample ratings of three stimuli against the T_{30} parameter extracted from each acoustic environment is depicted in Figure 8.7. The lowest blending impression among the three stimuli is consistently observed in the anechoic environment (R1A with T_{30}=0) irrespective of source-level blending variations. This observation suggests that room acoustic reflections generally enhance the blending impression. Additionally, the three stimuli follow the same order at R1A having approximately equally spaced ratings that are in agreement with their source-level blending ratings. The variance in the perceived blending impression of three stimuli undergoes minimal changes when the monophonic source signals are transformed to a spatially separated source-receiver scenario in a reflection-free environment, although the scale of blending evaluation differs in both contexts.

While the order of stimuli ratings in different acoustic environments mostly aligns with the source-level blending, stimulus B seems to outperform stimulus A in a few cases. Overall, the rating trend of stimulus B closely resembles that of stimulus A, while Stimulus C seems to be different which is consistent with the observations from

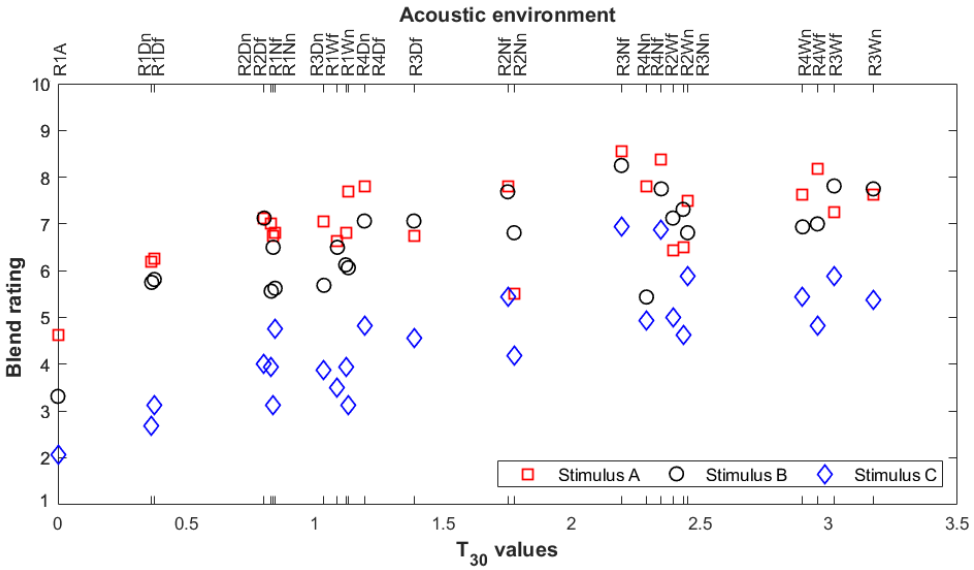


Figure 8.7: Variation of blending ratings of stimulus A (source-level blending of 7.9±1.6), stimulus B (5.5±2.1), and stimulus C (3.3±1.8) with respect to Reverberation time (T_{30}) exhibiting Spearman's correlation of 0.44, 0.60, and 0.78 respectively.

Figure 8.3. Although the blending impression generally improves from R1A with increasing T_{30}, the three stimuli appear to saturate in the blending ratings beyond which a minimal improvement is observed. Whereas stimuli A and B reach this asymptotic behavior at relatively lower T_{30} values, stimulus C exhibits a relatively stronger linear progression with T_{30} and reaches the asymptote at relatively higher T_{30} value. This phenomenon may be attributed to the larger headroom available for the improvement of perceived blending by the acoustic environment in samples with poor source-level blending as compared to a low headroom available in samples with high source-level blending. Therefore, the degree of source-level blending of the sound stimulus utilized appears to influence the alteration of final perceived blending caused by the acoustic environment, which agrees with the findings of the correlation analysis. However, it is worth noting that the highest rating of blending among the acoustic environments was consistently observed at 'R3Nf' (room with 1000^3 with normal absorption and far listener position) for the three stimuli, regardless of the differences in their degrees of source-level blending. This particular observation demands further investigation. The variation plot of EDT is not depicted, as it exhibited a similar trend to the T_{30} plot.

Figure 8.8 demonstrates the variation of the sample ratings of three stimuli and the Definition parameter (D_{50}) extracted from the different acoustic environments. Given the fact that C_{80}, suitable for experiments involving music, becomes infinite in an anechoic setting, the parameter D_{50} is utilized here to illustrate how blending changes with the subjective sensation of clarity by incorporating the anechoic condition. A notable

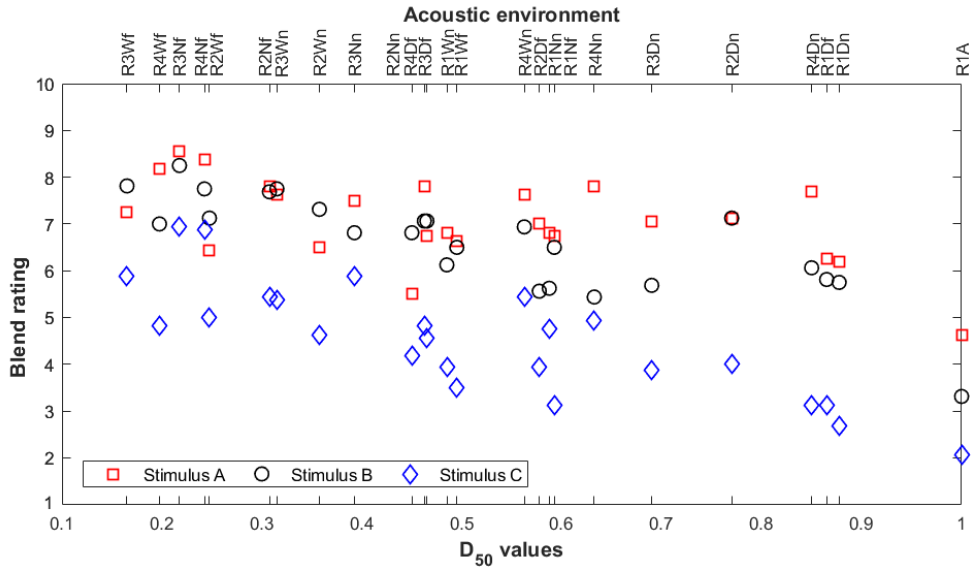


Figure 8.8: Variation of blending ratings of stimulus A (source-level blending of 7.9±1.6), stimulus B (5.5±2.1), and stimulus C (3.3±1.8) with respect to Definition parameter(D_{50}) exhibiting Spearman's correlation of -0.39, -0.77, and -0.78 respectively.

trend of decreasing blending ratings is observed with increasing D_{50} values, which is substantiated by a negative correlation between the derived T_{30} and D_{50} values (see Table 8.2). While stimuli A and B show a relatively lower degree of variation with D_{50}, this effect is particularly prominent in stimulus C, especially when excluding the anechoic condition. Similar trends were observed in the variation plots of C_{80} and STI.

The variation of blending ratings of three stimuli with the strength parameter estimated for the early part of RIR (G_{early}) is demonstrated in Figure 8.9. Unlike the earlier cases, the mean ratings of samples with three stimuli show a higher-order (i.e., non-linear) variation with G_{early} which is particularly evident in stimulus C. This trend of higher-order variation remains consistent, even when the anechoic condition is excluded as an outlier. Although the correlation is relatively weak, parameters like $G_{5\text{-}80}$ also show a similar higher-order variation with blending ratings.

Figure 8.10 illustrates the variation of blending ratings against the Late lateral sound level parameter (L_j). Given that the anechoic environment lacks a physically meaningful L_j value, it was omitted from the plot. While the perceived blending impression generally improves with increasing L_j values, it appears to have a greater impact on stimulus C, exhibiting a linear relationship. However, this linear relationship gradually diminishes towards stimulus B and then A. This trend aligns with the correlation analysis findings and reflects the observation of more headroom available for the improvement of perceived blending by the acoustic environment in samples with poor source-level blending as compared to a low headroom available in samples

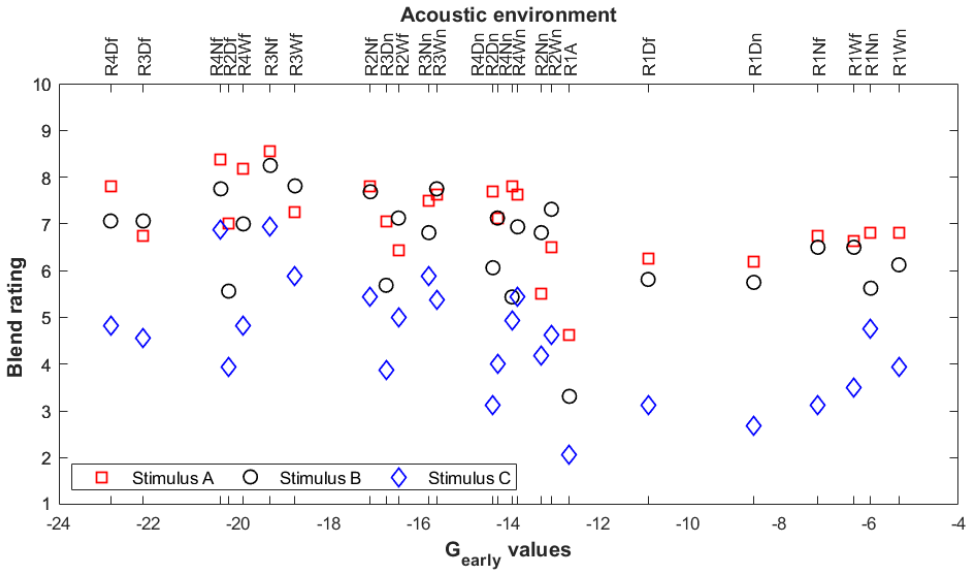


Figure 8.9: Variation of blending ratings of stimulus A (source-level blending of 7.9±1.6), stimulus B (5.5±2.1), and stimulus C (3.3±1.8) with respect to Strength parameter (G_{early}) exhibiting Spearman's correlation of -0.57, -0.45, and -0.51 respectively.

with high source-level blending. Therefore, analogous to the reverberance-related attributes, this spatial attribute also seems to demonstrate a systematic relationship with the perceived blending impression where its influence is determined by the degree of source-level blending in the stimuli.

8.2.4 Random forest modeling and feature importance

Modeling with separate test-train datasets

Random Forest regression modeling is performed to estimate the perceived degree of the overall blending of sound samples using the source-level blending ratings and room acoustic parameters. In the first phase of the modeling process, Random forest regression was carried out by randomly partitioning the samples into 70% training data (i.e., 50 samples) and 30% testing data (i.e., 22 samples) sets. Samples of anechoic condition were omitted from the regression modeling due to physically invalid values of certain parameters, and the remaining 72 samples were utilized for the modeling process. The modeling was executed using the Random Forest regressor function from the Scikit-learn machine-learning library package [185] in Python. The regressor in this study comprised 100 decision trees in which the nodes are expanded until all leaves are pure or contain only one element. A random bootstrap sampling method was utilized for assigning the input datasets to each decision tree. To ensure a reliable result that is

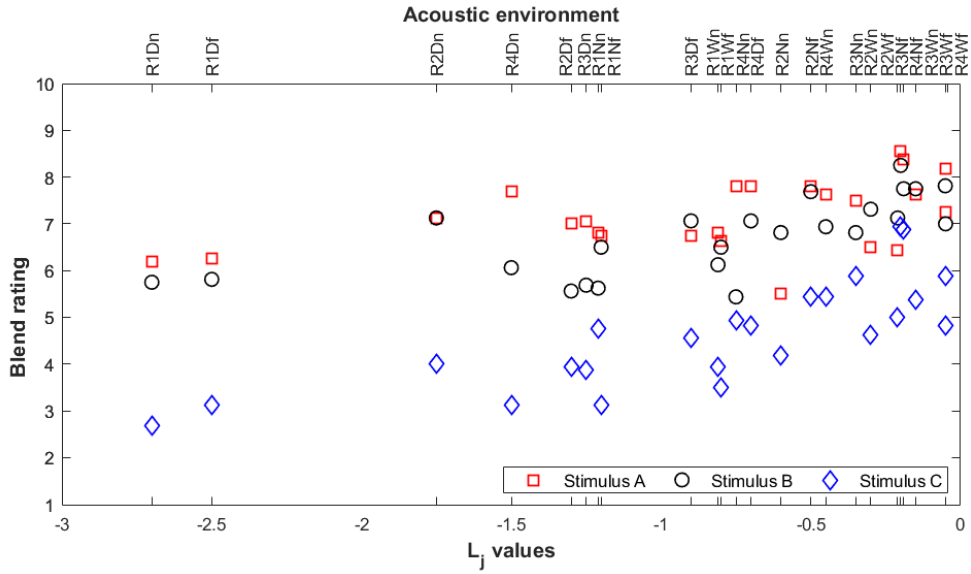


Figure 8.10: Variation of blending ratings of stimulus A (source-level blending of 7.9±1.6), stimulus B (5.5±2.1), and stimulus B (3.3±1.8) with respect to Late lateral sound level parameter (L_j) exhibiting Spearman's correlation of 0.44, 0.72, and 0.83 respectively.

free from bias or overfitting, the modeling process was repeated 20 times employing randomly assigned training and testing data in each iteration.

Figure 8.11 illustrates the performance of outputs of 3 randomly selected models from the mentioned 20 different ones, serving as an example to visualize the prediction accuracy of this modeling technique in predictions of blending ratings. The predicted blending ratings of randomly selected 22 test data samples in each model are plotted in the figure against the perceived blending ratings for comparison. The data points in the figure appear to be distributed closely to the 'y=x' line, signifying a strong agreement between the perceived and predicted ratings and thereby upholding the feasibility of the model. The remaining 17 models also exhibited the same trend, and therefore they were excluded from Figure 8.11 for visual clarity. The mean absolute error, defined as the average of the absolute difference between the predicted and perceived ratings, is calculated across 20 different models to be 0.59 with a standard deviation of 0.08. This indicates that the model is able to predict the perceived blending ratings within a deviation range of approximately ±6% in the prediction values.

To evaluate the conformity between perceived and predicted blending ratings, Lin's concordance correlation [150] is calculated between the perceived and predicted ratings of randomly selected 22 test samples in each of the 20 different models. Unlike the Pearson correlation which assesses the linearity between two variables, Lin's concordance correlation estimates the concordance or the level of agreement between a

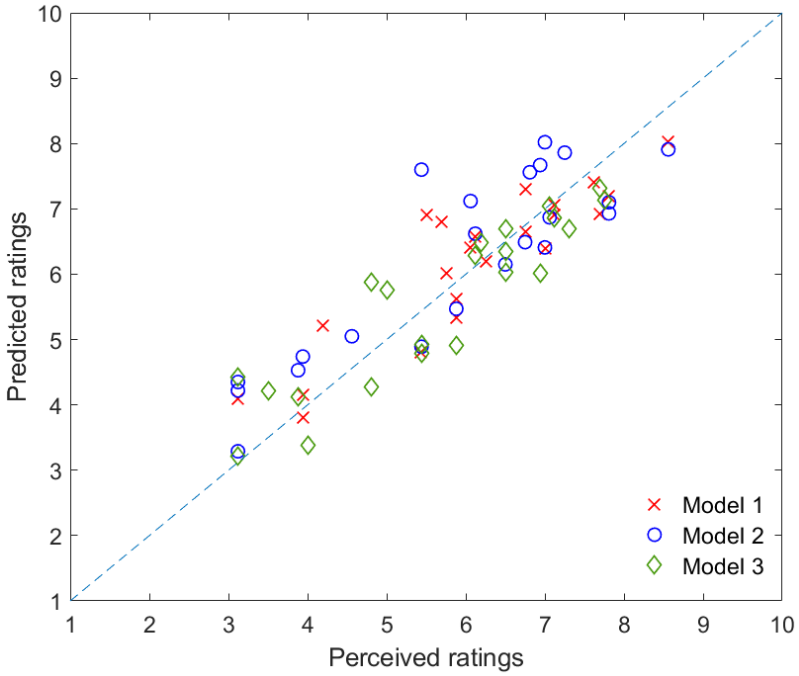


Figure 8.11: Blending ratings predicted by three different random forest models (having different test-train data sets) against the perceived blending ratings.

bivariate pair of observations of the same variable by examining the proximity of these observations to the identity line 'y=x' passing through the origin. The estimated concordance correlation coefficients from the 20 different models are distributed within the range of 0.83 to 0.88 with a mean value of 0.85, signifying a strong agreement between the perceived and predicted ratings and thereby underscoring the viability of the implemented Random Forest model.

Cross validation of the model

The results obtained from repeated modeling with separate test-train data sets may exhibit biases or randomness due to the limited sample size. To overcome this, following the same methodology used in Chapter 3, the accuracy of the model is further validated using Leave-One-Out Cross-Validation (LOOCV) [140]. LOOCV is implemented through an iterative modeling process, where all data points are used for training except one for prediction in each iteration, which resulted in 72 unique models in this evaluation using distinct data samples. Although it is computationally demanding, this approach guarantees a precise and unbiased assessment of the performance of the regression model. The cross-validation results reveal a mean absolute error of 0.56 out of 10, representing an approximate 6% deviation between the predicted and perceived ratings across the 72 models, which is consistent with earlier findings. The two-fold

evaluation of the Random Forest regression model, involving separate test-train data and LOOCV, demonstrates its resistance to biases, over-fitting, and chance predictions, and thereby substantiating its validity in predicting perceived blending ratings with reliability and accuracy.

Assessment of feature importance

The feature importance in the percentage scale of the involved parameters was estimated across the 20 distinct regression models created from diverse train-test data sets, and the distribution of the importance values of each individual parameter is visualized in Figure 8.12. The source-level blending ratings of the involved sound stimuli with feature importance values around 60% stand out as the primary contributor to the overall perceived blending impression while the room acoustic parameters selected for the analysis appear to collectively contribute to only 40% of the overall importance in explaining the perceived blending impression. Notably, the room parameters, T_{30}, EDT, D_{50}, L_j, and G_{early} along with source-level blending ratings account for the 85% variance of the estimated feature importance. This can be interpreted that the acoustic environments with enhanced reverberance and spatial envelopment, coupled with reduced clarity and weak early reflections, significantly increase the perception of blending.

The Random Forest model achieves good results by taking into account the multicollinearity among the room acoustic parameters and their non-linear relationship with the blending ratings. However, multicollinearity of involved variables may reduce the reliability of the feature importance values of the model, and its interpretability. For example, if two features are highly correlated, although they are randomly assigned for each decision tree, the impurity reduced by the first feature in the presence of the second feature may not be necessarily reduced again by the second feature. Considering the significant correlations among specific room acoustic parameters (see Table 8.2), it is necessary to carefully and critically analyze the feature importance values instead of interpreting the results strictly based on the estimated importance values. Therefore the feature importance values of the parameters have been grouped according to subjective sensations, as outlined in Section 8.1.2, and are presented in Table 8.4. Given the source-level blending ratings and the room acoustic parameters are mutually orthogonal, explaining the contribution of source-level blending and room acoustic support in the final perception of the blending as 60% and 40% respectively remains valid. But when it comes to room attributes, the perceived reverberance of performance spaces, encompassing parameters related to decay times, emerges as the most influential aspect of the room acoustic environment in shaping the perception of blending.

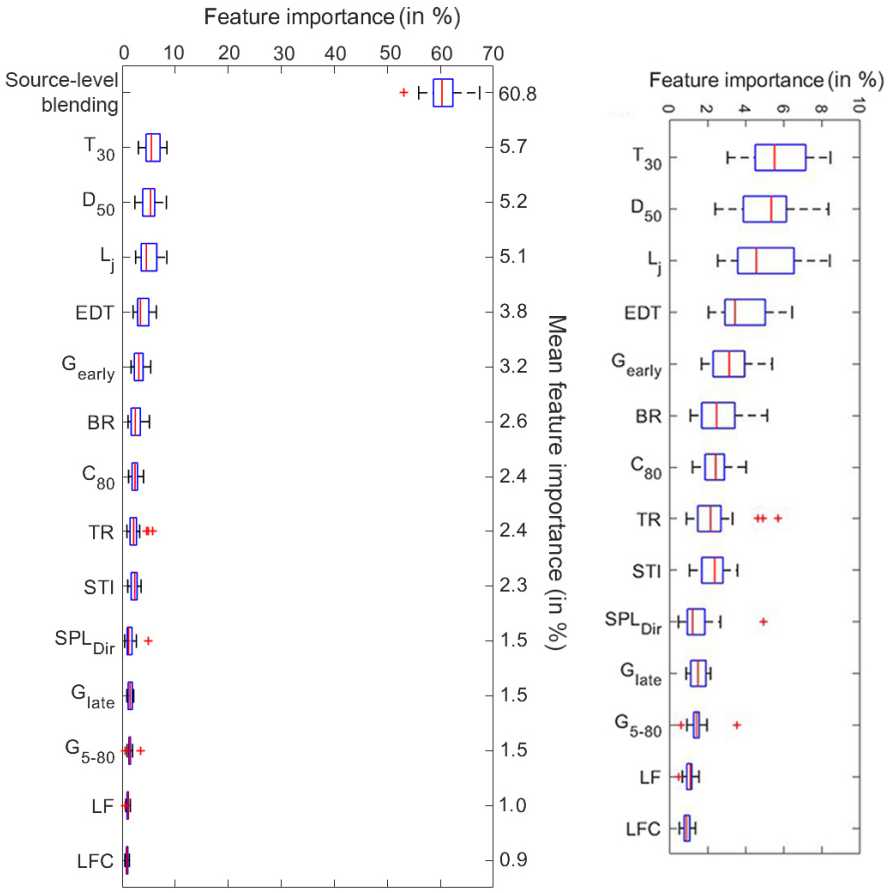


Figure 8.12: The distribution of feature importances of involved parameters (in percentage) assessed across 20 different models with mean feature importance denoted on the y-axis on the right (a zoomed version of the feature importance distribution of room acoustic parameters is given on the right side for a better-detailed view).

Subjective attribute	**Overall feature importance**
Source-level blend	60.8%
Reverberance (EDT, T_{30}, BR, TR)	14.5%
Clarity and intelligibility (C_{80}, D_{50}, STI)	9.9%
Sound strength(G_{early}, $G_{5\text{-}80}$, G_{late}, SPL_{Dir})	7.7%
Spatial parameters (J_{LF}, J_{LFC}, L_j)	7.0%

Table 8.4: Feature importance estimated for different subjective attributes of room acoustics in the blending perception.

8.3 Discussion

Although previous studies addressed blending from two distinct directions – one as a music-perception problem at the instrument level without an acoustic environment[14; 33; 30], and the other as a subjective attribute in the perceptual evaluation of acoustic environments[43; 7; 27] – this study integrates these two aspects for the first time by analyzing the overall perception of blending as a function of the perceived source-level blending ratings of the stimuli involved, as well as the room acoustic parameters extracted from the acoustic environments.

Since individuals from diverse backgrounds may have discrete conceptions and definitions about the blending of sounds, the two perceptual evaluations of blending (source-level and overall blending) discussed in this study were conducted 'exclusively' among critical listeners with musical ear training, with an expectation to get consistent ratings with good agreement due to their sensitivity to musical audio cues. To ensure the authentic and innate representations of constituent sound sources in a musically realistic joint performance, the in-situ close-mic recordings of ensemble performance, capturing the qualities of performance strategies and auditory-visual feedback, were utilized in this study for the creation of the test stimuli. Furthermore, employing room acoustic simulations for auralization of test samples better helped in measuring many of the chosen room acoustic parameters such as G, L_j, etc., with better precision and Signal-to-Noise Ratio (SNR), which are often challenging to estimate in real-life situations.

The correlation analysis conducted between room acoustic parameters and the ratings of the three test stimuli unveiled room acoustic parameters that exhibit a statistically significant influence on blending ratings including EDT, T_{30}, C_{80}, D_{50}, STI, G_{early}, SPL_{dir}, L_j, representing subjective sensation of reverberance, clarity, intelligibility, strength and envelopment. The correlation of EDT and T_{30} parameters with blending align with earlier findings [27]. While TR was identified as a significant parameter of blending in a previous study [27], it does not appear to have a significant influence in this study, which requires further investigation. Additionally, certain room acoustic parameters exhibited a trend where their magnitude of correlation was influenced by the degree of source-level blending rating of the stimuli, with a higher magnitude of correlation observed in samples with a poor source-level blend, and vice-versa. This observation can be interpreted that room acoustic assistance has a more pronounced impact on enhancing the perceived impression of blending in amateur performances compared to highly coordinated professional ones. However, the crucial role of room acoustic feedback in shaping the performance strategies of individual musicians[10; 11; 13] suggests that their performances may vary across different acoustic environments. Therefore, further investigation is needed to understand its impact on the results obtained in this study.

In this study, the perceived degree of source-level and overall blending is represented with one single value which does not include a temporal variation of blending and lacks information on the inherent musical/acoustic factor that caused the non-blending at the source-level (e.g. difference in onsets/offsets, spectral dissimilarity, variation in pitch and loudness, etc.). The way room acoustic reflections respond and interact with variation in source-level blending caused by individual factors such as pitch difference, timing asynchrony, loudness difference, spectral dissimilarity, etc., may not be necessarily similar. Hence, the source-level blending characteristics of the three stimuli utilized here might have an impact on the final perceived blending ratings. Therefore, a more controlled experiment is necessary in the future to examine how the room acoustic environment interacts with varying degrees of source-level blending, which arise from individual variations in the aforementioned factors within the musical stimulus. Moreover, the spatial attributes such as the position, spacing, and orientation of the sources on the stage shall also be scrutinized in detail to understand their impact on blending.

Though the selected room acoustic parameters are well-established in describing the perceptual attributes of acoustic environments, numerous studies have described the limitations of the existing parameters in fully explaining the subjective sensations [175; 186; 187], which extends the possibility of incorporating alternative parameters to try out the modeling approach. The future works can advance further by addressing the constraints of static binaural rendering of sound fields (such as front-back confusion) by performing a dynamic binaural rendering in which the movement and head-rotation of the listener are possible, or by utilizing measured/simulated Spatial Room Impulse Responses (SRIRs) and loudspeaker arrays for 3D sound field rendering in which the differences of HRTFs and limitations of binaural rendering are avoided.

The proposed RF regression prediction model serves as a tool for future research aiming to estimate blending across a large and diverse dataset encompassing various instruments and a broad spectrum of real and virtual room acoustic environments with a wide variety of architectural and geometrical variations. While the model presented in this study is constrained by its dependence on a single-valued perceived rating of source-level blending for sound stimuli, the ultimate goal is to develop a comprehensive model that accounts for all aspects of blending. This will be accomplished by developing a separate model for estimating time-varying source-level blending in dynamic musical signals and integrating it into the existing approach, enabling the estimation of variation in room acoustic contribution with changes in source-level blending.

Considering the feature importance of source-level blending and room acoustic parameters, it can be inferred that a non-symmetric relationship possibly exists between the source-level blending and the room acoustic contribution. In other words, when source-level blending is at moderate or high levels, room acoustics typically could enhance the perceived overall blending within the available headroom for improve-

ment until reaching a saturated level. Conversely, if the source-level blending is poor, the capacity of room acoustic support to enhance it is found to be constrained, even when there is ample headroom for improvement. Therefore, in reality, musical performances with a high source-level blend may not necessarily need the support from room acoustics to sound blended, whereas the well-regarded concert halls renowned for their acoustics may not exhibit optimal blending if performers fail to meet the requisite minimum source-level blending standards. Additionally, the findings from RF feature importance and correlation analysis suggest that acoustic environments featuring a better sensation of reverberance and enhanced spatial envelopment, along with reduced clarity and weak early reflections, significantly support the overall blending perception. When it comes to concert hall acoustic design, high clarity sensation and strong early reflections are generally preferred. Therefore, improving the blending must be carefully considered by achieving a sufficiently high reverberance and envelopment without affecting the so-desired clarity and reducing the early reflections.

8.4 Summary

This study showcases the viability of a computational modeling approach in evaluating the perceived blending impression in a musically realistic performance setting through the estimation of the distinct contributions made by source-level blending and the room acoustic environment. Room acoustic reflections are shown to enhance the perceived blending compared to a reflection-free environment, and the correlation analysis identified significant room acoustic parameters (T_{30}, D_{50}, G_{early}, L_j, etc.) that played a pivotal role in altering the final perceived blending impression. The degree of source-level blending of sound stimuli is shown to influence the alteration in perceived blending brought by the acoustic environment. Specifically, samples with poor source-level blending exhibited more significant potential for enhancement of the overall blending effectuated by the acoustic environment, whereas those with high source-level blending had limited room for improvement. This underscores the intricate relationship between source-level blending and room acoustics in shaping the perceived blending quality, emphasizing the necessity to model the overall blending as a function of both.

The proposed Random Forest regression model for predicting the overall perceived blending impression using source-level blending ratings and room acoustic parameters is meticulously tested and validated through a comprehensive two-fold evaluation, involving modeling with 20 different test-train datasets and leave-one-out cross-validation. The regression model demonstrates a significant accuracy, with a mean absolute error of 0.6 on a 10-point scale (i.e., a 6% deviation in predictions) and a mean concordance correlation of around 0.85 between the predicted and perceived ratings. This robust performance underscores the scope and applicability of this machine-

learning technique in analyzing complex psychoacoustic phenomena such as musical blending. Additionally, the relative importance of source-level blending rating and room acoustic contribution in the final perception of blending is estimated to be around 60% and 40% respectively, which suggests that the room contribution to overall blending impression is nearly as significant as source-level blending between the instruments in this controlled experimental setting.

While the perceptual rating scale of blending is relative and subject to influence from factors such as the background of the listeners involved and the characteristics of the sound stimuli utilized, the computational model presented in this work provides a foundation for future research on the comprehensive modeling of blending. Despite having limitations in assessing source-level blending features in this constrained experiment, the contributions of source-level blending and individual room acoustic attributes in the overall perception of blending are quantified in this analysis for the first time. As a result, it offers valuable insights into perceptually oriented room acoustic design, music performance, and music perception-related research.

Chapter 9

Conclusion

9.1 Overall summary

This study presents a foundational attempt to analyze and evaluate the perceptually relevant acoustic attributes of musical instruments in joint musical performances. This was performed by examining various acoustic and perceptual aspects involved in ensemble sound, including the blending between instruments at different levels, perceptual relevance of directivity of individual instruments, and representation of appropriate input source signals and directivity of instruments in the auralization of ensemble sound. The major content of this thesis is organized into three modules, with the primary results of each module summarized below.

The **first module** focussing on **musical performance-based representation of sound sources** included three studies. An initial exploration of ensemble sound and blending between instruments, conducted through a listening test associated with a live string ensemble performance, showed that the ability to predict the number of constituent sources reduces with an increase in the number of instruments in the ensemble and thereby supports the blending. This is shown to be influenced by the characteristics of the acoustic environment, where in specific acoustic conditions, no significant improvement in the prediction accuracy and the blending impression is observed when increasing the number of sources beyond a particular value. Additionally, the impression of ensemble sound improves with the increasing number of sources to an extent, beyond which no major change is observed. While this pilot study with a violin ensemble focuses only on the macroscopic perception of blending, it highlights the importance of blending between instruments at the source level and the influence of room acoustics in it. It also underscores need for microscopic examination of blending of individual musical samples in realistic settings.

Building upon these insights, blending between the instruments at the source level was investigated using sound samples of musically realistic score-independent monophonically-rendered unison performances of two violins in in-situ conditions. Based on the perceptual labeling of these samples from a listening experiment with expert listeners, this study illustrated the feasibility of a computational modeling approach to classify sound samples into blended or non-blended classes based on their overall perceived impression of source-level blending. Among the different dimensionality reduction techniques explored in that investigation, the Linear Discriminant Analysis (LDA) paired with the Euclidean distance measure performed on the Mel-Frequency Cepstral Coefficient (MFCC) features extracted from the sound samples is shown to be an effective method for the classification of samples based on source-level blending. This was tested and verified using a separate train-test data set, and leave-one-out cross-validation, showing an accuracy of 87.5%, and 87.1% respectively, indicating a promising method that considered sound samples with different musical content. In contrast to the previous research on the estimation of source-level blending impression which employed musically constrained sound samples (such as notes or chords) of instruments from sophisticated recording conditions, this study surpasses earlier limitations by classifying 'ecological' sound samples of joint performances, even without accessing the individual source recordings.

Considering the significance of room acoustic and inter-musician feedback in joint music performance, the following chapter analyzed the quality of close-microphone recordings from in-situ conditions for the auralization of the ensemble sound. This was done by comparing the similarity and naturalness of auralized sound samples of ensemble performances with different numbers of violins, using Binaural Room Impulse Responses (BRIRs) generated from in-situ measurements and room acoustic simulations, against binaural recordings of the actual performance. Although the real binaural recordings were not always rated to be highly natural, auralization from the clip-on microphone signals using in-situ measured BRIRs exhibited a similar distribution of naturalness impression. Moreover, the naturalness of these auralized samples seemed to improve with an increasing number of violins, masking deficiencies in the clip-on mic recordings in the auralized output. This demonstrates the applicability of such close microphone recordings for auralization. While the samples of measured BRIRs do not demonstrate significant similarity to the recorded samples, an improvement in similarity rating is observed with an increase in the number of violins. However, a consistently poorer naturalness impression and low similarity rating in comparison to the samples of measured BRIRs highlight the deficiencies of BRIRs of GA-based room acoustic simulations in the re-synthesis of complex acoustic sound fields.

In the **second module** on **directivity perception**, two separate investigations explored the directivity-related attributes of individual instruments and the role of room acoustics in a musically realistic performance context: the first study on the role of directivity in source orientation perception, and the second study on the perception of dynamically varying directivity, in in-situ acoustic environments. The study on orientation perception explored the prediction accuracy of sound source orientation across four cardinal facing angles (front, back, left, right) within the horizontal plane. This was done by utilizing recordings of five musical instruments (trumpet, trombone, violin, flute, saxophone) with distinct directivity profiles in three performance spaces characterized by contrasting acoustic features, under static binaural listening conditions. Perceptual evaluation with expert listeners showed that, although individual instruments achieved high prediction accuracies only for particular directions, no significant differences were observed in their prediction accuracy values across all orientations involved. However significant differences were observed in the prediction accuracies of three room acoustic environments, suggesting room acoustics play a more influential role in orientation perception than sound source directivity. Additionally, the study explored the role of potential parameters, extracted from measured BRIRs for each condition, in orientation perception within an 'ecological' performance context. While certain parameters influence particular directions – Interaural Level Difference (ILD) and Interaural Cross Correlation (IACC) significantly affect lateral (left, right) perception, and spectral centroid of direct sound and Direct-to-Reverberant Ratio (DRR) provide cues for medial (front, back) orientation – a multifaceted nature of these parameters was observed in orientation perception under in-situ conditions.

The second study on the perception of dynamic directivity in in-situ conditions was tested by comparing the binaural recording of real instruments against those generated by two electroacoustic sources (omnidirectional source and studio monitor). Along with binaural recordings of the sound fields from five different instruments in different orientations and room acoustic environments, the study utilized binaural recordings of the replicated performances by playing back the close-mic recordings of specific instrument orientation through the two electroacoustic sources. Perceptual comparison of these samples by expert listeners indicated that while real instruments were generally rated to be more natural, the electroacoustic sources showed comparable naturalness ratings to particular instruments in particular acoustic conditions. Even with their distinct radiation characteristics and potential spectral coloration from spot-mic recordings, the electroacoustic sources were observed to improve their similarity to the real instruments and each other, under specific acoustic conditions. Therefore, although a rudimentary approximation of a real instrument by an electroacoustic counterpart mostly does not achieve perceptual closeness to the real instrument, certain acoustic conditions–characterized by room acoustic attributes and relative source orientation–tend to obscure the large directivity differences between the sound sources. Addition-

ally, this study explored and presented an objective method for modelling the perceptual similarity of binaural audio samples by applying Principle Component Analysis (PCA) to MFCC features extracted from these samples, to analyze the perceived directivity differences.

Based on these insights, the perceptual importance of high spatial resolution of directivity of individual instruments in the auralizations of ensemble performances was analyzed by changing the number of sources from 1, 2, to 5. This was carried out by employing two extreme cases for the room acoustic environment (echoic and anechoic) as well as the instrument type (trumpet with 'unidirectional' characteristics, and violin with 'multi-directional' characteristics). In a MUSHRA test comparing the audio samples created with various degrees of 'detailedness' of the directivity of sound sources, generated with scalable directional complexity by Spherical Harmonics truncation of high-resolution reference data, it was noted that the samples with relatively low-resolution directivity are perceptually close to those with high-resolution directivity reference. Although an optimal Spherical Harmonics order for a perceptually plausible auralization requires further investigation, it is noted to be significantly controlled by the directivity characteristics of the instrument utilized and also influenced by the room acoustic characteristics. Interestingly, the relevance of directivity characteristics remains valid even with an increasing number of sources, and it holds true for the two kinds of instruments in both acoustic environments. This underscores the necessity of employing a directivity having a perceptually optimal level of detailing, for auralization of musical ensembles.

The **third module** focusing on **the importance of acoustic environments in ensemble sound** analyzed the role of room acoustic attributes in shaping the blending between instruments in a musically realistic performance setting. This was achieved by employing a computational modeling approach to evaluate the perceived overall blending between instruments by examining both the blending at the source level and its alteration due to room acoustics. Audio stimuli of two violins with varying degrees of source-level blending were auralized in diverse simulated room acoustic environments, and expert listeners assessed their overall blending. While room acoustic reflections typically enhance the blending impression, correlation analysis of room acoustic parameters revealed that their impact on overall blending depends on the source-level blending of the stimuli used. Samples with poor source-level blending exhibited more significant potential for enhancement of the overall blending by the acoustic environment, whereas those with high source-level blending had limited room for improvement. This underscores the complex interplay between source-level blending and room acoustics in shaping the overall blending and also highlights the necessity to model overall blending as a function of both factors. Random Forest regression model for predicting the overall perceived blending using source-level blending ratings and room acoustic parameters was tested and validated through a comprehensive

two-fold evaluation, including separate training and test datasets as well as Leave-one-out-cross-validation, with a mean absolute error of 6% in each case. Feature importance analysis showed that while source-level blending contributes 60%, the room acoustics contribute the rest 40% to the overall perceived blending ratings, with perceived reverberance being the primary contributor. This suggests that the room contribution to the overall blending impression is nearly as significant as source-level blending between the instruments in this controlled experimental setting.

By bringing together the observations and results from the individual investigations on different aspects of joint performances, musical blending, and directivity perception, this thesis expects to provide insights into some of the key aspects of ensemble sound formation. Investigations on different stages of the evolution of musical blending between instruments highlight the key factors involved in the evolution of blending and also demonstrate the applicability of computational models to analyze and quantify complex psychoacoustic phenomena like musical blending. The attempt to understand and assess the blending as a function of source-level blend and room acoustic attributes for the first time through incorporating 'ecologically' realistic samples makes this study unique in the field. Thus, the blending evaluations are expected to offer insights into music performance, music perception-related research, and perceptually oriented room acoustic designing. Investigations on directivity perception of diverse instruments in in-situ performance spaces reveal the influence of room acoustics, thereby providing insights into music recording, orchestral arrangement, and communication acoustics. Studies on capturing individual source recordings in joint performance and modeling their directivity in the simulation are expected to advance the understanding for the auralization of a perceptually plausible and musically authentic ensemble or orchestra performance. Specifically, by providing cues on perceptually plausible representations of individual instrument directivities in ensemble performance, the study is expected to contribute to the advancement of virtual orchestra simulations by optimizing the computational efforts on source modeling.

9.2 Future works

While the classification modeling of source-level blending from 'ecological' sound recordings utilized MFCC features, the significance and contribution of musically oriented, practically explainable acoustic parameters like pitch, spectral centroid, onset difference, loudness, and formant location on blending perception are subjected to research. In an advanced version of the current classification model trained with large and diverse datasets, introducing sound samples with controlled variation of these acoustic parameters could help in estimating the significance of these parameters and also their transition point, i.e., the point at which a blended sample becomes non-blended. Additionally, given that the directivity characteristics of instruments as

well as their spatial positioning and orientation alter the energy distribution in the room, these aspects could have a potential impact on the degree of blending, necessitating further exploration. Utilizing a wide range of data sets having different numbers and combinations of instruments, a potential future goal would be to develop a time-varying source-level blending prediction model using advanced machine learning tools for dynamic musical signals. This model can be integrated into the existing approach that calculates the overall blend as a function of source-level blending rating and room acoustic parameters, which can lead to a comprehensive assessment of the overall blending as a time-varying parameter.

The results derived from the investigations involved in this study are mostly based on static source and receiver conditions. Therefore, advanced studies should explore the role of movement and rotation of the source and receiver during the performance, as these factors are relevant in realistic conditions. Although being a widely used method, considering the limitations of static headphone-based binaural reproduction which is mostly utilized in this investigation, future studies could integrate other spatial audio reproduction methods by utilizing measured/simulated Spatial Room Impulse Responses (SRIRs) and loudspeaker arrays for 3D sound field rendering. This would enable the ability for head rotation which could be relevant for the upcoming advanced studies on directivity perception-related investigations such as source orientation perception. Alternatively, future studies can advance further by employing a real-time binaural rendering system with a head tracker, enabling both head rotation and movement of listeners and musicians. This approach would have more practical real-life applications compared to the spatial sound field creation in controlled laboratory conditions.

The formation and evolution of ensemble sound are complex and multifaceted. Moreover, extensive evaluations are necessary for a detailed understanding of the different aspects involved in ensemble sound. While this thesis only addressed certain aspects of ensemble sounds, several important areas remain underexplored, especially the impact of room acoustic feedback on the musicians in an ensemble and the resultant changes in the ensemble sound formation. While the variation in performance strategies of individual musicians with changes in acoustic environments have been analyzed, an extended investigation covering both objective and perceptual aspects is necessary, particularly in the context of joint performances.

Appendix A

Room acoustic parameters

Considering the room acoustic environment as a Linear Time-Invariant (LTI) system, the transformation of the input signal (sound radiated from the sound source) to the output (sound received by the listener) can be analyzed using Room Impulse Response (RIR), which serves as the transfer function of the system. The RIR illustrates how a room responds to a Dirac impulse generated from a source by showing its transfer to the receiver as direct sound as well as series of impulses as room reflections with decaying amplitude with time. Based on the perceptual aspects, the RIR can be classified into three regions: the direct sound part (i.e., 0 - 5 milliseconds), the early reflections (5 - 50 or 80 milliseconds), and late reverberation (80 milliseconds - end of RIR). The direct sound from the instrument is crucial in sound source localization and distance estimation. While the early reflections contribute to the perception of clarity and source width impression, the late reverberation influences the perception of spaciousness and envelopment. Numerous room acoustic parameters developed over the last century addressing different objective and subjective features of room acoustic environments have been widely utilized in a standardized manner to characterize acoustic environments. The major parameters discussed in the thesis, mostly defined from the ISO 3382-1 [37] and IEC 60268-16-2020 [38] are described below.

Reverberation time: The reverberation time is defined to be the time required for the sound energy to decay by 60 dB once the source stops radiating, and it is regarded as the major parameter that characterizes the room acoustics by analysing the sensation of 'reverberance'. It is one of the first parameters that was proposed to characterize the room acoustic attributes. The Reverberation time can be assessed using the Sabine's formula given below;

$$RT_{60(\text{Sabine})} = \frac{0.161V}{\sum_{i=1}^{n} S_i\alpha_i + 4mV} \tag{A.1}$$

where V is the volume, S_i and α_i represents the surface area with its corresponding

absorption coefficient, m denotes the air absorption coefficient. It can also be measured as the time for a 60 dB decay of sound from the integrated Schroeder curve[188] estimated from the RIR. In practical conditions, if it not possible to measure the 60 dB decay due to the background noise, etc., the RT_{60} is analyzed using the parameters T_{30} or T_{20}. These parameters estimate time required for the decay of sound from -5 to -35 dB (30 dB) or -5 to -25 dB (20 dB) respectively from the Schroeder curve, and linearly extrapolate it to find the time required for 60 dB decay.

Early Decay Time (EDT): It is another decay parameter estimated by calculating the decay of the Schroeder curve from 0 to -10 dB and extrapolating it to obtain the time for 60 dB decay. Since it includes the initial decay part of the reverberation tail, it is shown to better represent the subjective sensation of 'reverberance' than other RT measures like T_{30} [39; 40].

Clarity (C_{80}): It is is defined to be the ratio of early (0-80 ms) to late (80-∞) arriving energy in dB scale, and it is often referred to as an indicator of clarity of music.

$$C_{80} = 10 \lg \left(\frac{\int_0^{0.080} p^2(t)\, dt}{\int_{0.080}^{\infty} p^2(t)\, dt} \right) \text{ dB} \tag{A.2}$$

Definition (D_{50}): It is a similar energy ratio measure as C_{80}, that is defined to be the ratio of early (0-50 ms in this case) to total energies expressed as linear fraction or percentage which is better shown to represent clarity of speech.

$$D_{50} = \frac{\int_0^{0.050} p^2(t)\, dt}{\int_0^{\infty} p^2(t)\, dt} \text{ dB} \tag{A.3}$$

Speech Transmission Index (STI): Speech Transmission Index (STI), defined by [44], demonstrates the intelligibility of speech in rooms. It is calculated by evaluating the degradation of the modulation depth of an excitation signal, described as the Modulation Transfer Function, by the room acoustic reflections as it is transferred from the source to receiver.

Strength parameter (G): It is the ratio between the sound energy of the RIR measured with an omnidirectional source inside the room and the energy of the same source measured in a free field at a distance of 10 m. This parameter demonstrates the ability of the acoustic environment to amplify sound energy from the source, and it is often used to describe the subjective sensation of loudness [40].

$$G = 10 \lg \left(\frac{\int_0^\infty p^2(t)\, dt}{\int_0^\infty p_{10}^2(t)\, dt} \right) \text{ dB} \tag{A.4}$$

Studies demonstrate that analyzing the early and late arriving sound energy individually can be a useful way to better comprehend the subjective and objective characteristics of the acoustic environment [177; 175], although this method is not standardized in ISO. Therefore, in this thesis work, the strength parameter is estimated for three conditions; G_{early} (energy in the early part of RIR, i.e., in 0 - 80 ms, including direct sound and early reflections), $G_{5\text{-}80}$ (energy of early reflections in 5 - 80 ms excluding direct sound), and G_{late} (energy of late reflections in 80 ms - ∞).

Early Lateral Energy Fraction (J_{LF}): When it comes to spatial perception of acoustic environments, apparent source width (ASW), and listener envelopment (LEV) are the two different key concepts that contribute the most [178]. Apparent source width is related to the energy of early reflections from lateral sides and it is assessed using the parameters Early Lateral Energy Fraction (J_{LF}) and Early Lateral Energy Fraction Cosine (J_{LFC}). J_{LF} is defined to be the ratio of energy coming from the lateral directions measured using a figure-of-eight microphone (p_L) for 5-80 ms (excluding direct sound), and the energy received by an omnidirectional receiver at the same location.

$$J_{\text{LF}} = \frac{\int_{0.005}^{0.080} {p_{\text{L}}}^2(t)\, dt}{\int_0^{0.080} p^2(t)\, dt} \tag{A.5}$$

Early Lateral Energy Fraction Cosine (J_{LFC}): Since the figure-of-8 microphone already has cosine directivity, the LF described above is observed to vary with the square of the cosine of the angle of incident reflections. To compensate for this effect, J_{LFC} is proposed as the ratio between the energy as the dot product of a figure-of-eight microphone to the omni receiver for 5 to 80 ms, and the energy of the omnidirectional receiver for 0-80 ms (defined in [37]). The J_{LFC} is therefore expected to better represent the subjective sensation of perceived source width.

$$J_{\text{LFC}} = \frac{\int_{0.005}^{0.080} |p_{\text{L}}(t).p(t)|\, dt}{\int_0^{0.080} p^2(t)\, dt} \tag{A.6}$$

Late lateral energy level (L_j): L_j is typically used to better represent the listener's envelopment sensation by assessing the energy of the late part received from lateral directions. It is defined as the logarithmic ratio between the late arriving (80-∞) energy from lateral directions measured with a figure-of-eight microphone, and the total energy received in an omnidirectional receiver placed at a distance of 10 m in a free field.

$$L_J = 10 \lg \left(\frac{\int_{0.080}^{\infty} {p_L}^2(t)\, dt}{\int_{0}^{\infty} p_{10}^2(t)\, dt} \right) \text{dB} \tag{A.7}$$

Bass Ratio (BR): It is defined as the ratio of reverberation times of low-frequency bands to mid-frequency bands. Although it is not a standardized parameter, it has been widely utilized in room acoustics perception research to quantify the balance of reverberation between low-frqquency and mid-frequency bands.

$$BR = \frac{T_{30,125Hz} + T_{30,250Hz}}{T_{30,500Hz} + T_{30,1000Hz}} \tag{A.8}$$

Treble Ratio (TR): Similar to Bass Ratio, the Treble Ratio indicates the balance between the reverberation between high-frqquency and mid-frequency bands. It is estimated as the ratio of reverberation times of low-frequency bands to mid-frequency bands as given below.

$$BR = \frac{T_{30,2000Hz} + T_{30,4000Hz}}{T_{30,500Hz} + T_{30,1000Hz}} \tag{A.9}$$

Appendix B

Musical score

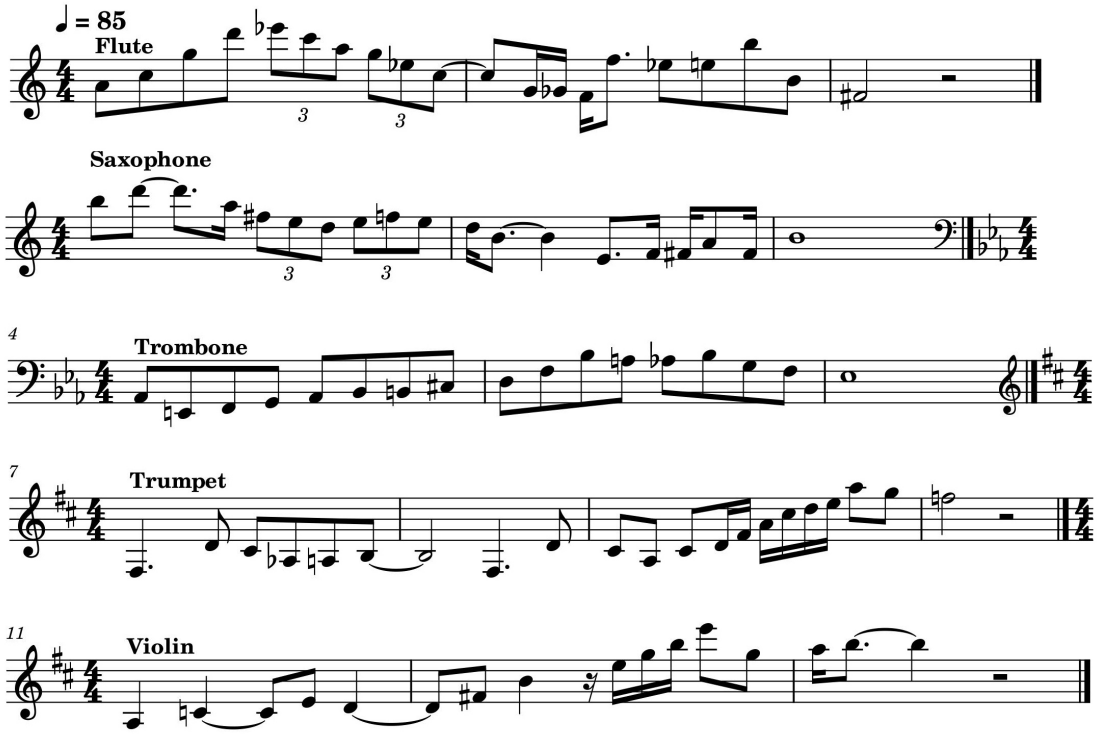


Figure B.1: Music scores for individual instruments from [152]: the score covers the full pitch range of instruments, essential for studying directivity perception of instruments through exciting diverse directivity shapes.

Appendix B. Musical score

Bibliography

[1] S. Ternström, "Preferred self-to-other ratios in choir singing," *The Journal of the Acoustical Society of America*, vol. 105, no. 6, pp. 3563–3574, 1999.

[2] J. F. Daugherty, "Choir spacing and formation: Choral sound preferences in random, synergistic, and gender-specific chamber choir placements," *International Journal of Research in Choral Singing*, vol. 1, no. 1, pp. 48–59, 2003.

[3] P. E. Keller, "Ensemble performance: Interpersonal alignment of musical expression," *Expressiveness in music performance: Empirical approaches across styles and cultures*, vol. 1, pp. 260–282, 2014.

[4] W. Goebl and C. Palmer, "Synchronization of timing and motion among performing musicians," *Music Perception*, vol. 26, no. 5, pp. 427–438, 2009.

[5] S.-A. Lembke, S. Levine, and S. McAdams, "Blending between bassoon and horn players: an analysis of timbral adjustments during musical performance," *Music Perception: An Interdisciplinary Journal*, vol. 35, no. 2, pp. 144–164, 2017.

[6] A. W. Goodwin, "An acoustical study of individual voices in choral blend," *Journal of Research in Music Education*, vol. 28, no. 2, pp. 119–128, 1980.

[7] A. Kuusinen and T. Lokki, "Wheel of concert hall acoustics," *Acta Acustica united with Acustica*, vol. 103, no. 2, pp. 185–188, 2017.

[8] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3148–3161, 2012.

[9] S. Weinzierl, S. Lepa, and D. Ackermann, "A measuring instrument for the auditory perception of rooms: The room acoustical quality inventory (raqi)," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1245–1257, 2018.

[10] S. Bolzinger, O. Warusfel, and E. Kahle, "A study of the influence of room acoustics on piano performance," *Le Journal de Physique IV*, vol. 4, no. C5, pp. C5–617, 1994.

[11] Z. Schärer Kalkandjiev and S. Weinzierl, "The influence of room acoustics on solo music performance: An experimental study.," *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 3, p. 195, 2015.

[12] K. Kato, K. Ueno, and K. Kawai, "Effect of room acoustics on musicians' performance. part ii: Audio analysis of the variations in performed sound signals," *Acta Acustica united with Acustica*, vol. 101, no. 4, pp. 743–759, 2015.

[13] S. V. A. Gari, M. Kob, and T. Lokki, "Analysis of trumpet performance adjustments due to room acoustics," in *Proceedings of International Symposium on Room Acoustics*, pp. 65–73, 2019.

[14] G. J. Sandell, "Roles for spectral centroid and other factors in determining" blended" instrument pairings in orchestration," *Music Perception*, vol. 13, no. 2, pp. 209–246, 1995.

[15] G. J. Sandell, *Concurrent timbres in orchestration: A perceptual study of factors determining "blend"*. PhD thesis, Northwestern University, 1991.

[16] J. Meyer, *Acoustics and the performance of music: Manual for acousticians, audio engineers, musicians, architects and musical instrument makers*. Springer Science & Business Media, 2009.

[17] D. R. Soderquist, "Frequency analysis and the critical band," *Psychonomic Science*, vol. 21, no. 2, pp. 117–119, 1970.

[18] B. C. Moore, *An introduction to the psychology of hearing*. Elsevier Ltd, London, 2004.

[19] B. N. Postma, D. Poirier-Quinot, J. Meyer, and B. F. Katz, "Virtual reality performance auralization in a calibrated model of notre-dame cathedral," in *Euroregio, Porto, Portugal*, 2016.

[20] J. H. Rindel and C. L. Christensen, "Room acoustic simulation and auralization–how close can we get to the real room," in *Proc. 8th Western Pacific Acoustics Conference, Melbourne*, 2003.

[21] J. Llorca-Bofí, C. Dreier, J. Heck, and M. Vorländer, "Urban sound auralization and visualization framework—case study at ihtapark," *Sustainability*, vol. 14, no. 4, p. 2026, 2022.

[22] A. Pedrero, A. Dıaz-Chyla, C. Dıaz, S. Pelter, and M. Vorländer, "Virtual restoration of the sound of the hispanic rite," in *Forum Acusticum*, vol. 5, 2014.

[23] B. N. Postma and B. F. Katz, "Creation and calibration method of acoustical models for historic virtual reality auralizations," *Virtual Reality*, vol. 19, pp. 161–180, 2015.

[24] R. A. Kendall and E. C. Carterette, "Identification and blend of timbres as a basis for orchestration," *Contemporary Music Review*, vol. 9, no. 1-2, pp. 51–67, 1993.

[25] M. De Francisco, M. Kob, J.-F. Rivest, and C. Traube, "Odessa – orchestral distribution effects in sound, space and acoustics: An interdisciplinary symphonic recording for the study of orchestral sound blending," in *Proceedings of the International Symposium on Music Acoustics (ISMA) 2019, Detmold, Germany*, 2019.

[26] S. Ioannou and M. Kob, "Investigation of the blending of sound in a string ensemble," in *Proceedings of International Symposium on Musical Acoustics (ISMA), Detmold, Germany*, pp. 42–49, 2019.

[27] D. Lincke, *Instrument blending in concert halls.* Master's thesis, Technische Universitat Berlin, Germany, 2020.

[28] J. Thilakan, O. C. Gomes, and M. Kob, "The influence of room acoustic parameters on the impression of orchestral blending," in *Proceedings of Euronoise 2021, Madeira, Portugal/Online*, 2021.

[29] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[30] S.-A. Lembke, K. Parker, E. Narmour, and S. McAdams, "Acoustical correlates of perceptual blend in timbre dyads and triads," *Musicae Scientiae*, vol. 23, no. 2, pp. 250–274, 2019.

[31] D. Tardieu and S. McAdams, "Perception of dyads of impulsive and sustained instrument sounds," *Music Perception*, vol. 30, no. 2, pp. 117–128, 2012.

[32] A. Antoine, P. Depalle, P. Macnab-Séguin, and S. McAdams, "Modeling human experts' identification of orchestral blends using symbolic information," in *Proceedings of the 14th International Symposium on Computer Music Multidisciplinary Research, Marseille, France*, pp. 544–555, 2019.

[33] S.-A. Lembke and S. McAdams, "The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds," *Acta Acustica united with Acustica*, vol. 101, no. 5, pp. 1039–1051, 2015.

[34] P. F. Assmann and Q. Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 680–697, 1990.

[35] A. de Cheveigné, H. Kawahara, M. Tsuzaki, and K. Aikawa, "Concurrent vowel identification. i. effects of relative amplitude and f difference," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2839–2847, 1997.

[36] A. Haapaniemi and T. Lokki, "The preferred level balance between direct, early, and late sound in concert halls.," *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 3, p. 306, 2015.

[37] ISO3382-1, *Acoustics: Measurement of Room Acoustic Parameters. Part 1: Performance Spaces*. International Organization for Standardization, Geneva, Switzerland, 2009.

[38] IEC60268-16, *Sound System Equipment-Part 16: Objective rating of speech intelligibility by speech transmission index*. International Electrotechnical Commission and others, 2020.

[39] V. Jordan, "A group of objective acoustical criteria for concert halls," *Applied Acoustics*, vol. 14, no. 4, pp. 253–266, 1981.

[40] G. A. Soulodre and J. S. Bradley, "Subjective evaluation of new room acoustic measures," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 294–301, 1995.

[41] A. P. Carvalho, A. E. Morgado, and L. Henrique, "Relationships between subjective and objective acoustical measures in churches," *Building Acoustics*, vol. 4, no. 1, pp. 1–20, 1997.

[42] A. Farina, "Acoustic quality of theatres: correlations between experimental measures and subjective evaluations," *Applied acoustics*, vol. 62, no. 8, pp. 889–916, 2001.

[43] R. Hawkes and H. Douglas, "Subjective acoustic experience in concert auditoria," *Acta Acustica united with Acustica*, vol. 24, no. 5, pp. 235–250, 1971.

[44] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[45] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2871–2876, 2003.

[46] T. Lokki, J. Pätynen, S. Tervo, A. Kuusinen, H. Tahvanainen, and A. Haapaniemi, "The secret of the musikverein and other shoebox concert halls," in *Ninth International Conference On Auditorium Acoustics, Paris, France, October 29-31*, Institute of Acoustics, 2015.

[47] G. Naylor, "A laboratory study of interactions between reverberation, tempo and musical synchronization," *Acta Acustica united with Acustica*, vol. 75, no. 4, pp. 256–267, 1992.

[48] J. Pätynen and T. Lokki, "Perception of music dynamics in concert hall acoustics," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3787–3798, 2016.

[49] R. Guski, "Auditory localization: Effects of reflecting surfaces," *Perception*, vol. 19, no. 6, pp. 819–830, 1990.

[50] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[51] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[52] S. Sutojo, J. Thiemann, A. Kohlrausch, and S. van de Par, "Auditory gestalt rules and their application," *The Technology of Binaural Understanding*, pp. 33–59, 2020.

[53] F. Otondo and J. H. Rindel, "The influence of the directivity of musical instruments in a room," *Acta acustica united with Acustica*, vol. 90, no. 6, pp. 1178–1184, 2004.

[54] J. Pätynen and T. Lokki, "Directivities of symphony orchestra instruments," *Acta Acustica united with Acustica*, vol. 96, no. 1, pp. 138–167, 2010.

[55] M. Pollow, *Directivity patterns for room acoustical measurements and simulations*, vol. 22. Logos Verlag Berlin GmbH, 2015.

[56] A. H. Benade, "From instrument to ear in a room: direct or via recording," *Journal of the Audio Engineering Society*, vol. 33, no. 4, pp. 218–233, 1985.

[57] D. Ackermann, F. Brinkmann, and S. Weinzierl, "Musical instruments as dynamic sound sources," *The Journal of the Acoustical Society of America*, vol. 155, no. 4, pp. 2302–2313, 2024.

[58] N. H. Fletcher and T. D. Rossing, *The physics of musical instruments*. Springer Science & Business Media, 2012.

[59] S. D. Bellows, *Acoustic Directivity: Advances in Acoustic Center Localization, Measurement Optimization, Directional Modeling, and Sound Power Spectral Estimation.* PhD thesis, Brigham Young University, 2023.

[60] S. D. Bellows, K. J. Bodon, and T. W. Leishman, "Baritone saxophone directivity," 2019. Brigham Young University, ScholarsArchive, Directivity. 7, *https://scholarsarchive.byu.edu/directivity/7*.

[61] S. D. Bellows, K. J. Bodon, and T. W. Leishman, "Flute directivity," 2023. Brigham Young University, ScholarsArchive, Directivity. 2, *https://scholarsarchive.byu.edu/directivity/2*.

[62] S. D. Bellows, K. J. Bodon, and T. W. Leishman, "Trombone directivity," 2020. Brigham Young University, ScholarsArchive, Directivity. 12, *https://scholarsarchive.byu.edu/directivity/12*.

[63] S. D. Bellows, K. J. Bodon, and T. W. Leishman, "Trumpet directivity," 2020. Brigham Young University, ScholarsArchive, Directivity. 13, *https://scholarsarchive.byu.edu/directivity/13*.

[64] S. D. Bellows, K. J. Bodon, and T. W. Leishman, "Violin directivity," 2020. Brigham Young University, ScholarsArchive, Directivity. 15, *https://scholarsarchive.byu.edu/directivity/15*.

[65] L. M. Wang and C. B. Burroughs, "Directivity patterns of acoustic radiation from bowed violins," *Architectural Engineering – Faculty Publications, 61*, 1999.

[66] T. Grothe and M. Kob, "High resolution 3d radiation measurements on the bassoon," in *International Symposium on Musical Acoustics (ISMA), Detmold, Germany*, pp. 139–145, 2019.

[67] A. C. Marruffo, J. Thilakan, A. Hofmann, V. Chatziioannou, and M. Kob, "High-resolution 3d directivity measurements of a trumpet," in *Fortschritte der Akustik-DAGA, Stuttgart, Germany*, pp. 131–133, 2022.

[68] L. M. Wang and M. C. Vigeant, "Evaluations of output from room acoustic computer modeling and auralization due to different sound source directionalities," *Applied Acoustics*, vol. 69, no. 12, pp. 1281–1293, 2008.

[69] A. Corcuera, V. Chatziioannou, and J. Ahrens, "Perceptual significance of tone-dependent directivity patterns of musical instruments," *Journal of the Audio Engineering Society*, vol. 71, no. 5, pp. 293–302, 2023.

[70] J. Pätynen, *A virtual symphony orchestra for studies on concert hall acoustics.* PhD thesis, Aalto University, 2011.

[71] H. Steffens, S. van de Par, and S. D. Ewert, "The role of early and late reflections on perception of source orientation," *The Journal of the Acoustical Society of America*, vol. 149, no. 4, pp. 2255–2269, 2021.

[72] D. P. Phillips and J. F. Brugge, "Progress in neurophysiology of sound localization," *Annual review of psychology*, vol. 36, no. 1, pp. 245–274, 1985.

[73] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[74] J. Blauert, *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

[75] J. G. Neuhoff, M.-A. Rodstrom, and T. Vaidya, "The audible facing angle," *Acoustics Research Letters Online*, vol. 2, no. 4, pp. 109–114, 2001.

[76] H. Kato, H. Takemoto, R. Nishimura, and P. Mokhtari, "Spatial acoustic cues for the auditory perception of speaker's facing direction," in *In Proc. of 20th International Congress on Acoustics (ICA) 2010, Sydney, Australia*, 2010.

[77] A. Y. Nakano, S. Nakagawa, and K. Yamamoto, "Auditory perception versus automatic estimation of location and orientation of an acoustic source in a real environment," *Acoustical Science and Technology*, vol. 31, no. 5, pp. 309–319, 2010.

[78] J. Edlund, M. Heldner, and J. Gustafson, "On the effect of the acoustic environment on the accuracy of perception of speaker orientation from auditory cues alone," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, pp. 1482–1485, Curran Associates, Inc., 2012.

[79] C. Imbery, S. Franz, S. van de Par, and J. Bitzer, "Auditory facing angle perception: The effect of different source positions in a real and an anechoic environment," *Acta Acustica united with Acustica*, vol. 105, no. 3, pp. 492–505, 2019.

[80] J. G. Neuhoff, "Perceiving acoustic source orientation in three-dimensional space," in *Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland*, 2001.

[81] B. B. Monson, J. Rock, A. Schulz, E. Hoffman, and E. Buss, "Ecological cocktail party listening reveals the utility of extended high-frequency hearing," *Hearing Research*, vol. 381, p. 107773, 2019.

[82] J. Edlund, M. Heldner, and J. Gustafson, "Who am i speaking at? perceiving the head orientation of speakers from acoustic cues alone," LREC, 2012.

[83] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization-an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[84] M. Vorländer, *Auralization: fundamentals of acoustics, simulation, algorithms and acoustic virtual reality*. Springer International Publishing, 2008.

[85] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[86] M. C. Vigeant, L. M. Wang, and J. Holger Rindel, "Investigations of orchestra auralizations using the multi-channel multi-source auralization technique," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 866–882, 2008.

[87] O. C. Gomes, W. Lachenmayr, J. Thilakan, and M. Kob, "Anechoic multi-channel recordings of individual string quartet musicians," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–7, IEEE, 2021.

[88] I. Witew, J. Paprotny, and G. Behler, "Auralization of orchestras in concert halls using numerous uncorrelated sources," *Proc. of the Institute of Acoustics*, vol. 28, no. 2, pp. 293–296, 2006.

[89] C. Böhm, D. Ackermann, and S. Weinzierl, "A multi-channel anechoic orchestra recording of beethoven's symphony no. 8 op. 93," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 977–984, 2021.

[90] M. Frank and M. Brandner, "Perceptual evaluation of spatial resolution in directivity patterns," in *DAGA Conf., Rostock, Germany*, pp. 18–21, 2019.

[91] A. Quélennec and P. Luizard, "Pilot study on the influence of spatial resolution of human voice directivity on speech perception," *Acta Acustica*, vol. 6, p. 10, 2022.

[92] M. R. Schroeder, "Novel Uses of Digital Computers in Room Acoustics," *The Journal of the Acoustical Society of America*, vol. 33, pp. 1669–1669, 11 1961.

[93] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[94] S. Weinzierl, P. Sanvito, F. Schultz, and C. Büttner, "The acoustics of renaissance theatres in italy," *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 632–641, 2015.

[95] D. Schröder, *Physically based real-time auralization of interactive virtual environments*, vol. 11. Logos Verlag Berlin GmbH, 2011.

[96] The user's manual of ODEON Room Acoustic Software version 16 is available at *https://www.odeon.dk/pdf/OdeonManual.pdf* (Last viewed February 1, 2024).

[97] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2746–2760, 2019.

[98] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[99] L. Shtrepi, A. Astolfi, S. Pelzer, R. Vitale, and M. Rychtáriková, "Objective and perceptual assessment of the scattered sound field in a simulated concert hall," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1485–1497, 2015.

[100] A. Taghipour, T. Sievers, and K. Eggenschwiler, "Acoustic comfort in virtual inner yards with various building facades," *International journal of environmental research and public health*, vol. 16, no. 2, p. 249, 2019.

[101] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[102] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[103] D. Schröder and M. Vorländer, "Raven: A real-time framework for the auralization of interactive virtual environments," in *Forum acusticum*, pp. 1541–1546, Aalborg Denmark, 2011.

[104] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, vol. 45, no. 6, pp. 456–466, 1997.

[105] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[106] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[107] K. I. McAnally and R. L. Martin, "Sound localization with head movement: implications for 3-d audio displays," *Frontiers in neuroscience*, vol. 8, p. 79254, 2014.

[108] G. Reardon, A. Roginska, P. Flanagan, J. S. Calle, A. Genovese, G. Zalles, M. Olko, and C. Jerez, "Evaluation of binaural renderers: a methodology," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[109] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2011–2023, 2017.

[110] E. A. Torres-Gallegos, F. Orduna-Bustamante, and F. Arámbula-Cosío, "Personalization of head-related transfer functions (hrtf) based on automatic photoanthropometry and inference from a database," *Applied Acoustics*, vol. 97, pp. 84–95, 2015.

[111] B. Zhi, D. N. Zotkin, and R. Duraiswami, "Towards fast and convenient end-to-end hrtf personalization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 441–445, IEEE, 2022.

[112] "Odessa – orchestral distribution effects in sound, space, and acoustics project, *www.actorproject.org/projects/funded-projects/strategic-projects/odessa*." Last viewed March 22, 2024.

[113] "Dpa 4099 specifications, *www.dpamicrophones.com/instrument/4099-instrument-microphone*." Last viewed March 22, 2024.

[114] "Head acoustics bhs ii binaural headset specifications, *https://cdn.head-acoustics.com/fileadmin/data/en/Data-Sheets/AH-BR/BHS-II-Binaural-Headset-3322-Data-Sheet.pdf*." Last viewed June 06, 2024.

[115] "Head acoustics hsu iii.2 dummy head microphone specifications, *https://cdn.head-acoustics.com/fileadmin/data/en/Data-Sheets/AH-BR/HSU-III.2-1391-Data-Sheet.pdf*." Last viewed June 06, 2024.

[116] "Neumann ku 100 dummy head microphone specifications, *https://www.neumann.com/en-in/products/microphones/ku-100/*." Last viewed June 06, 2024.

[117] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[118] J. Thilakan, B. B.T, J.-M. Chen, and M. Kob, "Sound samples for the evaluation of source-level blending between violins (1.1)," 2023. Zenodo *https://doi.org/10.5281/zenodo.8278236*.

[119] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[120] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[121] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of mfcc feature extraction accuracy using pca in indonesian speech recognition," in *International Conference on Information and Communications Technology (ICOIACT) 2018*, pp. 379–383, IEEE, 2018.

[122] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[123] S.-H. Chen and Y.-R. Luo, "Speaker verification using mfcc and support vector machine," in *Proceedings of the International Multiconference of Engineers and Computer Scientists (IMECS)*, vol. 1, pp. 18–20, Citeseer, 2009.

[124] B. Logan, "Mel frequency cepstral coefficients for music modeling.," in *Proceedings of International Symposium on Music Information Retrieval*, vol. 270, Plymouth, MA, 2000.

[125] R. Loughran, J. Walker, M. O'Neill, and M. O'Farrell, "The use of mel-frequency cepstral coefficients in musical instrument identification," in *Proceedings of International Computer Music Conference Proceedings (ICMC)*, 2008.

[126] M. Gilke, P. Kachare, R. Kothalikar, V. P. Rodrigues, and M. Pednekar, "Mfcc-based vocal emotion recognition using ann," in *Proceedings of International Conference on Electronics Engineering and Informatics (ICEEI) 2012*, vol. 49, pp. 150–154, 2012.

[127] S. Rajesh and N. Nalini, "Musical instrument emotion recognition using deep recurrent neural network," *Procedia Computer Science*, vol. 167, pp. 16–25, 2020.

[128] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE Region 10 conference*, pp. 1–5, IEEE, 2008.

[129] C. Richter, N. H. Feldman, H. Salgado, and A. Jansen, "A framework for evaluating speech representations," in *Proceedings of the Annual Conference of the Cognitive Science Society*, 2016.

[130] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," *Computer Science Review*, vol. 40, p. 100378, 2021.

[131] S. Dupont, T. Ravet, C. Picard-Limpens, and C. Frisson, "Nonlinear dimensionality reduction approaches applied to music and textural sounds," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2013.

[132] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[133] L. Van Der Maaten, E. O. Postma, H. J. van den Herik, *et al.*, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, pp. 66–71, 2009.

[134] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *Proceedings of International conference on computer graphics, simulation and modeling*, vol. 9, 2012.

[135] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI communications*, vol. 30, no. 2, pp. 169–190, 2017.

[136] J. Ye, T. Xiong, and D. Madigan, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis.," *Journal of Machine Learning Research*, vol. 7, no. 7, pp. 1183–1204, 2006.

[137] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[138] K. L. Elmore and M. B. Richman, "Euclidean distance as a similarity metric for principal component analysis," *Monthly weather review*, vol. 129, no. 3, pp. 540–549, 2001.

[139] M. K. Singh, N. Singh, and A. Singh, "Speaker's voice characteristics and similarity measurement using euclidean distances," in *Proceedings of International Conference on Signal Processing and Communication (ICSC), IEEE*, pp. 317–322, IEEE, 2019.

[140] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[141] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[142] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proceedings of 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE*, pp. 50–55, IEEE, 2019.

[143] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *Proceedings of the International workshop on Workshop on Multimedia Information Retrieval*, pp. 255–264, 2007.

[144] H. Terasawa, J. Berger, and S. Makino, "In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with mfcc-derived spectral envelopes," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 674–685, 2012.

[145] K. Siedenburg, I. Fujinaga, and S. McAdams, "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," *Journal of New Music Research*, vol. 45, no. 1, pp. 27–41, 2016.

[146] C. L. Christensen, G. Koutsouris, and J. H. Rindel, "The iso 3382 parameters: Can we simulate them? can we measure them," in *Proceedings of the International Symposium on Room Acoustics, Toronto, Canada*, vol. 910, 2013.

[147] S. A. Gari, T. Lokki, and M. Kob, "Live performance adjustments of solo trumpet players due to acoustics," in *International Symposium on Musical and Room Acoustics (ISMRA), Buenos Aires, Argentina*, Asociacion de Acusticos Argentinos, 2016.

[148] L. McCormack and A. Politis, "Sparta & compass: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, pp. 1–12, Audio Engineering Society, 2019.

[149] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[150] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[151] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (saqi)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[152] W. Buchholtzer, J. Thilakan, and M. Kob, "The impact of acoustic environments on the perception of directivity of musical instruments," *Fortschritte der Akustik-DAGA, Stuttgart, Germany*, pp. 856–859, 2022.

[153] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[154] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient (iacc e), lateral fraction (lf e), and apparent source width (asw) in concert halls," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, 1998.

[155] M. Skalevik, "The binaural signal from a symphony orchestra," in *Forum Acusticum, Lyon-France*, pp. 1265–1273, 2020.

[156] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer, "The ita-toolbox: An open source matlab toolbox for acoustic measurements and signal processing," in *Proceedings of the 43th Annual German Congress on Acoustics, Kiel, Germany*, vol. 2017, pp. 6–9, 2017.

[157] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[158] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[159] A. W. Mills, "Lateralization of high-frequency tones," *The Journal of the Acoustical Society of America*, vol. 32, no. 1, pp. 132–134, 1960.

[160] W. A. Yost and R. H. Dye Jr, "Discrimination of interaural differences of level as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 83, no. 5, pp. 1846–1851, 1988.

[161] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," *The Journal of the Acoustical society of America*, vol. 42, no. 2, pp. 489–493, 1967.

[162] "Outline globe source radiator manual, *https://outline.it/download/Documents/Manuals/Measurement/GlobeSource_eng.pdf/*."

[163] S. D. Bellows and T. W. Leishman, "Spherical harmonic expansions of high-resolution musical instrument directivities," in *Proceedings of Meetings on Acoustics*, vol. 35, AIP Publishing, 2018.

[164] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. A. Garí, "Spherical harmonic decomposition of a sound field based on microphones around the circumference of a human head," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 231–235, IEEE, 2021.

[165] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized hrtf fitting using spherical harmonics," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 257–260, IEEE, 2009.

[166] D. Ackermann, F. Brinkmann, F. Zotter, M. Kob, and S. Weinzierl, "Comparative evaluation of interpolation methods for the directivity of musical instruments," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, pp. 1–14, 2021.

[167] D. Ackermann, F. Brinkmann, and S. Weinzierl, "A database with directivities of musical instruments," *arXiv preprint arXiv:2307.02110*, 2023.

[168] F. Wefers, "A free, open-source software package for directional audio data," in *Proceedings of the 36th German Annual Conference on Acoustics (DAGA 2010), Berlin, Germany*, 2010.

[169] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "The fabian head-related transfer function data base," 2017.

[170] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st Web Audio Conference*, pp. 1–6, 2015.

[171] "Recommendation itu-r bs.1534-3 (method for subjective assessment of intermediate quality levels of audio systems), *https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf*." Last viewed June 13, 2024.

[172] C. Pörschmann, J. M. Arend, and R. Gillioz, "How wearing headgear affects measured head-related transfer functions," in *EAA Spatial Audio Signal Processing Symposium, Paris, France*, pp. 49–54, 2019.

[173] M. Kronlachner, "Mcfx plugin," 2020. Multichannel cross-platform audio plug-in suite *https://github.com/kronihias/mcfx* (Last viewed February 1, 2024).

[174] The user's manual of SQALA platform developed by Head acoustics is available at *https://cdn.head-acoustics.com/fileadmin/data/en/Data-Sheets/AS/ASM_15/D50500e-APR-500-ArtemiS-SUITE-Jury-Testing-SQala-Basic.pdf* (Last viewed February 14, 2024).

[175] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, 2011.

[176] L. L. Beranek, *Concert halls and opera houses: music, acoustics, and architecture*, vol. 2. Springer, 2004.

[177] J. S. Bradley, "Using iso 3382 measures, and their extensions, to evaluate acoustical conditions in concert halls," *Acoustical science and technology*, vol. 26, no. 2, pp. 170–178, 2005.

[178] J. S. Bradley and G. A. Soulodre, "The influence of late arriving energy on spatial impression," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2263–2271, 1995.

[179] H. Furuya, K. Fujimoto, C. Y. Ji, and N. Higa, "Arrival direction of late sound and listener envelopment," *Applied Acoustics*, vol. 62, no. 2, pp. 125–136, 2001.

[180] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[181] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble machine learning: Methods and applications*, pp. 157–175, 2012.

[182] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, pp. 1–11, 2008.

[183] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. van Hijum, "Data mining in the life sciences with random forest: a walk in the park or lost in the jungle?," *Briefings in bioinformatics*, vol. 14, no. 3, pp. 315–326, 2013.

[184] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[185] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[186] L. Kirkegaard and T. Gulsrud, "In search of a new paradigm: How do our parameters and measurement techniques constrain approaches to concert hall design," *Acoustics Today*, vol. 7, no. 1, pp. 7–14, 2011.

[187] W. Lachenmayr, *Perception and Quantification of Reverberation in Concert Venues*. PhD thesis, Hochschule für Musik Detmold, 2017.

[188] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6_Supplement, pp. 1187–1188, 1965.

List of Figures

List of Tables

DECLARATION

I declare that the work included in this thesis is my own work, completed solely and only with the help of the included references, except as stated otherwise in the text. This thesis has not been submitted for any other academic degree or professional qualification.

Jithin babu Pozhamkandath Thilakan	Detmold, Germany

List of publications

During the time of completion of this dissertation several articles were published, and some are under progress. The following articles which majorly contribute to form the thesis were written by the author.

- **J. Thilakan**, B. T. Balamurali, J. M. Chen, M. Kob, "Classification of the perceptual impression of source-level blending between violins in a joint performance," Acta Acustica 7, 62 (2023), (*https://doi.org/10.1051/aacus/2023050*).
- **J. Thilakan**, B. T. Balamurali, O. C. Gomes, J. M. Chen, M. Kob. "Exploring the role of room acoustic environments in the perception of musical blending," Journal of the Acoustical Society of America 157.2 (2025), 738-754 (*https://doi.org/10.1121/10.0035563*).
- **J. Thilakan**, O. C. Gomez, E. Mommertz, M. Kob, “Pilot study on the perceptual quality of close-mic recordings in auralization of a string ensemble”, Proceedings of Meetings on Acoustics vol. 49, Acoustical Society of America, (2023), (*https://doi.org/10.1121/2.0001686*).
- **J. Thilakan**, M. Kob, "Evaluation of subjective impression of instrument blending in a string ensemble", In Fortschritte der Akustik- DAGA, Vienna, (2021).
- **J. Thilakan**, O. C. Gomes, M. Kob , "The influence of room acoustic parameters on the impression of orchestral blending", Euronoise 2021, Portual (online), (2021).
- **J. Thilakan**, W. Buchholtzer, M. Kob, “Investigation of the perceptual relevance of directivity of real and virtual sources in different acoustic environments”, In Fortschritte der Akustik - DAGA, Stuttgart, (2022).
- **J. Thilakan**, B. T. Balamurali, W. Buchholtzer, J. M. Chen, M. Kob. "Source orientation perception; exploring the role of directivity of sound sources in diverse acoustic environments," (under preparation).
- **J. Thilakan**, A. C. Marruffo, L. R. Paz, D. Ackermann, T. Grothe, M. Kob. "Perceptual relevance of high-resolution directivity in the simulation of musical ensembles" (under preparation).

The author has written or contributed to the following articles which are not directly related to this thesis:

- **J. Thilakan**, B. T. Balamurali, P. M. Sarun, J. M. Chen, "Vocal Tract Resonance Detection at Low Frequencies: Improving Physical and Transducer Configurations," Sensors 23, no. 2:939, (2023), (https://doi.org/10.3390/s23020939).
- M. Kob, **J. Thilakan**, T. Grothe. "Simulation and interpretation of wind instrument directivity using a point source approach," (currently under review at Acta Acustica).
- **J. Thilakan**, B. T. Balamurali, J. M. Chen,"ACUZ-Lite: Ultra-Portable Real-Time Estimation of Vocal Tract Resonance", WESPAC, India(2018).
- **J. Thilakan**, B. T. Balamurali, P. M. Sarun, J. M. Chen, "Optimising 'ACUZ-LITE' for improved vocal tract estimation", ICSV, Prague (2021).
- W. Buchholtzer, **J. Thilakan**, T. Jench, D. Eddy, M. Kob, "Investigation of sound pressure near tone holes of a model pipe", In Fortschritte der Akustik - DAGA, Vienna, (2021).
- O. C. Gomes, W. Lachenmayr, **J. Thilakan**, M. Kob, "Anechoic Multi-Channel Recordings of individual string Quartet musicians", Proceedings of i3DA, Bologna, (2021).
- W. Buchholtzer, **J. Thilakan**, M. Kob ,"The impact of acoustic environment on the perception of directivity of musical instruments", In Fortschritte der Akustik - DAGA, Stuttgart, (2022).
- A. C. Marruffo, **J. Thilakan**, A. Hofmann, V. Chatziioannou, M. Kob, "Analysing musicians' acoustic shadowing on the directivity of the trumpet", In Fortschritte der Akustik - DAGA, Stuttgart, 131-133 (2022).
- M. Kob, **J. Thilakan**, N. H. Bernadoni, "Vocal-tract impedance at the mouth – from 1995 to today; A tribute to Joe Wolfe and John Smith", Stockholm Music Acoustic Conference (SMAC), Sweden (2023).

Curriculum Vitae

Personal information

Name :	Jithin babu Pozhamkandath Thilakan
Date of birth :	2nd December, 1995
Place of birth :	Thrissur, Kerala, India

Education

2011-2013 :	Higher secondary schooling, Kerala state board of examination, (aggregate score of 94%) Kerala, India
2013-2016 :	Bachelor of Science (B.Sc.) in Physics (CGPA: 8.9 /10) University of Calicut, Kerala, India
2017-2019 :	Master of Science (M.Sc.) in Physics (CGPA: 8.95 /10) Indian Institute of Technology (IIT-ISM) Dhanbad, India
2019-2024* :	Doctoral candidate in Music Acoustics at Erich Thienhaus Institute, Detmold University of Music, Germany.

Work experience

May 2018 – Aug. 2018 :	Research Internship at Singapore University of Technology and Design (SUTD), Singapore
Aug. 2019 – Feb. 2024 :	Early Stage Researcher in MSCA-funded Virtual Reality Audio for Cyber Environments (VRACE) project at Detmold University of Music, Germany.