Categorization, Differentiation, and Learning of Atypically Combined Mechanical Components of Musical Instruments

Erica Ying Huynh

Music Technology Area
Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

December 2023

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Contents

Abstract

Identifying sound sources would be impossible without assessing their timbres. In acoustical musical instruments, timbre provides information about two closely related mechanical components: the *excitation* which sets into vibration the *resonator*, a filter that amplifies, suppresses, and radiates sound components. Excitation-resonator interactions of musical instruments in the physical world are restrictive. For example, strings are bowed and struck but not blown. We used Modalys, a digital, physically inspired modelling platform, to combine three excitations (bowing, blowing, striking) with three resonators (string, air column, plate), simulating nine types of interactions. These interactions are either typical (e.g., bowed string) or atypical and physically impossible (e.g., blown plate). In three experimental studies, we examine whether categories of two mechanical components can be perceived and learned independently of one another beyond their typical interactions.

In the first experimental study, participants chose the excitation or resonator they thought produced each sound in separate blocks. Listeners categorized typical interactions accurately and made few confusions. Listeners correctly categorized either the excitations or resonators of atypical interactions, but never both. The mechanical component they incorrectly categorized was assimilated to the complementary mechanical component with which it typically interacts. For example, listeners correctly categorized the resonator of bowed air columns and the excitation of blown plates, so both were assimilated to blown air columns.

The second experimental study involved dissimilarity ratings of stimulus pairs. Multidimensional scaling (MDS) revealed a 3D timbre space. Dimension 1 showed a clear boundary between struck and sustained excitations and a subtle boundary between bowing and blowing. Positions of the sounds along Dimension 1 were best explained by changes in temporal centroid, which distinguishes impulsive from sustained excitations. Dimension 2 isolated plates from other resonators. Its acoustic correlates were a weighted sum of cues describing the global shape and tonal/noise content of the spectrum (e.g., spectral centroid, spectral crest, spectral crest, variability of spectral flatness). Strings

and air columns were further separated by Dimension 3, which was best explained by a weighted sum of audio descriptors related to the fine structure of the spectrum (e.g., tristimulus 2, harmonic spectral deviation).

The last experimental study comprised three separate learning tasks that trained listeners on the three excitation, three resonator, or nine interaction categories of the sounds. Training involved trial and error with corrective feedback and was followed by a categorization task without corrective feedback. Participants were less accurate at categorizing the atypical interactions than typical ones, but there was still a learning effect given that they made less confusions than in the first study.

Perception of atypical interactions changed depending on the task: categorical boundaries of excitation and resonator components were formed implicitly (study 2), but not made explicit (study 1) until training took place (study 3). The current studies demonstrate how timbre perception provides listeners with information about sound source mechanics to incorporate novel sounds into existing mental models or form new mental models for them. These studies emphasize timbre as a multidimensional attribute that contributes to the discernibility, identification, and learning of everyday sounds.

Résumé

Il serait impossible d'identifier les sources sonores sans évaluer leur timbre. Le timbre des instruments de musique fournit des informations sur deux composants mécaniques étroitement liés : l'excitation qui met en vibration le résonateur, un filtre qui amplifie, supprime et rayonne les composantes sonores. Les interactions excitation-résonateur des instruments de musique dans le monde physique sont restrictives. Par exemple, les cordes sont frottées et frappées, mais pas soufflées. Nous avons utilisé Modalys, une plateforme de modélisation numérique inspirée par la physique, pour combiner trois excitations (frotter, souffler, frapper) avec trois résonateurs (corde, colonne d'air, plaque), simulant ainsi neuf types d'interactions. Ces interactions sont soit typiques (par exemple, corde frottée), soit atypiques et physiquement inpossibles (par exemple, plaque soufflée). Dans trois études expérimentales, nous examinons si les catégories de deux composants mécaniques peuvent être perçus et appris indépendamment l'un de l'autre au-delà de leurs interactions typiques.

Dans la première étude, les participants ont choisi l'excitation ou le résonateur qui, selon eux, produisait chaque son dans des blocs séparés. Les auditeurs ont catégorisé les interactions typiques avec précision et ont fait peu de confusions. Ils ont correctement catégorisé soit les excitations, soit les résonateurs des interactions atypiques, mais jamais les deux. Le composant mécanique qu'ils ont incorrectement catégorisé a été assimilé à un autre avec lequel le composant correctement catégorisé interagit habituellement. Par exemple, les auditeurs ont correctement catégorisé le résonateur des colonnes d'air frottées et l'excitation des plaques soufflées, les assimilant tous deux à des colonnes d'air soufflées.

La deuxième étude portait sur les évaluations de dissimilarité des paires de stimuli. La mise à l'échelle multidimensionnelle (MDS) a révélé un espace de timbre en trois dimensions. La dimension 1 montrait une limite claire entre les excitations frappées et continues et une limite subtile entre les sons frottés et soufflés. La position le long de la dimension 1 s'explique le mieux par des changements dans le centroïde temporel, qui distingue les excitations impulsives et entretenues. La dimension 2 a

permis d'isoler les plaques des autres résonateurs. Ses corrélats acoustiques étaient une somme pondérée d'indices décrivant la forme globale et le contenu tonal/bruité du spectre. Les cordes et les colonnes d'air ont été séparées par la dimension 3, qui a été mieux expliquée par une somme pondérée de descripteurs audio liés à la structure fine du spectre.

La dernière étude comprenait trois tâches d'apprentissage distinctes qui formaient les auditeurs aux trois catégories d'excitation, aux trois catégories de résonateur ou aux neuf catégories d'interaction. L'entraînement comprenait des essais et des erreurs avec feedback et était suivi d'une tâche de catégorisation sans feedback. Les participants étaient moins précis dans la catégorisation des interactions atypiques que dans celle des interactions typiques, mais il y a eu un effet d'apprentissage étant donné qu'ils ont fait moins de confusions que dans la première étude.

La perception des interactions atypiques a changé en fonction de la tâche : les limites catégorielles des composants de l'excitation et du résonateur ont été formées implicitement (étude 2), mais n'ont pas été rendues explicites (étude 1) jusqu'à ce qu'un entraînement ait eu lieu (étude 3). Les études actuelles démontrent comment la perception du timbre fournit aux auditeurs des informations sur la mécanique des sources sonores afin d'incorporer les nouveaux sons dans les modèles mentaux existants ou de former de nouveaux modèles mentaux pour eux. Ces études soulignent que le timbre est un attribut multidimensionnel qui contribue à la discernabilité, à l'identification et à l'apprentissage des sons quotidiens et de leur identité.

Acknowledgements

Firstly, I would like to thank my supervisor, Stephen McAdams for helping me make all of this possible. Thank you for taking me on as a Master's student in 2017. I knew so little about timbre, but I was willing and excited to learn. I did not realize I would completely fall into the rabbit hole of timbre's role in sound source recognition, but I guess that is what research and a great supervisor does to someone! Stephen, you are clearly an inspiration to your students and collaborators, but you are also incredibly patient, kind, and overflowing with knowledge. My time in graduate school has been both thrilling and challenging, and you have made it all worth it. I feel very lucky to have you as a supervisor and mentor, and I cannot thank you enough for your understanding, moral support, and persistent belief in me. It really goes a long way to have a supervisor who is as considerate and encouraging as you.

I want to express my gratitude to my thesis examiners Gary Scavone and Stevan Harnad. Gary Scavone has seen every step of my project starting from my Master's work. I thank him for all his generous and helpful feedback and for his guidance in the more physics-based aspects of my research. After taking a course with Stevan Harnad, I became inspired to incorporate topics in cognitive science into my research. As a thesis examiner, I thank him for his insightful comments and questions during my defense, and for his kind compliments toward my project. I also thank my thesis committee member, Philippe Depalle, who has provided me with very useful comments during my comprehensive exam and thesis defense. He has an effective way of asking difficult questions by breaking them down into smaller, simpler questions to guide me to the answer. Additionally, I thank Helene Drouin for her patience and support in answering all my questions.

Bennett Smith, it has truly been a pleasure getting to know and work with you. Thank you for programming and helping me set up all my experiments. And thank you for all your help whenever I am facing a technological crisis. Some of my most fun memories during my time at McGill include working on experiments with you, having meetings that end up going on for hours, hiking up the

mountain (I really will miss those hikes), and the bantering that somehow becomes so therapeutic. You provide so much moral support to all the MPCL members—probably more than you will ever know. We have said this so many times, but truly, "every [research] lab needs a Bennett."

My current and former lab mates in the MPCL, it has been a gift to be among the presence of so many talented, hardworking, and helpful individuals coming from so many backgrounds of knowledge. Lena Heng, you have been here the entire time! I will never forget our discussions of your never-aging-most-beautiful-est-ness, the beer nights, and foosball matches. Thank you for all your help throughout the years and especially during the final stages of our degrees. To my amazing desk neighbours, Joshua Rosner and Andrés Gutiérrez Martinez, having your company and support in the last year has been extremely comforting. Jonas Regnier and Linglan Zhu, I am so grateful for the cooking nights and consistent moral support. Jade Roth, it has been so much fun having work sessions and playing with cats! Kit Soden, former Party Master and (ongoing) MPCL guru: thanks for passing on the Party Master role to me. Current Party Master and fellow Asian snack enthusiast, Ben Duinker, thanks for all the Toronto chats that reminded me of home! Corinne Darche, thank you for keeping me company during my "early" lunches in the lab (11:00 AM is a totally appropriate time for lunch, right?). Yuval Adler, thank you for your kindness and for sharing your incredibly delicious hummus at all the potlucks!!! Behrad Madahi, I could not have survived those audio computing courses without you. Huge thanks to Iza Korsmit for her guidance in all things SMACOF. To Yifan Huang, I really appreciate your help with running participants. It has been a true pleasure to witness your growing passion for research. Shoutout to Marcel Montrey, our statistics guru, for helping me with the analyses in my research. André Martins de Oliveira, thank you for your patience in booking all our travel and lodging accommodations during a hectic summer of conferences and dealing with that very long list of expense reports, all the while being so friendly and understanding. I really have learned so much from my former and current labmates. Thank you all for showing me so many different sides of timbre!

I am also grateful for my lovely pals who have been so supportive by endlessly cheering me on and up. Ricky Chow and Laagi Yoganathan, thank you for being my favourite study buddies since our undergrad days at McMaster University. My considerate, hard-working friend with the warmest heart, Mia Hershkowitz: I have really enjoyed plant shopping and our social distance walks. Kim Nguyen, thank you for always being there to listen and lift my spirits whenever I feel that I am lacking. And thank you, Erika Jang, for all the food adventures, plant trades, and for being willing to meet up on the rainiest, windiest days.

Family members: I finally made it!!! I am so grateful for your well wishes, encouragement, and for checking up on me. My grandparents, Kong Kong and Ma Ma, who are no longer with me physically, but will always be in my thoughts and heart; I thank you for all your love and care. To my loyal parents I owe the utmost appreciation and respect. My mother, Linda Chan, is my biggest role model. Her dedication, strength, and bravery are a tiny subset of the many qualities I look up to. My father, Tieu Binh Huynh, has always taught me to approach everything with confidence and to do what you love and love what you do. I am comforted by their company as I spent months at home analyzing data for the nth time and writing pretty much this entire dissertation. To me, they make the perfect team: my mother encourages me to do my best, while my father reminds me that it is okay to take breaks; I seek my mother for comfort during the most stressful times and I seek my father to give me a good laugh. My father has no idea what PhD stands for, so he has just been calling it the "Professor Huynh Degree." And now, I am hopefully one step closer to this title! I am so appreciative of the wonderful memories I have created with my parents and their immense, unconditional support. I love you Mama and Baba, and I sincerely thank you.

Contribution of Authors

This is a manuscript-based thesis. Each chapter in the body of this thesis (Chapters II, III, and IV) contains a manuscript that has been submitted or is prepared for submission to peer-reviewed scientific journals.

- Chapter II: Huynh, E. Y. and McAdams, S. (2023). Categorization of typical and atypical combinations of excitations and resonators of musical instruments: Assimilation of the unusual to the familiar. Manuscript submitted to *Music & Science*.
- Chapter III: Huynh, E. Y. and McAdams, S. (2023). Implicit differentiation of typically and atypically combined excitations and resonators of musical instruments. Manuscript intended for submission to *The Journal of the Acoustical Society of America*.
- Chapter IV: Huynh, E. Y. and McAdams, S. (2023). Learned categorization of atypically combined excitations and resonators of musical instruments. Manuscript intended for submission to *PLoS ONE*.

Stephen McAdams was the supervisor of this thesis and provided guidance in each research stage, from the experimental design, techniques of data analyses, and interpretation of the results. He also compensated participants for their time through his grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). Bennett Smith programmed each experimental interface and helped set up the technical aspects of the in-person (Chapters II and IV) and online (Chapter III) experiments. Marcel Montrey provided guidance in the statistical analyses of the experimental data. As a principal author, I was responsible for generating research questions, synthesizing stimuli, designing experiments, recruiting and running participants, analyzing data, interpreting results, and writing the manuscripts listed above.

List of Figures

List of Tables

Chapter I

Introduction

The ability to automatically recognize sound sources would be impossible without assessing their timbres. Timbre comprises a plethora of auditory attributes that bear perceptually useful information and influence the recognition of sound sources. Many previous studies have investigated how timbre perception correlates with the acoustical features of a sound wave (Caclin et al., 2005; Grey & Gordon, 1978; Lakatos, 2000; McAdams et al., 1995). These acoustical features do not arise spontaneously: they originate from sound sources, which have mechanical components. It is a common misunderstanding that mechanical components are secondary to acoustical features with respect to sound source identification, when it is in fact the primary cause of natural sound-producing events, such as those arising from musical instruments (Giordano & McAdams, 2010).

Two fundamental mechanical components of acoustic musical instruments are of interest to this dissertation: the excitation mechanism and resonant structure. They are closely related given that the excitation is a type of action that sets into vibration the resonator, which functions as a filter to amplify, suppress, and radiate sound components. Sounds produced by musical instruments are caused by interactions between excitations and resonators. In fact, the identification or recognition of a musical instrument relies heavily on the detection of these interactions.

Figure 1.1 outlines a schematic pathway describing how sound sources are recognized. This figure was initially presented by Li et al. (1991) and later supported by Giordano (2005) to explain the methodology used to study sound source recognition. Here, I will use it to describe the process of sound source recognition. The physical level features the mechanical components of sound sources and more specifically, how they interact with one another to produce a sound. An interaction is the coupling of a vibrating object to a controllable energy source. Interactions are defined in more detail in Section 1.3, but for now, simply consider that at the physical level, an excitation is applied to a resonator, and this interaction generates a sound. The generated sound has acoustic properties, forming the basis of the acoustical level. Some of these acoustic properties are called structural

invariants and others are called transformational invariants (McAdams, 1993). The structural invariants are the acoustic properties communicating the physical structure of the sound-producing object, such as its geometry and material. The transformational invariants are the acoustic properties that describe what is happening to the vibrating object or in other words, the sound-generating action. Detection of the appropriate structural and transformational invariants is what allows listeners to recognize the sound source. Recognition takes place at the perceptual level and can manifest in categorizing the sound or making judgments about it in isolation or relative to other sounds.

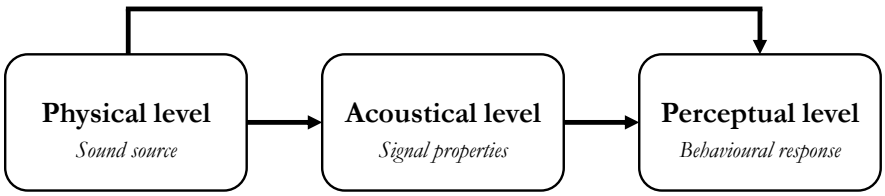

Figure 1.1 Framework of research design in sound source recognition, originally proposed by Li et al. (1991) and supported by Giordano (2005).

So, if a listener hears a bowed violin tone and can readily recognize it as such, it means they have identified the resonator to be a string and the excitation to be bowing through the detection of the structural and transformational invariants of the sound. With respect to the invariants of the sound, McAdams (1993) notes that "Each of these sets of properties remains constant in spite of variation in other properties" (p. 153). This means that the structural invariants of the violin string should remain constant regardless of whether it is bowed or plucked, i.e., differing in its transformational invariants. Of course, the *variation* that McAdams (1993) refers to primarily applies to typical interactions between excitations and resonators. One aim of this dissertation is to examine if the detection of structural and transformational invariants can drive the identification of excitation-resonator interactions in atypical contexts. That is, if structural invariants of a vibrating string are constant in spite of the variation in transformational invariants, then can a string still be identified as such if it is blown? Or is the detection of these invariants only helpful for identification when excitations and resonators are combined typically?

To investigate the perception of atypical excitation-resonator interactions, we used physically inspired modeling techniques to simulate these interactions. During my Master's research (Huynh, 2019), I used Modalys (Dudas, 2014) to simulate nine types of interactions between three excitations (bowing, blowing, striking) and three resonators (string, air column, plate). This generated four typical

interactions that are common to acoustic musical instruments: bowed string (BoSg), blown air column (BlAc), struck string (SkSg), struck plate (SkPl). The remaining five interactions were atypical and considered mechanically implausible (bowed air column [BoAc], blown string [BlSg], blown plate [BlPl]) or are rarely encountered (bowed plate [BoPl], struck air column [SkAc]). This dissertation presents three experimental studies that investigate the influence of atypically combined excitations and resonators on the categorization, dissimilarity perception, and learning of their mechanical components. The current chapter provides a review of the literature in sound source recognition, particularly focusing on sounds produced by musical instruments and impacted objects by comparing results from categorization and dissimilarity-rating tasks. An overview of categorization models and the relationship between categorization and learning is also discussed.

1.1 Timbre's Role in Musical Instrument Identification and Dissimilarity Perception

According to the American National Standards Institute, timbre is defined as the "attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are disimilar [sic]" (ANSI S1.1-1994, p. 34). This is a very misleading definition of timbre, describing it by what it is not instead of what it is. McAdams (2013) introduces two main facets of timbre. The first describes timbre as a plethora of auditory attributes, some of which continuously vary (e.g., auditory brightness, attack sharpness, nasality, roughness, richness, etc.), others of which are categorical (i.e., characteristic of a given musical instrument such as the pinched offset of a harpsichord sound). The second facet deals with the absolute categorization of a sounding object, such that timbre drives the identification and recognition of sound sources.

One misuse of timbre assumes that all notes played by a particular instrument, say a clarinet, for example, share a "clarinet timbre". However, McAdams and Goodchild (2017) explain that "a specific clarinet with a given fingering (pitch) at a given playing effort (dynamic) with a particular articulation and embouchure configuration produces a note that has a distinct timbre. Change any of these parameters and the timbre will change" (p. 129). So, it is better to think of the clarinet, or any musical instrument for that matter, as having a constrained universe of timbres rather than just a "clarinet timbre". Among the constrained universe of timbres of a specific instrument, there may be acoustic features that are maintained across all sounds that can be produced by it (McAdams & Goodchild,

2017). As mentioned, these acoustic features consequently drive the recognition of a musical instrument, usually by its mechanical or source components.

Timbres produced by musical instruments are affected by the gestures that generate the sounds. Although timbral features of a sound can be constrained by the physical properties of musical instruments, a variety of timbres can still emerge from a single musical instrument. Three methods of sound production are noted by Halmrast et al. (2010): moving strings, moving air directly, and striking plates or membranes. Specific control of certain parameters generates different gestures for each method of sound production. Timbres of sounds produced by moving strings with a bow will differ depending on the bow speed and bow pressure. Changes in breath pressure and embouchure pressure generate timbral changes in the movement of air. Furthermore, changing the force of a hammer on a plate or membrane will produce sounds with a variety of timbres. Changes in these parameters can also lead to variations in pitch and dynamics, which in turn, influence timbre perception.

1.1.1 Factors Influencing Musical Instrument Identification

Krumhansl and Iverson (1992) found that distinguishing between musical instruments was easier when their tones were produced by the same pitch. Nonmusician listeners recognize sounds as coming from the same instrument when their pitches are within an octave range (Handel & Erickson, 2001). This range extends to 2.5 octaves for musicians (Steele & Williams, 2006). McAdams et al. (2023) found that musical instrument identification improves when tones are played at pitches that are typical of the instrument. To minimize confusions between instruments resulting from pitch differences, the stimuli in the current studies are of the same pitch.

Fewer studies have investigated the role of musical dynamics on instrument identification. McAdams (2019) notes that a greater playing effort, such as playing in fortissimo, produces tones with greater energy at the higher frequencies than tones produced by less playing effort (e.g., pianissimo). Moreover, a greater playing effort causes more vibration modes to be excited, which generates a spectrum that spreads to higher frequencies. Fabiani and Friberg (2011) examined the influence of pitch, sound level, and timbre on the perception of dynamics. Their findings revealed that the identification of different dynamics was influenced by different sound levels and the timbres of different musical instruments, except for the flute, which did not impact the identification of dynamics.

Other studies have determined important portions of sounds for the identification of musical instruments. For example, the recognition of musical instruments worsens when attack portions are

removed from sounds (Berger, 1964; Elliott, 1975; Saldanha & Corso, 1964; Wedin & Goude, 1972). However, Saldanha and Corso (1964) found the presence of vibrato in the sustain portion of sounds improved categorization even when the attack was removed. Identification performance also decreased when sustain portions of sounds were removed. McAdams (1993) suggested that the information communicated by the attack portion is useful for the identification of a sound, such that it indicates how the instrument was set into vibration. When the attack is removed, listeners might rely on the patterns of change occurring in the sustain portion, such as vibrato, which can provide information about the instrument's resonant structure (McAdams, 1993; McAdams & Rodet, 1988).

Studies have also found that some instruments are more easily identifiable than others. Listeners in Saldanha and Corso's (1964) study more accurately identified the clarinet, oboe, and flute than the violin, cello, and bassoon. In Berger's (1964) identification task involving wind instruments, identification was better for the oboe, clarinet, cornet, and tenor saxophone than for flute, trumpet, alto saxophone, bassoon, French horn, and baritone. When a musical instrument was incorrectly identified, it was often confused for another instrument of the same family. Giordano and McAdams (2010) confirmed these findings in a review of musical identification studies. These researchers also found that confusions were made for sounds produced by similar excitation mechanisms. Extending these findings, McAdams et al. (2023) found that listeners confused harps and guitars (plucked strings), trombones and tubas (brass), and English horn with tenor saxophone and clarinet (woodwinds). Although these confusions were influenced by the registers of the instruments, they were still made for instruments that have similar resonant structures or excitation mechanisms.

1.1.2 Dissimilarity Perception and Acoustic Correlates of Musical Tones

The main purpose of obtaining dissimilarity ratings is to extract an interpretable set of salient dimensions underlying the perception of a set of stimuli (McAdams, 1993). Stimuli are presented in pairs, and listeners rate their perceived dissimilarity along a continuous scale ranging from "identical" to "very dissimilar". Dissimilarity ratings are then analyzed with multidimensional scaling (MDS), which assigns distances between the sounds in a timbre space model. The distance between any two sounds reflects their perceptual similarity or dissimilarity: sounds judged as similar will appear closer to each other and sounds judged as dissimilar will appear further apart (McAdams, 2013). Each dimension of the timbre space is often correlated with an acoustic property or group of acoustic properties that explain a proportion of its variance (McAdams 1993, 2013; McAdams & Giordano,

2015). It is important to note that these acoustic correlates only describe the most salient properties differentiating the sounds, but many descriptors can vary between any two sounds.

Table 1.1 Summary of the dimensions and their acoustic correlates of the timbre spaces reported in previous dissimilarity-rating experiments. Information about the stimulus set, such as whether the tones were recorded or simulated/synthesized, and the types of excitations and resonators involved are indicated. Acoustic correlates that were interpreted qualitatively are marked with an asterisk (*).

Study	Stimuli	Excitation	Resonator	No. of dim.	Spectral dimension	Temporal dimension	Other dimension
Gr77	Sim.	Bowing, Blowing	Chor., SR, DR, LV, AJ	3	Brightness*	Attack quality*	Spectral flux*
Kru89 [quantified by Kri94]	Sim.	Bowing, Blowing	Chor., SR, DR, LV	3	Spectral envelope* [Spectral centroid]	Temporal envelope* [Log attack time]	Spectral flux* [Spectral irregularity]
		Striking, Plucking	Chor., Idio.				
IvKru93	Rec.	Bowing, Blowing	Chor., SR, DR, LV, AJ	2	Spectral centroid	Amplitude envelope	
		Striking	Chor., Idio.				
Mc95	Sim.	Bowing, Blowing	Chor., SR, DR, LV	3	Spectral centroid	Log attack time	Spectral flux
		Striking, Plucking	Chor., Idio.				
Lak00	Rec.	Bowing, Blowing	SR, LV, AJ, Mem., Idio.	3	Log spectral centroid	Log attack time	Timbral richness*
		Striking, Plucking	Chor., Mem., Idio.				
Mar03	Rec., Sim.	Bowing, Blowing	Chor., SR, DR, LV	3	Spectral centroid, Spectral spread	Log attack time	
		Plucking	Chor.				

Note. Gr77 = Grey (1977); Kru89 = Krumhansl (1989); Kri94 = Krimphoff et al. (1994); IvKru93 = Iverson & Krumhansl (1993); Mc95 = McAdams et al., (1995); Lak00 = Lakatos (2000), mixed tones set; Mar03 = Marozeau et al. (2003), 2 semitones stimulus set. Sim. = simulated or synthesized; Rec. = recorded; Chor. = chordophones; SR = single reed aerophones; DR = double reed aerophones; LV = lip valve aerophones; AJ = air jet aerophones; Idio. = idiophones; Mem. = membranophones

There is no limit in terms of the number of dimensions that a timbre space can have, but most studies typically report two to four dimensions. A summary of six MDS studies and the acoustic properties associated with the dimensions that were either interpreted qualitatively (marked with asterisks) or determined quantitatively are reported in Table 1.1. General details pertaining to each stimulus set, such as whether the sounds were recorded or synthesized and the types of excitation mechanisms and resonant structures involved in the stimulus set, are also indicated. There are a few things to note about these studies. First, Krumhansl's (1989) stimulus set was later analyzed by Krimphoff et al. (1994), who quantitatively determined the acoustic properties of Krumhansl's timbre space dimensions (indicated in brackets in Table 1.1). All three stimulus sets (i.e., complete tones, onsets only, and onsets removed) of Iverson and Krumhansl's (1993) study are included in the summary table given that the three separate timbre spaces of the different stimulus sets were in agreement in terms of their dimensions and acoustic properties. Only the mixed tones set from Lakatos's (2000) study was considered as it comprised both harmonic and percussive tones which were most applicable to the stimulus set used in this dissertation. Lastly, only the stimulus set including two semitone pitch differences was included from the study by Marozeau et al. (2003) because the stimulus set with 11 semitone pitch differences resulted in a timbre space with a pitch related dimension, which would not be applicable to our stimulus set.

Each of these studies summarized in Table 1.1 found a temporal dimension and at least one spectral dimension. The temporal dimension in earlier studies seemed to be associated with the presence of low-amplitude, high-frequency energy, or inharmonic energy during the attack portion (Grey, 1977), or with changes in the amplitude envelope (Iverson & Krumhansl, 1993). More agreement, however, has been found for the log attack time in explaining the temporal dimension (Krimphoff et al., 1994; Lakatos, 2000; Marozeau et al., 2003; McAdams et al., 1995). The log attack time is known to distinguish impulsive from sustained excitations (Peeters et al., 2011). For the spectral dimension, almost all the studies reported its relation to the spectral centroid, which is the center of gravity of the energy distribution across frequencies and associated with auditory brightness (Peeters et al., 2011). Krimphoff et al. (1994) later confirmed that the spectral dimension in Krumhansl's (1989) timbre space—initially interpreted to be related to the spectral envelope—was highly correlated with the spectral centroid. In three-dimensional timbre spaces, the acoustic correlate of the third dimension appears to be less consistent across studies, with reports of: spectral flux (McAdams et al., 1995; and qualitatively by Grey, 1977), which is the fluctuation of the spectral envelope over time; spectral

deviation or the jaggedness of the spectral envelope (Krimphoff et al., 1994); and spectral spread (Marozeau et al., 2003), which measures the standard deviation of the spectrum around the spectral centroid. Moreover, the third dimension in Lakatos's (2000) timbre space did not appear to correlate with any acoustic properties, so it was qualitatively interpreted as timbral "richness". Overall, the acoustic correlates of the third dimension seem to depend on the type of sounds used in the stimulus set (McAdams, 1993; McAdams & Goodchild, 2017).

Knowledge of timbre space dimensions and their acoustic correlates is important for understanding the perceptual representation of a set of sounds. However, most studies to date have not directly linked the acoustic correlates of the dimensions to changes in the mechanical components of musical instrument tones. An exception was discussed in the findings of Grey's (1977) study, which suggested that the acoustic correlates of the second and third dimensions explain the clustering of the instrument families, distinguishing brass, woodwind, and string instruments from one another. Furthermore, Giordano and McAdams's (2010) review of previous dissimilarity-rating studies involving musical instrument tones found that tones produced by similar excitation types or instrument families appeared closer together in MDS space and occupied their own regions that were separate from other source types. The distinction between different excitation types, however, was more salient than that of the instrument families.

Of interest to the current research is to explore in further detail the influence of the acoustic correlates in differentiating mechanical components of musical instruments. For example, the log attack time and temporal centroid are known to distinguish impulsive from sustained excitations. However, both gross excitation categories each comprise different types of excitations, such as bowing and blowing for the sustained excitations and striking and plucking for impulsive ones. These more specific excitation categories, to the best of our knowledge, have not yet been directly linked to changes in log attack time, temporal centroid, or other temporal audio descriptors. Moreover, perceived brightness can be predicted by spectral centroid, but it is unknown if it can differentiate the instrument family categories or resonant structures. It is also possible that spectral centroid works in combination with other audio descriptors (e.g., those related to the spectrum fine structure) to differentiate the resonators or family categories. It is important to determine the acoustic correlates of dimensions in a timbre space, but it is equally important to explore their relationship to the sound-producing parameters of a signal.

1.2 Previous Studies on Impacted Materials

Although the current studies will concentrate on musical tones, many studies have focused on the direct manipulation of source components of impacted materials. In these studies, sounds are generally produced by impacting objects made of various materials (Giordano & McAdams, 2006; Klatzky et al., 2000; Lutfi & Oh, 1997; McAdams et al., 2004; McAdams et al., 2010). More recently, multiple impacts have been studied, such as bouncing or rattling (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012; Warren & Verbrugge, 1984). These studies are worth discussing, given that actions are synonymous with excitation mechanisms and materials are an aspect of resonant structures.

Sounds produced by different materials were perceived as more similar if they shared similar frequency components and decay contents (Klatzky et al., 2000). Most studies have found that categorization of materials is better across gross categories of metal-glass and wood-plastic (Giordano & McAdams, 2006; Hjortkjær & McAdams; Lemaitre & Heller, 2012). Lemaitre and Heller (2012) reported that there is no reliable acoustic property that differentiates the materials within gross categories, but a measure of the internal friction coefficient can differentiate between gross material categories. For solid objects, the coefficient of internal friction, $tan\theta$, can be related to the damping of vibration (Lemaitre & Heller, 2012). Distinctions within gross categories are more difficult, but not impossible. McAdams et al. (2010) found that distinctions across a continuum of simulated metal and glass plates (struck by mallets made of different materials) were linked to damping properties. Listeners considered the interpolations between models of thermoelastic damping for aluminum identification and viscoelastic damping for glass identification. Dissimilarity ratings, however, demonstrated that listeners considered differences in the wave velocity (related to material properties) in addition to the damping properties to differentiate the sounds. McAdams et al. (2010) therefore demonstrated that listeners made use of different acoustical properties, depending on the task. Research on action perception, however, is much less common. An exception was a study finding that participants can distinguish breaking versus bouncing glass (Warren & Verbrugge, 1984). Additionally, speeds of rolling balls can be determined by listeners (Houben et al., 2004). Determining the speed also depended on the size of the balls.

Recently, researchers have focused on the interaction between sound-producing actions and materials (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012). Lemaitre and Heller (2012) recorded wood, plastic, glass, and metal cylinders impacted by scraping, rolling, hitting, and bouncing. The first task of this study involved rating how well the stimuli resemble different action and material categories. In another task, participants categorized the stimuli based on their actions and materials.

Resemblance ratings and categorization performance revealed that action identification was more accurate and faster than material identification.

Hjortkjær and McAdams (2016) conducted studies on a set of recorded sounds that combined three actions (dropping, rattling, striking) with plates made of three materials (wood, metal, glass). The first experiment involved rating the dissimilarity of stimulus pairs. MDS based on the dissimilarity ratings revealed a timbre space with two dimensions. Dimension 1 separated gross material categories (wood versus metal-glass) and its acoustic correlate was the spectral centroid. Separation of the three actions was observed along Dimension 2, which was correlated with the temporal centroid (i.e., the center of gravity of the energy envelope). In a second experiment, the researchers synthesized two manipulations of the original stimulus set. The first manipulation, called spectral scrambling, removed spectral cues and preserved temporal cues. The second manipulation was called temporal scrambling, which removed temporal cues and preserved spectral cues. Participants in this task categorized sounds from the original and manipulated stimulus sets based on their actions and materials. Consistent with the findings in previous material categorization tasks (Giordano & McAdams, 2006; Lemaitre & Heller, 2012), identification of the different materials depended on whether they belonged to the same gross category or not. Identification of the different actions, however, was accurate regardless of gross category membership. Furthermore, material categorization was better for the temporally scrambled stimuli. As this manipulation preserved the spectral cues of the stimuli, it verifies the role of spectral centroid in differentiating gross material categories during the dissimilarity-rating task. On the other hand, action categorization was better for the spectrally scrambled stimuli, which confirms that temporal centroid best explains the differentiation among the actions. In summary, the findings from Lemaitre and Heller (2012) and Hjortkjær and McAdams (2016) demonstrated that listeners can differentiate sounds based on their actions and materials, but with greater sensitivity to action properties. Furthermore, listeners indeed prioritize the information pertaining to the mechanical components of sound sources in the processing of sounds.

1.3 Excitation-Resonator Interactions in Musical Tones

As mentioned, in musical instrument tones, there are two closely related macroscopic mechanical components involved in sound production: the excitation mechanism and resonant structure. These mechanical components interact to generate a sound. An interaction can be defined as a coupling process by which the excitation mechanism allows a controlled input of energy into the resonator to

set it into vibration. The three resonators of interest to this dissertation are the string, air column, and plate. Excitation mechanisms of interest are the frictional bow, (single) reed, and hammer mechanisms. Interactions that exist among acoustic musical instruments include bowed strings, blown air columns, struck strings, and struck plates. These interactions can generally be classified by their linearity or nonlinearity. Linearity refers to an increase in the output that is proportional to that of the input (Fletcher, 1999).

In a bowed string, the bow exerts a frictional force on the string. In a blown air column, the rate of air flow into the air column is controlled by the reed mechanism (McIntyre et al., 1983). The reed mechanism generates periodic vibrations only when it is attached to the air column. For the two sustained interactions, the bow or reed sets into vibration the resonator and the vibrations of the resonator in turn interact with the bow or reed. For the most part, sustained interactions rely on nonlinear couplings (McIntyre et al., 1983). The sounds produced by the nonlinear couplings in sustained periodic tones are harmonic, meaning the frequencies of the vibrating modes are integer multiples of the fundamental frequency. For a string that is fixed at the ends, the harmonic content will include both even and odd harmonics. An air column open at one end and closed at the other will produce odd harmonic content.

For impulsive interactions, such as the struck plate or string, the hammer sets into vibration the resonator instantaneously, so the vibrations of the resonator do not interact with the hammer. Impulsively excited instruments are primarily described as linear, and any nonlinearity is considered incidental (Fletcher, 1999). For the struck string, only the initial hammer contact is considered nonlinear. The coupling of a struck string produces nearly harmonic sounds. The coupling of struck plates does not correspond to any nonlinearity. Given that for most plate structures, the modes are not harmonically aligned (i.e., are not integer multiples of the fundamental), the sounds produced by it are usually inharmonic.

1.3.1 Typical and Atypical Interactions

The previous section discusses the couplings involved in typical excitation-resonator interactions. These interactions are considered typical because they are common to acoustic musical instruments and listeners, regardless of their musicianship, are quite familiar with them. Furthermore, they are mechanically plausible because listeners can conceptualize these interactions.

This dissertation is also concerned with the five atypical interactions previously synthesized by Huynh (2019): bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl),

struck air column (SkAc). They are atypical because they are physically impossible in the way that they are synthesized by Modalys. It is important to distinguish physical possibility from mechanical plausibility for two of these interactions. BoPl and SkAc can be considered mechanically plausible with respect to extended playing techniques in contemporary music. So, musicians will be more likely to conceptualize BoPl and SkAc than their nonmusician counterparts. A BoPl that is mechanically plausible would involve bowing a plate at its edge. However, Modalys simulates a plate that is fixed ats edges, so the bow passes through the plate to excite it. In a mechanically plausible SkAc, either a slap tongue technique can be applied to the mouthpiece of a clarinet or saxophone, or by removing the mouthpiece from their instrument, performers can slap the opening of the air column. Modalys, however, applies the striking excitation to the air molecules at a position along the length of the air column. So, for BoPl and SkAc, perceived mechanical plausibility will not coincide with physical possibility. On the other hand, physical possibility and mechanical plausibility are aligned in the context of the other three atypical interactions (BoAc, BlSg, and BlPl). They are considered mechanically implausible because listeners are likely unable to conceptualize their interactions in the physical world (i.e., there are likely no extended playing techniques resembling them) and they are also physically impossible in the way that Modalys simulates them. Findings from Huynh's (2019) study confirm that listeners have difficulty conceptualizing the atypical interactions, and they often perceive them as being produced by the typical interactions. This is likely because listeners have developed mental models for the typical interactions but not for the atypical ones.

1.4 A Mental Model of Sound Sources

A mental model is defined as an internal representation of how a system, such as a musical instrument, works or behaves in the world. Mental models are shaped by exposure: becoming more familiar with musical instruments allows one to have a better understanding of how they work and the sounds they can produce. Because musicians engage in sensorimotor activity with their musical instruments daily, they have a thorough understanding of the capabilities and physical limitations in terms of the sounds that can be produced by their instruments (Hajda, 2007). Playing in orchestral settings will also enhance musicians' understanding of the capabilities and limitations of other instruments. Nonmusicians, on the other hand, might have a general understanding of how musical instruments are played through passive exposure, but the specific techniques and restrictions of sound production may not be as well understood.

Evidence for mental models of sound sources is suggested in the findings of previous studies. Srinivasan et al. (2002) examined if experience with playing orchestral instruments affected musical instrument identification. Identification performance was tested with and without a short training session. Identification performance was no different between the presence or absence of a training session, but listeners who played orchestral instruments were better at instrument identification than those who did not. This suggests that long-term experience impacted listeners' mental models of musical instruments more so than short-term training. Neural evidence also demonstrates that mental models can be shaped by the type of instrument a musician is trained on. Trained violinists and trumpeters differ in their cortical representations of violin and trumpet tones even during passive listening (Pantev et al., 2001). Violinists have enhanced cortical representations for violin tones compared to trumpet tones, and the reverse is true for trumpeters. That is, trained musicians demonstrated a greater sensitivity to the timbres with which they have more experience. This is because they can better conceptualize how tones from their own instrument are produced. Evidence from mirror neurons support this explanation. Mirror neurons fire not only when a musician plays their own instrument, but also when they passively see or hear their instrument (or other instruments) being played (Cook et al. 2014). This is because seeing or hearing others play their instrument can activate regions in the motor cortex that allow one to carry out similar actions. So, mental models of musical instruments are shaped by the extent to which listeners can conceptualize the production of musical instrument tones.

1.5 A Review of Categorization Theories

Another interest of the current research pertains to the exploration of how categories are formed for musical instrument sounds. Research in the field of categorization takes a more general approach and is of course not limited to auditory stimuli. According to Rosch (1978), categories are formed on the foundation of two major principles. The first is based on cognitive economy, which means that categories should give us as much information about the world as possible with as little cognitive effort as possible. So, stimuli belonging in the same category are treated equivalently, whereas stimuli belonging in different categories are treated differently. This also means that it is beneficial to ignore certain properties of stimuli if they serve no purpose for categorization (Kruschke, 2005; Rosch, 1978). The second principle concerns how we structure the perceived world (Rosch, 1978). Humans are sensorimotor systems that interact with the world through what their sensory surfaces can afford

(Harnad, 2017). Not all the features of a given stimulus can be detected and an emphasis on some features can be placed over others. Through sampling and frequent exposure, the occurrences of some features of stimuli in the same category appear to be highly correlated. Stimuli of the same category should therefore share more co-occurring features than stimuli belonging to different categories. Detection of co-occurring features is enhanced by selective attention (Kruschke, 2005).

Most categorization theories discuss similarity-based and rule-based categorization. The former is automatic and holistic, whereas the latter is deliberative and reflective (Smith et al., 1998). Goldstone and Barsalou (1998) noted that similarity-based categorization relies on perceptual processes and rule-based categorization relies on conceptual processes and deeper meaning. The exact approach that categorizer adopts is unknown, but researchers have proposed that it incorporates some combination of similarity- and rule-based approaches (Erickson & Kruschke, 1998, 2002; Kruschke, 2005). We are additionally interested in the relationship between categorization and learning.

1.5.1 Rule-Based Versus Similarity-Based Categorization

Categories can be represented based on rules. For the most part, the rules are theory-driven, because they explain causal relationships underlying category membership. Items are categorized based on what best explains their attributes (Hampton, 1998). The rules can be featural, such that category membership depends on whether necessary and sufficient conditions are met (Kruschke, 2005). Once necessary and sufficient features are determined, they become the only basis for categorization, because they are the features that all members of the category have (Smith et al., 1998). Rules can also be based on boundaries that separate categories rather than the attributes of the category themselves. For example, Kruschke (2005) uses the example of skyscrapers, which are buildings that are taller than ten stories. Some theorists argue that the rules must be explicitly verbalized to have greater predictive power of category membership (Nosofsky et al., 1989). However, some rules are difficult to describe (Kruschke, 2005). The rule-based approach should also be applicable to items that are unfamiliar and abstract, not just ones that are familiar and concrete.

In similarity-based categorization, a novel stimulus to be categorized is compared to previously stored exemplars or a representative prototype in order to determine category membership. The more similar the item is to individual exemplars or a prototype of a category, the more likely it will be judged as belonging to that category (Goldstone, 1994; Krushke, 2005; Sloutsky, 2003). So, similarity-based categorization relies on the fact that items in the same category have more features in common. In exemplar theory, every encountered member of a category is stored as a representation in memory

(Kruschke, 2005; Goldstone et al., 2013). The similarity between a novel stimulus and every stored representation or exemplar of candidate categories is compared. The greater the similarity, the more likely the novel stimulus will be classified as a category member. Similarity assessment is mediated by selective attention (Nosofsky et al., 1989). Different attention weights may be placed on certain features, depending on how salient they are for optimizing categorization judgments (Estes, 1994). An argument against the exemplar theory is that it might be too tasking to store all known exemplars in memory (Kruschke, 2005). So, prototype theory suggests that category membership is decided based on an item's similarity to the most ideal category member, or prototype (Goldstone et al., 2013). The prototype can be defined as the most obvious case of the category or the member that best fits the category (Rosch, 1978). It can also be the average of every known case of the category, blending features of each of its encountered instances (Kruschke, 2005). Because the prototype comprises features that are representative of a category, these features are highly correlated with the attributes of other members in the same category and less correlated with features of members in different categories (Rosch, 1978).

Similarity-based and rule-based categorization suggest that different categories are independent of one another (Goldstone et al., 2013), when in fact some categories might be related to one another. With the number of stimuli we encounter daily and the different tasks we perform, it may be unrealistic to believe that categories are formed based on similarity or rules only. Some researchers have proposed a hybrid model of categorization that integrates exemplar theory and the rule-based approach (Erickson & Kruschke, 1998, 2002). The hybrid model comprises two simultaneously functioning modules. One module is rule-based, and the other is exemplar-based. Novel stimuli simultaneously activate both modules. The rule module determines where a stimulus falls within the boundaries separating category membership. The exemplar module determines the associations between the exemplars and categories. An additional component of the model, called the gate node (Erickson & Krushke, 1998), receives input from the exemplars. The gate node consequently decides which model will be used for categorization, primarily depending on the strength of the associations between the exemplars and categories. Stronger associations between exemplars and categories predict that categorization will rely more on the exemplar module and less on the rule module.

1.5.2 Similarity Perception and Categorization

Regardless of the use of exemplars or prototypes to decide category membership, there is a general principle that similarity can predict categorization (Goldstone, 1994; Sloutsky, 2003). In the context

of musical instrument tones, previous findings generally show that tones that are rated as more similar tend to be confused for one another during categorization (Giordano & McAdams, 2010; McAdams, 1993). In Section 1.2, we discussed Hjortkjær and McAdams's (2016) study, in which the results from the dissimilarity ratings and categorization of multiple impact sounds complemented one another. However, in another study by McAdams et al. (2010), listeners prioritized different acoustic features, depending on whether they rated the dissimilarity of the simulated plates or categorized their material.

Harnad (1990) argues that the methodology and task involved in testing dissimilarity perception are independent of those involved in testing categorization performance. Dissimilarity ratings rely on iconic representations, which are analog projections of the sensory input. Instances of iconic representations, such as the notes played by a musical instrument, vary in terms of the instrument's constrained universe of timbres. They also vary in terms of how they are presented; for example, they can be heard live or as recordings. Dissimilarity perception merely concerns the degree to which these instances are different. Listeners can consequently judge the dissimilarity of sounds without knowing which instrument played them. So, iconic representations are not governed by category boundaries. Identification, on the other hand, involves an absolute judgment of the category membership of a sound. This relies on categorical representations, such that categorization relies on the detection of invariant features that differentiate members from nonmembers. Moreover, invariants also distinguish nonmembers that might be confused for one another.

Given the independent methods of dissimilarity ratings and categorization, we might find discrepancies between how the same set of sounds is processed across the two types of tasks. This might be especially true for unfamiliar or novel sounds, much like those produced by the atypical excitation-resonator interactions used in the current studies. Identification of the mechanical components of the atypical interactions may demonstrate some confusions, whereas assessing the dissimilarity of pairs of them can rely on any features that are uncommon between them without explicitly having to name the features or the sounds. So, much like the study by McAdams et al. (2010) the findings from a dissimilarity-ratings task might not always predict those from a categorization task.

1.5.3 Categorization Theory Based on Learning

Because categorization involves the detection of category-distinguishing features, Harnad (1990) additionally notes that these features can be learned, and learning facilitates and maintains categorization. In fact, most categories are learned rather than innate (Harnad, 2017). Through evolution, we acquire the innate feature-detectors that reliably differentiate categories. Learned

categories, however, involve the detection of invariant features through reinforcement training or supervised learning. Therefore, we are interested in categorization theories that highlight the role of learning.

As sensorimotor systems, humans interact with sensory input based on the affordances of their sensory surfaces (Harnad, 2017). The auditory system, for example, allows humans to hear frequencies between 20 to 20,000 Hz. This is different from the auditory system of bats, for example, which allow them to hear frequencies between 2,000 to 110,000 Hz (Strain, 2017). So, there are sounds at frequencies higher than 20,000 Hz that can be detected by bats because their auditory system affords it, whereas the auditory system of humans does not. In the constrained universe of timbres of a given musical instrument, there might be invariant features that allow them to be identified as being produced by the same instrument, or more generally, the same instrument family set into vibration by similar types of excitations, as demonstrated by previous categorization tasks (Berger, 1964; Giordano & McAdams, 2010; McAdams et al., 2023). Detection of these invariant features allows listeners to produce the same output (i.e., identify a violin/bowed string) when presented with the same kind of input (i.e., notes of a violin or bowed string that vary in many features).

Harnad (2017) points out that sensorimotor detection of invariances is not enough. There needs to be an additional process that indicates which invariants reliably determine category membership. This additional process is learning, which plays a large role in categorization. Inputs are sampled and outputs are produced based on trial and error. Furthermore, corrective feedback is crucial for informing the categorizer whether the output was right or wrong. Therefore, corrective feedback facilitates the selective filtering of invariants that reliably determine category membership and the selective ignoring of the remaining variation. Supplementing trial and error with corrective feedback is called supervised learning (Harnad, 2017). Learning is successful when categorization improves and when fewer errors are made.

Several studies have implemented short-term supervised learning tasks to examine categorization performance in various fields (Goudbeek et al., 2009; Lupyan, 2006; McAdams et al., 2023). These studies usually involve a training phase which is based on supervised learning. The maintenance of learning is determined by a testing phase. Each of the studies discussed in this section contribute important findings that are considered for the design of the learning tasks in Chapter IV.

Lupyan (2006) found that individuals can classify images of aliens as ones to be approached versus ones to be avoided with supervised category training. Furthermore, the findings conclude that learning is faster and more robust when unfamiliar categories are learned with labels than without labels. This

finding speaks to the usefulness of labels, given that they allow participants to associate multiple characteristics of a category with a simple verbal distinction.

In a study by Goudbeek et al. (2009), listeners learned to categorize auditory stimuli varying along one dimension in duration or formant frequency or varying in both dimensions. In the training phase, corrective feedback was either present or absent. Categorization during the testing phase improved for stimuli varying in one dimension when there was corrective feedback during the training phase. It seemed difficult for participants to learn the categories of the stimuli varying in two dimensions, even with supervised learning. However, Goudbeek et al. (2009) noted that learning depended on whether the variability of the stimuli in each category was presented in both the training and testing phases. Given that the training phase presented stimuli that represented the variability of each category, learning was more successful when the stimuli in the testing phase also represented the variability of each category than when it did not. This study therefore highlights the importance of training and testing participants on the variability of each category, as it has the potential to improve categorization learning of complex, multidimensional stimuli.

McAdams et al. (2023) investigated whether listeners can learn to identify musical instruments across changes in pitch using supervised learning. Identification performance varied across musical instruments, and the detection of invariants seemed to be context dependent. For example, the cello and tubular bells were identifiable across pitches because they each had specific acoustic features (i.e., bow noise for cello, inharmonicity for tubular bells) that might have been detected as invariants. For other instruments, such as English horn and vibraphone, supervised learning was effective when the pitches of the tones that listeners were tested on were consistent with the ones they were trained on. So, for these instruments, listeners focused on the shared invariants in the pitches of their tones across the two phases. On the other hand, listeners were able to identify the tuba even when the pitches of its tones in the testing phase were lower than during training. As tubas are known for their lower register, listeners might have detected the shared invariant features between tuba tones of the testing set and tuba tones from previous exposure. Confusions within instrument families can be explained by the fact that instruments of the same family generally cover different pitch registers (i.e., high, middle, low), so identification might have been based on the typical pitches of a given instrument. Additionally, it might have been easier for listeners to detect the invariants within different pitches that were played by instruments of the same family than the invariants across pitch differences of a particular instrument. Listeners appeared to have adopted different strategies to try to identify the instruments producing the tones. In any case, Harnad (2017) states that it is easier to determine *what*

listeners can learn to categorize, but explaining *how* they manage to detect invariants to reliably categorize stimuli proves to be a more difficult task.

1.6 Thesis Overview

The next three chapters of this dissertation present manuscripts that have been submitted or are intended for submission to scientific journals. Each manuscript is based on one of three scientific experiments that test:

1) The categorization of nine types of excitation-resonator interactions based on their excitations or resonators (Chapter II);
2) Dissimilarity perception of pairs of sounds to model a perceptual representation of the nine interactions (Chapter III); and
3) Whether the nine interactions can be learned based on their excitation, resonator, or interaction categories (Chapter IV).

First, a summary of the stimulus design and experimental findings of two previous experiments based on my Master's research will be provided. Then, an outline of the current experimental studies along with the corresponding research questions and hypotheses will be presented.

1.6.1 Summary of the Stimulus Design

Stimuli used in the studies of Chapters II and III were the same as those synthesized in previous experiments (Huynh, 2019). Some stimuli in Chapter IV were slightly modified but the general synthesis approach was similar to the previous experiments. The purpose of the stimulus design was to simulate combinations between three excitation mechanisms (bowing, blowing, striking) and three resonating structures (string, air column, plate). This was done with Modalys (Dudas, 2014), which is a physically inspired modeling synthesizer that was developed at the Institut de recherche et coordination acoustic/musique (IRCAM). Modalys allows the user to operate as a (digital) instrument designer by employing modal synthesis to simulate the behaviour of a structure in reaction to an external excitation that is applied to it (Dudas, 2014). Modal synthesis estimates the properties of a resonator by decomposing a vibrating structure into the weighted sum of constituent modes (Ellis et al., 2005). Each mode can be defined by modal shape (i.e., amplitude), eigen frequency, and loss factor (i.e., decay rate). Properties of an excitation are estimated by solving time equations that describe the temporal evolution and interactions of each mode (Ellis et al., 2005). An exploratory approach was

implemented to simulate the nine excitation-resonator interactions under the manipulation of two selected parameters for each interaction. Twenty values were tested for each parameter. The 20 values of each of the two manipulated parameters were combined in each interaction, generating a total of 400 (20×20) stimuli for each interaction. The main aim of this synthesis design was to have some variability within each interaction type, particularly in the timbres that they would produce.

1.6.1.1 Resonant Structures

Three completely different models were used to simulate the resonators. For the air column, we used Modalys's tube object to simulate the modes representing the air particles within the air column. Given that a conical air column and string excited at a short distance from the bridge can be modeled similarly, we used a cylindrical rather than a conical air column. The air column was also open at one end and closed at the other, whereas the string was fixed at both ends. Furthermore, the string and air column are modeled after different systems. The string represents a mechanical system. The air column, on the other hand, is itself a mechanical structure but functions as an acoustical system in which the air can vibrate. The string and air column are also expected to differ in their harmonic content. The string will produce modes vibrating at frequencies that are integer multiples with respect to the fundamental frequency (i.e., even and odd harmonic content). The air column, on the other hand, will have primarily odd harmonic content, such that its odd harmonics will have much greater energy than its even harmonics. Modalys's plate object was rectangular, thin, and fixed at its edges. It is expected to generate modes vibrating at frequencies that are not harmonically related to the fundamental frequency (i.e., inharmonic content). The exception would be if a sustained excitation is applied to plate: a fundamental frequency will correspond to one of the modes and the remaining modes will be close to harmonically aligned to the fundamental. Parameters of each resonator were set to produce the lowest mode of vibration at 155 Hz, corresponding to a pitch of E-flat-3. Parameters of each resonator were kept as consistent as possible across each type of excitation that was applied to them.

1.6.1.2 Excitation Mechanisms

As much as possible, the same temporal envelope of each excitation was applied to each resonator. Two chosen parameters of each excitation were manipulated based on their ability to influence the resulting timbres of the produced sounds (Halmrast et al., 2010). The bowing mechanism is a mechanical system. The two manipulated parameters were the bow speed and bow pressure, which

are modeled as the velocity and the vertical displacement of the resonator by the bow, respectively. Manipulating the bow speed will primarily influence the loudness of the resulting sound and manipulating the bow pressure should influence its brightness (Halmrast et al., 2010; Rossing & Hanson, 2010). The bowing excitation was first applied to the string because bowed strings are typical of acoustic musical instruments. Time values of the temporal envelopes of the bow speed and pressure were adjusted to make the bowing sound as realistic as possible. Then, these temporal envelopes were applied to the air column and plate. A combination of 20 values for both the maximum bow speed and maximum bow pressure were used for initial synthesis when bowing was applied to the three resonators.

Blowing was simulated with a mouth and reed in Modalys. The simulation involved an oscillating flow of air that was activated by a vibrating reed. So, the mouth and reed system is mechanical, but controlled partly by acoustic pressure. One of the parameters of interest was the breath pressure. The other parameter was described by the pressure of the lips on the reed which controls the physical resting position or opening of the reed and is hereafter referred to as the embouchure pressure (Coyle et al., 2015). Together, the breath pressure and embouchure pressure influence the timbres of the resulting sounds such that a higher breath pressure and tighter embouchure pressure (i.e., smaller reed tip opening) should produce brighter sounds. As blowing typically interacts with an air column, the blown air columns were synthesized first, and the time values of temporal envelopes of the breath pressure and embouchure pressure were modified to produce a realistic blowing sound. Then, the same temporal envelopes for each parameter were applied to the other resonators. A combination of 20 values of the maximum breath pressure and of the maximum embouchure pressure were used for the initial synthesis when blowing was applied to the three resonators.

The striking excitation made use of Modalys's hammer object. It was first applied to the string and plate, given that these were simulations of typical interactions. When applied to the string, we manipulated the force of the hammer, which is modeled as the displacement of the resonator by the hammer on Modalys. Slight adjustments were made to the temporal envelope of the hammer force to generate a realistic striking sound. Then, the same temporal envelope could be applied to the air column. In the previous studies (Huynh, 2019) and the studies presented in Chapters II and III, we also manipulated the position at which the output was recorded on the string or air column. For the study in Chapter IV, we instead manipulated the position along the length of the resonator at which it was struck. We went with this modification for Chapter IV because the modes that are activated by striking a resonator at one position will differ from the modes that are activated from striking the same

resonator at a different position. This should produce a greater variability in the timbres of the synthesized sounds compared to the previous manipulation in which the resonator was struck at one position and the output is recorded at different positions. For the struck string and struck air column, we tested a combination of 20 values of the maximum hammer force (all studies) and 20 values corresponding to the position at which the output was recorded (in Huynh, 2019, and Chapters II & III) or 20 values corresponding to the positions at which the resonator was struck (Chapter IV) in the synthesis of these interactions.

When striking was applied to the plate, we used the same temporal envelope for the control of the hammer force as when the string and air column were struck. Changing the maximum force of the hammer striking the plate, however, did not appear to impact the resulting sounds as Modalys normalizes their amplitude. Instead, in Huynh (2019) and Chapters II and III, we tested a combination of 20 horizontal coordinates and 20 vertical coordinates on the plate at which the output was recorded. In Chapter IV, we instead tested the combination of the 20 horizontal and 20 vertical coordinates of the plate at which it was struck. Again, changing the position at which the plate is struck will cause the activation of different modes of vibration, which would generate sounds with a greater variability in their timbres.

1.6.1.3 Interaction Exemplars

Given that there were 400 versions of each of the nine excitation-resonator interactions, we first took note of which of them produced an audible output. Of the stimuli with audible outputs, we informally chose three (Huynh, 2019, and Chapters II & III) or seven (Chapter IV) exemplars to represent each interaction. The chosen exemplars for each interaction were perceived to be produced by the same source components and conveyed the variability in their timbres. This was supposed to represent the fact that any set of sounds produced by the same source can vary in their timbres (McAdams & Goodchild, 2017). Stimulus sets for each chapter can be accessed here.

1.6.2 Perceived Resemblance of Excitations and Resonators

In two previous experiments (Huynh, 2019), two groups of listeners rated the extent to which the exemplars of each interaction resembled each of the excitations (group 1) or each of the resonators (group 2). For each interaction type, the mean rating of the resemblance to each excitation and resonator category is presented in Figure 1.2. For the typical interactions—bowed string (BoSg), blown

air column (BlAc), struck string (SkSg), and struck plate (SkPl)—resemblance ratings were highest for the excitations and resonators that produced the sounds.

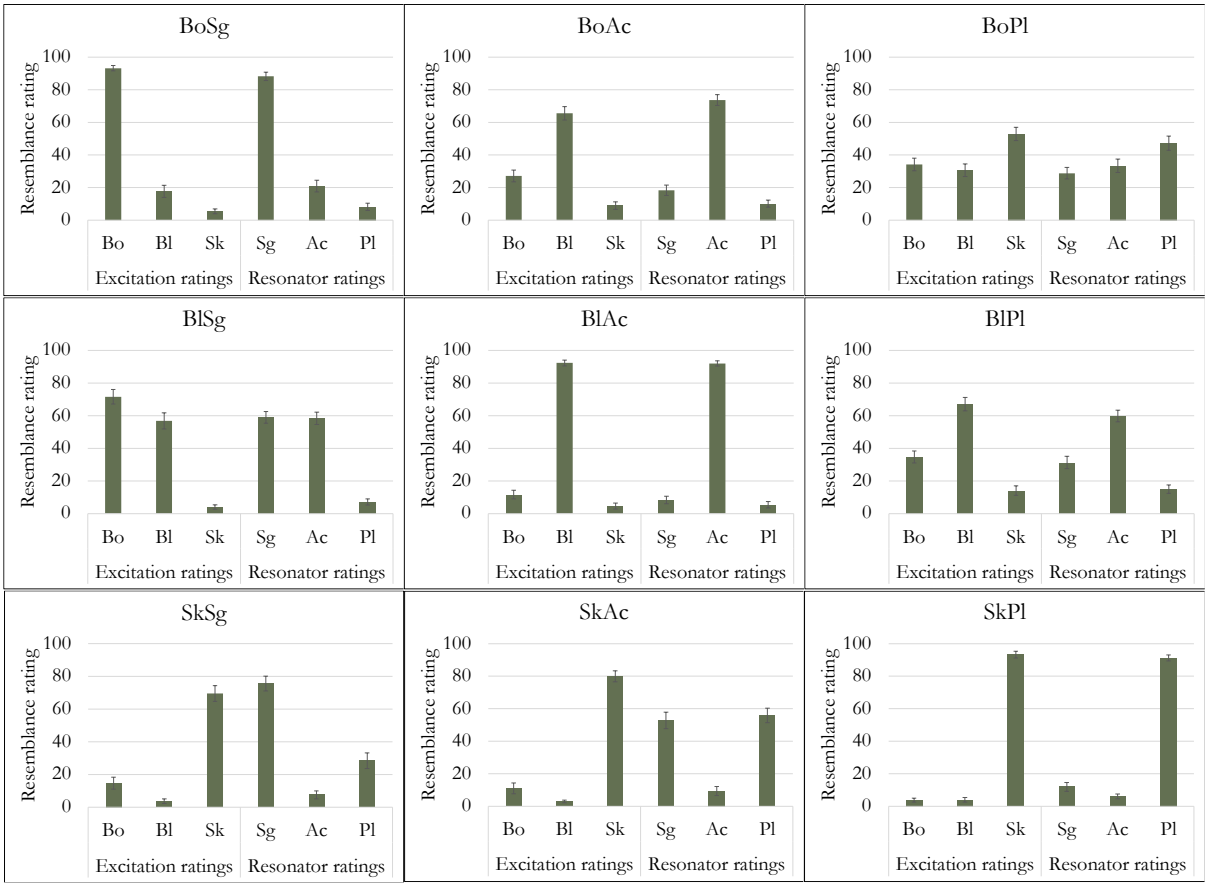

Figure 1.2 Mean bowing (Bo), blowing (Bl), striking (Sk), string (Sg), air column (Ac), and plate (Pl) resemblance ratings for each type of excitation-resonator interaction. Error bars represent the standard error of the mean.

For the atypical interactions—bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc)—resemblance ratings indicated that the listeners confused some excitations and resonators for others. Participants rated BoAc and BlPl as resembling blowing and the air column. For BoAc, listeners thought they resembled the air column, so their perception of the excitation was assimilated to something that is typically applied to the air column. The reverse was apparent for BlPl: listeners thought they resembled blowing, so they perceived the resonator as assimilating to an air column. Furthermore, when a sustained excitation such as blowing is applied to a plate, the harmonic content is nearly harmonic, which might also explain the confusion if listeners associate plates with inharmonic content. For BoPl, the resemblance ratings were a bit

higher for striking than the other excitations and for the plate than the other resonators. BoPl had an artifact at the beginning of the sound that resulted from the bow making contact with the plate. This is because almost all of the plate's modes will be initially excited when the bowing starts. The bowing excitation might have been vague to listeners, so the artifact became more salient in their resemblance ratings. For BlSg, listeners' ratings for bowing were a bit higher than for blowing, but the mean resemblance ratings for the string and air column were nearly equal. Confusing blowing for bowing was expected, as they are both continuous excitations. Moreover, physical models for bowing and blowing have been assumed to be interchangeable (Ollivier et al., 2004). Perceived resemblance of BlSg to the string might have depended on whether listeners thought the sounds resembled bowing; likewise, resemblance ratings to the air column might have depended on the perceived resemblance to blowing. However, the string and air column should differ in their harmonic content, so it is interesting that listeners were sometimes unable to distinguish the string with even and odd harmonics from the air column with primarily an odd harmonic timbre. SkAc were rated as resembling striking more than the other excitations, but listeners thought they resembled the string and plate more than the air column. Both the string and plate are involved in typical interactions with striking, which explains these ratings. These results generally demonstrated that listeners were unable to isolate the excitations and resonators from one another outside of their typical interactions. So, excitations and resonators seem to be closely related in the mental models of sound sources.

1.6.3 Framework for the Research Design

Figure 1.1 was introduced as a pathway to describe the process of sound source recognition. It was originally presented by Li et al. (1991), who stated the importance of considering all the pairwise relationships between the three stages for a complete research design. The relationship between the physical and acoustical levels can be studied by analyzing acoustic waveforms. In doing so, the acoustic properties that reliably differentiate the sounds based on their mechanical components can be determined. For example, we might find an increase in bowing pressure of a violin tone generates an acoustic signal with greater energy concentrations towards the upper harmonics (i.e., an increase in spectral centroid, or a brighter sound). Consequently, the relationship between the physical and acoustical levels is causal because changes in the mechanical components undoubtedly generate changes in the acoustic properties of the resulting sounds (Giordano, 2005). Although the physical-acoustical relationship will not be directly studied in this dissertation, it will be indirectly discussed in

Chapter III in the determination of the acoustic properties (that arise from simulated mechanical interactions) correlating with the salient dimensions that perceptually represent the stimulus set.

The relationship between the physical and perceptual levels are often studied by using categorization tasks. The task is straightforward, such that listeners are presented with sounds and asked to identify the instruments or mechanical components that produced them. I argue that this relationship must also take into consideration the acoustical level, given that the acoustic properties can communicate what mechanical components were involved in producing the sound. A comparison of how categorization without corrective feedback (Chapter II) and with corrective feedback (Chapter IV) influences the physical-perceptual relationship will be of interest to this dissertation.

Lastly, dissimilarity-rating tasks can be used to analyze the relationship between the acoustical and perceptual levels. Participants are presented with pairs of sounds within a stimulus set and asked to rate the dissimilarity of the pairs. Participants are usually not confined to rate the dissimilarity of sounds based on any criteria; instead, they can base their ratings on anything that varies among the sounds' timbres. The ratings are then analyzed with multidimensional scaling (MDS) to determine the underlying continuous, perceptual dimensions that best represent the sounds within the stimulus set. These dimensions can then be correlated with the acoustic properties that best explain a proportion of the variance among the dimensions. Investigating the acoustical-perceptual relationship will be the main focus of Chapter III.

1.6.4 Research Questions and Methods

Chapter II of this dissertation is based on an experiment that tested the categorization of the nine excitation-resonator interactions based on their excitations or resonators in separate tasks. One of the main purposes of this study was to see whether categorization performance reflected the resemblance ratings of the previous experiments (Huynh, 2019). A few analyses were performed on the data. First, two binomial logistic regressions with mixed effects modeling were computed on the correct response data for both excitation categorization and resonator categorization. We aimed to guard against Type I errors (Schielzeth & Forstmeier, 2009) by controlling for random effects (Barr et al., 2013). We also examined the percent response of choosing each excitation and resonator category for a given interaction. This will show which excitations or resonators are confused for one another within each interaction type. Hierarchical cluster analyses were used to determine the groupings of the interactions based on excitation categorization and resonator categorization. Consequently, we can determine whether the interactions cluster differently depending on the categorization task. The main research

question of this experiment was: How do the sounds produced by atypical interactions fit with the mental models of sound sources? We hypothesized that atypical interactions would be assimilated to typical interactions because the mental models of typical interactions already exist. The mental models for typical interactions are quite restrictive, which makes it difficult to perceive excitations and resonators independently of one another outside of their typical contexts.

The experiment presented in Chapter III involved rating pairs of stimuli based on their perceived dissimilarity. We were interested in the perceptual organization of sounds produced by typical and atypical interactions. Another aim was to determine the number of salient dimensions that might be required to differentiate the sounds. Consequently, we performed MDS using the SMACOF algorithm (Scaling by MAjorizing a COmplicated Function; first developed by de Leeuw & Mair, 2009). SMACOF computes the most optimal multidimensional configuration by implementing a function that reduces stress (i.e., error). The number of dimensions is determined by a cross-validation technique that compares the average R^2 and AIC (Akaike Information Criterion) for configurations with one to ten dimensions. We were also curious about potential acoustic properties or combinations of them that could best explain a proportion of variance along each revealed dimension, and whether they could be interpreted in terms of the structural or transformational invariants that differentiate the resonators or excitations, respectively. The audio descriptors of each sound were analyzed with the Timbre Toolbox Version R1.0 (Kazazis et al., 2022; originally developed by Peeters et al., 2011). A subset of the audio descriptors was regressed onto the positions of each dimension of the timbre space. The subset of audio descriptors was idiosyncratic to each dimension and chosen using model selection techniques in an exploratory approach. Model selection for each dimension involved a backward stepwise regression to obtain a model with the lowest BIC (Bayesian Information Criterion). In line with previous MDS studies on timbre perception, we predicted that the generated timbre space would have two or three dimensions (Grey, 1977; Hjortkjær & McAdams, 2016; Iverson & Krumhansl, 1993; Lakatos, 2000; Marozeau et al., 2003; McAdams et al., 1995). Furthermore, we expected that two of the dimensions would be correlated with an audio descriptor distinguishing continuous from impulsive excitations (e.g., log attack time, temporal centroid) and spectral centroid, as these acoustic correlates are consistent with previous studies. The research question of interest was: Can the excitations and resonators of musical instruments be perceived independently of one another beyond their typical interactions? If the different excitations and different resonators can be differentiated on separate dimensions, then this would suggest that they can be perceived independently. Furthermore, the acoustic correlates of the dimension(s) differentiating the excitations

would be confirmed as transformational invariants, and the acoustic correlates of the dimension(s) differentiating the resonators would be confirmed as structural invariants. However, an alternative finding might be that the atypical interactions appear closer to the typical interactions to which they were assimilated during the categorization tasks of Chapter II. If this were the case, it would support categorization theories claiming that perceived similarity predicts categorization (Goldstone, 1994).

In Chapter IV, we tested whether the atypical interactions can be learned based on their excitations, resonators, or interactions in three separate learning tasks. The procedure of each learning task was inspired by McAdams et al. (2023). Each learning task implemented supervised category training which comprised trial and error with corrective feedback. Successful completion of the supervised learning task meant that listeners reached a passing threshold by identifying the correct category of at least 75% of the sounds for a certain number of training blocks. Then, listeners were tested on their ability to categorize the sounds without corrective feedback. The correct response data of categorizing excitations, resonators, or interactions—when corrective feedback was removed—were analyzed in three separate binomial logistic regressions with generalized mixed effects modeling. The percent response of choosing each excitation, resonator, or interaction category was also examined for each type of interaction to see if any confusions were made. The main research question concerned whether listeners could learn the excitation, resonator, and/or interaction categories of the atypical interactions with supervised learning. If learning is not possible, we expected categorization performance to resemble that of the experiment in Chapter II. If learning is possible and categorization performance improves, then listeners might be able to detect the structural and/or transformational invariants allowing them to reliably categorize the different excitations, resonators, and interactions. This would imply that the structural and transformational invariants can be detected independently of one another, such that listeners can perceive one as constant despite variation in the other. This would also imply that new mental models are formed for the atypical interactions.

Finally, a summary and discussion of all the findings will be presented in Chapter V. It will additionally discuss the relationship between the findings of each experiment as well as the roles of perceived mechanical plausibility and selective attention during each task. The main limitations, future directions, and additional questions are addressed, with a conclusion on the role of timbre in the recognition of sound sources.

Chapter II

Categorization of typical and atypical combinations of excitations and resonators of musical instruments: Assimilation of the unusual to the familiar

This chapter is based on the following research article:

Huynh, E. Y., and McAdams, S. (2023). Categorization of typical and atypical combinations of excitations and resonators of musical instruments: Assimilation of the unusual to the familiar. Manuscript submitted to *Music & Science*.

Abstract. Sound categorization is automatic, yet very little is known about how this process works. Physical sound sources such as musical instruments generate sounds that carry timbral information about two mechanical components: The excitation sets into vibration the resonator, which acts as a filter to amplify, suppress, and radiate sound components. Given that excitation-resonator interactions are quite limited in the physical world, Modalys, a digital, physically inspired modeling platform, was utilized to simulate the combinations of three excitations (bowing, blowing, striking) and three resonators (string, air column, plate). This formed nine types of interactions, which are either typical (e.g., struck string) or atypical (e.g., blown plate). In two separate categorization tasks, participants chose either the excitation or resonator they thought produced each interaction. For the typical interactions, participants accurately categorized their excitations and resonators. Atypical interactions were assimilated to typical ones and listeners identified either the correct excitation or the correct resonator but not both. Hierarchical clustering revealed that interactions were perceived differently depending on the categorization task. These findings suggest that unfamiliar sound sources were interpreted as conforming to familiar sound sources for which mental models exist. These studies consequently highlight the importance of timbre in sound source recognition.

Keywords: Timbre perception, sound source recognition, music perception, categorization, musical instruments

2.1 Introduction

On a daily basis, humans encounter many different sounds from voices to musical instruments. The ability to recognize sound sources is automatic yet impossible without considering the timbres they emit. Timbre is a set of auditory attributes that carry musical qualities and contribute to the recognition of sound sources (McAdams & Goodchild, 2017). It also bears perceptually useful information about the mechanical components of sound sources (Giordano & McAdams, 2010). A large body of literature has investigated how timbre perception is associated with the acoustical features of a sound wave (Caclin et al., 2005; Grey & Gordon, 1978; Lakatos, 2000; McAdams et al., 1995). These acoustical features, however, do not arise spontaneously: they originate from sound sources, which have mechanical components. A common misunderstanding is that mechanical components are secondary to acoustical features with respect to sound source identification, when it is in fact the primary cause of natural sound-producing events, such as those arising from musical

instruments (Giordano & McAdams, 2010). The two mechanical components of acoustic musical instruments of interest to the current study are the excitation mechanism and resonant structure, which interact to create sound.

2.1.1 Musical Instrument Identification

Timbre has been defined as the "attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are disimilar [sic]" by the American National Standards Institute (ANSI S1.1-1994, p. 34). This definition is misleading, tells us very little about timbre, and only describes it as what it is not, rather than what it is. McAdams (2013) explained that timbre perception is composed of two characteristics. The first is that timbre carries a plethora of auditory attributes that either continuously vary over the duration of a sound (e.g., brightness, attack sharpness, nasality, etc.) or are categorical (i.e., characteristic of a specific sound). The second characteristic describes timbre as the driving force of sound source identification such that sounding objects can be classified into categories.

As sound source identification is an automatic process, listeners have the tendency to assign a label to everyday sounds, including those produced by musical instruments. However, instruments tend to be confused for one another if they belong in the same family. Some instruments are more easily identifiable than others (Saldanha & Corso, 1964). Berger (1964) found that listeners more easily identified the oboe, clarinet, cornet, and tenor saxophone compared to the flute, trumpet, alto saxophone, bassoon, French horn, and baritone. Musical instrument identification also depends on register: nonmusicians judged tones separated in pitch by more than an octave as originating from different instruments even when they were played by the same instrument (Handel & Erickson, 2001, 2004; McAdams et al., 2023). Additionally, Elliott (1975) found that musical instrument identification was more difficult when the attacks and releases of recorded sounds were removed. Identification performance for unaltered recordings was much better, except for the cello, which was commonly mistaken for the violin.

2.1.2 Identification of Impacted Materials

Although the primary focus of the current research will be on musical tones, the majority of previous studies have focused on impacted materials (i.e., mostly nonmusical sounds, except McAdams et al., 2004, which used synthesized xylophone bars). Many studies have examined listeners' abilities to identify the material of an impacted object (Aramaki et al., 2009; Giordano & McAdams,

2006; Klatzky et al., 2000; Lutfi & Oh, 1997; McAdams et al., 2004; McAdams et al., 2010); fewer studies have examined the role of multiple actions such as bouncing (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012; Warren & Verbrugge, 1984). These studies are important to consider, given that actions are synonymous with excitation mechanisms, and materials are an aspect of resonant structures. Listeners are generally good at differentiating between the materials of impacted objects across gross categories of metal-glass and plexiglass-wood even across different sizes of objects (Giordano & McAdams, 2006). Distinction within gross categories of materials becomes confusing (Lemaitre & Heller, 2012). McAdams et al. (2010), however, examined material distinctions within the gross category of metal-glass with dissimilarity ratings and a categorization task. The researchers simulated impacted plates that varied in two material properties: wave velocity (related to elasticity and mass density) and damping properties (thermoelastic and viscoelastic damping). Because of the interpolations between the two damping models, the plates represented a continuum of materials between aluminum and glass. Dissimilarity ratings depended on the information corresponding to wave velocity and damping, but categorization performance depended on damping properties alone. Density and elasticity change the modal frequencies, which can also vary with object size. So, the modal frequencies were not considered a reliable cue for material categorization. Therefore, listeners based their judgments on acoustical features that were relevant to the perceptual task (McAdams et al., 2010). Less common is research on how sound-producing actions affect timbre perception, even though listeners are likely more sensitive to the actions rather than the materials or objects that produce a sound (Lemaitre & Heller, 2012; McAdams, 2019). One exception includes a study by Warren and Verbrugge (1984), which found that listeners could accurately distinguish between breaking and bouncing glass. A second exception includes a study demonstrating that listeners can distinguish between different speeds of rolling balls, but determining the speed depended on the size of the ball (Houben et al., 2004).

More recently, research has focused on sounds that are produced by combining different actions and materials. In Lemaitre and Heller's (2012) study, four different actions (scraping, rolling, hitting, and bouncing) were applied to cylinders made of four different materials (wood, plastic, metal, and glass). In one experiment, listeners rated sounds based on how well they conveyed different materials and actions. The ratings demonstrated that listeners confused materials within gross categories of wood-plastic or metal-glass, but not between gross categories. On the other hand, listeners consistently rated the resemblance of the actions correctly regardless of gross material category membership (i.e., sustained versus discrete actions). The second experiment was an identification task,

which also measured reaction times. Action identification was faster and more accurate than material identification. These findings demonstrate the informational value of the mechanical components of sound sources and a greater sensitivity to actions than materials.

Hjortkjær and McAdams (2016) employed a similar stimulus design as that of Lemaitre and Heller (2012). The stimuli combined three actions (drop, rattle, and strike) with plates made of three materials (wood, metal, and glass). Dissimilarity ratings of the stimuli were explained by two dimensions, as revealed by multidimensional scaling (Hjortkjær & McAdams, 2016). One dimension separated materials (wood versus metal-glass) and was correlated with changes in spectral centroid (i.e., spectral center of gravity). The second dimension separated the three actions and was correlated with variability in temporal centroid (i.e., centre of gravity of the energy distribution across time). Consistent with Lemaitre and Heller's (2012) findings, the identification of materials was better between than within gross categories (i.e., wood versus metal-glass) and distinctions among actions were clear regardless of gross action category membership (i.e., single versus multiple impacts). However, confusions were made for materials with similar spectral centroids and actions with similar temporal centroids. Hjortkjær and McAdams's (2016) findings highlight the influence of sound source mechanics on acoustical properties, which in turn contribute to the timbres of generated sounds.

2.1.3 Interaction of Excitations and Resonators in Musical Tones

Musical instruments contain interactions between two closely related macroscopic mechanical components: the excitation mechanism and the resonant structure. For example, a clarinet consists of an interaction between blowing (excitation) and an air column (resonator). An interaction can be considered a coupling process, such that an excitation mechanism allows a controllable source of energy into the resonant structure (e.g., string, air column, plate), setting it into vibration (Fletcher, 1999). The excitation mechanisms of interest are the frictional bow, single reed, and hammer mechanisms. Interactions of sustained sounds are nonlinear (McIntyre et al. 1983), meaning that an increase in the input disproportionally increases the output (Fletcher, 1999). Struck strings have been described as nearly linear given that only the initial hammer contact is considered nonlinear. Furthermore, the coupling of struck plates does not correspond to any nonlinearity (Fletcher, 1999).

Excitation-resonator interactions of musical instruments are very specific and perceived as restrictive. This was demonstrated in a recent study that used Modalys (Dudas, 2014), a physically inspired modeling synthesizer, to simulate combinations between three types of excitations (bowing, blowing, striking) and three types of resonators (string, air column, plate), forming nine classes of

excitation-resonator interactions (Huynh, 2019, reports a detailed explanation of the synthesis design). Four of these interactions were typical of acoustic musical instruments (bowed string, blown air column, struck string, struck plate). These interactions are deemed typical as they are frequently encountered. Accordingly, typical interactions are mechanically plausible because they are physically modeled after acoustic musical instruments and listeners can conceptualize these interactions. The remaining five interactions were atypical: bowed air column, bowed plate, blown string, blown plate, struck air column. Atypical interactions can be described by their familiarity and mechanical plausibility as well. Bowed plates and struck air columns might be more familiar to musicians than nonmusicians, considering musicians' exposure to extended playing techniques of musical instruments (e.g., a bowed xylophone bar or a slap tongue on the clarinet or saxophone). So, musicians might be more likely to say the bowed plate and struck air column are mechanically plausible in comparison to nonmusicians. However, the five atypical interactions are physically impossible in terms of their synthesis in Modalys. Modalys synthesizes the bowed plate such that the bow passes through the plate to excite it, whereas a plate would have to be bowed at its edge for the interaction to be mechanically plausible and physically possible. A slap tongue technique is normally applied at the mouthpiece of a clarinet or saxophone, but Modalys applies the striking excitation somewhere along the length of the air column (which is not connected to a mouthpiece). The other three atypical interactions (bowed air column, blown string, and blown plate) are even less likely to be considered familiar and mechanically plausible, and they are also physically impossible.

In two experiments by Huynh (2019), listeners rated the extent to which exemplars of the nine interactions resembled the three excitations (Experiment 1) or resonators (Experiment 2). For the stimuli produced by typical interactions, listeners assigned higher resemblance ratings for both the excitations and resonators that actually produced them. However, for the atypical interactions, listeners' ratings reflected confusions among different excitations or resonators. For example, listeners assigned higher blowing and air column ratings for the bowed air column and blown plate. In general, listeners had difficulty isolating the excitations and resonators from one another for the atypical interactions. These findings suggest that these two mechanical components are closely related in their mental models of sound sources.

2.1.4 A Mental Model of Musical Sound Sources

McAdams and Goodchild (2017) argued that listeners form mental models of sound sources, even when their timbral characteristics vary with changes in pitch, dynamics, and other parameters. A

mental model is an internal representation of how a system (such as a musical instrument) behaves in the world. Mental models are acquired through exposure and shaped by learning, such that one acquires an understanding of the features of a sound source by becoming familiar with it. For example, musicians interact with their instruments daily, allowing them to understand the techniques and restrictions of sound production. Nonmusician listeners can generally understand how a musical instrument is played through passive exposure, but the specific techniques or restrictions of sound production may not be as well understood. Researchers have found that listeners categorize musical sound sources very quickly, suggesting evidence for mental models of musical sound sources (Agus et al., 2012). It is important to note that the stimulus set of the current study will be exploring excitation-resonator interactions outside their typical contexts. Listeners have likely developed the mental models for typically combined excitations and resonators over frequent exposure. However, it is possible that the atypical interactions are too complex for listeners to form new mental models for them. This was demonstrated in previous experiments by Huynh (2019), during which listeners confused certain excitations or resonators of the atypical interactions for one another.

2.1.5 The Current Study

Previous studies have highlighted the role of sound source mechanics in timbre perception of mostly nonmusical sounds (Aramaki et al., 2009; Giordano & McAdams, 2006; Hjortkjær & McAdams, 2016; Klatzky et al., 2000; Lemaitre & Heller, 2012; Lutfi & Oh, 1997; McAdams et al., 2004; McAdams et al., 2010; Warren & Verbrugge, 1984). A recent study, however, examined how listeners perceived sounds produced by typical and atypical excitation-resonator interactions that were synthesized with physically inspired modeling (Huynh, 2019). Participants rated how well the stimuli resembled three excitations and three resonators on a continuous scale from "not at all" to "completely". This rating indicated whether they thought the stimuli resembled excitations and resonators that did or did not produce them. However, a categorization task might be more direct in determining whether and how the atypical interactions fit into listeners' mental models of sound sources. The task of the current study involves categorizing the nine digital combinations of three excitations and three resonators in Huynh (2019) based on either their excitations or their resonators. Categorization is "doing the right thing with the right kind of thing" (Harnad, 2017, p. 22). The right kind of thing means sorting items into groups and assigning them names. A category is defined by the features that members of the same category possess (Harnad, 2017). Members of different categories therefore possess different features. Features are sensory properties of the instances of categories. The

nine interactions in this experiment each comprise two features: an excitation and a resonator. Pérez-Gay et al. (2017) note that features can potentially be categories with respect to which members and non-members are defined by higher-order features, such as the acoustic properties of the sounds produced by the interactions. Consequently, the different excitations and resonators will hereafter be referred to as categories.

Four hypotheses are proposed for the current study. First, the easiest excitation and resonator to identify are striking and the string, respectively. Striking is the only impulsive excitation involved in the stimuli of the current study, and impulsive excitations are known to be easily distinguished from sustained ones (Giordano & McAdams, 2010; Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012). The string is the only resonator involved in two typical interactions as it is typically struck or bowed. The air column and plate are each associated with one typical interaction, which will make them harder to identify across all interactions. The second hypothesis proposes that excitations and resonators might not be processed independently of one another in the way that actions and materials are when dealing with impacted materials. Mechanical plausibility and familiarity through exposure play a large role in the perception of natural sound-producing events. The sounds produced by impacted materials are mechanically plausible, meaning listeners can easily interpret their sound production without knowing the exact actions and materials involved. The atypical interactions in the current study are physically impossible in the way that Modalys synthesizes them, so their sound production will be ambiguous to most listeners. This leads to the third hypothesis: the perception of atypical excitation-resonator interactions will conform to existing mental models of sound sources by assimilating them to typical interactions. This prediction would align with the findings from the previous experiments by Huynh (2019) by using a categorization task rather than a resemblance rating task. For example, the bowed air column and blown plate might be categorized as being produced by blowing an air column, since these atypical interactions were previously associated with higher blowing and air column resemblance ratings. Furthermore, assimilation to typical interactions indicates that listeners seek invariant features between an ambiguous sound source and one for which a mental model exists. Finally, the fourth hypothesis is that atypical interactions will be categorized and perceived differently depending on whether listeners' attention is directed to the excitations or resonators. For example, the bowed plate might assimilate to mental models of either: (1) the bowed string when sounds are categorized based on their excitations; or (2) the struck plate when categorization is based on resonators. As it will be difficult for listeners to conceptualize the sound production of the atypical

interactions, they might in turn interpret them as something that is known or familiar. This interpretation, however, will depend on what their focus is directed to.

Thus, this paper aims to investigate how sounds produced by atypical excitation-resonator interactions are incorporated into mental models of sound sources by comparing their perception to the typical interactions. The current study tests whether listeners can detect the invariant features among the different excitations or resonators regardless of the complementary mechanical property they are applied to; or if their interpretations of the stimuli conform to what listeners are already familiar with. The experiment involves a two-part categorization task. Participants listened to each sound in the stimulus set and categorized them based on their excitations in one part and resonators in another part. For the purpose of the analyses, the responses will be analyzed in terms of categorization accuracy. The confusions that listeners make will also be analyzed to speculate which typical interactions the atypical ones generally conform to and whether it depends on excitation or resonator categorization.

2.2 Method

2.2.1 Participants

Forty-seven participants (35 females, 11 males, 1 self-identified as "other") were recruited from either a mailing list or web-based advertisement certified by McGill University. They reported having normal hearing, which was confirmed by a pure-tone audiometric test with octave-spaced frequencies from 125 to 8,000 Hz at a hearing threshold of 20 dB HL relative to a standardized hearing threshold (ISO 398-8, 2004; International Organization for Standardization, 2004; Martin & Champlin, 2000). They had an average age of 22.8 (*SD*=2.7) and a range of 0 to 21 years of formal musical training (Mean=10.2, *SD*=6.8). All participants provided informed consent and were compensated for their time with cash or course credit. This study was certified for ethical compliance by the McGill University Research Ethics Board II.

2.2.2 Apparatus

Listeners completed the audiometric test and main experiment in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). The experiment ran on a Mac Pro computer running OSX (Apple Computer, Inc., Cupertino). The stimuli were presented over Seinnheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany) and were amplified

through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA). The physical levels of the sounds were measured by coupling the headphones to a Bruel and Kjær Type 4153 Artificial Ear connected to a Type 2205 sound-level meter (A-weighting) (Bruel & Kjær, Nærum, Denmark). The sounds varied in level from 57.8 to 72.8 dB SPL. The experimental task was programmed in the PsiExp computer environment (Smith, 1995).

2.2.3 Stimuli

A digital, physically inspired modeling synthesizer, Modalys, was developed by The Musical Acoustics Team at the Institut de recherche et coordination acoustique/musique (IRCAM) in Paris, France. Modalys has the advantage of simulating different excitations and resonators without the resulting sounds necessarily being perceived as an existing musical instrument (Dudas, 2014; Eckel et al., 1995). It implements modal synthesis to isolate parameters of an excitation and a resonator and predicts the acoustical outcome between their interaction.

Nine classes of interactions between three excitations (bowing, blowing, striking) and three resonators (string, air column, plate) were simulated. Four of these interactions are considered typical as they can be produced in the physical world: bowed string (BoSg), blown air column (BlAc), struck string (SkSg), and struck plate (SkPl). The remaining five interactions—bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc)—either cannot be physically produced and are mechanically implausible (BoAc, BlPl, BlSg) or are rarely encountered by listeners (BoPl, SkAc), so they are considered atypical interactions. Given that the atypical interactions are rarely or never encountered in everyday music listening, it was difficult to anticipate how they would sound. Moreover, physically inspired modeling of atypical interactions is quite uncommon (for notable exceptions, see Böttcher et al., 2007 for musical sounds; and Conan et al., 2014, for continuously excited objects). Consequently, an exhaustive approach was implemented to synthesize 400 versions of each excitation-resonator interaction type, such that a variety of timbres could be chosen for each interaction type. The stimulus design was described in detail in Huynh (2019), but a summary will be described here.

To maintain experimental control that is necessary for the experiments, physical parameters pertaining to each resonator and the temporal envelope of each excitation were kept as consistent as possible. For example, the same string was used for each excitation applied to it and the same temporal envelope for bowing was applied to each resonator. Given that Modalys provides users with default examples of typical interactions upon installation, those of the bowed string, blown air column, and

struck plate were used first. The goal was to isolate the parameters of each excitation from those of each resonator in their typical interactions before simulating them in atypical contexts. In general, Modalys estimates the resonator properties by computing the modes that would be present during vibration. Excitation properties were estimated by solving a time equation that predicts the temporal evolution corresponding to its movement.

The bowing excitation was primarily made up of two temporal envelopes; one controlled the bow speed and the other controlled the bow pressure. Variations in these two parameters are known to produce considerable changes in timbre: loudness is mainly controlled by the bow speed and brightness is mainly controlled by the bow pressure (Halmrast et al., 2010). Minor adjustments were made to these temporal envelopes during their interaction with the string to make the bowing sound as realistic as possible. Then, the bowing excitation could be applied to the air column and plate (which have also been isolated from their interactions with blowing and striking, respectively). Twenty values of the maximum bow speed were combined with 20 values of the maximum bow pressure. This formed 400 versions for each of the BoSg, BoAc, and BoPl interactions.

Blowing through a model of a clarinet mouthpiece with a single reed was controlled by two temporal envelopes for the breath pressure and embouchure pressure. The embouchure pressure is defined by the pressure of the lips on the reed which subsequently controls the height of the opening of the reed (Coyle et al., 2015). Varying breath pressure and embouchure pressure is known to change the resulting timbre, such that a higher breath pressure and a smaller reed opening generate brighter sounds (Coyle et al., 2015). Adjustments to the temporal envelopes of the breath pressure and embouchure pressure were initially made when they were applied to an air column. Once a realistic blowing sound was achieved, the two temporal envelopes were then applied to the string and plate. Twenty values for both the maximum breath pressure and maximum embouchure pressure were combined with one another, generating 400 versions for each of the BlAc, BlSg, and BlPl interactions.

The striking excitation was involved in two typical interactions. When applied to the string and plate, the temporal envelope of the hammer force was adjusted to generate a realistic sound. Then, the same temporal envelope was applied to the air column. Twenty values of the maximum hammer force were combined with 20 values corresponding to the position at which the sound output was recorded along the length of the string or air column. Consequently, 400 versions of each of the SkSg and SkAc interactions were synthesized. For SkPl, Modalys normalizes the output of the sounds, so changing the hammer force did not generate considerable changes in the resulting timbre. Instead, 20 values of both the horizontal and vertical coordinates at which the sound output was recorded,

emphasizing different vibration modes, were combined in the initial synthesis of SkPl, forming 400 versions of this interaction.

In the final stimulus set, three exemplars out of the 400 versions were selected for each interaction. These 27 exemplars were chosen in the previous study by Huynh (2019), and the same exemplars are used in the current study. The general process of selecting three exemplars first involved noting which of the 400 versions of each interaction produced an audible output. This generated a space of perceptible sounds for the physically inspired models of each interaction. The three exemplars were then informally chosen on the basis that they were among the perceptible sounds and conveyed a variability of the timbres of that interaction. Each sound had a duration of 2 s and a lowest vibration mode of 155 Hz, corresponding to a pitch of E-flat-3.

2.2.4 Procedure

Prior to beginning the experiment, listeners were given the definitions of each excitation and resonator property (Table 2.1). Excitations were defined as types of actions performed to set a resonator into vibration. Resonators were defined as types of objects that vibrate and radiate sound components. After receiving written and verbal instructions of the main experiment, participants completed two practice blocks, one for excitation categorization and one for resonator categorization. The practice blocks were based on three stimuli that were produced by typical interactions: BoSg, BlAc, and SkPl. They were synthesized to have a lowest vibration mode of 220 Hz, corresponding to an A3 pitch. These interactions were chosen as they each represented one of the three excitations and one of the three resonators. Each practice block comprised three trials, one for each relevant excitation or resonator category. The practice blocks were completed with the experimenter so that the participant was able to ask questions about the instructions or interface before beginning the main experiment. The practice blocks followed the same format and paradigm as the experimental blocks.

At the beginning of the experimental blocks, participants listened to the full range of stimuli to get a sense of the variability of the excitations and resonators producing the sounds. The sounds were presented with an inter-onset interval (IOI) of 2 s in a pseudo-random order such that two successive sounds were not produced by the same excitation-resonator interaction (e.g., a BoSg was not presented after another BoSg). The experiment was divided into two blocks, each concerning either excitation or resonator categorization of each stimulus. The order of the two blocks was counterbalanced across participants. In each block, there were 27 trials (one for each stimulus), giving a total of 54 trials in the entire experiment. In each trial, participants played a stimulus, which they were able hear only once.

Depending on whether the excitations or resonators were being categorized, there were three boxes on the screen representing the names of the three corresponding categories. The positions of the three boxes were randomized for each participant. Participants were instructed to click the box corresponding to the excitation or resonator they thought produced the sound, depending on the task. We did not provide corrective feedback to participants following their response in each trial. The order of stimulus presentation within each block was pseudo-randomized such that two stimuli produced by the same interaction were not presented successively. Once participants were satisfied with the excitation or resonator they had chosen, they proceeded to the following trial.

Table 2.1 The definitions of each excitation and resonator category.

	Category	Definition
Excitation	Bowing (Bo)	The action of rubbing a bow on an object to make it vibrate.
	Blowing (Bl)	The action of blowing into a mouthpiece to make an object vibrate.
	Striking (Sk)	The action of using a mallet to hit an object to make it vibrate.
Resonator	String (Sg)	An object that is a thin wire fixed at its endpoints.
	Air column (Ac)	An object that is a tube[a], sealed at one end and open at the other.
	Plate (Pl)	An object that is thin, rectangular, flat, and rigid.

[a] The experimenter verbally clarified that the air column refers to air molecules within the tube rather than the tube itself.

2.3 Results

2.3.1 Categorization Accuracy

The current study examined how well participants categorized the excitations and resonators of the nine types of interactions without knowing how they were produced. Figure 2.1 shows the mean percent correct of excitation and resonator categorization for each interaction type. Participants were best at identifying striking regardless of the type of resonator on which it acted. This was not surprising because striking was the only impulsive excitation whereas there were two types of sustained excitations. Accuracy of bowing and blowing categorization depended on the resonator of the interaction. Listeners were generally better at identifying the string than the air column or plate. The string was more versatile and associated with two typical interactions in the stimulus set (e.g., bowed and struck), but the air column and plate were each associated with one typical interaction. A general pattern of excitation and resonator categorization among the atypical interactions was observed: participants were better at categorizing either the excitation or the resonator but not both. Participants

more often identified the correct excitations than the correct resonators for the struck air column (SkAc) and blown plate (BlPl). On the other hand, the resonators of the bowed air column (BoAc), bowed plate (BoPl), and blown string (BlSg) were more correctly categorized than their excitations.

To analyze the correct response data for both categorization tasks, a binomial logistic regression using generalized linear mixed effects modeling was computed. The type of excitation, type of resonator, and their interaction were included as fixed effects in separate models for excitation categorization and resonator categorization. To find the maximal random effects structure justified by the correct responses, the approach proposed by Barr et al. (2013) was implemented. A random intercept for participant and random slopes for each of the within-groups factors of excitation type and resonator type were initially fit onto the regression model. If the model with all random effects resulted in a singular fit, the random slope that either had a very low variance (i.e., close to 0) or was highly correlated with another random slope was removed. Random slopes that contributed to the singular fit were removed one by one until the model no longer had a singular fit. According to Schielzeth and Forstmeier (2009), this technique guards against high Type I errors that intercept-only mixed effects models are prone to. For the correct response of excitation categorization, the random slopes for the resonator categories were dropped from the model, whereas none of the random slopes were dropped for the correct response of resonator categorization.

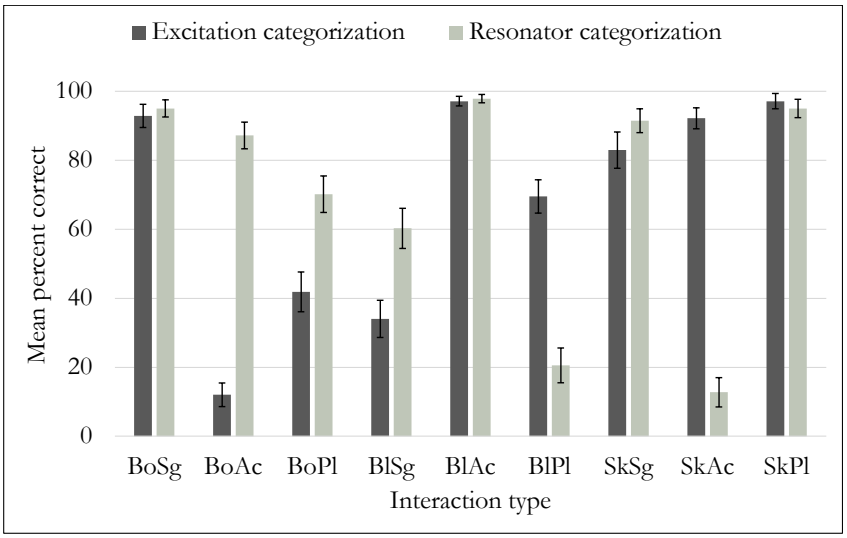

Figure 2.1 Mean percent correct of categorizing the excitations and resonators of each interaction type. Error bars represent standard error of the mean. Bo = bowing, Bl = blowing, Sk = striking, Sg = string, Ac = air column, Pl = plate.

Selected fixed effects of excitation categorization correct responses are reported in Table 2.2. By running the same model and changing the reference excitation and resonator categories, log odds ratios (i.e., the regression coefficients) were computed for selected fixed effects. These fixed effects were chosen because they directly compared the effects of the different resonators on the correct categorization of each excitation. Odds ratios were calculated by exponentiating the corresponding regression coefficients. So, when the reference excitation and resonator were bowing (Bo) and the air column (Ac), respectively, the fixed effect of the string (Sg)—meaning that the comparison is between BoSg relative to BoAc—yields a regression coefficient or log odds ratio of 5.56. Exponentiating this value generates an odds ratio of 294.04, which means the odds of correctly categorizing bowing were 259.04 times greater when stimuli were produced by BoSg than when they were produced by BoAc. Generally, an odds ratio greater than 1 indicates that the odds of correctly categorizing the excitation of the comparison interaction were greater than those of the reference interaction; an odds ratio between 0 and 1 means that the odds of correct categorization were less for the comparison interaction than for the reference. As expected, participants also had greater odds of correctly identifying bowing for BoSg than for BoPl. So, the odds of correctly categorizing bowing were Sg > Pl > Ac (as confirmed by Figure 2.1). Among the blown (Bl) stimuli, the odds of correctly categorizing the excitation from greatest to smallest were Ac > Pl > Sg. Because BlAc is a typical interaction, it was expected that it would be more correctly categorized than BlPl and BlSg. Although the odds ratios imply that there were significant differences in the odds of correctly categorizing striking (Sk) among the different resonators (i.e., Pl > Ac > Sg), the mean percent correct scores show that listeners categorized striking accurately overall (Figure 2.1).

Selected fixed effects of the different types of excitations on the correct categorization of the resonators are shown in Table 2.3. These selected effects were obtained by running the same regression model and changing the reference excitation and resonator categories. The odds ratios generally illustrate that the odds of correctly categorizing the resonator were greater when sounds were produced by typical interactions than atypical ones. For example, the odds of correctly categorizing the string were greater when it was bowed or struck than when it was blown. The regression coefficient for the fixed effect of bowing (Bo) relative to SkSg was non-significant, meaning that there were no differences in the odds of accurately choosing the string (Sg) between BoSg and SkSg. As mentioned, these are both typical interactions, so resonator categorization was quite accurate for them (Figure 2.1). For air column (Ac) categorization, the odds of correct categorization were Bl > Bo > Sk. As expected, the air column was categorized most accurately when it was blown because it is a typical

interaction. But the air column was also identified quite accurately when it was bowed (Figure 1). Among the sounds produced by the plate (Pl), the odds of its correct categorization from greatest to smallest were Sk > Bo > Bl. So, these patterns of results support the hypothesis that resonator categorization would be more accurate for typical than atypical interactions.

Table 2.2 Selected fixed effects (β) and the corresponding odds ratios of correctly categorizing a type of excitation between stimuli produced by different resonators.

	Reference	Comparison	β	SE	p	Odds ratio
Bowing categorization	BoAc	Bo**Sg**	5.56	0.56	<0.001	259.04
	BoPl	Bo**Sg**	3.54	0.46	<0.001	34.61
	BoPl	Bo**Ac**	–2.01	0.36	<0.001	0.13
Blowing categorization	BlSg	Bl**Ac**	4.85	0.60	<0.001	128.23
	BlPl	Bl**Ac**	2.99	0.56	<0.001	19.91
	BlPl	Bl**Sg**	–1.86	0.31	<0.001	0.16
Striking categorization	SkSg	Sk**Pl**	3.61	0.81	<0.001	36.81
	SkAc	Sk**Pl**	1.63	0.74	0.027	5.13
	SkAc	Sk**Sg**	–1.97	0.64	0.002	0.14

Note. The Reference column indicates the reference excitation and resonator categories. Fixed effects are in boldface in the Comparison column.

Table 2.3 Selected fixed effects (β) and the corresponding odds ratios of correctly categorizing a type of resonator between stimuli produced by different excitations.

	Reference	Comparison	β	SE	p	Odds ratio
String categorization	SkSg	**Bo**Sg	–0.10	1.02	0.918	0.90
	BlSg	**Bo**Sg	4.04	0.75	<0.001	56.95
	BlSg	**Sk**Sg	4.15	0.92	<0.001	63.20
Air column categorization	BoAc	**Bl**Ac	2.62	0.83	0.002	13.73
	SkAc	**Bl**Ac	9.60	1.49	<0.001	14767.83
	SkAc	**Bo**Ac	6.98	1.24	<0.001	1075.62
Plate categorization	BoPl	**Sk**Pl	4.11	0.92	<0.001	61.09
	BlPl	**Sk**Pl	7.24	1.01	<0.001	1399.4
	BlPl	**Bo**Pl	3.13	0.47	<0.001	22.91

Note. The Reference column indicates the reference excitation and resonator categories. Fixed effects are in boldface in the Comparison column.

In Tables 2.2 and 2.3, some of the odds ratios comparing stimuli produced by atypical interactions reflect large differences in how accurately participants categorized their excitations and resonators. Furthermore, the mean percent correct scores of Figure 2.1 reveal that participants were more accurate at categorizing either the excitation or the resonator but not both among the atypical interactions. They also seemed to vary in how well they categorized the mechanical components, depending on the interaction. For example, even though resonator categorization was better than excitation categorization for BoAc, BlSg, and BoPl, participants were much better at identifying the air column in BoAc sounds than the string in BlSg and the plate in BoPl sounds. To further dissect the categorization performance, confusion data were analyzed to examine which excitations or resonators were mistaken for one another among the atypical interactions that could not be explained by the percent correct scores and linear mixed effects model analyses alone.

2.3.2 Confusion Analyses

The previous section discussed whether participants categorized the excitations and resonators of each interaction correctly or incorrectly. For the atypical interactions, participants achieved higher percent correct scores for either the excitation or the resonator. If they were better at identifying the excitation, for example, of interest to the current study is why they performed worse for resonator categorization. Did they confuse the resonator that produced the sound for a different one? If so, which one? And what does this confusion imply about the mental models for novel or unfamiliar sound sources? Similar questions for the excitations that were confused were proposed as well. The mean percent response of choosing each excitation and resonator category for each interaction is shown in Figure 2.2. The nine interactions are separated into different graphs and the different excitation and resonator categories are labeled on the horizontal axis. The height of each bar reflects how often the corresponding category was chosen. Black and grey colorations indicate correct and incorrect responses, respectively. For the typical interactions (e.g., BoSg, BlAc, SkSg, and SkPl) participants most often chose the excitation and resonator that were actually involved in the interactions. This aligns with the hypothesis that listeners would not confuse the mechanical components of these interactions for others. Participants were more likely to say BoAc and BlPl were produced by the blowing excitation and air column resonator, suggesting that they assimilated these interactions to BlAc. This was confirmed by computing two hierarchical cluster analyses, one for excitation categorization (Figure 2.3a), and another for resonator categorization (Figure 2.3b). Between-groups linkage was used as the cluster method. As seen in Figure 2.3, BlAc, BoAc, and BlPl

clustered regardless of whether categorization was based on excitations or resonators. A possible explanation for this can be attributed to the modal frequencies of the resonators. The cylindrical air column should produce modes that have more energy at odd harmonic ratios than at even harmonic ratios with respect to the lowest vibrating mode (i.e., fundamental frequency). This might have been detectable for BoAc, explaining its assimilation to BlAc. For BlPl, when a plate is driven by a sustained excitation, such as blowing, the resulting vibrations lock into a periodic pattern. There will be a fundamental frequency that likely corresponds to the frequency of one of the modes, and then the other modes will be nearly harmonically aligned to the fundamental. So, BlPl was perhaps not categorized as a plate because listeners expected it to have inharmonic rather than harmonic content.

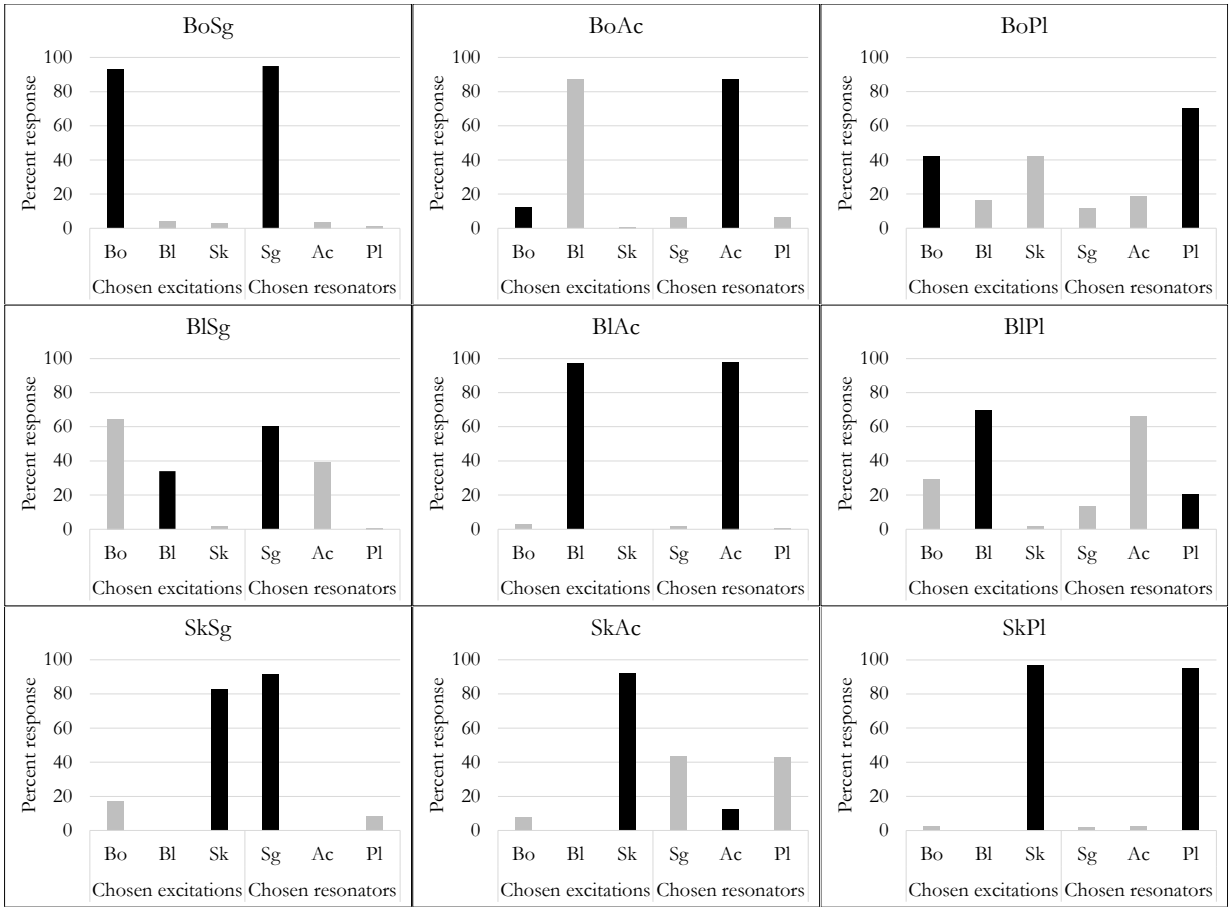

Figure 2.2 Percent response of choosing each excitation and resonator category (horizontal axis) for each of the nine interaction types (separated by different graphs). Black bars represent correct categorization (i.e., percent correct scores).

The remaining atypical interactions were also assimilated to typical ones, but the type of typical interaction they were assimilated to depended on the categorization task. Looking at excitation

categorization only in Figure 2.2, there was a clear assimilation of SkAc to SkSg and SkPl. The hierarchical cluster analysis based on excitation categorization in Figure 2.3a further confirmed this: struck sounds clustered together and were easily distinguished from the sustained sounds. Participants chose bowing more often than blowing for BlSg (Figure 2.2); so, not only did they confuse blowing for bowing, but they assimilated BlSg to BoSg when they were instructed to focus on the excitations. As for BoPl, participants were indecisive between bowing and striking. This confusion may be due the decay times for the modes of the plate being quite long, so it might have been difficult for listeners to differentiate between a sustained and impulsive excitation in this case. The dendrogram for excitation categorization in Figure 2.3a, however, generated a cluster for BoSg, BoPl, and BlSg, suggesting that the latter two interactions were assimilated to the former.

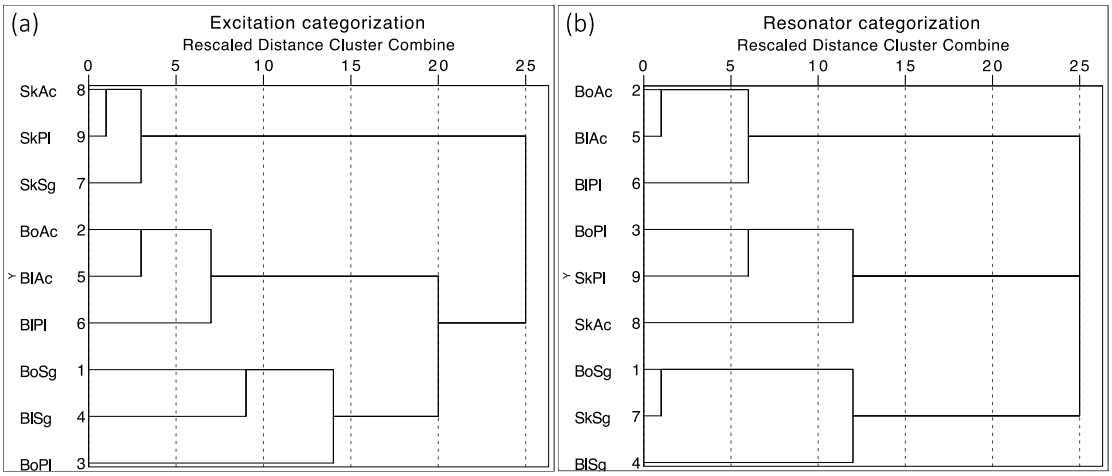

Figure 2.3 Dendrograms showing the clustering of the nine interactions based on (a) excitation categorization and (b) resonator categorization.

Focusing on resonator categorization in Figure 2.2, listeners sometimes confused BlSg for the air column. However, they chose the string more often, generating a percent response pattern that reflects that of BoSg and SkSg. The dendrogram generated from the hierarchical cluster analysis (Figure 2.3b) supports the assimilation of BlSg to BoSg and SkSg based on resonator categorization. Although participants were divided when deciding whether BoPl was bowed or struck, they were more certain that it was produced by a plate (see Figure 2.2). Lastly SkAc was categorized as a string or plate almost equally but was hardly categorized as an air column. The dendrogram in Figure 2.3b, however, revealed that BoPl and SkAc clustered with SkPl. These results suggest that there were more confusions between the air column and plate, but the string was easily discernable from them.

2.4 Discussion

In two categorization tasks, participants chose either the excitation or resonator that they thought produced each of the sounds in the stimulus set. Each sound was produced by one of nine excitation-resonator interactions. Four interactions were typical: bowed string (BoSg), blown air column (BlAc), struck string (SkSg), and struck plate (SkPl). The remaining five interactions were atypical: bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc). For the most part, the results of the current study converge with previous experiments by Huynh (2019). The findings from Huynh's (2019) experiments showed that listeners assigned the highest resemblance ratings to the excitations and resonators that actually produced the typical interactions, but the resemblance ratings for atypical interactions reflected some confusions. Likewise, in the current experiment, listeners were most accurate at identifying the excitations and resonators of the typical interactions, whereas categorization performance for the atypical interactions demonstrated listeners' confusions. The confusions depended on whether excitations or resonators were being categorized.

In line with the first hypothesis, striking and the string were the most accurately categorized excitation and resonator, respectively. This was reflected in previous experiments as well: all struck sounds had the highest striking resemblance ratings and, apart from BlSg, the string sounds had higher string resemblance ratings (Huynh, 2019). In the current study, listeners categorized BlSg as a string more often than as an air column. BoSg and SkSg were most often categorized as the string. As mentioned, these were the two typical interactions involving the string, which increased the chances of its correct identification. Striking was the most correctly categorized excitation because it was the only impulsive one, whereas there were two sustained excitations.

The two sustained excitations, bowing and blowing, were often confused for one another, especially when they were involved in atypical interactions. A possible explanation for this is that the physical models of these two excitation mechanisms have been considered analogous (Ollivier et al., 2004). Two significant parameters contributing to the bowing excitation are the pressure of the bow on the string and the speed of the bow. These parameters are analogous to the embouchure pressure on the mouthpiece (which controls the height of the initial reed opening) and the breath pressure of the blowing excitation (Ollivier et al., 2004). Another explanation of this confusion corresponds to how listeners assimilated them to the mental models of typical interactions. For BoAc and BlPl,

listeners were more decisive in categorizing them as blown over other excitations; this converged well with the resemblance ratings of previous experiments (Huynh, 2019). Listeners of the current study were also a bit more decisive in saying that BlSg sounds were produced by bowing rather than by blowing. For BoPl sounds, listeners were confused between bowing and striking instead of bowing and blowing. If it were the case that half of the time these atypical interactions were categorized as bowing and categorized as blowing the other half of the time, then it is likely that the primary reason for the confusion concerns the potential interchangeability between bowing and blowing as mechanical processes. However, because listeners were more decisive in choosing one excitation over the other, these findings suggest that an assimilation is what is primarily at play, whereas the potential interchangeability between bowing and blowing plays a secondary role. The role of assimilation is more obvious because the atypical interactions that bowing and blowing were involved in conformed to interactions that are familiar and perceived as mechanically plausible.

The explanation of assimilation also supports the second, third, and fourth hypotheses of the current study. The second hypothesis states that an excitation mechanism cannot be teased apart or perceived independently from a resonator it typically interacts with and vice versa. The current findings imply that listeners perceived typical interactions even when they were listening to the atypical ones. In acoustic musical instruments, the perception of an excitation is very much tied to the resonator to which it is applied. Sound production in musical instruments is restrictive in the sense that, even with the many extended techniques one could perform on a stringed instrument, for example, no one thinks to blow a string. The breath would not create enough force on the string to set it into vibration, so it would not create much sound. Due to this mechanical implausibility, it is difficult to conceptualize this interaction along with other atypical interactions. In the third hypothesis, the assimilation of atypical to typical interactions was predicted. This was partly discussed in the case of bowing and blowing confusions, but SkAc is also a notable example. In the current study, SkAc was almost equally categorized as the string and plate and was also previously equally rated as resembling them (Huynh, 2019), suggesting that the listeners assimilated it to SkSg or SkPl. In other words, listeners denied that SkAc was produced by the air column and conformed the atypical interaction to two typical ones. Another clear example was the assimilation of BoAc and BlPl to BlAc. Interestingly, it was different mechanical components to which BoAc and BlPl were assimilated: the excitation of the former was assimilated to blowing and the resonator of the latter was assimilated to an air column. However, these assimilations were consistent across both excitation and resonator categorization. The same cannot be said, however, for the remaining atypical interactions. This aligns

with the fourth hypothesis that the type of categorization task would change how the interactions were perceived, and consequently they could assimilate differently. The hierarchical cluster analyses demonstrated different patterns of clustering depending on the categorization task, with the exception of the BlAc-conforming cluster that included BlAc, BoAc, and BlPl. For excitation categorization, there was a striking cluster featuring all struck sounds and a BoSg-conforming cluster with BoSg, BoPl, and BlSg. On the other hand, for resonator categorization, there was a string cluster (SkSg, BoSg, BlSg) and a SkPl-conforming cluster comprising SkPl, BoPl, and SkAc. Therefore, assimilations of the atypical interactions to typical ones differed across categorization tasks. The assimilations observed in the current experiment were likely formed based on unsupervised learning. It is considered unsupervised because the assimilations were learned through repeated exposure and the detection of co-occurring features (Harnad, 2017) between the unfamiliar, atypical interactions and the familiar, typical ones. This was context-dependent, as listeners sought different co-occurring features depending on whether categorization was based on the excitations or resonators of the stimuli. A possible approach to test the improvement of both types of categorizations and the reduction of assimilations would be a supervised learning task involving trial and error and corrective feedback (Harnad, 2017). This would train listeners to detect the features that are common among sounds produced by the same excitations or resonators, rather than the features shared between atypical interactions and the typical ones to which they conformed. Consequently, it would be of interest to examine the role of supervised learning in the categorization of the atypical interactions.

One question these findings have yet to answer deals with the efficacy of Modalys as a physically inspired modeling synthesizer. With the atypical interactions, we could not anticipate how they would sound prior to their synthesis. Even after their synthesis, we could not conclude whether the different excitations and resonators were accurately conveyed by Modalys because the atypical interactions are physically impossible. It is important to note that BoPl and SkAc can be considered mechanically plausible as they are representative of extended techniques such as bowed xylophone bars and slap tongue on a clarinet or saxophone, respectively. However, the way these two interactions were synthesized in Modalys does not reflect the mentioned extended techniques. Instead, the bow passes through the plate to simulate BoPl, and a hammer strikes the air molecules within the air column at a position along its length rather than at a mouthpiece for SkAc. So, these two atypical interactions along with the other three are not physically possible. Moreover, each excitation-resonator interaction was simulated with very controlled approaches, with manipulations applying to only two parameters of each interaction. Although a performer does not manipulate only two parameters at a time when

they play single notes on their instrument, McAdams and Goodchild (2017) have noted that changing just one parameter, such as the embouchure configuration of a clarinet, can alter the timbre of produced sounds. Additionally, the parameters manipulated in the current stimulus set were among those that are commonly manipulated in the physical modeling of acoustic musical instruments: bow speed and pressure (Halmrast et al., 2010); breath pressure and embouchure pressure (Dalmont et al., 2005); and a variety of options for percussive sounds (Halmrast et al., 2010). Manipulating other types of parameters would be of interest to test the efficacy of Modalys in the synthesis of typical and atypical interactions.

The current study showed that excitation and resonator categorization of nine types of excitation-resonator interactions depended on the familiarity with the interactions and their perceived mechanical plausibility. Contrary to previous studies (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012), excitation categorization was not any better than resonator categorization, but the excitations used in the current study sometimes interacted with resonators atypically. Therefore, the accuracy of excitation and resonator categorization depended on the typicality of the interactions. For the atypical interactions, listeners interpreted them as conforming to an interaction for which they already have the mental models. It would be interesting to see how these findings relate to dissimilarity ratings of the same set of sounds and whether atypical interactions can be learned. Successful learning of the atypical interactions would provide insight on how and whether new mental models are formed for them. Taken together, the current study reveals how timbre perception provides listeners with information about sound source mechanics and how novel and unfamiliar sound sources conform to existing mental models of typical sound sources. These studies highlight the complex role of timbre in a process that is essential to human behavior: identifying the source of a sound.

Chapter III

Implicit differentiation of typically and atypically combined excitations and resonators of musical instruments

This chapter is based on the following research article:

Huynh, E. Y., and McAdams, S. (in preparation). Implicit differentiation of typically and atypically combined excitations and resonators of musical instruments. Manuscript intended for submission to *The Journal of the Acoustical Society of America*.

Abstract. Two mechanical components pertaining to acoustic musical instruments are the excitation mechanism and resonant structure. The former sets into the vibration the latter, which is a filter that radiates, amplifies, and suppresses sound components. A previous study simulated nine types of interactions by combining three excitations (bowing, blowing, striking) with three resonators (string, air column, plate). In the current study, listeners rated the dissimilarity of interaction pairs. Multidimensional scaling (MDS) of the ratings using the SMACOF algorithm revealed a three-dimensional timbre space. Dimension 1 showed a clear boundary between struck and sustained excitations and an ambiguous boundary between bowing and blowing. Its primary acoustic correlate was the temporal centroid. Dimension 2 isolated plates, whereas Dimension 3 further separated strings and air columns. Dimension 2 correlated with a weighted sum of acoustic properties related to the global spectral shape and energy modulation of the signal. Dimension 3 was correlated with a weighted sum of acoustic properties describing spectrum fine structure and variability of the signal's energy. The timbre space reveals that listeners differentiated the excitations and resonators on the basis of acoustic cues. These findings imply that excitations and resonators can be perceived independently of one another.

Keywords: Timbre perception, sound source recognition, music perception, psychoacoustics, acoustics, musical instruments

3.1 Introduction

The primary source of natural sound-producing events comprises the interaction of two mechanical components. First, there is an object made of some material that has its own size and geometry and is free to vibrate once excited in some way. Second is an action that sets the object into vibration. In musical instruments, the action and object are analogous to the excitation mechanism and resonant structure, respectively. Neither can be explained without the other, demonstrating their close relationship. In acoustic musical instruments, resonators are coupled to a controllable energy source, serving as the basis of an excitation-resonator interaction. Sustained sounds are primarily defined by nonlinear couplings (i.e., an increase in the input increases the output, disproportionately) (McIntyre et al., 1983). Struck sounds, on the other hand, are broadly defined by linear couplings, but struck strings are nonlinear only in the initial hammer contact (Fletcher, 1999). Excitation-resonator interactions are restrictive in the physical world, such that bowing can be applied to strings but not air

columns. In fact, interactions between excitations and resonators beyond their typical contexts influence perception of the resulting sounds (Huynh, 2019; Huynh & McAdams, 2023a). Huynh (2019) previously simulated nine interactions between three excitations (bowing, blowing, striking) and three resonators (string, air column, plate) and asked listeners to rate their resemblance to each excitation and resonator category. In another study, listeners categorized the sounds based on their excitations and resonators (Huynh & McAdams, 2023a). In both the resemblance ratings and categorization tasks, listeners perceived the atypical interactions as being produced by typical ones. For example, in the case of the bowed air column and blown plate, they assigned higher resemblance ratings for blowing and the air column, and more often categorized them as being produced by these mechanical components. One interest of the current study is to investigate whether listeners detect commonalities in certain characteristics of the interactions that explain the assimilation of atypical excitation-resonator interactions to typical ones. Also, these detectable characteristics might predict whether it is the excitation or resonator of the atypical interaction that becomes assimilated. Changes in mechanical components can directly modify the timbres of produced sounds (Halmrast et al., 2010), but it cannot be denied that acoustic features also play a role in the differentiation of sounds. Acoustic features vary depending on the type of mechanical components and their changes are detectable by listeners (Kazazis et al., 2021a, 2021b).

Previous studies have primarily focused on how the acoustic properties, or audio descriptors, correlate with timbre perception. These studies involve dissimilarity ratings of pairs of sounds within a stimulus set. Listeners are not told the identities of the sounds and are instructed to rate the pairs based on the differences in their timbres on a scale from "identical" to "very dissimilar". The dissimilarity ratings are then analyzed with multidimensional scaling (MDS) to map them into a perceptual distance model, referred to as a timbre space (McAdams, 1993). Each point on the timbre space represents a sound and the distances between them reflect the sounds' perceived dissimilarities. Timbre spaces computed in these studies generally have two to four perceptual dimensions. Acoustic correlates are then determined to inspect the acoustic properties that explain a proportion of the variance in each dimension (McAdams, 1993).

Regardless of the number of dimensions generated by previous studies, two audio descriptors consistently reported as acoustic correlates are the spectral centroid and (log) attack time (Krimphoff et al., 1994; Lakatos, 2000; McAdams et al., 1995; with a confirmatory study by Caclin et al., 2005). The spectral centroid is the spectral center of gravity and describes whether there are greater energy concentrations towards the lower, middle, or upper partials of a signal. The attack time measures the

duration of the attack portion of the signal and is often associated with determining whether the excitation of a sound is impulsive or sustained. Other studies revealed that one of their dimensions correlated with the distinction between impulsive and sustained sounds even when onsets are removed from the signal (Iverson & Krumhansl, 1993) or across different fundamental frequencies (Marozeau et al., 2003). Another audio descriptor distinguishing impulsive from sustained excitations other than the log attack time is the temporal centroid, which is the center of gravity of the energy envelope (Hjortkjær & McAdams, 2016; Kazazis et al., 2021a). As reported in Caclin et al. (2005), a third dimension, if apparent in the timbre space, has been linked to a variety of audio descriptors, namely spectral flux (McAdams et al., 1995); however, other descriptors, such as spectral spread (Marozeau et al., (2003) and harmonic onset asynchrony (Grey, 1977) have also been found. Caclin et al. (2005) determined acoustic properties pertaining to the fine structure of the spectrum (i.e., attenuation of even relative to odd harmonics), or spectral irregularity, to be of relevance to the dissimilarity ratings of synthetic tones in a confirmatory study. It is important to note that the acoustic correlates, especially of the third dimension, may depend on the types of sounds used. In the mentioned studies, the stimuli comprised musical instrument tones that were recorded (Iverson & Krumhansl, 1993; Lakatos, 2000; Marozeau et al., 2003) or synthesized (Caclin et al., 2005; Grey, 1977; McAdams et al., 1995). Moreover, stimuli either included only sustained excitations (Grey, 1977) or both sustained and impulsive excitations (Iverson & Krumhansl, 1993; Lakatos, 2000; Marozeau et al., 2003; McAdams et al., 1995), and varied across instrument families. Furthermore, one stimulus set used by Hjortkjær & McAdams (2016) included recordings of three actions (dropping, striking, rattling) on objects made of three different materials (wood, metal, glass). Interestingly, regardless of the stimuli used, the two most common acoustic correlates across these experiments were associated with the availability of temporal information (particularly focusing on impulsiveness of the attack) and availability of spectral information (notably spectral centroid). Yet no study has directly associated these dimensions and their acoustic correlates with the changes in the mechanical components of musical instrument tones. Although the log attack time and temporal centroid are known to distinguish impulsive from sustained excitations, there are different types of excitations within these gross categories that have not been directly associated with log attack time, temporal centroid, or other temporal audio descriptors. Additionally, spectral centroid is associated with perceived brightness, but it is unknown if that perceptual quality allows listeners to distinguish between different instrument families or resonant structures. It could also be the case that other acoustic properties are associated with potential differentiations among the excitations or resonators. Determining these audio descriptors as acoustic

correlates is important, but even more so is tying them back to the sound-producing parameters of the signal.

Giordano and McAdams (2010) conducted a review to survey previous dissimilarity-judgment studies dealing with musical instrument tones. Their datasets of interest included those mentioned in the previous paragraph except for Caclin et al. (2005) and Hjortkjær and McAdams (2016), as their stimuli did not deal with musical instrument tones; but they included additional datasets from other studies. The distance-based and regional separation analyses on the previous dissimilarity data generally revealed that tones produced by similar excitation types or instrument families (i.e., sharing similarities in resonant structures) appeared closer together in the MDS space and occupied separate regions from other source types. Distinctions between different types of excitations seemed to be more perceptible than distinctions between instrument families. However, the review by Giordano and McAdams (2010) highlighted an association between the mechanical components of sound sources and their perceptual judgments.

The study by Hjortkjær and McAdams (2016) did not comprise musical instrument sounds, but their stimuli involved direct manipulations of the interactions between three actions and three materials. Listeners rated the dissimilarity of stimulus pairs and a two-dimensional timbre space was revealed by MDS. As mentioned, one of their dimensions was best explained by changes in temporal centroid and the other dimension was correlated with spectral centroid. The researchers further determined that sounds produced by similar actions and materials shared similar temporal and spectral centroids, respectively. Hjortkjær and McAdams's (2016) study is one of the few that directly related the changes in acoustic properties reflected by the perceptual dimensions to mechanical components of the sounds. The current study aims to extend these findings to musical instrument tones.

The main purpose of the current study is to determine the perceptual organization of sounds produced by typical and atypical excitation-resonator interactions. This paper is interested in what qualities in the mechanical components, if any, can be differentiated by listeners. The same stimulus set was used as in recent experiments by Huynh (2019) and Huynh and McAdams (2023a), which simulated interactions between three excitations and three resonators using a physically inspired synthesis algorithm called Modalys (Dudas, 2014). Of the nine classes of interactions, five are atypical: bowed air column, bowed plate, blown string, blown plate, and struck air column. They are considered atypical because they are physically impossible in the way that they are synthesized in Modalys. In the physical world, bowed plates and struck air columns are technically mechanically plausible as listeners can conceptualize their interactions in acoustic musical instruments. They are also more familiar to

musicians than nonmusicians in contemporary music and through extended playing techniques. For example, a xylophone bar can be bowed, and a slap tongue technique can be applied to the mouthpiece of a clarinet or saxophone. However, Modalys synthesizes the bowed plate such that the bow passes through the plate, rather than exciting it at its edge. As for the struck air column, Modalys applies the striking excitation to the air molecules somewhere along the length of the air column instead of striking it at a mouthpiece attached to the top of the air column. So, even though the bowed plate and struck air column can be considered mechanically plausible, they are simulated in ways that are physically impossible. The remaining three atypical interactions (bowed air column, blown string, and blown plate) are both mechanically implausible and physically impossible. The purpose of implementing atypical interactions was to compare how they are perceived relative to their typical counterparts: struck and bowed strings, blown air column, and struck plate. These interactions are typical because they are common to acoustic musical instruments, and listeners, regardless of musical background, are familiar with them. The task of the current study involved dissimilarity ratings of pairs of sounds produced by the nine different interactions. An MDS algorithm called Scaling by MAjorizing a COmplicated Function (SMACOF) was used to convert the dissimilarity ratings into a timbre space that reflects the perceived dissimilarity between the stimuli (de Leeuw & Mair, 2009). Then, model selection using backward stepwise regression was implemented in an exploratory approach to determine the acoustic correlates of each dimension of the revealed timbre space.

One question of interest pertains to how the results from dissimilarity ratings correspond to previous tasks involving resemblance ratings to each type of excitation and resonator (Huynh, 2019) and categorization based on excitations and resonators (Huynh & McAdams, 2023a). Because these previous studies demonstrated an assimilation of atypical to typical interactions, the current study will determine whether the stimuli would be clustered in the timbre space based on these assimilations. Researchers have argued that similarity predicts categorization (Goldstone, 1994; Sloutsky, 2003). Additionally, similarity can also predict potential confusions between items not belonging in the same category (Goldstone et al., 2001). If these theories are true, then the dissimilarity data should predict how the excitation-resonator interactions were categorized in Huynh and McAdams's (2023a) experiment. Consequently, the atypical interactions would be expected to occupy similar regions on the generated timbre space as the typical interactions to which they were assimilated.

Harnad (1990) proposes that the methods of testing dissimilarity perception are independent of those of categorization. Dissimilarity ratings are relative judgments concerning the degree to which two things are different. Listeners can consequently judge the dissimilarity of sounds without knowing

what they are, or which excitations and resonators produced them. Categorization, on the other hand, is an absolute judgment: category membership is determined by whether a class of inputs are treated as having shared invariant features. Given the independent methods of testing dissimilarity perception and categorization, listeners might process the same set of sounds differently, depending on the task. Findings from McAdams et al.'s (2010) study support this hypothesis. They synthesized plates on a mechanical continuum between metal and glass, which were struck by mallets made of different materials. Material categorization of the plates depended on just the damping properties distinguishing them, whereas dissimilarity ratings depended on information corresponding to both damping properties and wave velocity (related to elasticity and mass density). Therefore, listeners based their judgments on acoustical features relevant to the perceptual task. If the stimuli of the present study involving dissimilarity ratings are processed differently than they were in a previous categorization study (Huynh & McAdams, 2023a), then the clustering of the stimuli in a timbre space might not predict the assimilations observed in the categorization task using the same stimuli. Instead, sounds produced by the same excitations and resonators might occupy similar regions in timbre space. This would imply that listeners can attend to potential invariant features among them.

Another question of interest concerns the number of dimensions in the timbre space that will be uncovered. Consistent with the findings from previous MDS studies, two to three salient dimensions are expected to be revealed in the timbre space of the current study. One of those dimensions is expected to differentiate the excitations based on temporal changes of the stimuli's acoustic properties, as this would align with previously mentioned timbre spaces with a temporal dimension differentiating impulsive and sustained excitations. However, it might require one or two salient dimensions to differentiate the resonators. Material differentiation was previously attributed to one spectral dimension best explained by changes in spectral centroid (Hjortkjær & McAdams, 2016), but materials are only one aspect of resonators. Another dimension might be expected to differentiate other aspects of resonators (e.g., geometry or other properties of their physical structure). This third dimension, if apparent in the timbre space, could be associated with spectral flux or audio descriptors related to spectrum fine structure (Caclin et al., 2005). Consequently, the main goal is to more directly conclude whether different mechanical components can be perceptually differentiated and additionally acknowledge how changes in the mechanical components correspond to changes in the sounds' acoustical features.

3.2 Method

3.2.1 Participants

Eighty-three participants were recruited through an online platform called Prolific. As the experimental task involved rating the dissimilarity of pairs of sounds on a scale from "identical" to "very dissimilar," the ratings therefore ranged continuously from 1 to 9. The dissimilarity ratings of identical stimulus pairs should ideally have a rating very close to 1. Consequently, the ratings of identical pairs were used to determine whether any participants' data should be discarded from the analyses: if participants rated over half of the identical pairs higher than 2.6 (i.e., 20% along the scale from identical to very dissimilar), their data were removed from the analyses as it indicated they did not properly understand the use of the rating scale. By this criterion, three participants' data were discarded. The remaining 80 participants (30 females, 50 males) had a mean age of 27.7 (SD=10.3). Participants provided consent and were compensated for their time through Prolific. This study was certified for ethical compliance by the McGill University Research Ethics Board II.

3.2.2 Apparatus

The experiment ran during the COVID-19 lockdown, so data were collected from participants via an online experiment on Prolific. Given that online experiments have constraints on experimental control, they were accounted for as much as possible in the following ways. Participants were instructed to run the experiment on a laptop or computer and not on a cellular device. Additionally, they were asked to use headphones for the experiment instead of the built-in speakers of their computers. They reported the model and make of their headphones. Participants were strongly advised to complete the experiment in a quiet environment with very little background noise. Participants adjusted their volume to a comfortable listening level using three sample stimuli (bowed string, blown air column, struck plate); once this level was determined by playing all three stimuli at least twice, they were instructed not to adjust the volume for the duration of the experiment. Although it could not be confirmed that participants had normal hearing with a pure-tone audiometric test, they self-reported normal hearing in the questionnaire. The experimental task was programmed with JavaScript in the PsiExp computer environment (Smith, 1995).

3.2.3 Stimuli

The same set of 27 stimuli were used as in previous studies (Huynh, 2019; Huynh & McAdams, 2023a). The stimuli were generated with Modalys, a digital, physically inspired modeling synthesizer

(Dudas, 2014). Using modal synthesis, Modalys isolates the parameters of excitations and resonators and models the acoustical outcome between their interactions, even without the resulting sounds being perceived as existing musical instruments (Eckel et al., 1995). Nine types of interactions were generated by combining three excitations (bowing, blowing, striking) with three resonators (string, air column, plate). Huynh (2019) intended for each of the excitations and resonators to have different acoustical and perceptual outcomes and to be represented by different physically inspired models.

Bowing was controlled by two temporal envelopes for the bow speed and bow pressure. Changes in the bow speed and pressure have been associated with changes in loudness and brightness, respectively (Halmrast et al., 2010). Blowing was applied through the mouthpiece with a single reed and was controlled by temporal envelopes for the breath pressure and embouchure pressure (i.e., the pressure of the lips on the reed which controls the height of the initial reed opening). These two parameters are known to change the resulting timbre: e.g., a higher breath pressure and smaller reed opening can generate brighter sounds (Coyle et al., 2015). Striking was represented by a temporal envelope controlling the force of a hammer on an object.

The string was simulated as a thin wire fixed at its edges and should vibrate at modes that are integer multiples of the fundamental frequency. The air column represents the air molecules within a cylindrical tube that is closed at one end and open at the other: it should produce modes with greater energy at odd-harmonic ratios than at even-harmonic ratios with respect to the fundamental frequency. The plate was rectangular, flat, rigid, and fixed at its edges. When excited by striking, it should have inharmonic modal content. However, when a sustained excitation is applied to it, all its modes would initially be excited; then, it should have a fundamental frequency corresponding to the lowest mode of vibration and the higher modes should vibrate at frequencies that are close to harmonically related to the fundamental.

The general procedure for the stimulus design involved isolating the parameters of an excitation from those of a resonator it typically interacts with. In the example of a bowed string, the bowing parameters were isolated from the string parameters. Similarly, parameters of a blowing excitation were isolated from those of an air column resonator. Then, the bowing parameters were applied to the air column parameters. This process was repeated to synthesize each type of atypical interaction. Given that modifying any parameter can change the timbre of the resulting sound, we synthesized 400 versions of each excitation-resonator interaction using a controlled and exhaustive approach to manipulate specific parameters of each interaction (refer to Huynh, 2019, for more details). Three exemplars of each interaction were informally selected to represent the most variability in their timbres

while being perceived as produced by the same source components. Four types of these interactions are deemed typical because they are common to acoustic musical instruments that can be produced in the physical world: bowed string (BoSg), blown air column (BlAc), struck string (SkSg), and struck plate (SkPl). The five other interactions—bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc)—are considered atypical because they are mechanically implausible (BoAc, BlSg, and BlPl), or are rarely encountered by listeners (BoPl, SkAc). None of the atypical interactions can be produced in the physical world, even the rarely encountered ones. BoPl is simulated in Modalys such that the bow passes through the plate to excite it, whereas musicians may have encountered a metal bar that is bowed at its edge. The simulation of SkAc involves striking a hammer on the air molecules at some point along the length of an air column, whereas musicians may have some familiarity with a slap tongue being applied to the mouthpiece of a clarinet or saxophone. Each sound had a duration of 2 s and the lowest vibrating mode was 155 Hz, which is closest to the pitch of E-flat-3. Stimuli were already matched for loudness in Huynh's (2019) previous experiments. Because participants rated stimulus pairs, the 27 stimuli therefore generated 378 pairs of sounds corresponding to a lower triangular matrix, including the main diagonal (i.e., identical pairs).

3.2.4 Procedure

Prior to beginning the experiment there was a practice block consisting of dissimilarity ratings among three stimuli, which formed six pairs. The three practice stimuli included three typical interactions that were simulated using Modalys: bowed string, blown air column, struck plate. They each had a lowest vibration mode of 220 Hz (i.e., A3 in pitch). The practice block followed the same procedure as the main experiment, and none of the practice stimuli were used in the main experiment. The purpose of the practice block was to allow participants to familiarize themselves with the online experimental interface.

Before participants performed the ratings in the main experiment, they played the full range of the 27 stimuli, presented with an inter-onset interval of 2 s. They were only allowed to play the full range once. The purpose of hearing the full range was for participants to grasp the variability in the timbres among the sounds. The sounds were presented in a pseudo-random order such that two successive sounds were not produced by the same interaction type (i.e., a BoPl could not be played before or after another BoPl).

Given that there were 378 pairs of sounds (i.e., lower triangle plus main diagonal of an $N \times N$ matrix, hereafter referred to as "lower triangular matrix"), the experiment would be extremely tasking

if each participant rated every pair. Thus, the 80 participants were divided into 20 "super-participants," which meant that each participant completed one-fourth of the lower triangular matrix of ratings, similar to the approach of Elliott et al. (2013). That is, the lower triangular matrix of ratings was randomly divided into four groups of 94 or 95 ratings, and each participant within a given super-participant was randomly assigned their own unique group of ratings. So, no two participants within or across super-participants would rate the exact same pairs of sounds. Every super-participant (i.e., every four participants) therefore completed one full set of ratings of the lower triangular matrix.

For each participant, the 94 or 95 stimulus pairs were divided into three blocks of 31 or 32 trials. In each trial, participants listened to a pair of sounds. If necessary, they could replay the pair only once. They were then asked to rate the dissimilarity of the pair of sounds based on their timbres. In the experiment, timbre was defined as such: "Timbre is the auditory quality that distinguishes musical tones played on different instruments. Sounds can differ in brightness, sharpness of attack, roughness, richness, and any number of qualities." The slider was continuous and the rating scale was labeled "identical" and "very dissimilar" on the extreme left and right ends, respectively. Participants were instructed to click anywhere within the slider to make a yellow diamond cursor appear. The yellow diamond could be dragged anywhere along the slider to indicate the position of the rating. Dragging the slider all the way to the left meant that the pair comprised identical sounds. Dragging the slider all the way to the right indicated that the participant heard the biggest difference in the given pair out of the whole set of sounds. If participants thought the pair of sounds was somewhat dissimilar or similar, they placed the diamond somewhere on the slider reflecting their judgment. Participants proceeded to the following trial once they were confident in their rating. The order in which the stimulus pairs were presented was randomized for each participant. Within a given pair, the order of stimulus presentation was randomized across participants. Throughout the entire experiment, the names of the interactions of the stimuli were never revealed to participants.

3.3 Results

3.3.1 Analyses Overview

The dissimilarity ratings were first converted into a timbre space using MDS. The number of dimensions was selected using cross-validation and model selection techniques involving the comparison of R^2 and the Akaike Information Criterion (AIC), which were implemented by the SMACOF algorithm (Elliott et al., 2013). After retrieving the positions of each sound along each

dimension in the timbre space, the audio descriptors that best explained the variance in each dimension were analyzed using an exploratory rather than confirmatory model selection method. To do this, the audio descriptors of each stimulus were computed using the Timbre Toolbox Version R1.0 (Kazazis et al., 2022). This paper focused on a subset of audio descriptors. The criteria for the selection of the subset will be explained in Section 3.3.3. The selected audio descriptors were then regressed onto the positions of each dimension using backward stepwise regression. This further reduced the subset of audio descriptors considered as potential predictors of the positions along a given dimension (Thayer, 2002). This technique was necessary as each outcome variable had only 27 observations (corresponding to each sound), so there would be too many predictors in the model. The model selection technique therefore selected different predictors for each dimension. These selected predictors were then regressed onto the corresponding dimension using multiple regression with standardized coefficients.

3.3.2 The Timbre Space

The timbre space was generated with the MDS algorithm SMACOF, which was first developed by de Leeuw & Mair (2009). It was later modified by Elliott et al. (2013), who implemented cross-validation by computing the transformations of the timbre space with left-out participants. Classical MDS generates a group configuration depicting the position of n items (where n corresponds to the number of sounds in this case) in p dimensions. The Euclidian distance between the positions of any two items in the space reflects their perceptual dissimilarity. The most optimal space is computed by minimizing a stress function: the weighted sum of square errors is minimized between the distances in the configuration and their corresponding observed dissimilarities. The SMACOF algorithm accounts for missing ratings, which is applicable to the current analyses, given that the participants of each super-participant only rated about one-fourth of the stimulus pairs. Each participant's ratings were represented in an $n \times n$ symmetrical matrix. The value of any given cell in the matrix indicated the rating between two corresponding sounds on a continuous scale, ranging from 1 to 9. If a cell had a rating of 0, it meant that the corresponding rating was not performed. An additional $n \times n$ weight matrix was computed for each participant, which determined whether a given rating was performed: the absence and presence of a rating was coded by 0 and 1, respectively.

A diagonal linear transformation, otherwise known as INDSCAL (INdividual Differences in SCALing; Carroll & Chang, 1970), was also employed in the algorithm. In summary, INDSCAL computes a configuration for each participant and the same dimensions are assumed for all

participants, but the dimensions might be weighted differently for each participant (de Leeuw & Mair, 2009). SMACOF then implements an iterative procedure deemed majorization to minimize the stress function for both the group configuration and the linear transformation (Elliott et al., 2013). This is done by first computing the group configuration with the classical MDS estimate. Majorization aims to improve the configurations for each participant by updating the group configuration and individual configurations in each iteration. It computes the optimal linear transformations for each participant by scaling and rotating the transformations of the previous group configuration and obtaining the best projection. The updated transformations are then inverted before being applied to the individual configurations, thus generating the updated group configuration. Each updated individual configuration is consequently a multiplication of the linear transformation and the updated group configuration. These iterations continue until the stress function is optimally minimized.

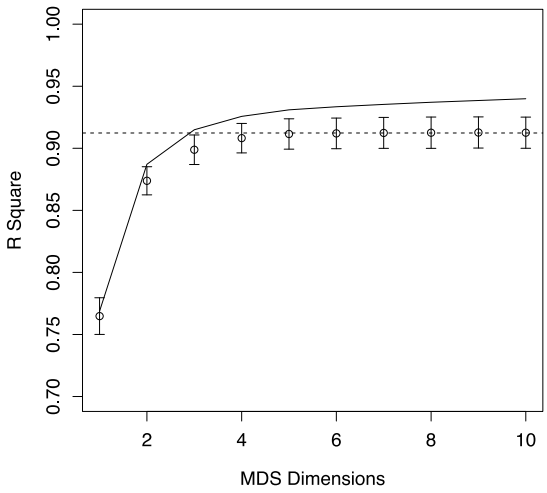

Figure 3.1 Average R^2 values at each number of dimensions for left-out participants are represented by the circles, with the standard error represented by the error bars. The dashed line shows the average R^2 across the solutions with five to ten dimensions. The R^2 of the fitted participants at each number of dimensions based on the MDS analysis is illustrated by the solid line.

The cross-validation technique was also applied to the SMACOF algorithm introduced by Elliott et al. (2013). This technique tests the predictive power of including each additional dimension, given that increasing the number of dimensions in the model will consequently increase R^2. Cross-validation primarily involves jackknifing, which obtains a group configuration (with SMACOF) leaving out each participant in turn and then computes the average R^2 and standard error based on the configurations

generated with each left-out participant. The average R^2 obtained using the jackknifing procedure begins to plateau after five dimensions as seen in Figure 3.1. The AIC of the MDS solutions with one to ten dimensions was also evaluated (Figure 3.2). AIC is useful in model selection as it estimates the predictive error of the MDS solution with each additional dimension. The solution with the lowest AIC is considered the best model. For the current study, the MDS solution with the lowest AIC contained four dimensions (AIC=303.20), and the next lowest AIC was observed in the solutions with five dimensions (AIC=313.70) and then three dimensions (AIC=314.52); however, the difference in AIC between the three solutions was quite minimal. This implies that the optimal balance between goodness of fit and parsimony is seen in solutions with three to five dimensions.

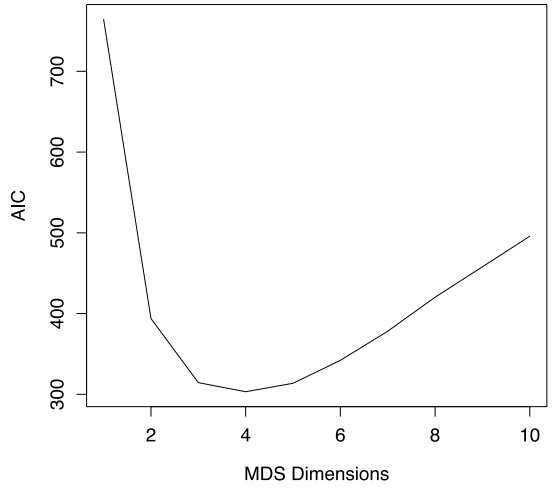

Figure 3.2 Akaike Information Criterion (AIC) values computed for the MDS solution at each number of dimensions.

To further examine which MDS solution produced the most interpretable timbre space, the positions of each stimulus were plotted along each dimension. The comparison was narrowed down to the three- and four- dimension solutions for two reasons. First, the 4D solution was associated with the lowest AIC. Second, AICs for the 3D and 5D solutions were nearly identical and the solution with fewer dimensions would be more parsimonious. The 4D solution conveyed a clustering of struck sounds along one dimension and plates along another, but the clustering of the remaining excitation-resonator interactions across the other dimensions was vague or nonexistent. The 3D solution, on the other hand, showed that sounds were differentiated based on their excitations and resonators. As seen in Figure 3.3a, the different interactions generally occupied different regions in the 3D space. The first dimension isolated the struck sounds (dark red points) from the sustained excitations; however, there

was additionally a subtle separation between the blown (brown points) and bowed (light grey points) sounds. The distinction between impulsive and sustained excitations is consistent with previous findings (Giordano & McAdams, 2010).

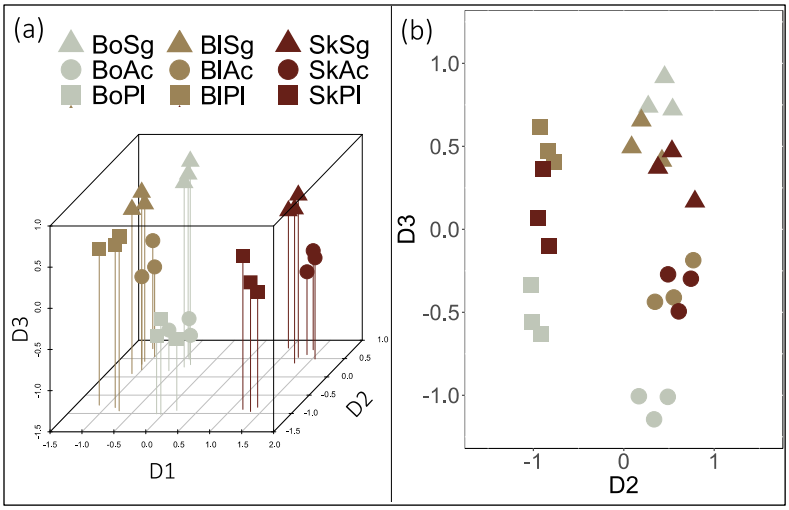

Figure 3.3 The timbre space generated from MDS using SMACOF algorithm with the INDSCAL linear transformation. Each point represents an exemplar of an excitation-resonator interaction. Bowing (Bo), blowing (Bl), and striking (Sk) excitations are represented by the light grey, brown, and dark red colors, respectively; string (Sg), air column (Ac), and plate (Pl) resonators are represented by triangle, circle, and square symbols, respectively. (a) The three-dimensional timbre space. (b) A closer look at Dimension 2 versus Dimension 3 to show resonator distinction.

A clearer relationship between Dimensions 2 and 3 is shown in Figure 3.3b. Sounds produced by the plate (square points) were very clearly separated from sounds produced by the other resonators on Dimension 2. The plate also spanned a considerable range along Dimension 3, with the lowest to highest regions occupied by the bowed plate, struck plate, then blown plate. It is likely that the plate produced the most variability in terms of its harmonic content, depending on whether an impulsive or sustained excitation was applied to it (see Section 3.2.3 for a comparison of the vibrating modes resulting from the different excitations). The harmonic content of the string and air column was more consistent regardless of the excitations applied to each of them, with the string having both even and odd harmonics and the air column having primarily odd harmonics. The differences in their harmonic content might be a possible explanation for their differentiation: at higher values of Dimension 2, the string sounds (triangle points) occupied the space at higher values of Dimension 3, whereas the air column sounds (circle points) generally occupied the space at lower values of Dimension 3.

The results of this timbre space imply that the perception of excitations was likely associated with one dimension, whereas resonator perception was likely associated with two dimensions. This also suggests that the perception of the differences in the resonators was more complex than that of excitations. Furthermore, this timbre space was generated from the dissimilarity ratings between sound pairs based on participants' arbitrary criteria. They were not required to judge dissimilarity based on anything specific and there was no mention that the sounds were produced by combining different excitations and resonators during the experimental task. The participants, however, were still able to separate the sounds, albeit implicitly, based on the excitations and resonators that produced them. Therefore, the next question of interest involves finding if any acoustical properties were correlated with each of the three dimensions.

3.3.3 Acoustic Correlates of the Dimensions

The Timbre Toolbox Version R1.0 (Kazazis et al., 2022; originally developed by Peeters et al., 2011) was used to extract the acoustical properties, otherwise known as audio descriptors, that could potentially characterize the timbres of the stimuli. There are three categories of audio descriptors: spectral, temporal, and spectrotemporal. Temporal audio descriptors were computed using the temporal energy envelope (TEE) of the audio signal. Spectral and spectrotemporal descriptors were analyzed using the power spectrum of the Short-Time Fourier Transform (STFT). A subset of the spectral descriptors, more specifically referred to as harmonic descriptors, were computed on the harmonic and inharmonic partials. Given that the spectral, spectrotemporal, and harmonic descriptors are time-varying, median and interquartile range (IQR) values were computed, i.e., robust measures of central tendency and variability over the duration of the sound. Audio signal descriptors were analyzed using the temporal waveform; the median and IQR were calculated for them as well. Given that over 55 audio descriptors were computed, and the stimulus set comprised only 27 sounds, a subset of 12 audio descriptors was selected to simplify the analyses. The subset was chosen based on two criteria. The first criterion was based on the statistically independent clusters of audio descriptors found in Peeters et al. (2011). Each cluster revealed descriptors that are generally correlated across a stimulus set of over 6,000 musical instrument sounds. Consequently, inferences made about one descriptor in a cluster can be loosely applied to other descriptors in the same cluster. The second criterion for selecting the subset was based on the descriptors that were qualified as relevant to the current stimulus set even if they were in the same cluster as the audio descriptors chosen based on the first criterion.

The definitions of the selected set of audio descriptors based on the Timbre Toolbox Version R1.0 Manual (Kazazis et al., 2022) are provided in Table 3.1.

Table 3.1 Selected audio descriptors, their definitions, and input representations from which they are derived. Values computed from the Timbre Toolbox (i.e., median and/or IQR) are indicated in parentheses. If no values are indicated in parentheses, that means the Timbre Toolbox computed only one value for the corresponding audio descriptor. TEE = temporal energy envelope, STFT = Short-Time Fourier Transform (power spectrum), Harm = harmonic, AS = audio signal.

Audio descriptor	Abbreviation	Definition	Input representation
Temporal centroid	TC	The center of gravity of the energy envelope.	TEE
Amplitude of energy modulation	AM	Amplitude of the energy modulation during the sustained portion of the sound.	TEE
Frequency of energy modulation	FM	Frequency of the energy modulation during the sustained portion of the sound.	TEE
Spectral centroid (median, IQR)	SCent	The spectral center of gravity.	STFT
Spectral flatness (IQR)	SFlat	Flatness or noisiness of the spectrum.	STFT
Spectral crest (median)	SCrest	"Peakiness" and tonalness of the spectrum.	STFT
Spectral flux (median)	SFlux	Degree of variation of the spectrum over time.	STFT
Harmonic spectral deviation (median)	HDev	A measure of the deviation of the partial's amplitudes from a global spectral envelope.	Harm
Odd-to-even ratio (median)	OER	Ratio of the energy of odd harmonics relative to the energy of even harmonics.	Harm
Tristimulus 2 (median)	TS2	Amplitude of the second to fourth harmonics relative to the sum of amplitudes of all harmonics.	Harm
Root mean square energy (IQR)	RMS	The root mean of the signal's frame energy.	AS

Some of the most common acoustic correlates of timbre space dimensions are the log attack time, spectral centroid, and either spectral flux (McAdams et al., 1995) or spectrum fine structure (Krimphoff et al., 1994). Thus, the median and IQR of spectral centroid as well as the median spectral

flux were included in the selected subset of audio descriptors as they each appeared in separate statistically independent clusters (Peeters et al., 2011). Although log attack time is notable for distinguishing impulsive from sustained excitations, it also clusters with temporal centroid (Peeters et al., 2011). The attack temporal centroid has been implicated in having slightly higher explanatory power than the attack time (Kazazis et al., 2021a); so, temporal centroid was included instead of log attack time in the selected subset of audio descriptors. Although harmonic spectral deviation was in the same cluster as temporal centroid and log attack time (Peeters et al., 2011), it highlights details pertaining to the spectra of the sounds rather than their temporal attributes. Harmonic spectral deviation is sometimes referred to as spectral irregularity (Krimphoff et al., 1994) and can be calculated for inharmonic sounds, such as those produced by vibrating plates.

It would be ideal to include inharmonicity in the subset of audio descriptors, but the median and IQR of inharmonicity were very similar across all the sounds. It might be the case that the way this audio descriptor is calculated by the Timbre Toolbox does not generate the expected values. So, other audio descriptors that clustered with inharmonicity and that might be associated with describing the spectrum fine structure were considered. Median odd-to-even ratio, for example, is a special case of harmonic spectral deviation and might distinguish the air column from the string and plate. Furthermore, median tristimulus 2 might reveal specific details regarding the prominence of the second to fourth harmonics in the sounds. Tristimulus 2 also correlates with tristimulus 1 and 3, which describe the prominence of the first harmonic and the prominence of the harmonics above the fourth harmonic, respectively (Pollard & Jansson, 1982).

Frequency of energy modulation (FM) and the IQR of spectral flatness both appeared in individual statistically independent clusters (Peeters et al., 2011) and were included in the selected subset. Clustering among the following pairs of audio descriptors was also revealed: spectral crest (median) and spectral flatness (median); amplitude of energy modulation (AM) and spectral crest (IQR); and frame energy (IQR) and root mean square energy (IQR). Spectral crest and flatness are often strongly negatively correlated because the former assesses the "peakiness" of the spectrum, whereas the latter measures its noisiness. These descriptors can be discussed as opposites if either of them happens to explain the variance in any of the three dimensions. AM was chosen over the IQR of spectral crest as both describe changes in the signal, but AM provides a clearer indication of tremolo effects (Kazazis et al., 2022). The IQR of the root mean square (RMS) energy was arbitrarily chosen over the IQR of frame energy; both are related, as the RMS energy is the root mean of the frame energy.

If the 12 selected audio descriptors of Table 3.1 were regressed onto the positions of each dimension computed in Section 3.3.2, there would still be too many predictors for the small number of observations. This could introduce problems of overfitting, multicollinearity, and difficulty of interpreting the relationship between salient descriptors and each dimension. Moreover, it is uncertain in general which audio descriptors might be working in combination to predict the positions of a given dimension. Using a model selection technique, three separate backward stepwise regressions were performed for the positions of the stimuli along each of the three dimensions. Backward stepwise regression can be favorable in an exploratory rather than confirmatory analysis (Thayer, 2002; Ruengvirayudh & Brooks, 2016). Given that any two sounds can differ in a multitude of audio descriptors, it is unrealistic to assume that each perceptual dimension of a timbre space is related to a single audio descriptor. Using an exploratory analysis allows for an open-ended interpretation of the acoustic properties that can be associated with each dimension and does not limit the interpretation to just one descriptor per dimension. Although Huynh (2019) intended for the different resonators to produce different harmonic content (i.e., even and odd harmonics for the string, primarily odd harmonics for the air column, and higher-frequency and inharmonic content for the plate), it cannot be known for sure that listeners prioritized these differences when making their dissimilarity ratings. As mentioned, the calculation of inharmonicity by the Timbre Toolbox did not produce the expected values, so it was also of interest to examine if other audio descriptors were related to the isolation of the plate on Dimension 2. Furthermore, although audio descriptors that could be potentially associated with each dimension were hypothesized based on previous findings (e.g., temporal centroid, spectral centroid, spectral flux, harmonic spectral deviation), it is possible that other audio descriptors that were not considered can contribute to the variability of the dimension positions as well.

Therefore, an exploratory analysis was implemented using backward stepwise regression. It starts with a full model including 12 audio descriptors as predictors and removes a predictor in each step based on a penalization criterion (e.g., *p* value, AIC, or Bayesian Information Criterion [BIC]) until the removal of a predictor no longer meets the criterion. Backward stepwise regression can be prone to generating larger models than necessary (Thayer, 2002). To compensate for this, the removal of predictors was determined by the BIC which selects less variables and generates a more parsimonious model than AIC (Zhang, 2016). The backward stepwise regression implemented in the current study therefore compares the BIC between the model with and without the predictor in question in each step. The removal of predictors stops when the final model is computed to have the lowest BIC. An alpha value (i.e., significance level) of 0.01 was used in the stepwise regression. Additionally, a

bootstrapping method was implemented such that the regression was specified to repeat 100 times by using sampling with replacement to simulate a new dataset in each iteration (Rizopoulos, 2022). This allows for the user to determine the percentage of times each predictor: (1) was selected in the model; (2) was significant in the model; and (3) had a positive or negative sign for its regression coefficient.

Three separate backward stepwise regressions were computed with the 12 audio descriptors as predictors. The dependent variable of each model was the position of each stimulus along a given dimension. So, a different reduced set of predictors will be selected depending on the dimension. The reduced predictors were then regressed onto the positions of their corresponding dimension using multiple linear regression. Standardized regression coefficients were computed to compare the relative contribution of each predictor to each dimension. The confidence intervals for the standardized regression coefficients in each model were bootstrapped using 1,000 samples with replacement. Consequently, bias-corrected and accelerated (bca) 95% confidence intervals (CI) were reported for each contributing predictor.

3.3.3.1 Dimension 1

The 12 selected audio descriptors were regressed onto the positions of Dimension 1 with a backward stepwise regression. Using BIC as a means for model selection and 100 bootstrap samples, four predictors were removed from the model. The remaining eight predictors were TC, FM, SCent (med), SCent (IQR), SFlat (IQR), TS2 (med), HDev (med), and RMS (IQR). As mentioned above, the bootstrapping procedure of the backward stepwise regression can determine the percentage of times a predictor was selected in the final model and the percentage of times it was significant (based on the alpha of 0.01). Among the eight predictors remaining in the final model, TS2 (med) was selected 76.00% of the time and SFlat (IQR) was selected 99.00% of the time. The percentage of times the other six predictors were selected falls within this range. Additionally, the percentage of times each of the predictors in the final model was significant ranged from 62.82% (for the IQR of SCent) to 100.00% (for TC). The eight predictors were regressed onto the positions of Dimension 1 in a multiple linear regression. The model had an $R^2_{Adjusted}$=0.970 [F(8, 18)=105.90, p<.001, R^2=0.979], indicating that the data strongly fit the model. Table 3.2 reports the standardized coefficient estimates for all involved audio descriptors, ranked by largest to smallest absolute value.

Table 3.2 The standardized coefficients and corresponding 95% bias-corrected and accelerated (bca) confidence intervals from the multiple regression of eight audio descriptors onto the positions of Dimension 1. The audio descriptors are ranked by their standardized coefficients from largest to smallest absolute value.

Audio descriptor	β	SE	t	p	95% bca CI
TC	−0.72	0.056	−12.77	<0.001	[−0.89, −0.61]
HDev (med)	−0.22	0.046	−4.80	<0.001	[−0.38, −0.10]
SFlat (IQR)	0.20	0.049	4.08	<0.001	[0.08, 0.30]
TS2 (med)	−0.18	0.045	−4.10	<0.001	[−0.28, −0.08]
FM	−0.15	0.046	−3.20	0.005	[−0.26, −0.01]
RMS (IQR)	−0.10	0.040	−2.57	0.019	[−0.21, 0.08]
SCent (med)	0.09	0.043	2.04	0.056	[−0.04, 0.17]
SCent (IQR)	−0.07	0.046	−1.60	0.127	[−0.16, 0.04]

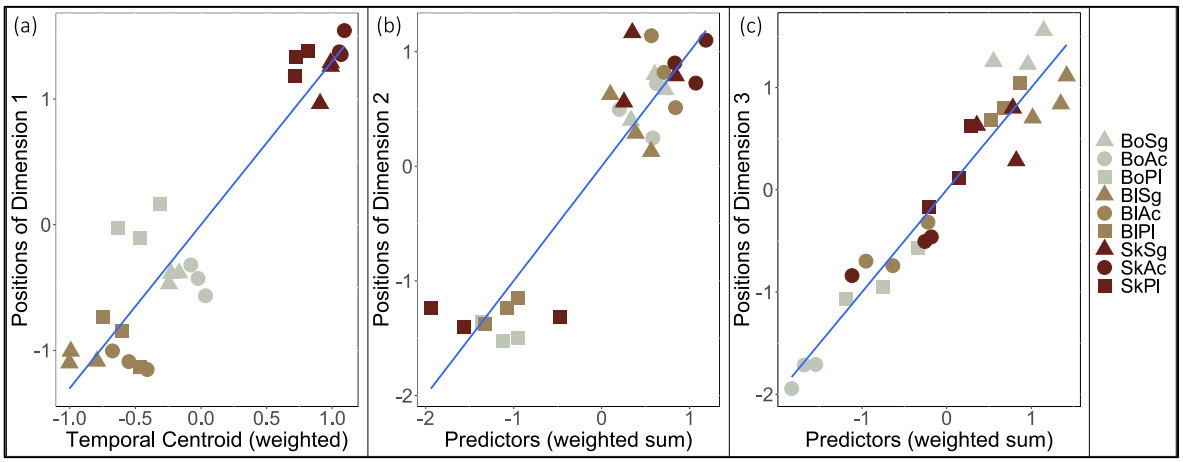

Figure 3.4 Scatter plots representing the relationship between audio descriptors weighted by standardized regression coefficients and the MDS dimensions whose variance they best explain. (a) Temporal centroid versus Dimension 1. (b) Weighted sum of acoustic predictors versus Dimension 2. (c) Weighted sum of acoustic predictors versus Dimension 3. For each scatter plot, the individual points represent one exemplar of a type of excitation-resonator interaction. The regression line for each scatter plot is indicated in blue.

Based on the bca 95% CIs, TC, HDev (med), SFlat (IQR), TS2 (med), and FM were significant predictors of Dimension 1. Temporal centroid was the most contributing predictor and had a negative relationship with the positions along Dimension 1. Temporal centroid is known to distinguish impulsive and sustained excitations (Hjortkjær & McAdams, 2016; Kazazis et al., 2021a; Peeters et al., 2011). In Section 3.3.2, Dimension 1 positions clearly separated struck from sustained excitations and vaguely separated bowed and blown sounds (Figure 3.3a). Figure 3.4a shows a scatterplot of the

temporal centroid—weighted by its standardized regression coefficient—versus the Dimension 1 positions of each sound. The points fall closely on the regression line (i.e., blue line), and temporal centroid does a fairly good job of separating the excitations, indicating that it is the primary acoustic correlate of Dimension 1.

3.3.3.2 Dimension 2

The general procedure of the analyses for Dimension 2 was the same as for Dimension 1. The 12 selected audio descriptors were regressed as predictors onto the positions of each sound on Dimension 2 in a backward selection model. Six predictors were selected in the final model: FM, AM, SCent (med), SFlat (IQR), SCrest (med), and HDev (med). These predictors were selected in the model between 61.00% (for median HDev) to 97.00% (for median SCent) of the time. They were significant between 72.13% (for median HDev) and 95.88% (for median SCent) of the time. The six predictors were then regressed onto the positions of each sound on Dimension 2 in a multiple linear regression. Standardized coefficients are reported and ranked by their magnitudes in Table 3.3. The data fit well in the model, $R^2_{Adjusted}$=0.808 [$F(6,20)$=19.18, p<.001, R^2=0.852]. Based on the bca 95% CIs, each predictor had a significant effect on Dimension 2. The median spectral centroid and median spectral crest had the two largest effects, followed by the amplitude of energy modulation, and then the IQR of spectral flatness. The weighted sum of the six predictors was plotted against the positions of Dimension 2 in Figure 3.4b. The weights assigned to each predictor were their standardized regression coefficients. The isolation of the plate from the other two resonators could therefore be predicted by a weighted combination of primarily spectral and energy modulation descriptors.

Table 3.3 The standardized coefficients and corresponding 95% bias-corrected and accelerated (bca) confidence intervals from the multiple regression of six audio descriptors onto the positions of Dimension 2. The audio descriptors are ranked by their standardized coefficients from largest to smallest absolute value.

Audio descriptor	β	SE	t	p	95% bca CI
SCent (med)	−1.05	0.146	−7.19	<0.001	[−1.53, −0.73]
SCrest (med)	−0.89	0.145	−6.14	<0.001	[−1.29, −0.54]
AM	0.56	0.121	4.63	<0.001	[0.14, 0.94]
SFlat (IQR)	−0.53	0.114	−4.60	<0.001	[−0.76, −0.11]
HDev(med)	0.34	0.115	2.93	0.008	[0.08, 0.62]
FM	0.28	0.096	2.89	0.009	[0.09, 0.50]

To further assess the contribution of each weighted audio descriptor to Dimension 2 positions, six additional scatterplots were generated, each leaving out a weighted audio descriptor (Appendix, Figures 3.5 to 3.10). A comparison of each of these scatterplots with Figure 3.4b examines how well a particular audio descriptor contributed to the points falling on the regression line. Median spectral centroid seemed the likely candidate for globally separating the plate from the air column and string, especially given that it had the largest regression coefficient. When the predictors did not include median spectral centroid (Figure 3.5), none of the points representing the plate fall on the regression line. This implies that listeners generally perceived the plate as differing in auditory brightness from sounds produced by the string and air column. Median spectral crest brought the bowed plate, blown string, and struck string closer to the regression line (Figure 3.6). That is, when considering the differentiation between the plate and string in these interactions, listeners might have relied on differences in the "peakiness" of their spectra: the bowed plate would have more high-frequency content than the blown and struck string, but there might have been a particular mode of the bowed plate dominating in its sound. Median harmonic spectral deviation brought the blown air column closer to the regression line (Figure 3.9), implying that its hollowness allowed listeners to differentiate it from the plate. The IQR of spectral flatness, amplitude of energy modulation, and frequency of energy modulation were responsible for bringing many of the points a bit closer to the regression line; so, perception of small differences in the variability of the noisiness of the spectrum, depth and rate of tremolo, respectively, allowed listeners to further isolate the plate from the string and air column.

3.3.3.3 Dimension 3

A final backward stepwise regression was implemented to find the most contributing predictors of Dimension 3. The final model reduced the 12 predictors to six: FM, AM, SFlat (IQR), TS2 (med), HDev (med), and RMS (IQR). Interestingly, the median odd-to-even ratio (OER) was not one of the selected predictors, even though the air column would be expected to have modes vibrating primarily at odd-harmonic ratios relative to the fundamental frequency. Further inspection of the Timbre Toolbox's calculation of median OER showed that its values were much higher for air column sounds (ranging from 94.06 to 6.72e14) than for string (range: 0.93 to 17.26) and plate (range: 0.46 to 41.67) sounds. So, median OER might have been excluded by the backward regression model because of its nonlinearity. However, because OER is a special case of harmonic spectral deviation (HDev) which was selected in the final model, the predominance of odd harmonics that is characteristic of air column sounds might be captured by median HDev.

According to the bootstrapped backward selection, AM was selected in the final model 71.00% of the time and SFlat (IQR) was selected 97.00% of the time; the percentage of times that each of the other four predictors was selected was within this range. Furthermore, the six predictors of the final model were significant between 73.24% (for AM) and 97.94% (for the IQR of SFlat) of the time. Regressing the six predictors onto the positions of Dimension 3 generated the standardized coefficients presented in Table 3.4. The fit of the data on the model was $R^2_{Adjusted}$=0.891 [F(6, 20)=36.43, p<.001, R^2=0.916]. All six predictors were deemed significant by the bca 95% CIs of their regression coefficients. TS2 was the most contributing predictor to the variability in Dimension 3 positions, followed by RMS (IQR), SFlat (IQR), and HDev (med). Figure 3.4(c) is a scatterplot showing the relationship between the weighted sum of the significant predictors and Dimension 3. In Section 3.3.2, Dimension 3 distinguished the string from the air column. In Figure 3.4c, a similar pattern is observed. Disregarding the plate (square points), the string (triangle points) has higher values of the weighted sum than the air column (circle points), but the plate spanned the range of the other two resonator types. So, the acoustic correlates of Dimension 3 are a weighted combination of audio descriptors associated with specific details of the spectrum (median TS2 and HDev), the variability of the signal's energy (RMS) and noisiness (SFlat), and the frequency and depth of energy modulation (FM and AM, respectively).

Table 3.4 The standardized coefficients and corresponding 95% bias-corrected and accelerated (bca) confidence intervals from the multiple regression of six audio descriptors onto the positions of Dimension 3. The audio descriptors are ranked by their standardized coefficients from largest to smallest absolute value.

Audio descriptor	β	SE	t	p	95% bca CI
TS2 (med)	0.60	0.071	8.41	<0.001	[0.41, 0.75]
RMS (IQR)	−0.50	0.074	−6.73	<0.001	[−0.58, −0.12]
SFlat (IQR)	−0.41	0.086	−4.82	<0.001	[−0.75, −0.25]
HDev (med)	−0.34	0.086	−3.98	<0.001	[−0.60, −0.04]
FM	−0.27	0.075	−3.64	0.002	[−0.66, −0.08]
AM	0.21	0.084	2.53	0.020	[0.05, 0.44]

In a similar approach to Dimension 2, the contribution of each weighted significant descriptor was assessed for predicting the positions of the stimuli along Dimension 3 (Appendix A, Figures 3.11 to 3.16). Because Dimension 3 was responsible for further separating the string from the air column, the contributions of the significant descriptors will be discussed with respect to this distinction. By

comparing Figure 3.11 (removal of weighted median tristimulus 2) to Figure 3.4c, median tristimulus 2 seems to bring most of the stimuli closer to the regression line. Its regression coefficient was the largest, highlighting the importance of the second to fourth harmonics' prominence for globally differentiating the string from the air column: the string has more prominent second to fourth harmonics in comparison to the air column. Air column sounds would have lower amplitudes at the second and fourth harmonics relative to the third harmonic. The IQR of RMS energy primarily brings the bowed air column closer to the regression line (Figure 3.12), so distinguishing it from a string is related to the perception of differences in the variability of signal energy. As with Dimension 2, median harmonic spectral deviation appears to bring the blown air column closer to the regression line for Dimension 3 (Figure 3.14), implying that its hollowness distinguishes it from a string. Because harmonic spectral deviation is measured over a running average of the amplitudes of three adjacent harmonics in a spectrum (Kazazis et al., 2022), it might be capturing deviations of odd harmonics from a smooth spectral envelope, which would be characteristic of blown air column sounds. The IQR of spectral flatness and frequency and amplitude of energy modulation seem to bring most stimuli closer to the regression line, implying that minor differences in these descriptors allow listeners to fine-tune the separation of the string and air column stimuli.

3.4 Discussion

The current study examined the perceptual dissimilarity of sounds produced by interactions between three excitations and three resonators. The MDS analysis used the INDSCAL constraint in the SMACOF algorithm to transform dissimilarity ratings into an interpretable timbre space. The generated timbre space was best explained by three dimensions, with the exemplars of each interaction occupying distinct regions in the timbre space. One dimension differentiated striking from the sustained excitations noticeably and bowing from blowing vaguely. The second dimension isolated the plate from the other resonators and Dimension 3 further separated the string and air column.

The general positions that each excitation and resonator occupied on the timbre space dimensions are summarized in Table 3.5. Several audio descriptors had a significant effect on Dimension 1, but the temporal centroid (TC) was certainly the most salient. The direction of the relationship between TC and Dimension 1 was negative. Because Dimension 1 differentiated the excitations, struck sounds were predicted by lower TC values than bowed and blown sounds. The average TC among the bowed sounds was slightly lower than that of the blown sounds. So, differentiating the excitations was

associated with the perception of the attack's impulsiveness. This finding is consistent with previous MDS studies, which determined the acoustic correlate of one dimension to be related to the impulsiveness of the attack (Hjortkjær & McAdams, 2016; Iverson & Krumhansl, 1993; Krimphoff et al., 1994; Lakatos, 2000; Marozeau et al., 2003; McAdams et al., 1995). Furthermore, TC could be acting as what McAdams (1993) refers to as a transformational invariant. Transformational invariants are considered acoustic properties that describe what happens to a sounding object or the way it sets said object into vibration. Thus, the detection of the transformational invariant in the current stimulus set is unidimensional.

Table 3.5 Summary of the general positions occupied by each excitation and resonator on each perceptual dimension and their acoustic correlates. The most contributing audio descriptors are reported along with the direction (i.e., positive [+] or negative [−]) of their relationship to the corresponding dimension.

Dim	Acoustic correlates		Excitation			Resonator		
	Audio descriptor	Correlation (+/−)	Bo	Bl	Sk	Sg	Ac	Pl
D1	TCent	−	Lower–middle	Lower	Higher			
D2	SCent (median)	−				Higher	Higher	Lower
	SCrest (median)	−						
	SFlat (IQR)	−						
	AM	+						
	HDev (median)	+						
	FM	+						
D3	TS2 (median)	+				Higher	Lower	Spread out based on excitations: Bl>Sk>Bo
	RMS (IQR)	−						
	SFlat (IQR)	−						
	HDev (median)	−						
	FM	−						
	AM	+						

In Dimension 2, the sounds produced by the plate corresponded to a lower value of the weighted acoustic correlates than those produced by the air column and string. Due to the direction of the relationship between each acoustic correlate and Dimension 2 (Table 3.5), the plate was perceived as brighter (higher SCent), having a peakier spectrum (higher SCrest), having less energy modulation in terms of depth and rate of tremolo (lower AM and FM), more variable in noisiness (higher IQR of SFlat), and less hollow (lower HDev) than the string and air column. It appeared that SCent (med) was responsible for global distinctions between the plate and the other two resonators, SCrest (med) and HDev (med) accounted for distinctions of specific excitation-resonator interactions, and small changes in AM, SFlat (IQR), and FM contributed further guided the separation of the string and air column from the plate. Focusing only on the distinctions between the string and air column in Dimension 3, sounds produced by the string were perceived as having more prominent second to fourth harmonics (higher TS2), less hollow (lower HDev), less variable in noisiness (lower IQR of SFlat) and in the signal's energy (lower IQR of RMS), and having a slower rate and higher depth of tremolo (lower FM and higher AM, respectively) in comparison to sounds produced by the air column. TS2 (med) seemed to account for global differentiations between the string and air column, RMS (IQR) and HDev (med) explained distinctions of specific interactions, and much like Dimension 2, listeners might have used SFlat (IQR), FM, and AM to fine tune distinctions between the string and air column. Together, Dimensions 2 and 3 differentiated the resonators of the current stimulus set. The acoustic correlates associated with Dimensions 2 and 3 can be considered the structural invariants that provide information about a sounding object's physical structure (McAdams, 1993). So, the detection of structural invariants appears to be two-dimensional, with each dimension encompassing its own weighted combination of invariants. This not only demonstrates that the detection of structural invariants is more complex than the detection of transformational invariants in this stimulus set, but that listeners were more sensitive to the temporal properties than the spectral properties of the sounds. This is consistent with previous findings demonstrating a greater sensitivity to excitations or actions relative to instrument families or materials (Giordano & McAdams, 2010; Lematire & Heller, 2012).

In a review of dissimilarity perception of musical instrument tones, Giordano and McAdams (2010) found a common pattern among previous data. Tones produced by instruments with similar excitations or in the same family were perceived as more similar and clustered together in timbre space. The current study directly tested these patterns with musical tones that were synthesized by manipulating the interactions between three excitations and three resonators. So, the patterns

observed in Giordano and McAdams's (2010) review were extended to atypical interactions between excitations and resonators, given that the different types of excitations and resonators were differentiated irrespective of the typicality of the interaction. Furthermore, the timbre space generated in the current study extends previously generated timbre spaces obtained from impacted materials to musical instrument tones. McAdams et al. (2004) found that listeners were sensitive to the changes in the damping properties and pitches of impacted bars, as revealed by a two-dimensional timbre space. The damping- and frequency-related dimensions corresponded to the manipulations in the physical model of viscoelastic damping and mass density/bar length, respectively. In McAdams et al.'s (2010) simulation of struck plates representing a continuum of materials between glass and metal, MDS on the dissimilarity ratings produced a two-dimensional timbre space. Changes in wave velocity were differentiated based on a frequency-related dimension, whereas variation in the damping properties (i.e., interpolations between viscoelastic [glass] and thermoelastic [metal] damping) was detected based on a dimension related to timbre and duration perception. Hjortkjær and McAdams (2016) also reported two dimensions in their timbre space of sounds produced by combining three actions to objects made of three materials. One dimension distinguished the actions and was correlated with temporal centroid, while the other differentiated materials and correlated with spectral centroid. The aforementioned studies along with the current study directly link timbre space dimensions to the varying mechanical components of sound sources, but the current study is the first to demonstrate the association using musical tones.

Interestingly, the discernability of the sounds based on excitations and resonators did not predict the assimilation patterns observed in previous categorization tasks by Huynh and McAdams (2023a). Harnad (1990) has explained that the methods involved in measuring dissimilarity perception are independent of those that measure categorization performance. Comparison of the findings from Huynh and McAdams's (2023a) study and the current study demonstrates that the different methods of measuring dissimilarity and categorization can lead to different perceptual and behavioral responses to the same stimulus set. Unlike what was proposed by Goldstone (1994) and Sloutsky (2003), the categorization performance seen in Huynh and McAdams (2023a) was not necessarily a function of similarity perception in the current study because the two processes seemed to rely on different information, depending on what was relevant for the task. Perceived mechanical plausibility interfered with categorization in the experiment by Huynh and McAdams (2023a) as implied by the assimilations of atypical to typical interactions. It was difficult for listeners to categorize the excitations and resonators of interactions that could not be conceptualized by listeners. However, perceived

mechanical plausibility had very little to do with dissimilarity ratings in the current study, because dissimilarity ratings involve relative judgments that are local, holistic, and based on whatever is available to discern the sounds; this judgment did not concern knowing how the sounds were produced or if the ways that they were produced were mechanically plausible. Pérez-Gay et al. (2017) proposed that similarity perception predicts categorization performance once categories are learned. Categorization involves the detection of invariant features that reliably determine category membership. Learning can enhance the detection of invariant features, which would then improve categorization performance. Consequently, items belonging in the same category will be perceived as more similar, whereas items belonging to different categories will be perceived as more dissimilar (Pérez-Gay et al., 2017). The current study shows that the invariant features that dictate membership to excitation, resonator, and interaction categories can be detected implicitly. So, it would be of interest to examine whether a learning task encourages the explicit detection of these invariants to improve categorization performance and reduce the previously observed assimilations of Huynh and McAdams (2023a).

The fact that different excitations and resonators occupied separate regions on the timbre space suggests that the different mechanical components were perceived differently from one another. There was even a vague distinction between bowing and blowing, which are both sustained excitations that are often confused for one another. Moreover, Ollivier et al. (2004) compared the physical models for these excitations and demonstrated their interchangeability. In the process of synthesizing the atypical interactions in the original study (Huynh, 2019), it was difficult to judge how well the excitations and resonators were conveyed given that these interactions cannot be reproduced in the physical world. That different categories of the mechanical components occupied their own regions in the timbre space potentially speaks to the efficacy of Modalys in producing distinctive results in the simulation of the atypical interactions. Furthermore, it validates Modalys's use of different physically inspired models to simulate bowing and blowing. The results from the current study can bridge the gap between physically inspired modeling approaches and human perception of the synthesized sounds. Perception of sounds produced by physically inspired modeling can inform its efficacy and limitations.

3.5 Conclusion

In the current study, different excitation mechanisms and resonant structures were discernable by listeners even when they were combined in atypical contexts. The analysis of the dissimilarity data generated a timbre space comprising three dimensions. Exemplars produced by the same excitation-resonator interactions occupied their own regions in the timbre space. Additionally, distinctions among excitations and resonators required one and two dimensions, respectively. This suggests that resonator distinction is more complex than excitation distinction, at least for the current stimulus set. Still, the current findings imply that listeners can detect the transformational or structural invariants among sounds produced by the same excitations or resonators, respectively. These invariants are the audio descriptors correlated with each dimension. Given that the acoustic correlates were estimated based on model selection of a subset of audio descriptors, it is possible that the right combination of audio descriptors was not determined to fully capture the acoustic distinctions among the different mechanical components. Despite this, the current findings highlight the usefulness of dissimilarity perception because listeners can rate the dissimilarity of sounds without knowing their identities or recognizing them. The comparison of sounds can rely on characteristics that listeners are unable to describe explicitly. Listeners might have formed mental models or internal representations of the atypical interactions implicitly and on the basis of acoustic cues, as reflected by the organization of sounds in the timbre space. Listeners might be able to explicitly categorize the atypical interactions based on their excitation, resonator, and interaction categories if they are trained to learn them. Our study therefore emphasizes timbre as multidimensional attribute contributing to the discernability everyday sounds.

3.6 Appendix

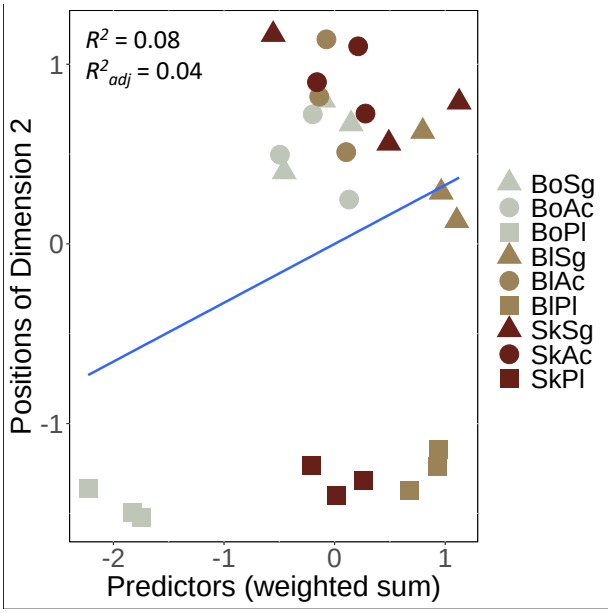

Figure 3.5 Scatter plot showing the relationship between the weighted sum of acoustic properties without median spectral centroid and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

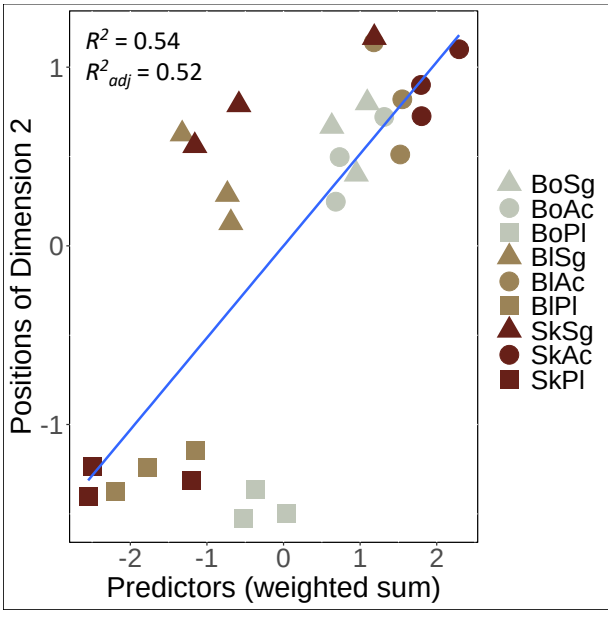

Figure 3.6 Scatter plot showing the relationship between the weighted sum of acoustic properties without median spectral crest and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

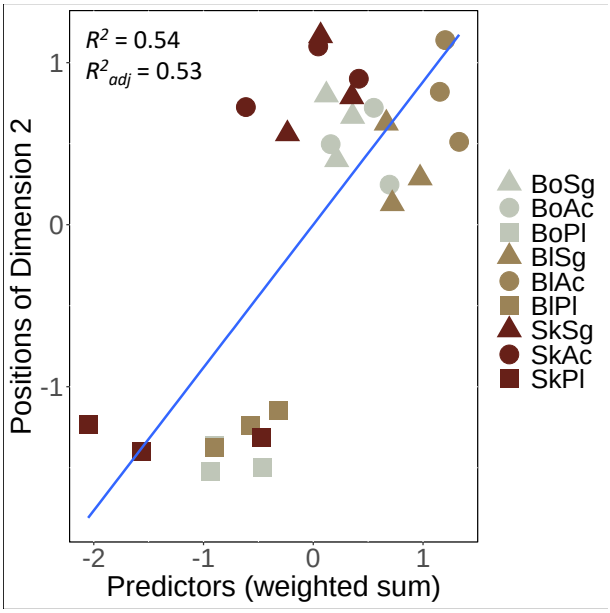

Figure 3.7 Scatter plot showing the relationship between the weighted sum of acoustic properties without the amplitude of energy modulation and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

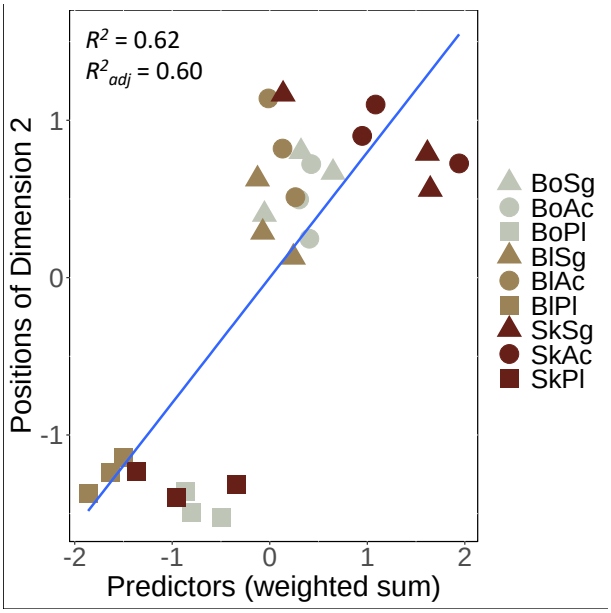

Figure 3.8 Scatter plot showing the relationship between the weighted sum of acoustic properties without the IQR of spectral flatness and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

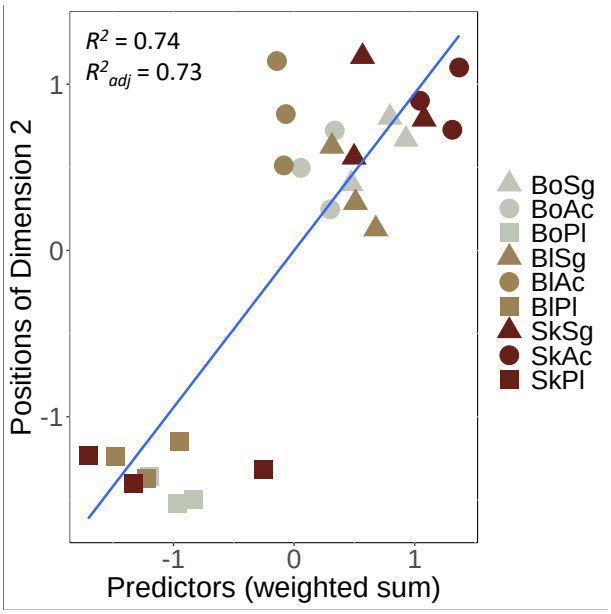

Figure 3.9 Scatter plot showing the relationship between the weighted sum of acoustic properties without median harmonic spectral deviation and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

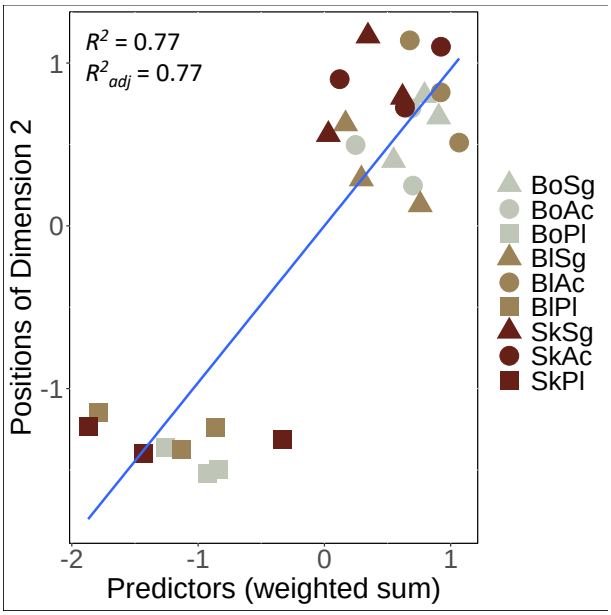

Figure 3.10 Scatter plot showing the relationship between the weighted sum of acoustic properties without the frequency of energy modulation and the positions of the stimuli on Dimension 2. The blue line represents the regression line.

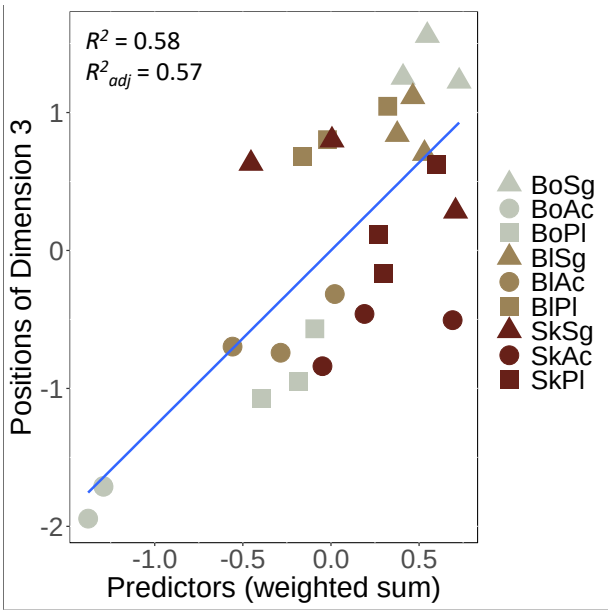

Figure 3.11 Scatter plot showing the relationship between the weighted sum of acoustic properties without median tristimulus 2 and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

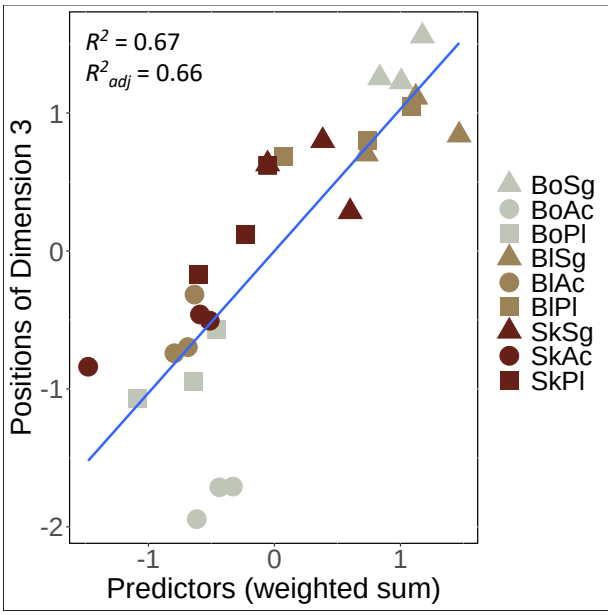

Figure 3.12 Scatter plot showing the relationship between the weighted sum of acoustic properties without the IQR of root mean square energy and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

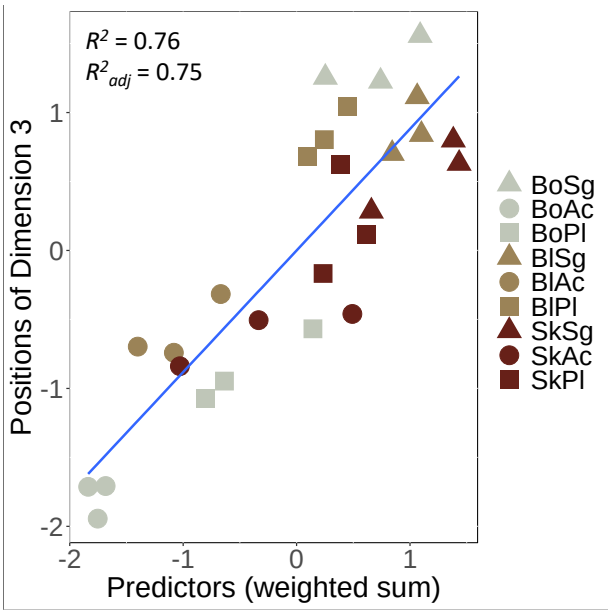

Figure 3.13 Scatter plot showing the relationship between the weighted sum of acoustic properties without the IQR of spectral flatness and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

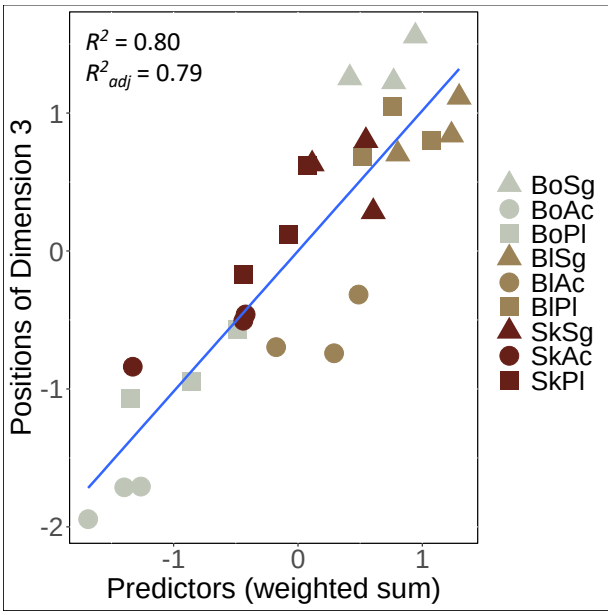

Figure 3.14 Scatter plot showing the relationship between the weighted sum of acoustic properties without median harmonic spectral deviation and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

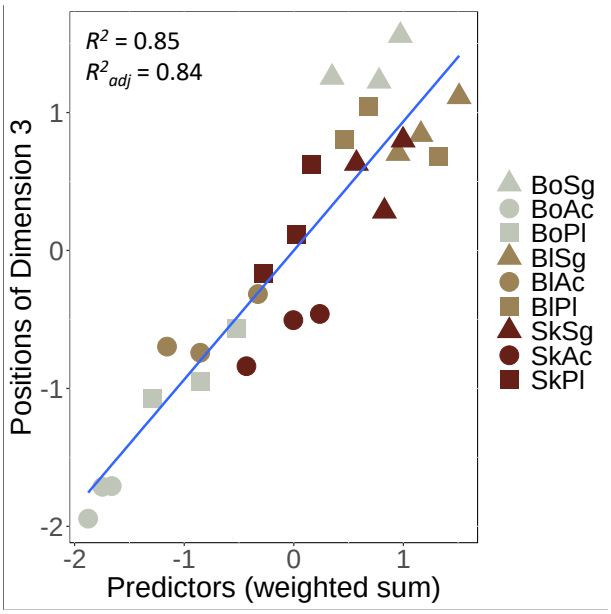

Figure 3.15 Scatter plot showing the relationship between the weighted sum of acoustic properties without the frequency of energy modulation and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

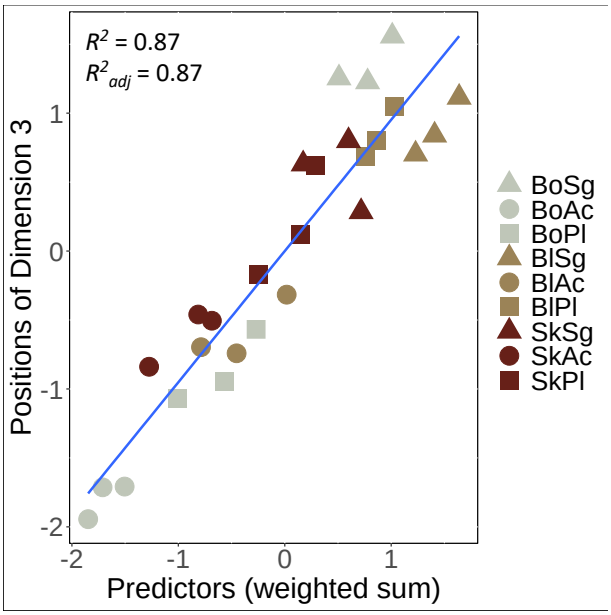

Figure 3.16 Scatter plot showing the relationship between the weighted sum of acoustic properties without the amplitude of energy modulation and the positions of the stimuli on Dimension 3. The blue line represents the regression line.

Chapter IV

Learned categorization of atypically combined excitations and resonators of musical instruments

This chapter is based on the following research article:

Huynh, E. Y., and McAdams, S. (in preparation). Learned categorization of atypically combined excitations and resonators of musical instruments. Manuscript intended for submission to *PLoS ONE*.

Abstract. The goal of this study is to determine whether listeners can create new mental models of unfamiliar sound sources. Synthesized stimuli comprised nine types of simulated interactions between three excitations (bowing, blowing, striking) and three resonators (string, air column, plate). Three groups of participants were each trained on the excitation, resonator, or interaction categories of the interactions in separate three-phase learning paradigms. The familiarization phase allowed listeners to associate category names with their sound examples. The training phase comprised a series of blocks of nine trials, each corresponding to one type of interaction. Listeners had to choose the correct category in at least 75% of the trials for four blocks in a row within a maximum of 23 blocks. Corrective feedback was provided in each trial. Participants that successfully completed the training phase continued to the testing phase, which involved categorization without corrective feedback. Categorization performance was above chance for both typical interactions (e.g., bowed string) and atypical interactions (e.g., blown plate) across excitation, resonator, and interaction categorization. Listeners consequently transferred short-term learning effects from training to testing. Instead of assimilating atypical interactions to the mental models of typical interactions, listeners may have formed new mental models for the atypical interactions.

Keywords: Timbre perception, sound source recognition, music perception, categorization, learning, cognitive science

4.1 Introduction

The recognition of everyday sounds, such as those produced by musical instruments, would be impossible without considering their timbres. Timbre is a multidimensional attribute of sound comprising a plethora of acoustic properties that allow listeners to identify sound sources. For acoustic musical instruments, timbre provides listeners with information about two interacting mechanical components: the *excitation mechanism* and *resonant structure*. Interactions between excitations and resonators are limited in the physical world, reflecting their close relationship. For example, strings can be bowed and struck, but are rarely blown. An interaction is therefore defined by a coupling process: an excitation mechanism sets into vibration a resonant structure by allowing a controlled input of energy into it. The resonator functions as a filter that radiates, suppresses, and amplifies sound components. The resonators of interest to the current study are a string, air column, and plate. The excitations of interest are bowing a frictional bow, blowing a vibrating single reed, and striking with a

hammer. Interactions that listeners are familiar with—as they are common to those of acoustic musical instruments—are bowed strings, blown air columns, and struck strings and plates. In this study, these interactions are deemed typical interactions. The two typical interactions produced by sustained excitations (i.e., bowed string and blown air column) have couplings that are nonlinear, such that an increase in the input disproportionately increases the output (McIntyre et al., 1983). The couplings of struck plates and strings are not considered nonlinear, with the exception of the initial hammer contact of the struck string (Fletcher, 1999). The specific ways in which excitations and resonators interact influence their identification (Huynh & McAdams, 2023a).

4.1.1 Sound Source Recognition

The recognition of musical instruments has been studied for some time. Some instruments are more easily recognizable than others. For example, Saldanha and Corso (1964) found that the identification of clarinets, oboes, and flutes was more accurate than that of the violin, cello, and bassoon. In Berger's (1964) study, which focused on the recognition of wind instruments, listeners more accurately identified the oboe, clarinet, cornet, and tenor saxophone than the flute, trumpet, alto saxophone, bassoon, French horn, and baritone. Confusion data revealed that participants occasionally made confusions between instruments belonging to the same wind instrument family (e.g., lip valves, single reeds, etc.). Recognition accuracy increased if these within-family confusions were noted as correct responses. Similarly, in a review of musical instrument identification studies, Giordano and McAdams (2010) found that participants confused instruments belonging in the same family or played by similar types of excitations. Furthermore, McAdams et al. (2023) tested the categorization of tones and sequences from 11 musical instruments. In line with previous studies, confusions were made within excitation categories of impulsive and sustained excitations; confusions were never made across excitation categories. Confusions within instrument families were also reported: English horn with tenor saxophone and clarinet (woodwinds); tuba and trombone (brass); harp and guitar (plucked strings). These confusions also depended on the registers of the instruments, demonstrating that instrument identification is best when the pitch of the tones reflects what is typical of the musical instrument. For this reason, the stimuli in the current study will be of the same pitch and rather than identifying specific instruments, listeners will learn to identify more general categories such as types of resonators and the excitations that set them into vibration.

4.1.2 Material and Action Perception of Impacted Objects

Previous studies have also focused on the identification and dissimilarity perception of materials in impacted sounds (Aramaki et al., 2009; Giordano & McAdams, 2006; Klatzky et al., 2000; Lutfi & Oh, 1997; McAdams et al., 2004; McAdams et al., 2010). Fewer studies have examined the identification of multiple impacts such as bouncing (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012; Warren & Verbrugge, 1984), despite action identification being more robust than material identification (Lemaitre & Heller, 2012). These studies are of interest to us given that actions are synonymous with excitations and materials are an aspect of resonators.

Differentiation of materials across gross categories of metal-glass and plexiglass-wood is usually quite accurate (Giordano & McAdams, 2006). However, listeners have more difficulty differentiating materials within the same gross category (Lemaitre & Heller, 2012). A notable exception was observed in a study by McAdams et al. (2010), which investigated the distinction between simulated plates representing a continuum of materials between metal and glass that were struck by mallets made of different materials. In one of the tasks, listeners categorized the material of the sounds. Listeners seemed to rely on damping properties of the sounds, attending more to the interpolations between thermoelastic (aluminum) and viscoelastic (glass) damping models than to other varying parameters such as the wave velocity or those distinguishing the material of the mallet. With respect to action identification, Warren and Verbrugge (1984) found that listeners were able to distinguish glass objects that were broken and bounced. Listeners can also differentiate larger rolling balls from smaller ones (Houben et al., 2004). Different speeds can also be determined, depending on the size of the balls.

More recently, research has focused on material and action perception in sounds that are produced by applying different types of actions to objects made of different materials. Lemaitre and Heller (2012) recorded cylinders made of four materials (wood, plastic, metal, glass) that were scraped, rolled, hit, and bounced. Listeners rated each interaction based on how well they conveyed the different action and material categories. They also categorized sounds based on materials or actions. Action identification was accurate as indicated by higher resemblance ratings to the actions that actually produced the sounds, as well as correct and faster categorization. Resemblance ratings and categorization of the materials demonstrated that listeners confused materials within gross categories (metal-glass vs. wood-plastic), but less confusions were made between gross categories. Material categorization was also slower than action categorization. In a similar synthesis design, Hjortkjær and McAdams (2016) recorded combinations of three actions (strike, drop, rattle) and three materials (wood, metal, glass). For the dissimilarity ratings of the sounds, multidimensional scaling (MDS)

revealed a two-dimensional space. Dimension 1 separated the different materials (wood vs metal-glass), and its acoustic correlate was the spectral centroid (i.e., center of gravity of frequency distribution). The three actions were separated by Dimension 2 and its acoustic corelate was the temporal centroid (i.e., center of gravity of the energy distribution across time). Two digital manipulations of the original stimulus set were also generated: one manipulation preserved spectral cues and removed temporal cues, and the second manipulation preserved temporal cues and removed spectral cues. Stimuli from the original and each manipulated set were categorized based on their actions or materials. Consistent with Lemaitre and Heller (2012), the identification of actions was more accurate than material identification. Material identification was better between than within gross material categories. Furthermore, action identification was less accurate when temporal cues were removed, whereas material identification was less accurate when spectral cues were removed. Categorization performance therefore verified the acoustic correlates of the two-dimensional timbre space observed from dissimilarity perception. Hjortkjær and McAdams (2016) demonstrated that listeners were able to isolate the invariant features (i.e., acoustic properties) that differentiate the action categories and material categories. Overall, these studies demonstrate a greater sensitivity to actions than materials of sound sources as well as the informational value of the acoustic cues of timbre in sound source recognition.

4.1.3 Perception of Mechanical Components in Musical Sounds

A previous study (Huynh, 2019) extended the stimulus design of Lemaitre and Heller (2012) and Hjortkjær and McAdams (2016) to the context of excitations and resonators of acoustic musical instruments. Huynh (2019), using a physically inspired sound synthesis platform called Modalys (Dudas, 2014), simulated interactions between three excitations (bowing, blowing, striking) and three resonators (string, air column, plate), which formed nine types of interactions much like the stimuli used in the current study. Huynh and McAdams (2023a) classified four of the nine excitation-resonator interactions as typical interactions: bowed string (BoSg), blown air column (BlAc), struck string (SkSg), and struck plate (SkPl). These interactions are typical because they are representative of acoustic musical instruments and listeners are familiar with them. They are also perceived as mechanically plausible because listeners can conceptualize: (1) the sensorimotor activity that is required to excite the resonator and (2) the timbres of the sounds that can be produced by the interactions. The remaining five interactions—bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc)—were classified as atypical by Huynh and McAdams

(2023a). They note that BoPl and SkAc can be more familiar and considered mechanically plausible to musicians in comparison to nonmusicians, especially in the context of extended playing techniques in contemporary music. However, a plate would have to be bowed at its edge for the interaction to be considered mechanically plausible; Modalys instead simulates the interaction by passing the bow through the plate. SkAc would be mechanically plausible, much like a slap tongue technique, if striking was applied to a mouthpiece that is connected to the top of the air column; but Modalys strikes the air column at a point along its length and not through a mouthpiece. So, the way that these two atypical interactions, along with BoAc, BlSg, and BlPl, are simulated in Modalys are physically impossible. Listeners would consequently be very unfamiliar with them and unlikely to be able to conceptualize the sensorimotor activity involved in sound production and the resulting timbres.

Listeners rated the resemblance of the nine interactions to each excitation and each resonator in one study (Huynh, 2019). In another study, listeners categorized the interactions based on the excitations and resonators they thought produced the sounds (Huynh & McAdams, 2023a). Typical interactions were assigned higher resemblance ratings to the excitations and resonators that actually produced the sounds. Categorization performance was also very accurate for the typical interactions, with correct excitations and resonators chosen more often than incorrect ones. For the atypical interactions, resemblance ratings and categorization performance revealed that they were assimilated to typical interactions. Huynh and McAdams (2023a) suggested that listeners learned to assimilate the atypical interactions to typical ones in an unsupervised manner. That is, through repeated exposure and because of the lack of corrective feedback, listeners detected similar features between the atypical interactions and the typical ones to which they were assimilated in order to make their categorization judgments. One of the aims of the current study is to examine the effect of supervised learning on the categorization of the same interactions.

In a separate study using the same stimulus set, listeners rated the dissimilarity of pairs of sounds (Huynh & McAdams, 2023b). MDS based on the dissimilarity ratings revealed a three-dimensional timbre space. Dimension 1 separated striking from the two sustained excitations and was correlated with temporal centroid. So, temporal centroid was proposed to be the primary transformational invariant that distinguished between the different excitations. Dimension 2 isolated the plate from the other two resonators and was associated with global spectral shape, tonal content, and noise content of the spectrum (i.e., spectral centroid, spectral crest, variability of spectral flatness) in addition to the energy modulation of the signal (e.g., depth and frequency of amplitude modulation). Dimension 3 further separated the string and air column and was best explained by the finer details of the spectrum

(i.e., tristimulus 2, harmonic spectral deviation), the variability of the signal's energy (i.e., root mean square energy), the variability of the spectrum's noisiness (i.e., spectral flatness), and the energy modulation of the signal. Huynh and McAdams (2023b) suggested that there were two sets of structural invariants involved in the differentiations of resonators. The motivation of the current study is to examine whether these transformational and structural invariants can be detected and learned through supervised training, to reduce the assimilations observed in the previous tasks that involved categorization and resemblance ratings.

4.1.4 Categorization and Supervised Learning

Additionally, we are interested in investigating the formation of categories with respect to musical instrument sounds. According to Rosch (1978), listeners form categories on the foundation of two major principles. The first is based on cognitive economy: gaining the most information from the environment with as little cognitive effort as possible. This means things belonging in the same category are perceived as equivalent and those not in the same category are perceived as different. Moreover, differentiations between stimuli can be ignored if the differentiation has no purpose for categorization. The second principle has to do with structuring the perceived world. This concerns what can be perceived in the world, given that humans are sensorimotor systems interacting with the world through what their sensory surfaces can afford (Harnad, 2017). Sensorimotor systems may not detect all features of a given stimulus, and they may place an emphasis on some features over others. When these features co-occur and are eventually considered highly correlated through frequent exposure, listeners detect these patterns of co-occurring features as invariants (Harnad, 1987b; 2017). In other words, invariant features distinguish members of one category from another. Invariant features of a category are therefore present in all members of that category and absent in members of different categories (Harnad, 2017).

Harnad (2017) describes categorization as being closely related to learning, given that most categories are learned rather than innate. Innate feature-detectors that determine category membership are acquired through evolution. For learned categories, however, invariants are detected through supervised learning, which involves trial and error with corrective feedback. This means that, through exposure, the invariant features of a category are detected and selectively attended to, whereas non-distinguishing features are ignored or given less attention. Additionally, knowing the correct output when an incorrect output is produced is important for categorization. Categorization is both defined and learned by determining right from wrong through corrective feedback. Furthermore, Kruschke

(2005) explains that what we do with an encountered item or instance is determined by its category. So, to categorize is to "do the right thing with the right kind of thing" (Harnad, 2017, p. 22). Supervised learning can also be enhanced by assigning labels or words to the category. Faster and more robust learning was observed when novel categories were given labels than when they were not (Lupyan, 2006). The categories of the current study will be labeled based on the mechanical components that were simulated to produce the sounds and the interactions between them. We note that the two mechanical components (i.e., excitations and resonators) are technically features of the nine interactions. However, features can potentially be categories with respect to which members and non-members are defined by higher-order features (Pérez-Gay et al., 2017). These higher-order features include acoustic properties of sounds produced by different interactions. Accordingly, the excitations, resonators, and interactions will henceforth be defined as categories and the labels of the sounds will differ depending on the categories that the learning tasks of the current study are based on.

4.1.5 The Current Study

We are interested in the role of supervised learning, if any, in improving categorization performance of the atypical interactions. Successful learning would imply that new mental models are formed for the atypical interactions. Mental models are internal representations of how systems such as musical instruments work in the world. They are shaped by exposure. Musicians will be more likely than nonmusicians to be familiar with the many types of sounds that can be produced by different musical instruments. This familiarity is acquired through sensorimotor interactions with their own instrument and by listening attentively to other instruments around them in practice or performance settings. Nonmusicians' mental models of musical instruments are therefore shaped more passively given that they would be less familiar with the restrictions or extended techniques of musical sound production.

The main question of interest in the current paper is whether categorization learning theory can apply to mechanically implausible sounds and therefore reduce previously observed confusions. We compared categorization performance based on supervised learning of the three excitations (Experiment 1), three resonators (Experiment 2), and nine interactions (Experiment 3) of the stimuli. Each experiment comprised a three-phase learning paradigm adapted from McAdams et al. (2023). In the first phase, called the familiarization phase, listeners heard examples of labeled sounds to become familiar with their category membership. Then, the training phase contained a series of blocks. Listeners had to reach a passing threshold of 75% correct for four blocks in a row within a maximum

of 23 blocks. The training phase was based on supervised learning proposed by Harnad (2017) and employed by McAdams et al. (2023). It involved trial and error and corrective feedback, such that for every response a participant made, they were told the correct answer. Participants who passed the training phase then completed a testing phase. The testing phase comprised a task that was similar to the training phase, but corrective feedback was no longer provided to examine if learning was maintained from the training phase.

We propose a few possible predictions, some of which rely on whether the atypical interactions can be learned based on their excitations, resonators, and/or their combination. First, we expect fewer participants to fail the training phase for excitation categorization. Supervised learning for acoustic stimuli was shown to be most effective for changes within one dimension in comparison to multidimensional changes (Goudbeek et al., 2009). Because excitation differentiation was mapped onto a single dimension correlating primarily with one acoustic property (Huynh & McAdams, 2023b), excitation training might be more successful than resonator or interaction training. We also expect fewer participants to fail the training phase based on resonator learning relative to interaction learning. Goudbeek et al. (2009) found supervised learning to be less effective for stimuli varying along two dimensions, and the differentiation of resonators was previously associated with two dimensions (Huynh & McAdams, 2023b). Moreover, each of these dimensions was correlated with a weighted sum of multiple audio descriptors, which might make the resonators more difficult to learn. Consequently, more participants will be expected to fail interaction training than excitation or resonator training, given that interaction differentiation involves the combined dimensions of excitation and resonator differentiation, which means three dimensions in total.

We expect categorization performance in the testing phases of each learning task to reflect that of the training phases: overall categorization accuracy will be best for excitations, followed by resonators, and lastly interactions. However, based on the categorization performance in Huynh and McAdams (2023a), even without knowing how the atypical interactions were produced, listeners identified either their excitations or resonators more correctly. For example, resonator categorization was better than excitation categorization for the bowed air column, bowed plate, and blown string. In contrast, listeners categorized the excitations more accurately than the resonators of the blown plate and struck air column. So, we expect these patterns to be maintained in the current study. Of interest is whether excitation categorization will improve for the bowed air column, bowed plate, and blown string and whether resonator categorization will improve for the blown plate and struck air column.

One possibility in the findings is that the atypical interactions cannot be learned. If this were the case, then even if participants pass the training phase, testing phase performance will reflect that the atypical interactions are assimilated to the typical ones, similar to the findings of our previous study (Huynh & McAdams, 2023a). This would imply that there are short-term learning effects during training, but when corrective feedback is removed during the testing phase, perceived mechanical plausibility interferes with any learning effects. That is, particularly with interaction categorization, the typical interactions that the atypical ones were assimilated to (in Huynh & McAdams, 2023a) will be chosen more often in the testing phase: blown air column for bowed air column and blown plate; bowed string or struck plate for bowed plate; bowed string for blown string; and struck string or struck plate for struck air column. Moreover, for excitation categorization and resonator categorization, this would mean that the percent response data would be similar to Figure 2.2 in Huynh and McAdams (2023a). These results would be inconsistent with the categorization learning theory stating that categories can be learned with trial and error and corrective feedback (Harnad, 2017).

However, if the atypical interactions can be learned, then we would expect that the percent correct scores in the testing phase would be greater than chance performance of each categorization task. The percent response data would reflect very minimal confusions and the correct excitations, resonators, and interactions would be chosen more often than incorrect ones. The possibility of learning can also depend on the interaction, as some might be easier to learn if they have distinct or characteristic timbres (McAdams, 1993). Furthermore, learning might be predicted by participants' musical experience. To account for various behaviours or skills that can indicate one's musical expertise, each participant completed the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefan et al., 2014). We chose not to categorize participants as musicians and nonmusicians as the criteria defining musicianship in scientific studies are quite restrictive and difficult to generalize to the population. Participants often qualify as musicians if they have a certain level of formal musical training, which does not always account for self-taught musicians. Nonmusicians are usually defined as individuals who have no more than one year of formal musical training before the age of 12 and have not played a musical instrument since then. It is quite uncommon to find participants who meet these criteria, and it might not be representative of nonmusicians in the general population (i.e., those who have briefly played musical instruments after the age of 12 can still be self-reported nonmusicians). On the other hand, the Gold-MSI contains a general musical sophistication score as well as scores for subscales of active musical engagement, perceptual abilities, musical training, singing abilities, and sophisticated emotional engagement (Müllensiefen et al., 2014). The musical training and general

musical sophistication scores are of particular interest to us. The musical training subscale includes items pertaining to years of formal music training, number of hours of practice, number of instruments played, years of music theory training, and so on. The general musical sophistication scale considers the variability of individual experiences in the assessment of musical experience; the calculation of its score includes items from each subscale. We did not consider the other subscales as they did not seem relevant to the differentiation of the mechanical components of musical instruments: perceptual abilities refer to the accuracy of listening skills (i.e., discussing differences between performances, recognizing if someone is singing out of tune, etc.); active engagement is about the time and resources one spends on music; singing abilities indicate one's singing accuracy; and emotional engagement concerns the discussion of emotional expression in music. Consequently, we examine whether higher musical sophistication and/or musical training scores will predict better categorization performance. If learning does transfer from the training phase to the testing phase, we would argue that the interference of perceived mechanical plausibility becomes diminished with supervised learning. Either one of these findings would provide insight on whether new mental models are generated for the atypical interactions and highlight timbre's role in sound source recognition.

4.2 General Method

4.2.1 Participants

Across the three experiments, 132 participants were recruited from either a mailing list or web-based advertisement certified by McGill University. Each participant completed a pure-tone audiometric test with octave-spaced frequencies from 125 to 8,000 Hz at a hearing threshold of 20 dB HL relative to a standardized hearing threshold (ISO 398-8, 2004; International Organization for Standardization, 2004; Martin & Champlin, 2000). Of the 132 recruited participants, 127 met the hearing thresholds of the audiometric test. They had an average age of 23.3 years (SD=4.4). Participants provided written informed consent and were compensated for their participation. This study was certified for ethical compliance by the McGill University Research Ethics Board II.

Participants were randomly assigned to one of three experiments upon signing up for the study. Random assignment to Experiment 1, 2, or 3, determined whether participants learned to categorize the stimuli based on their excitations, resonators, or interactions, respectively. Over the course of the study, there were less and less individuals taking interest in signing up for participation, so we asked participants who completed one version of the experiment to return and participate in a version they

had not done. All participants who met the following two criteria were invited back to participate in another version of the experiment. First, their previous participation had to be over four months ago. Four months was considered enough time for participants to remember very little of the experimental task and stimuli. Second, the participants must have passed the training phase of the version of the experiment in which they participated. If participants agreed to return, they were randomly assigned to one of the other two versions of the experiment they had not completed. A summary of the information about the returning participants is presented in Table 4.1. The participant with the shortest duration between participations in Experiments 1 and 2 did not pass the training phase of Experiment 2. So, the next shortest duration between participations was 150 days (4 months 28 days); this was for a participant who also participated in Experiments 1 then 2.

Table 4.1 The number of returning participants based on the order in which they completed two versions of the experiment and how many days passed between their participations.

First participation	Second participation	Number of returning participants	Number of days between participations
Experiment 1	Experiment 2	3[a]	140–182
	Experiment 3	1[b]	163
Experiment 2	Experiment 1	2	171–181
	Experiment 3	1	154
Experiment 3	Experiment 1	2	165–191
	Experiment 2	3	156–184

[a] One of the three participants did not pass the training phase in Experiment 2.
[b] This participant did not pass the training phase in Experiment 3.
Note. First and second participation indicate which experiment they originally participated in and the experiment for which they returned, respectively.

4.2.2 Apparatus

Listeners completed the experiment in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). The experiment ran on a Mac Pro computer running OSX (Apple Computer, Inc., Cupertino). Stimuli were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented over Seinnheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The physical levels of the sounds were measured by coupling the headphones to a Bruel and Kjær Type 4153 Artificial ear connected to a Type 2205 sound-level meter (A-weighting; Bruel & Kjær, Nærum, Denmark). The levels of the sounds ranged

from 61 to 67 dB SPL. The experimental interface was programmed in the PsiExp computer environment (Smith, 1995).

4.2.3 Stimuli

The experiments contained seven exemplars for each of nine types of excitation-resonator interactions, forming a total of 63 stimuli. We detail the synthesis paradigm in a previous study (Huynh, 2019), but a summary and any modifications are presented here. We used Modalys (Dudas, 2014), a digital physically inspired modeling software developed by The Music Acoustics Team at the Institut de recherche et coordination acoustique/musique (IRCAM) in Paris, France. Modalys can simulate different excitation mechanisms and resonant structures without the resulting sound necessarily being perceived as an existing musical instrument (Eckel et al., 1995; Dudas, 2014). Modalys uses modal synthesis to model the acoustical outcome between an excitation and a resonator.

We generated nine classes of interactions between three excitation mechanisms (bowing, blowing, striking) and three resonant structures (string, air column, plate). Given that the atypical interactions are physically impossible and mechanically implausible (e.g., bowed air column, blown string, blown plate), or rarely encountered (e.g., bowed plate, struck air column) in everyday musical experiences, it was difficult to anticipate how they would sound. Moreover, physically inspired modeling of atypical interactions is quite uncommon (for notable exceptions, see Böttcher et al., 2007 for musical sounds; and Conan et al., 2014 for continuously excited objects). Consequently, we used an exhaustive approach to synthesize 400 versions of each type of excitation-resonator interaction, which allowed us to choose from a variety of timbres that could be perceived as conveying the source components that produced the sounds.

4.2.3.1 Resonant Structures

The three resonators—string, air column, and plate—were made up of completely different models. For the air column, Modalys's tube object simulates the acoustic modes of an air column with particular boundary conditions at its ends. We took into consideration the assumption that a string excited at a short distance from the bridge and a conical air column can be modeled similarly. Therefore, the string was fixed at both ends, whereas the air column was cylindrical (not conical), open at one end, and closed at the other. This should guarantee that these two resonators will sound different, even when the same excitation is applied to each of them. The string and air column additionally differ in their harmonic content: the former has modes that vibrate at integer multiples of

the fundamental frequency (i.e., even and odd harmonic content), and the latter has modes vibrating with greater energy at odd harmonic ratios than at even harmonic ratios with respect to the fundamental. The plate was thin, rectangular, and fixed at its edges. Its harmonic content should primarily be inharmonic unless a sustained excitation is applied to it. Each resonator was synthesized to produce a lowest vibration mode of 155 Hz, corresponding to a pitch of E-flat-3. Consequently, the parameters chosen for each resonator ensured vibration at this pitch and were kept as consistent as possible across all types of excitation mechanisms that were applied to them. Only minor exceptions were made to achieve a better sound quality; they were explained in our previous study (Huynh, 2019).

4.2.3.2 Excitation Mechanisms

Three excitation mechanisms were simulated: bowing, blowing, and striking. As much as possible, the same temporal envelopes controlling the parameters of each excitation were applied to each resonator. For each type of excitation, we manipulated two parameters that have been known to influence the resulting timbre significantly (Halmrast et al., 2010). The bowing excitation was primarily made up of the control of the bow speed and bow pressure (i.e., modeled as the vertical displacement of the string by the bow). We first applied the bowing excitation to the string and modified the time values of separate temporal envelopes for the bow speed and bow pressure to make the bowing sound as realistic as possible. We worked with the string first, given that bowed strings are typical of acoustical musical instruments. Once realistic temporal envelopes of the bow's speed and pressure were obtained for the string, we applied the same temporal envelopes to the air column and plate. We tested and combined 20 values each for the maximum bow speed and maximum bow pressure when bowing was applied to each of the three resonators.

For blowing, we simulated a mouth and reed with Modalys, such that vibration of the reed resulted in an oscillating flow of air, which could then set a resonator into vibration. We controlled separate temporal envelopes of the breath pressure and valve-zeta, which describes how the pressure applied by the lips controls the physical resting position or opening of the reed and is hereafter referred to as the embouchure pressure (Coyle et al., 2015). We synthesized blown air columns first and then applied the same temporal envelopes of breath and embouchure pressure to the string and plate. Minor changes to the temporal envelope of breath pressure were made to prevent the resulting sound from resembling a squeak. We tested and combined 20 values each for the maximum breath pressure and maximum embouchure pressure when blowing was applied to each resonant structure.

We synthesized a hammer for the striking excitation. Struck sounds did not correspond to an auto-oscillating coupling that is an attribute of the synthesis of bowed and blown sounds. Perceptual outcomes of auto-oscillation sounds depend on the manipulation of parameters given to nonlinear coupling. For struck sounds, these parameters do not exist. Consequently, we manipulated parameters that affected the resulting timbre significantly, such as the force of the hammer, i.e., modeled as the vertical displacement of the resonator by the hammer. In our original synthesis of the struck sounds (Huynh, 2019), we also manipulated the output positions of the sound from the resonator. However, for the current study, this manipulation was changed to the position on the resonator at which the hammer comes to contact with it. For the plate, it was the normalized horizontal and vertical coordinates of the access position, whereas for the string and air column, it was the normalized length along the string or air column. The timbres produced by this modification did not differ drastically from when the output positions were manipulated; however, this modification was more compatible with theories explaining how timbre varies with respect to where the resonator is struck due to the activation of different modes at different striking positions (Halmrast et al., 2010). No additional changes were considered for the striking excitation. We used the same temporal envelope for the control of the hammer force to strike each resonator. Struck strings and plates were synthesized first because they are more common with respect to acoustic musical instruments. Then the same temporal envelope was applied to the air column. For the string and air column, we tested and combined 20 values for the maximum hammer force and 20 values corresponding to the normalized position at which they were struck. For the struck plate, we tested and combined 20 horizontal coordinates and 20 vertical coordinates at which the plate was struck. Changing the maximum hammer force for the plate did not significantly impact the timbre of the output as Modalys normalizes its amplitude.

4.2.3.3 Final Stimulus Set

The authors informally recorded the perceptual outcomes of the 400 sounds (i.e., 20 values of one parameter × 20 values of another parameter) of each interaction and whether an audible output was produced. Additionally, the seven chosen exemplars for each interaction type were perceived to be produced by the same source components and the most variable in their timbres among the sounds that also conveyed the same excitation and resonator. The process of selecting exemplars represented the fact that a performer can play a single note of a musical instrument in different ways, and there will be variability in the timbre of each of the sounds (McAdams & Goodchild, 2017). Each sound had a duration of 2 s.

4.2.4 Procedure

After obtaining informed consent from participants, they were seated in the audiometric booth in front of a computer. Depending on the version of the experiment the listener participated in, we introduced the definitions of the three excitations (Experiment 1), three resonators (Experiment 2), or both (Experiment 3). The definitions are presented in Table 4.2. Before beginning the main experiment, listeners were given written and verbal instructions of the experimental task. They were encouraged to ask questions for clarification given that this experiment did not have a practice block. There were three phases in the main experiment: (1) familiarization, (2) training, and (3) testing. The structure of the experiment was largely inspired by the experimental design presented in McAdams et al. (2023). The general procedure of the three phases is presented first, with specific details pertaining to each experiment reported in their respective sections.

Table 4.2 The definitions of each excitation and resonator category.

	Category	Definition
Excitation	Bowing (Bo)	The action of rubbing a bow on an object to make it vibrate.
	Blowing (Bl)	The action of blowing into a mouthpiece to make an object vibrate.
	Striking (Sk)	The action of using a mallet to hit an object to make it vibrate.
Resonator	String (Sg)	An object that is a thin wire fixed at its endpoints.
	Air column (Ac)	An object that comprises the air molecules within a tube that is sealed at one end and open at the other end.
	Plate (Pl)	An object that is rectangular, flat, and rigid.

4.2.4.1 Familiarization Phase

This phase presented participants with one example of each excitation-resonator interaction, organized by their category names, depending on the learning task, i.e., excitation categories for Experiment 1, resonator categories for Experiment 2, and interaction categories for Experiment 3. For each participant, the example of each interaction was chosen at random among its seven exemplars. Each example had its own sound box. The arrangement of the sound boxes for each experiment will be described in its corresponding Method section. Clicking any of the sound boxes played its example. There was no limit as to how many times each example could be played, but listeners had to play them at least three times each. They were encouraged to play each example until they felt comfortable associating its sound with the corresponding category and its description. Once they felt familiar with the categories of the sounds, they proceeded to the next phase of the experiment.

4.2.4.2 Training Phase

The training phase comprised a series of blocks, each with nine trials. Each trial concerned one of the nine types of excitation-resonator interactions chosen at random among the seven exemplars of each of them. The order in which the stimuli were presented in the trials of each block was randomized. Participants heard a sound in each trial and chose the category they thought was associated with it, depending on the experiment. A one-time replay button for the sound was provided in each trial. Corrective feedback was given following every response. If the response was correct, the chosen category flashed in green. If the response was incorrect, the chosen category flashed in red and the correct category simultaneously flashed in green. To move on to the next phase, participants were required to obtain a passing threshold of at least 75% correct (i.e., choosing the correct category of at least seven out of nine sounds) in a block for four consecutive blocks. The first two blocks did not count towards the passing threshold as they were supposed to help participants become familiar with the interface. The minimum number of blocks that could be completed during the training phase, granted that the passing threshold was met, was six blocks (i.e., 54 trials). If the passing threshold was not reached within 20 blocks, the experiment would end, and the participant would not proceed to the testing phase. The training phase would only exceed 20 blocks if participants obtained at least 75% of correct categorizations on the 18th to 20th block. Listeners proceeded to the next block if they continued to obtain at least a 75% score, but if they failed one block at or beyond the 20th block, the experiment would terminate. The training phase never exceeded 23 blocks (i.e., 207 trials).

4.2.4.3 Testing Phase

Participants who passed the training phase continued to the final stage of the experiment. In the testing phase, there were 63 trials, one for each stimulus. The task in each trial was similar to that of the training phase. Participants heard a sound in each trial and a one-time replay button was included. After hearing the sound, they chose the category that they thought produced it, depending on the experiment. Unlike the training phase, there was no corrective feedback following each response. The stimuli were presented in a pseudo-random order such that two successive sounds were not produced by the same interaction (i.e., a struck plate was not presented before or after another struck plate).

4.3 Experiment 1: Learning Excitations

This experiment tested whether the excitations of the nine excitation-resonator interactions could be learned. The three excitations (Bo, Bl, Sk) were each involved in three interactions: BoSg, BoAc, BoPl; BlSg, BlAc, BlPl; and SkSg, SkAc, SkPl, respectively.

4.3.1 Method

4.3.1.1 Participants

Forty-two participants were recruited for Experiment 1. One participant did not pass the audiometric screening test. The participants who completed the experiment filled out the Gold-MSI (Müllensiefan et al., 2014) and we report their mean general musical sophistication scores and mean musical training scores. The minimum and maximum scores that can be obtained on the general musical sophistication scale are 18 and 126, respectively. For musical training, the minimum and maximum scores are 7 and 49, respectively. One participant did not pass the training phase (female, age 19). This participant had a general musical sophistication score of 29.0 and a musical training score of 20.0 on the Gold-MSI. Of the 40 participants who passed the training phase (27 females, 12 males, 1 preferred not to disclose their sex), their ages ranged from 17 to 36 years old, with a mean age of 22.6 years (SD=2.5). Their mean general musical sophistication score was 73.0 (SD=20.8) and ranged between 34 and 116. The mean musical training score was 26.1 (SD=12.1), ranging from 7 to 47.

4.3.1.2 Procedure

In the familiarization phase, an interface was created using a 3×3 grid of sound boxes, each representing one type of excitation-resonator interaction. Participants played each sound by clicking on its corresponding box. The sound boxes in the same column had the same excitations and the sound boxes in the same row had the same resonators. The category names—'bowed sounds', 'blown sounds', and 'struck sounds'—were presented above the grid of sound boxes and arranged based on the type of excitation that was represented by each column. The order of the arrangement of excitations in columns and resonators in rows was randomized across participants.

In each trial of the training and testing phases, after listening to the sound, participants were asked: 'Which excitation produced this sound?' There were three response boxes presented on screen, each corresponding to one of the excitation categories (i.e., 'bowing', 'blowing', and 'striking'). The arrangement of the response boxes in columns in the training and testing phases was the same as that

of the familiarization phase for each participant. Participants then chose one response by clicking its corresponding box before proceeding to the next trial.

4.3.2 Results

4.3.2.1 Training Performance

For the passing participants, the average number of blocks it took for them to pass the training phase was 9.05 blocks. The least number of blocks it took to pass was 6 blocks and the most was 23. The failing participant completed 22 blocks. We calculated the participants' categorization performance in the training phase. Each participant completed a different number of training blocks, so the mean proportion correct was obtained by dividing the number of trials for which a participant got a correct answer by the total number of trials completed. These proportions were then converted to percent correct scores. The mean score of successful participants during the training phase was 84.61% (*SD*=7.79) and the failing participant obtained a score of 73.23%. All participants correctly categorized the excitations of the typical interactions, except for the failing participant who incorrectly categorized the excitation of SkSg (Figure 4.1). For atypical interactions that involved bowing (BoAc, BoPl), the successful participants performed better than the failing participant; but for the atypical interactions produced by blowing (BlSg, BlPl), the failing participant correctly categorized their excitation more often than the successful participants.

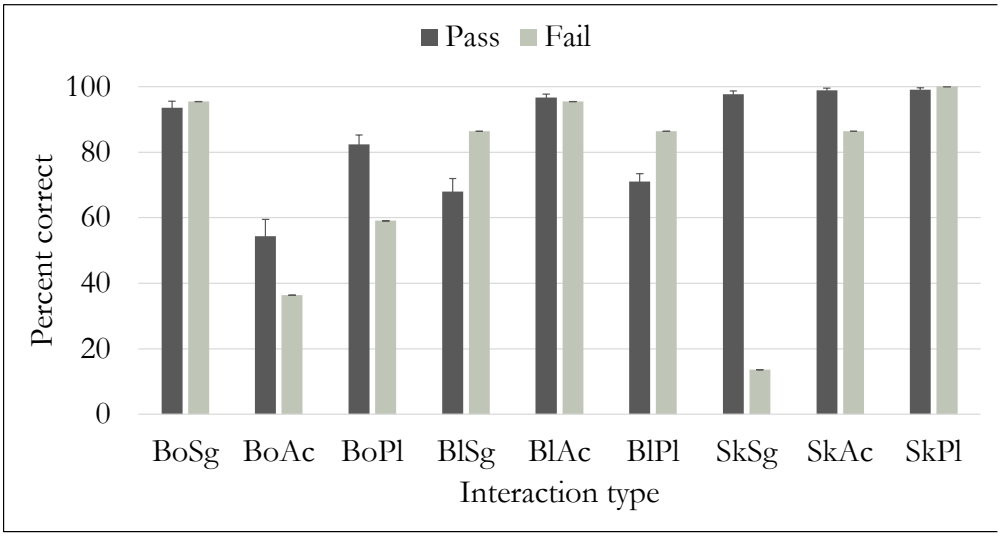

Figure 4.1 Mean percent correct of excitation categorization between passing participants (*n*=40) and the failing participant (*n*=1) (different colors) during the training phase for each type of interaction (horizontal axis). Error bars represent the standard error of the mean.

4.3.2.2 Testing Performance

We first examined whether the general musical sophistication score and musical training subscore of the Gold-MSI (Müllensiefan et al., 2014) would predict the correct response data during the testing phase. A binomial logistic regression was computed using generalized mixed effects modeling to control for the by-participant random intercept. Neither of the two scores significantly predicted the number of correct excitation categorizations: β=0.36, z=1.49, p=0.136, 95% CI [−0.13, 0.86] for general musical sophistication; β=−0.29, z=−1.21, p=0.226, 95% CI [−0.79, 0.20] for musical training. So, participants with more musical experience did not necessarily categorize excitations more correctly than those with less musical experience.

We analyzed the accuracy of excitation categorization in the testing phase where there was no corrective feedback. We ran a binomial logistic regression using generalized linear mixed effects modeling with excitation categories, resonator categories, and the interaction between these two predictors included as fixed effects. A statistical approach aiming to obtain the maximal random effects structure justified by the correct responses was implemented (Barr et al., 2013). We initially included a random intercept for participant and random slopes for each excitation and resonator category. If this model generated a singular fit, random slopes were removed one by one until there was no longer a singular fit. A random slope was removed if it was highly correlated with another random slope or if its variance was close to 0. Because intercept-only models are prone to higher Type I errors, controlling for random effects in this manner guards against them (Schielzeth & Fortsmeier, 2009). Consequently, random slopes for each resonator category were removed. Additionally, the random slope for striking was removed when the reference excitations were bowing and blowing, and the random slope for blowing was removed when the reference excitation was striking.

Selected fixed effects of excitation categorization correct responses are reported in Table 4.3. These fixed effects were obtained by running the same model and changing the reference category of the excitations and resonators. The regression coefficients (β) correspond to log-odds ratios. When exponentiated, they become odds ratios. These fixed effects were chosen because they demonstrated the effect of resonators on the correct categorization of each excitation. For example, if the reference excitation was bowing (Bo) and the reference resonator was an air column (Ac), the fixed effect of the string (Sg) meant that the odds ratio compared correct excitation categorization of the bowed string (BoSg) relative to the bowed air column (BoAc). The odds ratio was 36.02, so the odds of correctly categorizing the excitation of BoSg were 36.02 times the odds of correctly categorizing the excitation

of BoAc. For bowed and blown stimuli, the odds of correctly categorizing excitations were greater for typical interactions relative to atypical interactions: BoSg > BoAc and BoPl; BlAc > BlSg and BlPl. Among the struck stimuli, none of the fixed effects were significant, suggesting that the odds of identifying striking did not differ between the resonators to which it was applied. Given that striking was the only impulsive excitation, its categorization was most accurate in comparison to bowing and blowing.

Table 4.3 Selected fixed effects (β) and corresponding log odds ratios of correctly categorizing a type of excitation between interactions involving different resonators during the testing phase.

	Reference	Comparison	β	SE	p	Odds ratio
Bowing categorization	BoAc	Bo**Sg**	3.58	0.370	<0.001	36.02
	BoPl	Bo**Sg**	1.24	0.381	0.001	3.45
	BoPl	Bo**Ac**	–2.34	0.266	<0.001	0.10
Blowing categorization	BlSg	Bl**Ac**	2.40	0.321	<0.001	11.03
	BlPl	Bl**Ac**	1.92	0.325	<0.001	6.85
	BlPl	Bl**Sg**	–0.48	0.201	0.018	0.62
Striking categorization	SkSg	Sk**Pl**	–0.70	1.229	0.570	0.50
	SkAc	Sk**Pl**	–0.70	1.229	0.570	0.50
	SkAc	Sk**Sg**	0.00	1.418	1.000	1.00

Note. The 'Reference' column indicates the reference excitation and resonator categories. Fixed effects are in boldface in the 'Comparison' column.

To observe if participants confused different excitations for one another, we calculated the mean percent response of choosing each excitation category for each interaction (Figure 4.2). Overall, listeners chose the correct category more often than the other categories, regardless of the interaction. As the black bars reflect the percent correct of excitation categorization, bowing was most often correctly categorized when it was applied to the string, then the plate, and then the air column. BoAc was occasionally confused for blowing. However, this confusion was much less common in comparison to Huynh & McAdams's (2023a) study, during which listeners most often categorized BoAc as blown. Among the interactions involving blowing, the percent correct of excitation categorization was highest for BlAc, BlPl, then BlSg. Percent correct for categorizing striking was consistently very high and there were hardly any confusions for these impulsive sounds.

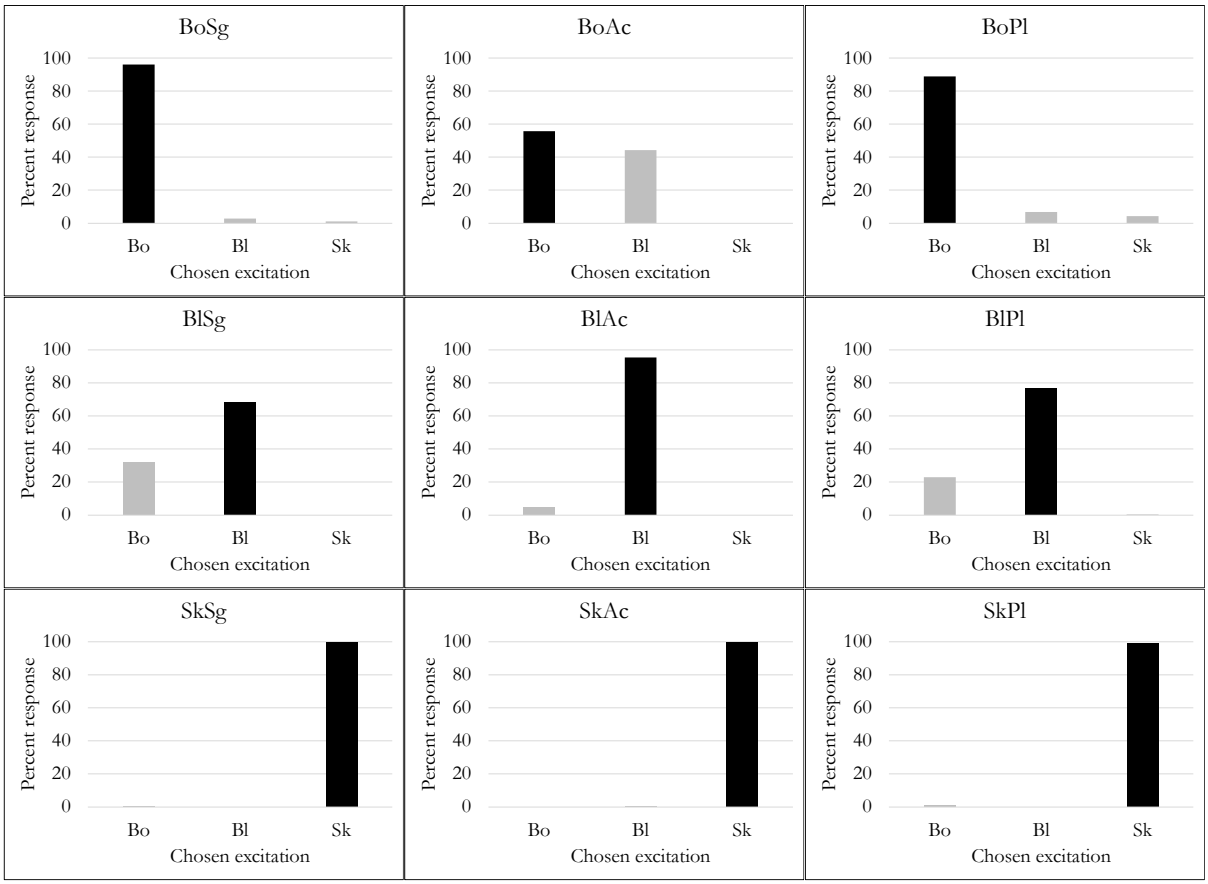

Figure 4.2 Mean percent response of choosing each excitation category (horizontal axis) for each of the nine interaction types (separated by different graphs) during the testing phase. Correct categorization (i.e., mean percent correct) is represented by black bars.

4.3.3 Discussion

The results of Experiment 1 demonstrate that the atypical interactions can be learned based on their interactions. Only one participant was unsuccessful during the training phase. This participant, however, was unable to identify the excitation of a struck string (Figure 4.1), consistently categorizing it as bowed rather than struck. However, categorizing the struck string should not require any learning given that it is a typical interaction, and striking was the only impulsive excitation. Nearly all participants in Experiment 1 were able to pass the training phase, so we concluded that the excitations of the atypical interactions can indeed be learned. The categorization performance of the passing participants during the testing phase was generally above chance for all interaction types. Among the atypical interactions, the most confusions were made for BoAc, followed by BlSg. However, participants made fewer confusions than in a previous categorization task of sounds produced by the same interactions (Huynh & McAdams, 2023a). So, comparing categorization performance between

Experiment 1 of the current study and excitation categorization in Huynh & McAdams's (2023a) study, supervised training indeed improved categorization of the excitations. Corrective feedback during the training phase likely guided the detection of transformational invariants, which are acoustic properties characterizing the sound-generating action (McAdams, 1993). The detection of transformational invariants was consequently useful during the testing phase to maintain categorization performance even with the removal of corrective feedback.

4.4 Experiment 2: Learning Resonators

In this experiment, listeners were trained to categorize the resonators that produced the sounds. The three resonators (Sg, Ac, Pl) were each involved in three interactions: BoSg, BlSg, SkSg; BoAc, BlAc, SkAc; and BoPl, BlPl, SkPl, respectively.

4.4.1 Method

4.4.1.1 Participants

We recruited 49 participants for Experiment 2. Three participants did not meet the hearing thresholds. Six participants (4 females, 2 males) did not pass the training phase; they had a mean age of 25.2 years (*SD*=4.3) ranging from 19 to 32 years old. Their general musical sophistication scores were between 41 and 77 with a mean of 58.8 (*SD*=11.7). The musical training scores ranged from 8 and 28, and the mean was 15.3 (*SD*=7.7). The 40 participants who passed the training phase (27 females, 12 males, 1 self-identified as "other") had a mean age of 23.1 years (*SD*=3.6) spanning from 18 to 33 years. Their general musical sophistication scores ranged from 37 to 116, with a mean of 76.2 (*SD*=21.5). The mean musical training score was 26.1 (*SD*=12.6), and the range was 7 to 46.

4.4.1.2 Procedure

In this experiment, the category names were based on the resonators of the stimuli. Like Experiment 1, the sound boxes in the familiarization phase were arranged in a 3×3 grid. The sounds in the same column had the same resonators and the sounds in the same row had the same excitations. Category names ('string sounds', 'air column sounds', 'plate sounds') were positioned above the grid, corresponding to the type of resonator that was represented by each column. The arrangement of resonators in columns and excitations in rows was randomized for each participant.

In the training and testing phases, three response boxes corresponding to each of the resonator categories were presented on the screen. The arrangement of the resonator categories in columns during the training and testing phases was the same as that of the familiarization phase for each participant. In each trial of the training and testing phases, participants were asked 'Which resonator produced this sound?' after listening to a sound. They chose the resonator they thought produced the sound by clicking on its corresponding response box before moving on to the next trial.

4.4.2 Results

4.4.2.1 Training Performance

Listeners who passed the training phase based on resonator categorization completed between 6 to 20 blocks with an average of 9.33 blocks. The mean percent correct score of the successful participants during the training phase was 82.60% (*SD*=7.30). The six failing participants each completed 20 blocks and had a mean percent correct score of 62.31% (*SD*=7.68).

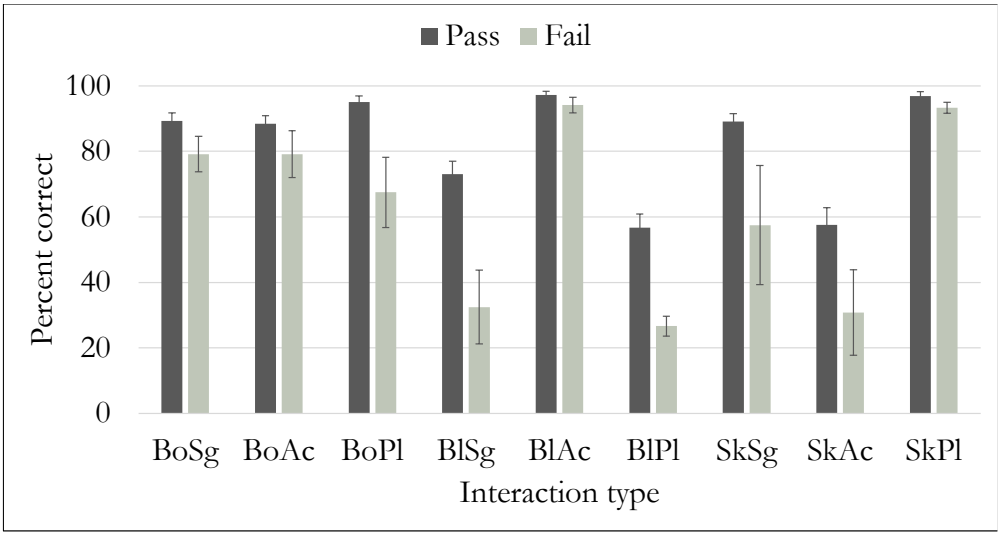

Figure 4.3 Mean percent correct of resonator categorization between passing participants (*n*=40) and failing participants (*n*=6) (different colors) during the training phase for each type of interaction (horizontal axis). Error bars represent the standard error of the mean.

A comparison of the mean percent correct scores between successful participants and failing participants for each interaction type is shown in Figure 4.3. Successful participants were generally more accurate at categorizing the resonators in the training phase compared to the failing participants, especially for the bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), struck air column

(SkAc), and struck string (SkSg). Although SkSg is a typical interaction, the failing participants were less correct in categorizing its resonator compared to the successful participants. However, there was a lot of variability in the failing participants' responses as indicated by the standard error bars. For the successful participants, resonator categorization was least correct for BlPl and SkAc, suggesting that confusions were occasionally made for these interactions during training.

4.4.2.2 Testing Performance

To test the effects of musical experience on categorization performance, a binomial logistic regression with generalized mixed effects modeling was computed. We controlled for the by-participant random intercept in the regression model. Neither of the Gold-MSI scales of interest significantly predicted the correct response data: β=0.33, z=1.42, p=0.156, 95% CI [−0.14, 0.81] for general musical sophistication; β=−0.18, z=−0.76, p=0.446, 95% CI [−0.66, 0.29] for musical training. So, participants with more musical experience did not categorize the resonators more correctly than participants with less musical experience.

We regressed the excitation categories, resonator categories, and their interaction onto the correct responses of resonator categorization using a binomial logistic regression. The regression incorporated generalized linear mixed effects modeling. Much like the aim of protecting against Type I errors in Experiment 1, we added random slopes for excitation and resonator categories in addition to the random intercept for participant. No random slopes were removed from the model, as they did not generate a singular fit. Selected fixed effects for the correct response of resonator categorization are shown in Table 4.4. We ran the same model and changed the reference excitation and resonator categories to obtain these selected fixed effects. In this case, we were interested in the effects of the different excitations on resonator categorization accuracy. For example, when the reference excitation and resonator categories were striking and the string, respectively, the effect of bowing compares the odds of correctly categorizing the resonator of BoSg to the odds of correct resonator categorization of SkSg. The log odds ratio was nonsignificant, which meant there were no differences in the odds of correctly identifying the string between either of them. This comes as no surprise given that both BoSg and SkSg are typical interactions. In general, the odds of correctly categorizing the resonator were significantly greater for typical interactions than for atypical ones: BoSg and SkSg > BlSg, BlAc > SkAc, SkPl > BlPl. The odds of correct resonator categorization were not significantly different between BlAc and BoAc. Although BoAc is an atypical interaction, it was assigned higher resemblance ratings to the air column (Huynh, 2019) and was most often categorized as an air column in a previous

categorization task (Huynh & McAdams, 2023a). The odds of correct resonator categorization of BoPl did not differ significantly from those of SkPl. Listeners also categorized BoPl as a plate more often than as the other resonators in Huynh and McAdams's (2023a) categorization task.

Table 4.4 Selected fixed effects (β) and corresponding log odds ratios of correctly categorizing a type of resonator between interactions involving different excitations during the testing phase.

	Reference	Comparison	β	SE	p	Odds ratio
String categorization	SkSg	**Bo**Sg	–0.47	0.558	0.397	0.62
	BlSg	**Bo**Sg	1.83	0.468	<0.001	6.20
	BlSg	**Sk**Sg	2.30	0.483	<0.001	9.95
Air column categorization	BoAc	**Bl**Ac	1.18	0.635	0.063	3.26
	SkAc	**Bl**Ac	3.65	0.571	<0.001	38.43
	SkAc	**Bo**Ac	2.47	0.434	<0.001	11.79
Plate categorization	BoPl	**Sk**Pl	0.42	0.730	0.567	1.52
	BlPl	**Sk**Pl	4.34	0.618	<0.001	76.70
	BlPl	**Bo**Pl	3.92	0.569	<0.001	50.48

Note. The 'Reference' column indicates the reference excitation and resonator categories. Fixed effects are in boldface in the 'Comparison' column.

The mean percentage of times that each interaction was categorized as each resonator is shown in Figure 4.4. In general, listeners chose the correct resonator of each interaction more often than other resonators. Listeners sometimes confused BlSg and BlPl for the air column and SkAc for the string and plate. However, listeners made less confusions compared to the previous categorization task of the same interactions (Huynh & McAdams, 2023a). Thus, supervised training improved resonator categorization performance of the current stimulus set.

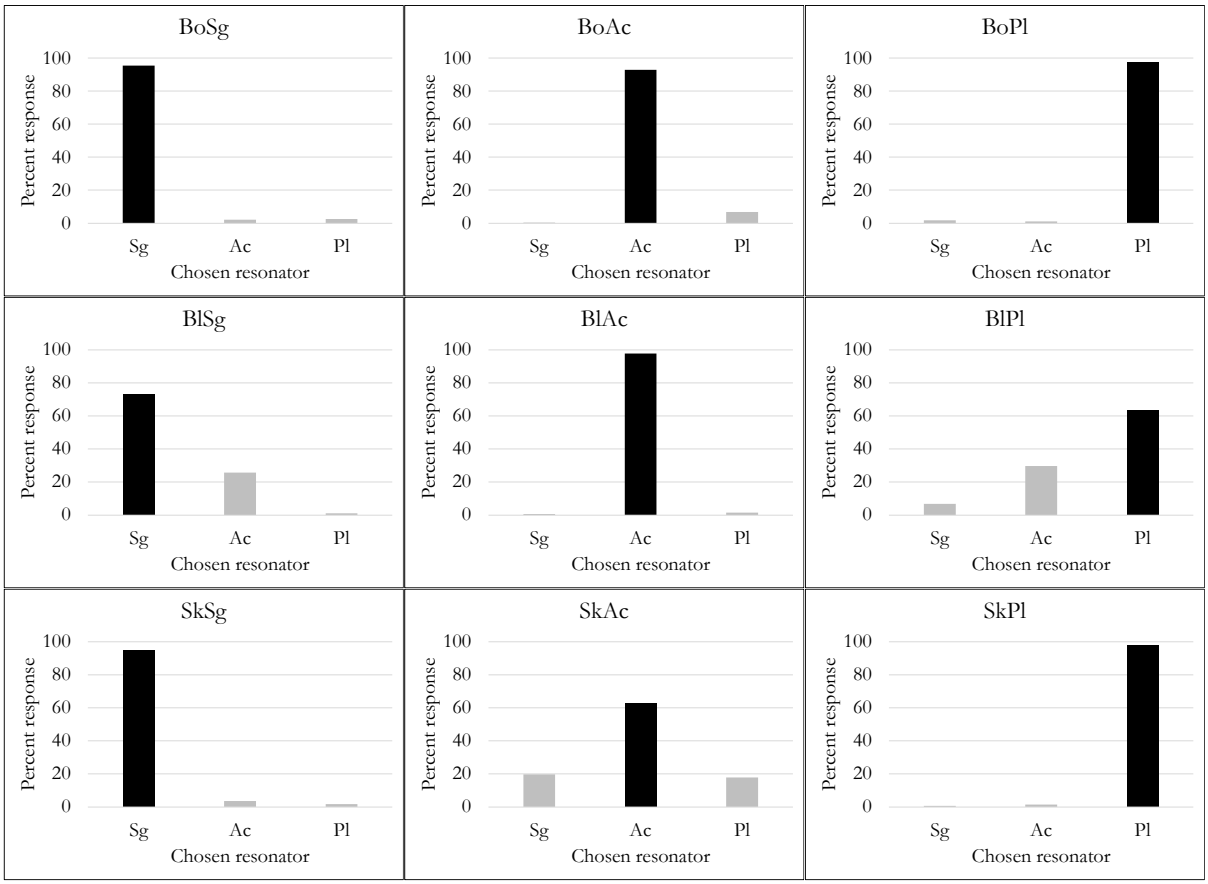

Figure 4.4 Mean percent response of choosing each resonator category (horizontal axis) for each of the nine interaction types (separated by different graphs) during the testing phase. Correct categorization (i.e., mean percent correct) is represented by black bars.

4.4.3 Discussion

The majority of participants who were trained to categorize nine types of excitation-resonator interactions based on their resonator categories were successful at learning. However, there were six participants who were unable to pass the training phase. There was a considerable amount of variability in the categorization performance among the six failing participants, but they were generally unable to categorize the resonators of BlSg, BlPl, and SkAc. So, the failing participants were unable to apply the corrective feedback effectively to detect the relevant structural invariants that differentiate the resonators. Structural invariants are acoustic properties that describe the physical structure of the sound-generating object (McAdams, 1993). In contrast, the 40 out of 46 participants who passed the training phase were more able than their failing counterparts to detect the structural invariants with the guidance of corrective feedback. As categorization performance was maintained during the testing

phase, these findings suggest that the atypical interactions can indeed be learned based on the resonators that produced them.

4.5 Experiment 3: Learning Interactions

In this experiment, we examined the learning of nine categories, each corresponding to one type of excitation-resonator interaction.

4.5.1 Method

4.5.1.1 Participants

Fifty-four participants signed up for participation in this experiment. One participant did not pass the audiometric screening, one participant was feeling unwell and opted out, and 11 participants did not pass the training phase. Among the failing participants (8 females, 3 males), their mean age was 26.3 (*SD*=6.1) and they were between 19 to 39 years of age. The mean general musical sophistication score was 57.7 (*SD*=18.4) and ranged from 29 to 93. Their scores on the musical training subscale ranged from 8 to 43, with a mean of 16.5 (*SD*=11.5). The ages of the 41 successful participants (28 females, 12 males, 1 self-identified as "other") ranged from 17 to 39 years old with a mean of 23.2 years (*SD*=5.0). The mean general musical sophistication score of the successful participants was 83.37 (*SD*=20.9) and ranged from 47 to 116. The musical training scores were between 9 and 47 with a mean of 30.2 (*SD*=12.9).

4.5.1.2 Procedure

The procedure of the three phases was similar to Experiments 1 and 2, except that learning was based on the interactions. Throughout the three phases of the experiment, the boxes of the category names were arranged in a 3×3 grid. For half of the participants, the boxes in the same column had the same excitations and the boxes in the same row had the same resonators. For the other half of the participants, it was the other way around. The order of the arrangement of excitations and resonators in columns or rows was randomized across participants, but the nine boxes had the same arrangement for each participant across the three phases of the experiment. During the familiarization phase, the boxes of the category names corresponded to their own sound examples. During each trial of the training and testing phases, listeners were presented a sound and asked 'Which interaction produced

this sound?' Participants chose between one of the nine interactions by clicking the box corresponding to its category.

4.5.2 Results

4.5.2.1 Training Performance

A mean of 11.93 blocks was completed by the passing participants during the training phase. The least number of blocks completed was 6 and the most was 23. The mean percent correct during the training phase for the successful participants was 75.87% (*SD*=8.13). The failing participants completed 20 to 21 blocks with an average of 20.27 blocks. Their mean percent correct score was 51.38% (*SD*=11.48). Figure 4.5 compares the successful participants' and failing participants' mean percent of correct categorization for each interaction. Successful participants were more correct than the failing participants across each interaction type. It seemed that the failing participants made some confusions, even for the typical interactions with which they should have been familiar. Failing participants were most incorrect for BlPl, whereas the successful participants were most incorrect for BlPl and BlSg.

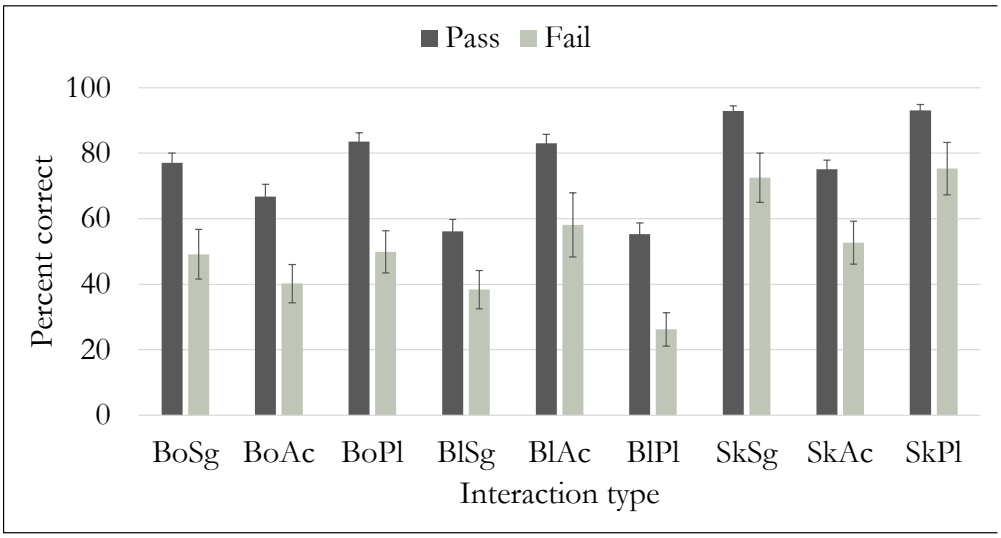

Figure 4.5 Mean percent correct of interaction categorization between passing participants (*n*=41) and failing participants (*n*=11) (different colors) during the training phase for each type of interaction (horizontal axis). Error bars represent the standard error of the mean.

4.5.2.2 Testing Performance

The general musical sophistication and musical training scores of the Gold-MSI were regressed onto the correct response data using a binomial logistic regression. We used a generalized mixed effects model approach and controlled for the by-participant random intercept. Neither of the Gold-MSI scores contributed significant effects to the model: β=0.01, z=0.03, p=0.973, 95% CI [−0.32, 0.69] for general musical sophistication; β=0.20, z=0.75, p=0.456, 95% CI [−0.34, 0.75] for musical training. Thus, the degree of musical experience did not predict categorization accuracy of the interactions.

Table 4.5 Selected fixed effects (β) and corresponding log odds ratios of correctly categorizing the different types of interactions during the testing phase. The effect of resonators on each excitation type and the effect of excitations on each resonator type are shown.

Simple main effect	Reference	Comparison	β	SE	p	Odds ratio
Resonators on bowing	BoAc	Bo**Sg**	0.48	0.354	0.172	1.62
	BoPl	Bo**Sg**	–0.82	0.394	0.037	0.44
	BoPl	Bo**Ac**	–1.31	0.383	0.001	0.27
Resonators on blowing	BlSg	Bl**Ac**	1.12	0.335	0.001	3.07
	BlPl	Bl**Ac**	1.03	0.321	0.001	2.82
	BlPl	Bl**Sg**	–0.08	0.291	0.773	0.92
Resonators on striking	SkSg	Sk**Pl**	–0.23	0.552	0.677	0.79
	SkAc	Sk**Pl**	1.13	0.488	0.020	3.11
	SkAc	Sk**Sg**	1.36	0.494	0.006	3.91
Excitations on strings	SkSg	**Bo**Sg	–1.70	0.488	<0.001	0.18
	BlSg	**Bo**Sg	1.27	0.290	<0.001	3.56
	BlSg	**Sk**Sg	2.97	0.476	<0.001	19.55
Excitations on air columns	BoAc	**Bl**Ac	0.34	0.295	0.254	1.40
	SkAc	**Bl**Ac	–0.50	0.405	0.229	0.61
	SkAc	**Bo**Ac	–0.82	0.381	0.031	0.44
Excitations on plates	BoPl	**Sk**Pl	0.65	0.501	0.193	1.92
	BlPl	**Sk**Pl	2.66	0.471	<0.001	14.28
	BlPl	**Bo**Pl	2.01	0.339	<0.001	7.44

Note. The 'Reference' column indicates the reference excitation and resonator categories. Fixed effects are in boldface in the 'Comparison' column.

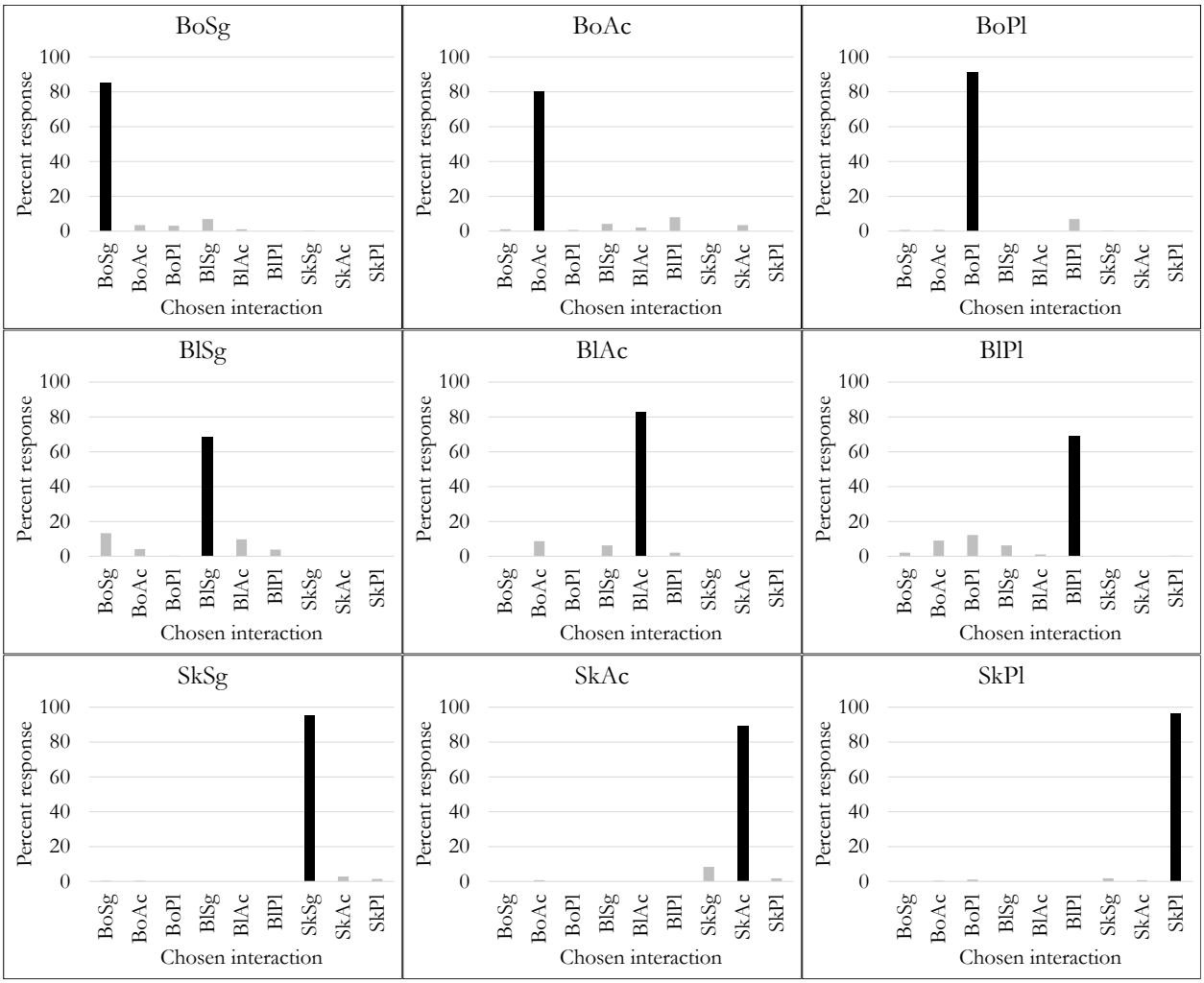

Figure 4.6 Mean percent response of choosing each interaction category (horizontal axis) for each of the nine interaction types (separated by different graphs) during the testing phase. Black bars represent correct categorization (i.e., percent correct scores).

We analyzed the correct response of interaction categorization with a binomial logistic regression. The same predictors as Experiments 1 and 2 were used: excitation categories, resonator categories, and their interaction. We implemented the same technique for the generalized mixed effects modeling as in Experiments 1 and 2 and controlled for random effects. No random slopes were removed from the model. Regression coefficients and odds ratios based on the effects of the different resonators on each excitation and the effects of the different excitations on each resonator are reported in Table 4.5. Overall, the odds of correctly categorizing the interactions were significantly greater for typical interactions than for atypical interactions that share the same excitations or resonators: BoSg > BlSg, BlAc > BlSg, BlAc > BlPl, SkSg > SkAc, SkSg > BlSg, SkPl > SkAc, SkPl > BlPl. There were some

exceptions. The first one was BoPl, for which the odds of correct categorization were not significantly different from SkPl. Moreover, the odds ratio comparing BoSg to BoPl was 0.44; because this odds ratio is less than one, it means that the odds of correctly categorizing BoSg were significantly less than the odds of correctly categorizing BoPl. So, BoPl was surprisingly more correctly or just as correctly categorized compared to two of its typical counterparts. For BoAc, listeners' odds of correct categorization did not differ significantly from BoSg or BlAc, suggesting that they were just as accurate at categorizing BoAc compared to its typical counterparts. Lastly, the odds of correct categorization were no different between SkAc and BlAc, implying that listeners were mostly correct when they categorized SkAc.

To observe if there were any confusions between any of the interactions, we compared the mean percent response of choosing each interaction category for each interaction type in Figure 4.6. The grey bars represent confusions, and the black bars are equivalent to mean percent correct scores. Listeners were quite accurate overall and made few confusions. The most confusions were made for BlSg and BlPl. For these interactions, the other bowed and blown sounds were chosen at least once. In general, listeners improved from their performance in the training phase (Figure 4.7). Categorization was more accurate for all the atypical interactions as well as BoSg in the testing phase compared to the training phase. So, listeners retained what they had learned during the training phase even when there was no corrective feedback in the testing phase.

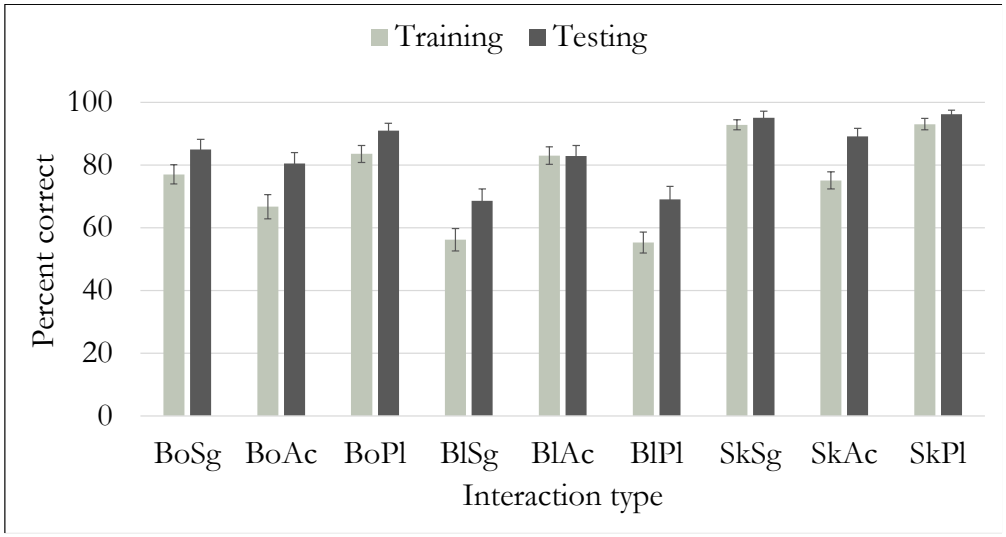

Figure 4.7 Mean percent correct of interaction categorization based on interaction type between successful participants of the training phase and their performance in the testing phase. Error bars represent standard error of the mean.

4.5.3 Discussion

This experiment had the largest number of failing participants in comparison to the other two experiments. Categorization performance of each interaction type was worse for the failing participants than for the successful participants during the training phase. The standard error bars for the failing participants in Figure 4.5 suggest that there was some variability in their categorization performance. Upon further inspection of each failing participant's categorization performance, most of them were unable to improve the number of correct categorizations between successive blocks. Furthermore, the types of interactions that they had difficulty learning seemed to depend on the participant. For example, BoAc was confused for BlAc by two participants, whereas two different participants confused it for BoSg, and two other participants sometimes confused it for SkAc. So, the types of confusions made were consistent for a given participant but varied across participants. Given that 41 out of 52 participants passed the training phase, they were able to apply the corrective feedback to learn the categories of the interactions. Furthermore, in this particular version of the experiment, successful participants improved in the categorization of the atypical interactions between training and testing (Figure 4.7). This suggests that the atypical interactions can be learned and that listeners might have detected the invariant features that reliably determine category membership.

4.6 General Discussion

In three separate learning tasks, listeners were trained to categorize the excitations, resonators, or interactions of nine types of interactions that were produced by digitally combining three types of excitations (bowing, blowing, striking) and three types of resonators (string, air column, plate). As listeners were already familiar with the typical interactions—bowed string (BoSg), blown air column (BlAc), struck string (SkSg), and struck plate (SkPl)—from everyday music listening, we were curious whether they could learn the atypical interactions: bowed air column (BoAc), bowed plate (BoPl), blown string (BlSg), blown plate (BlPl), and struck air column (SkAc). Overall, percent correct scores in the testing phase were above chance for each categorization task, suggesting that short-term learning effects from supervised training were maintained in the testing phase.

One of our predictions was that the number of failing participants would be the lowest for excitation training, then resonator training, and greatest for interaction training. One out of 41, 6/46, and 11/52 participants failed excitation, resonator, and interaction training phases, respectively. The rate of failures and difficulty of each type of categorization task might be predicted by the number of

dimensions that best explain their differentiation: one dimension for excitations, two for resonators, and three for interactions (Huynh & McAdams, 2023b). Overall, the failing participants had lower mean general musical sophistication and musical training scores than the successful participants. However, there were not enough failing participants for us to obtain a representative sample to statistically compare their Gold-MSI scores to those of the successful participants. It is also worth noting that the general musical sophistication and musical training scores of the successful participants spanned considerable ranges, with certain individual scores being lower than those of the failing participants. Across the different learning tasks, we found that general musical sophistication and musical training scores did not predict the accuracy of categorizing the excitations, resonators, or interactions. So, categorization performance during the testing phase had very little to do with musical experience.

Among the testing phases, the overall mean percent scores of excitation, resonator, and interaction categorization were 86.63%, 86.11% and 84.17%, respectively. We initially expected categorization performance in the testing phase to be best for excitations and least accurate for interactions, but the percent scores of the overall performance in each task reflected minimal differences. Mean percent correct scores of each interaction across the three types of categorization tasks during the testing phase are compared in Figure 4.8. For the typical interactions, each type of categorization was accurate overall, but listeners made slightly more errors in categorizing the interactions than the excitations or resonators of BoSg and BlAc. However, the typical interactions did not seem to require learning, given that performance was quite consistent with that of Huynh and McAdams (2023a).

Interestingly, categorization of BoPl was comparable to that of the typical interactions across the three types of categorizations (Figure 4.8). Mostly consistent with the findings from the previous categorization task by Huynh and McAdams (2023a), resonator categorization was better than excitation categorization for BoPl and BoAc and this difference in categorization accuracy was greater for BoAc. Especially for BoAc, interaction categorization was better than excitation categorization, although not as good as resonator categorization (Figure 4.8). More importantly, fewer confusions were made for excitation categorization of BoAc, which was previously rated as resembling blowing more than bowing (Huynh, 2019) and categorized as blown more often than bowed (Huynh & McAdams, 2023a). During the excitation learning task (Experiment 1) of the current study, however, bowing was chosen more often than blowing (Figure 4.2), demonstrating a reduction in the confusion previously observed.

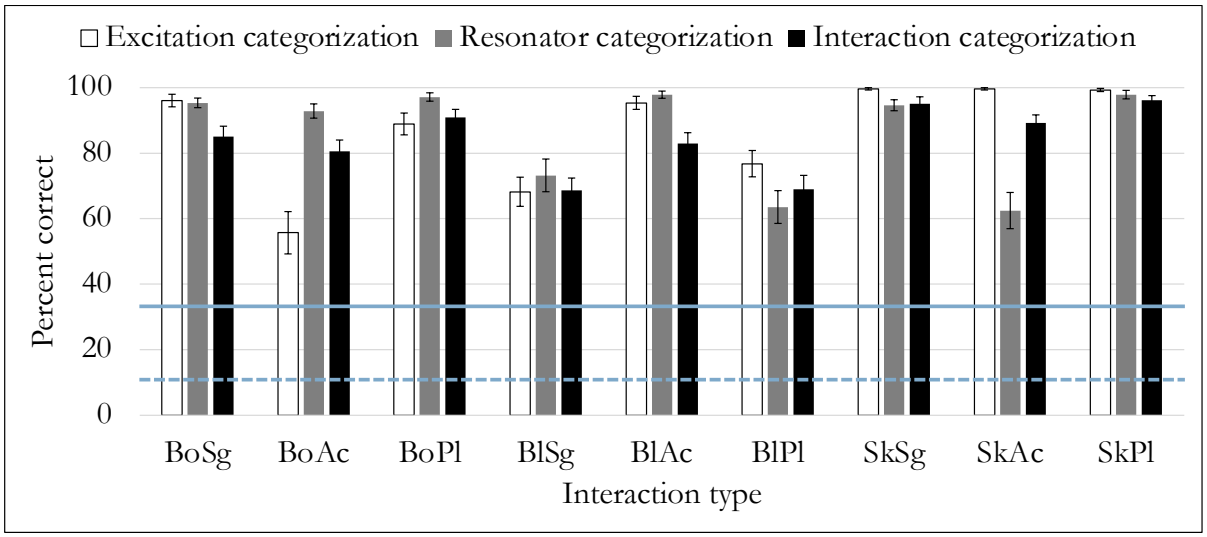

Figure 4.8 A comparison of the mean percent correct scores of categorizing the excitations (Experiment 1), resonators (Experiment 2), and interactions (Experiment 3) for each interaction during the testing phases. The solid line indicates chance performance of correct excitation and resonator categorization, and the dashed line indicates chance performance of correct interaction categorization. Error bars reflect standard error of the mean.

For BlPl and SkAc, excitation categorization was better than resonator categorization (Figure 4.8), and even more so for SkAc. Interaction categorization was also better than resonator categorization for SkAc. In the testing phase of the resonator learning task (Experiment 2), the plate was chosen more often than the other resonators for BlPl (Figure 4.4). Similarly, for SkAc, the air column was chosen more often than the plate and string. These results contrast those of previous findings that reported confusions of BlPl with the air column and confusions of SkAc with the string and plate in resemblance rating and categorization tasks (Huynh, 2019; Huynh & McAdams, 2023a). Categorization accuracy for BlSg was quite consistent across each type of categorization task. BlSg was previously categorized as bowed more often than blown and as the string more often than the air column (Huynh & McAdams, 2023a). In the current study, however, fewer confusions were made as blowing, the string, and blown string were chosen most often during excitation, resonator, and interaction categorization, respectively.

Categorization performance for the testing phase of the interaction learning task (Experiment 3) reflected less confusions overall (Figure 4.6). There were likely no patterns of assimilation, since it was not the same types of confusions being made for each atypical interaction. In other words, the black bars were much higher than any of the grey bars, implying that listeners were able to learn the atypical

interactions quite accurately. We have now demonstrated that the atypical interactions can be learned, but the 'categorization problem' explains that it is difficult to understand *how* categories are learned (Harnad, 2017): explaining *what* listeners do in order to learn categories is easier than explaining *how* they do it. The explanation of what listeners do involves a discussion of attention.

Humans are exposed to a plethora of information every day. If we attended and responded to the information pertaining to every input, we would not be able to categorize so accurately (Harnad, 2017; Kruschke, 2005). The role of attention is important for amplifying and suppressing the processing of appropriate features. When the correct features are amplified by attention, learning categories becomes easier, faster, and more accurate (Kruschke, 2005). The features that require amplification are called invariants, as they distinguish members of a category from non-members. Therefore, invariants that distinguish between members of different categories are given more attention, whereas the features that distinguish members of the same category are given less attention (Harnad, 1990, 2017; Kruschke, 2005). In the case of the interactions of the current study, BoSg, BoAc, and BoPl, for example, belong in the same category with respect to excitation categorization and therefore do not differ in their transformational invariants. However, they belong to different categories during resonator categorization, which means they differ in their structural invariants. Accordingly, for the nine interactions of the current study, the transformational invariants that differentiate them would be attended to in the learning of the excitations, whereas the structural invariants would be ignored. During resonator learning, on the other hand, the structural invariants differentiating the interactions would be attended to, and transformational invariants would be ignored as much as possible.

As demonstrated in Huynh and McAdams (2023b), the invariants are likely the weighted sums of audio descriptors that vary with each dimension of the timbre space. For excitation differentiation, the transformational invariant was likely the temporal centroid, which distinguishes between impulsive (i.e., struck) and sustained (i.e., bowed and blown) excitations. For resonator differentiation, one set of structural invariants comprised a weighted combination of the spectrum's global shape, tonal content, variability of noise content, and the signal's modulation energy. This first set of structural invariants allowed listeners to isolate the plate from the other resonators. A weighted sum of acoustic properties that explain the finer details of the spectrum, the variability of the signal's energy and noisiness, and the modulation energy of the signal make up the second set of structural invariants that allowed listeners to further differentiate the string from the air column. Consequently, we would expect interaction categorization to be better than excitation and resonator categorization if the attention to transformational and structural invariants has an additive impact on categorization

performance. However, interaction categorization performance was worse than excitation categorization, resonator categorization, or both in some cases. This implies that the combined information pertaining to transformational and structural invariants is not additive and can sometimes be destructive. It is also possible that listeners might have ignored the transformational and structural invariants altogether and instead attended to the features of each interaction that distinctly characterized them from the others. Regardless of the strategy, selection of the relevant features that separate category members from non-members generates categorical representations of the interactions (Harnad, 1990).

Kruschke (2005) additionally mentioned that frequent categories are learned before infrequent or unfamiliar ones. This explains why categorization performance for the typical interactions was very accurate across excitation, resonator, and interaction categorization. They were already learned as they are frequent and familiar to everyday listening. So, excluding the typical interactions, listeners had to learn the categories of five atypical interactions, which are novel and unfamiliar to daily exposure. Categorization of atypical interactions had the most improvement between training and testing during interaction learning (Figure 4.7). Listeners might have attended more to the category-distinguishing features among the atypical interactions, which led to said improvement. Moreover, they attended to the relevant features of each atypical interaction that reliably distinguished them from the typical ones to which they would have been assimilated without supervised learning.

In summary, listeners were able to learn to correctly categorize the excitations, resonators, and interactions of nine types of excitation-resonator interactions. Participants were trained based on supervised learning which implemented trial and error with corrective feedback. Even when corrective feedback was absent in the testing phase, listeners were able to retain the categories of the sounds that they were trained on. When listeners learned the interaction categories, categorization accuracy improved for the atypical interactions from training to testing. In the percent response data, the correct categories were chosen more often than their incorrect counterparts regardless of the categorization task. Therefore, atypical interactions were not assimilated to typical interactions. Interestingly, the perceived mechanical implausibility of the atypical interactions did not interfere with categorization performance. Mechanical plausibility refers to how well an excitation-resonator interaction can be conceptualized as existing in the physical world. The typical interactions are physically possible and are modeled after acoustic musical instruments, so listeners can conceptualize these interactions. The atypical interactions, on the other hand, are physically impossible in the way that they are digitally simulated, so we expected listeners to have difficulty learning the categories of sounds of which their

production would be difficult to conceptualize. Instead, we found that supervised learning diminished the interference of mechanical plausibility and listeners likely formed new mental models for the atypical interactions based purely on acoustic properties. Furthermore, supervised learning allowed listeners to attend to the invariant features (i.e., acoustic properties) that reliably distinguished the categories depending on the type of categorization task. For example, listeners might pay more attention to the invariants pertaining to temporal centroid in the categorization of excitations and interactions, but not in the categorization of resonators. Because the excitation, resonator, and interaction categories of the atypical interactions can be learned based on supervised training, we might infer that the categories of the typical interactions were learned in a similar manner at some point in our early development. The findings of the current study can therefore serve as a case study to examine the formation of categories of novel sounds in our acoustic environment. These studies taken together emphasize the role of timbre in sound source recognition, sound category learning, and the formation of mental models of musical instruments.

Chapter V

Conclusion

The final chapter of this dissertation summarizes and connects the findings from the experimental studies reported in Chapters II, III, and IV. A discussion of this dissertation's contribution to the research on sound source recognition and categorization theory will be provided. Furthermore, this chapter reviews the limitations, proposes future research, and concludes with timbre's role in the identification of sound sources and the formation of mental models for novel sound sources.

5.1 Summary of Findings

Given that novel sounds are encountered daily and listeners tend to categorize them, we explored whether we can form mental models for unfamiliar or novel sound sources. This dissertation focused primarily on musical tones and was inspired by the synthesis design used in previous studies that applied different actions to objects made of various materials (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012). We simulated combinations of two macroscopic mechanical components of musical instruments: three different excitations (bowing, blowing, striking) were each applied to three different resonators (string, air column, plate), forming nine interactions. We modeled a large part of our research design after Figure 1.1 (Chapter 1), which was introduced by Li et al. (1991) as an effective strategy for studying sound source recognition. Li et al. (1991) explained that all pairwise comparisons between the physical, acoustical, and perceptual levels should be studied. Accordingly, we conducted three sets of experiments to examine the categorization (Chapter II), dissimilarity perception (Chapter III), and learning (Chapter IV) of the atypical interactions (bowed air column [BoAc], bowed plate [BoPl], blown string [BlSg], blown plate [BlPl], struck air column [SkAc]) in comparison to the typical interactions (bowed string [BoSg], blown air column [BlAc], struck string [SkSg], struck plate [SkPl]). Figure 5.1 presents a modified version of Figure 1.1 as it pertains to the research design and findings of this dissertation. Chapters II and IV were designed to examine the relationship between the physical

and perceptual levels in unsupervised learning and supervised learning contexts, respectively. As such, the path in green will be referred to as the categorization or unsupervised learning path (Chapter II) and the path in brown is the supervised learning path (Chapter IV). Chapter III was designed to examine the relationship between the acoustical and perceptual levels and in turn, speculate on the relationship between the physical and acoustical levels. It is represented by the orange path in Figure 5.1. We will summarize Chapters II, III, and IV individually in terms of their respective paths. Then, the connections between the different findings will be discussed along with their contribution to the shaping of mental models.

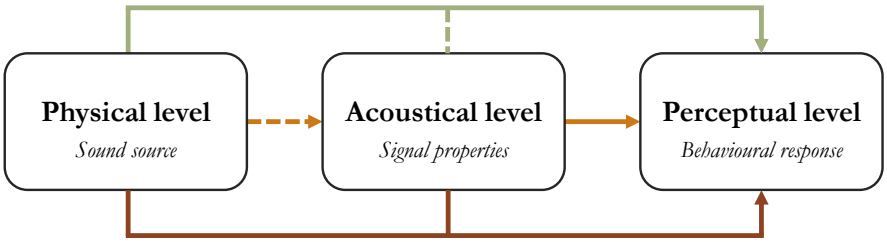

Figure 5.1 Illustration of the relationships between the physical, acoustical, and perceptual levels in the research design and findings of Chapters II (green), III (orange), and IV (brown). Solid lines represent direct influences, and dashed lines represent weak or implicit influences.

5.1.1 Findings from the Categorization Tasks (Chapter II)

For Chapter II, we will refer to the green path (i.e., the top pathway) of Figure 5.1. The experiment of Chapter II involved two categorization tasks. Each task concerned either excitation categorization or resonator categorization of the nine interactions. So, these categorization tasks investigated the direct relationship between the physical and perceptual levels in Figure 5.1. There is some influence from the acoustical level during categorization because the acoustical properties of the sounds should communicate the mechanical components that produced them. However, this influence is weak (as indicated by the dashed green line), given that categorization accuracy was worse for the atypical interactions than the typical ones. So, listeners did not seem to detect some of the reliable acoustic properties that would help them categorize the atypical interactions accurately. For each type of atypical interaction, listeners were better at categorizing its excitation or resonator but seldom both. In general, the atypical interactions were assimilated to typical ones, suggesting that these assimilations were learned through mere exposure (i.e., unsupervised learning) because they were not provided with corrective feedback following the categorization responses. Hierarchical cluster analyses revealed

different groupings of the interactions depending on the categorization task. Cluster analysis of excitation categorization revealed a cluster of struck sounds that were separated from the sustained excitations. Within sustained excitations, there were two clusters: a BlAc-assimilated cluster with BlAc, BoAc, BlPl; and a BoSg-assimilated cluster with BoSg, BlSg, BoPl. These clusters represented confusions between sustained excitations (blowing and bowing here) that are consistent with previous findings (Giordano & McAdams, 2010). During resonator categorization, the hierarchical cluster showed three groups of interactions: a string cluster with BoSg, SkSg, BlSg; a BlAc-assimilated cluster with BlAc, BoAc, BlPl; and a SkPl-assimilated cluster comprising SkPl, BoPl, SkAc. The BlAc-assimilated cluster was consistent in both categorization tasks, but BoPl was assimilated to BoSg in excitation categorization and to SkPl in resonator categorization. There are two main implications of the observed confusions. First, listeners had difficulty perceiving excitations and resonators independently of one another beyond their typical interactions. This result suggests that the features pertaining to the two mechanical components are not perceived as separable in unfamiliar combinations. Second, mental models were not formed for the atypical interactions and listeners instead perceived their sounds as conforming to the existing mental models of the typical interactions. The next experiment was conducted to see if categorization performance could be predicted by similarity in line with previous studies (Giordano & McAdams, 2010; Hjortkjær & McAdams, 2016) and theories (Goldstone, 1994; Nosofsky, 1986, 1989).

5.1.2 Findings from the Dissimilarity-Rating Task (Chapter III)

Participants in the experimental study of Chapter III rated stimulus pairs based on their perceived dissimilarity. These ratings were analyzed with MDS using SMACOF (de Leeuw & Mair, 2009; Elliott et al., 2013) to illustrate the underlying perceptual representation of the nine interactions. Consequently, this experiment examined the relationship between the acoustical and perceptual level as indicated by the solid orange arrow in Figure 5.1. Dissimilarity perception was best explained by three dimensions. Dimension 1 separated the three excitations. The main acoustic correlate of this dimension was the temporal centroid, which is known to distinguish impulsive from sustained excitations (Hjortkjær & McAdams, 2016; Kazazis et al., 2021a; Peeters et al., 2011). Among the two remaining dimensions, Dimension 2 isolated the plate and Dimension 3 further separated the string and air column. Dimension 2 was best explained by a weighted sum of the following acoustic correlates: median spectral centroid, median spectral crest, depth of amplitude modulation, IQR of spectral flatness, rate of amplitude modulation, and median harmonic spectral deviation. Most of these

acoustic correlates generally describe the global shape and tonal/noise content of the spectrum as well as the energy modulation of the signal. For Dimension 3, the acoustic correlates were a weighted sum of median tristimulus 2, IQR of root-mean-square energy, IQR of spectral flatness, median harmonic spectral deviation, and the depth and rate of amplitude modulation. These audio descriptors are associated with fine spectral details, the variability of the noise and energy content, as well as the signal's energy modulation. Table 3.5 (Chapter III) summarizes the general positions that the interactions occupy along the timbre space dimensions based on their excitations and resonators. Acoustic correlates and whether they were positively or negatively correlated to the corresponding dimension are also indicated. For example, BlPl occupies lower positions on D1, lower positions on D2, and higher positions on D3. Because D2 is negatively correlated with median spectral centroid, the plate can be interpreted as having higher spectral centroids than the other resonators. Given that these inferences can be made about the mechanical components of the interactions and the audio descriptors that differentiate them, we argue that Chapter III also indirectly examines a relationship between the physical and acoustical levels (i.e., dashed orange arrow of Figure 5.1). The relationship cannot be concluded as direct because the acoustic correlates were determined using an exploratory model selection approach, and we did not directly manipulate the mechanical components to generate the differences in the audio descriptor values. Overall, the findings from this experiment suggest that there were salient acoustical features guiding the differentiations among the excitations and resonators, even if participants were not explicitly aware of how the sounds were produced. To investigate if listeners can learn to detect the acoustic features reliably determining category membership, a learning paradigm was conducted for the next experiment.

5.1.3 Findings from the Learning Tasks (Chapter IV)

The three experiments in Chapter IV concerned the learning of the excitations, resonators, or interactions of the sounds. Learning involved trial and error with corrective feedback and was subsequently tested using a categorization task without corrective feedback. Chapter IV therefore examined the direct relationship between the physical and perceptual levels, as represented by the solid brown arrow in Figure 5.1. The training phase of each experiment took on a supervised learning approach: after listening to each sound, listeners chose its category, depending on the task, and were provided with corrective feedback. If they correctly identified the categories of at least seven out of nine sounds for four blocks in a row up to a maximum of 23 blocks, then they moved on to the testing phase. Learning was mostly successful given that 40/41, 40/46, and 41/52 participants passed the

training phases when learning was based on the excitations, resonators, and interactions, respectively. During the testing phases, listeners identified the correct excitations, resonators, or interactions even with the removal of corrective feedback. Confusions were much less frequent than in the categorization tasks of Chapter II, and there were no clear patterns showing that the atypical interactions were assimilated to typical ones. These results imply that listeners categorized the source components based on learning the invariant features that differentiated them. Consequently, we speculate that the acoustical level strongly influences the physical-perceptual relationship (i.e, brown solid line in Figure 5.1). As mentioned, the acoustical features undoubtedly communicate information about the mechanical components involved in sound production. In contrast to Chapter II, participants of Chapter IV were able to pick up on these acoustic features to accurately determine category membership of the atypical interactions. This therefore characterizes the brown path of Figure 5.1 as a supervised learning path and distinguishes it from the unsupervised learning path of Chapter II in green. Furthermore, the findings from Chapter IV suggest that new mental models may have been formed for the atypical interactions based on the detection of reliable acoustic properties.

5.1.4 Detection of Structural and Transformational Invariants

Given the perceptual organization of the interactions in Chapter III, we propose that the successful participants in Chapter IV used the timbral information conveyed by the audio descriptors (reported in Table 3.5 in Chapter III) to differentiate the categories based on excitations, resonators, or interactions. As mentioned, one salient dimension that was correlated with the temporal centroid differentiated excitations in Chapter III. So, listeners in Chapter IV might have learned to detect temporal centroid as a transformational invariant. Transformational invariants are the acoustic properties that describe the sound-generating action that causes an object to vibrate (McAdams, 1993). Detection of transformational invariants therefore guides the successful categorization of excitations. Two salient dimensions, each correlating with a weighted sum of spectral and spectrotemporal audio descriptors, differentiated the resonators in Chapter III. So, with supervised learning during Chapter IV, listeners who were trained to learn the resonator categories might have used the combined acoustic correlates of the two dimensions as structural invariants. Structural invariants are the acoustic cues that communicate the physical structure of a sound-generating object, such as its material or geometry (McAdams, 1993). Detection of the structural invariants consequently improved the categorization of the resonators. Furthermore, participants who successfully learned the interaction categories reliably detected the combination of transformational and structural invariants across the three dimensions.

Because these participants learned nine categories that were labeled by the combinations of excitations and resonators, they were likely able to detect the structural invariants that remained constant regardless of the variation in the transformational invariants and vice versa.

Listeners who did not pass the training phase in Chapter IV demonstrated that they were unable to use corrective feedback to guide the detection of transformational and structural invariants. Interestingly, the number of participants who failed the training phase in each type of learning task reflected the number of dimensions that differentiated the source components and their interactions. One, six, and 11 participants failed the training phases of the excitation, resonator, and interaction learning tasks, respectively; and excitation, resonator, and interaction differentiation appeared to map onto one, two, and three dimensions, respectively. So, the difficulty in detecting reliable invariants was associated with the perceived complexity of the category.

5.1.5 Improvements in Categorization Performance

The ability to successfully detect transformational and structural invariants to improve categorization performance can be observed by comparing categorization patterns across Chapters II and IV. The main difference between these two chapters was that Chapter IV incorporated supervised learning. Chapter II, however, was likely based on unsupervised learning given that the atypical interactions were consistently assimilated to typical ones through mere exposure. Because the typical interactions did not require learning and listeners identified their source components correctly across the different categorization tasks, this discussion focuses on the improvement of categorization performance for the atypical interactions.

Figure 5.2 shows the percent correct of excitation categorization for the atypical interactions across three categorization tasks: categorization during Chapter II (white bars) and the testing phase performance for excitation and interaction categorization during Chapter IV (grey and black bars, respectively). Note that for interaction categorization (Chapter IV), we initially calculated the percentage of times each interaction was correctly identified (e.g., in Figure 4.6). For Figure 5.2, we recorded the percentage of times the excitation was correctly chosen during interaction categorization: e.g., if a participant heard BoAc but categorized it as BoSg, it still counted as a correct categorization of the excitation. The percent correct of excitation categorization during Chapter II was lower for BoAc, BlSg, and BoPl in comparison to BlPl and SkAc. As discussed in the previous section, excitation differentiation relied on the detection of transformational invariants. So, listeners in Chapter II had difficulty detecting the transformational invariants for BoAc, BlSg, and BoPl, but they were able to

detect those invariants for BlPl and SkAc through mere exposure. Thus, listeners seemed to learn the excitations of these two atypical interactions in an unsupervised manner.

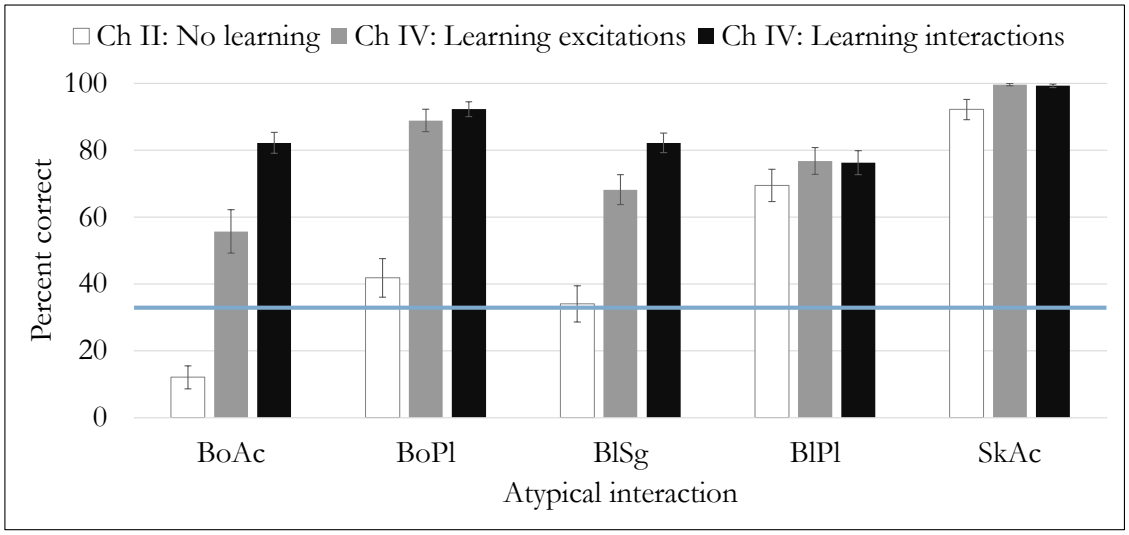

Figure 5.2 Mean percent correct of excitation categorization for each atypical interaction across the categorization tasks of Chapter II (no learning involved) and Chapter IV (testing phases of excitation and interaction learning tasks). Note that for interaction learning (Chapter IV), listeners chose the interaction categories rather than individual excitations. For comparison purposes in this figure, the percentage of times the correct excitation was chosen was recorded (e.g., hearing BoAc but choosing BoPl still counted as a correct categorization of the excitation). Chance performance is indicated by the blue line. Error bars represent the standard error of the mean.

Focusing just on BoAc, BlSg, and BoPl, we can see that learning to categorize them based on their excitations or interactions in a supervised manner (i.e., with error-corrective feedback) (Chapter IV) improved listeners' ability to accurately categorize the excitations. Supervised learning was successful at helping listeners detect the transformational invariants of these interactions. More interestingly, the participants who were successful at interaction learning were better at categorizing the excitations of these atypical interactions than those who were successful at excitation learning. For example, compare the height of the grey and black bars in Figure 5.2 for BoAc: listeners were more accurate at identifying its excitation during the interaction learning task (black bar) than during the excitation learning task (grey bar). A similar pattern can be observed for BlSg, but only very slightly for BoPl. During the interaction learning task, it seemed that even if listeners did not always correctly identify these atypical interactions, they were still able to identify the correct excitation through successful detection of transformational invariants. On the other hand, listeners in the excitation learning task

occasionally confused BoAc and BlSg for other excitations. In either case, the detection of the transformational invariants consequently diminished the previously observed assimilations in Chapter II. Overall, these findings show that the excitations of the atypical interactions can be learned.

In Figure 5.3, the mean percent correct of resonator categorization across three categorization tasks (i.e., categorization in Chapter II [white bars] and resonator and interaction learning tasks in Chapter IV [grey and black bars, respectively]) for each atypical interaction is shown. During the categorization task of Chapter II, listeners were better at categorizing the resonators of BoAc, BoPl, and BlSg than of BlPl and SkAc. This result suggests that unsupervised learning drove the detection of the structural invariants of the former three atypical interactions, whereas detecting the structural invariants of the latter two interactions seemed to be difficult for listeners.

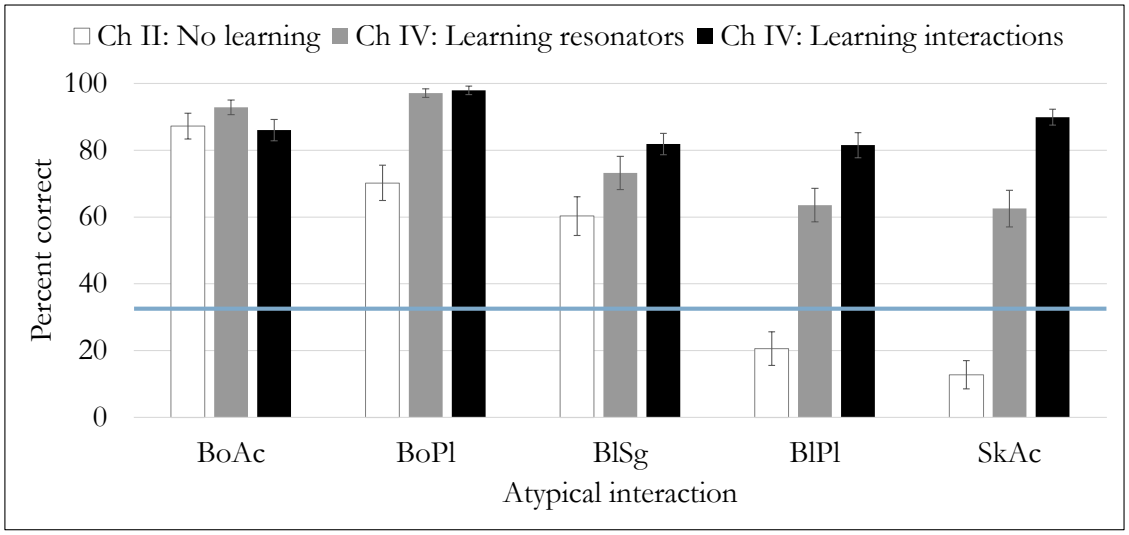

Figure 5.3 Mean percent correct of resonator categorization for each atypical interaction across the categorization tasks of Chapter II (no learning involved) and Chapter IV (testing phases of resonator and interaction learning tasks). Note that for interaction learning (Chapter IV), listeners chose the interaction categories rather than individual resonators. For comparison purposes in this figure, the percentage of times the correct resonator was chosen was recorded (e.g., hearing BoPl but choosing BlPl still counted as a correct categorization of the resonator). Chance performance is indicated by the blue line. Error bars represent the standard error of the mean.

The training phases based on resonator and interaction learning in Chapter IV contributed to the improvement in categorizing the resonators of BlPl and SkAc, suggesting that supervised learning motivated the detection of the structural invariants. Resonator categorization was more accurate for BlPl and SkAc when listeners learned to categorize their interactions than when they learned to

categorize their resonators (i.e., compare grey versus black bars for these two interactions in Figure 5.3). This suggests that the detection of structural invariants was more difficult during resonator learning than during interaction learning for BlPl and SkAc. It also suggests that if listeners did not identify the correct interaction of BlPl and SkAc during interaction learning, they still correctly identified the resonator. Furthermore, listeners were more likely to confuse the resonators when learning was based on resonators than when it was based on interactions. However, given that resonator categorization was well above chance in Chapter IV, as well as the considerably large improvement from Chapter II to IV, the assimilations were reduced, and performance during Chapter IV demonstrates that listeners can indeed learn the resonators of the atypical interactions.

5.2 Contributions to Knowledge

Previous studies investigating sound source recognition have used a synthesis design that manipulates mechanical components of mostly nonmusical sounds. This involved combining different actions (e.g., scraping, dropping, rolling, striking, etc.) with objects made of different materials (e.g., glass, metal, wood, plastic) and testing the recognition of source components using various perceptual tasks (Hjortkjær & McAdams, 2016; Lemaitre & Heller, 2012). Other studies have synthesized hybrids such as a *guitarnet* to capture the perceptual qualities of the clarinet and guitar (McAdams et al., 1995) or chimeric sounds that combine the spectrotemporal fine structure (i.e., source) of one instrument tone with the time-varying spectral envelope (i.e., filter) of a different instrument tone (Siedenburg et al., 2016). Rather than synthesizing hybrids or chimeras, we implemented a synthesis design that simulates the combinations of excitations and resonators in ways that deviate from how they typically behave in acoustic musical instruments (Huynh, 2019). To the best of our knowledge, we might be the first to use Modalys for this type of synthesis design (apart from the "simple blow", an example of a blown string simulation that is provided upon installation of Modalys). With this unique sound synthesis design, we tested how timbre perception plays a role in various perceptual tasks and in the shaping of mental models for sound sources.

5.2.1 Acoustic Correlates of Perceptual Dimensions

Many studies have generated timbre spaces using dissimilarity ratings of musical instrument tones (Grey, 1977; Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Lakatos, 2000; Marozeau et al., 2003; McAdams et al., 1995). Studies using sounds produced by impacted materials have more directly

investigated the relationship between the acoustic correlates of timbre space dimensions and the differences in the source components (Hjortkjær & McAdams, 2016; McAdams et al., 2010). Additionally, Giordano and McAdams (2010) reviewed previous dissimilarity-rating studies and found connections of the acoustic correlates of the perceptual dimensions to the excitation mechanisms and instrument families of musical instruments. The experimental study of Chapter III tests the integration of these findings in a single study. The acoustic correlates of the perceptual dimensions were associated with the perceived differences between the excitation and resonator types. Although the categories of the excitations and resonators were not added as predictors to the regression analyses, these findings suggest that, in line with Siedenburg et al. (2016), categorical information pertaining to mechanical components might inform dissimilarity ratings in combination with acoustic properties.

We found the most salient acoustic correlates of the three dimensions were the temporal centroid (Dimension 1), spectral centroid (Dimension 2), and tristimulus 2 (Dimension 3). Both the log attack time and temporal centroid distinguish impulsive from sustained excitations and are highly correlated with one another (Peeters et al., 2011), but the latter has been proposed to have slightly more explanatory power than the former (Kazazis et al., 2021a). Dimensions 2 and 3 were each associated with other audio descriptors listed in Table 3.5. The weighted combinations of audio descriptors contributing to Dimensions 2 and 3 imply that perception of the different resonators is more complicated than excitation perception. Not only does resonator differentiation require two dimensions, but even within each dimension, there were several audio descriptors that each had their own relative contribution to its respective dimension. We note that there were four audio descriptors that were common among the two resonator dimensions: IQR of spectral flatness, median harmonic spectral deviation, depth of amplitude modulation, and rate of amplitude modulation. However, because the perceptual dimensions underlying a timbre space are orthogonal (McAdams, 2019) and the combinations of acoustic correlates in Dimensions 2 and 3 are not identical, it suggests that different aspects of the spectromorphology of the sounds are being captured by the different combinations and their respective weights. Given that timbre is a complex phenomenon, it is unsurprising that several acoustic properties contribute to the perception of source components. It might be more unrealistic to assume that a given dimension can be correlated to a single, most salient audio descriptor, when in fact, sounds differ based on a plethora of acoustic properties.

5.2.2 Categorization, Similarity, and Learning

One of the goals of this dissertation was to explore the relationship between similarity and categorization across the three experimental studies. There is a common belief that similarity predicts categorization. This is intuitive, as things that are similar often belong in the same category (Goldstone, 1994; Sloutsky, 2003). Similarity also predicts patterns of confusion because members of different categories can be confused for one another if they are perceived to be similar to one another (Nosofsky, 1986). In support of this view, previous studies have found that sounds produced by the same instrument family or similar types of excitations (as in Giordano & McAdams, 2010), or by the same actions and materials (as in Hjortkjær & McAdams, 2016), were often confused for one another and rated as more similar.

Harnad (1990) notes that, although similarity and categorization are correlated, the methods and tasks that measure them are independent. Dissimilarity ratings involve relative judgments: two items are compared based on any information that is available to differentiate them (Harnad, 1987a). Moreover, one does not have to know the identities of two things to judge the dissimilarity between them (Harnad, 1990). Categorization, on the other hand, is an absolute judgment. It is "doing the right thing with the right kind of thing" (Harnad, 2017, p. 22). It involves detecting the features that reliably distinguish category members from nonmembers. Given the independence in their tasks, we hypothesized that listeners might rely on different acoustic information, depending on whether they are categorizing or rating the dissimilarity of sounds. We thought this would especially be true for unfamiliar or novel sounds. A study showing at least partial independence of the acoustic cues prioritized between dissimilarity-rating and categorization tasks was conducted by McAdams et al. (2010). They simulated impacted plates that represented a continuum between glass and aluminum. Although impacted plates are not necessarily unfamiliar or novel to listeners, they likely have not encountered plates that were graded hybrids of two materials. McAdams et al. (2010) reported that listeners relied on damping cues and wave velocity (i.e., pitch differences) to rate the dissimilarity of pairs of sounds but prioritized completely damping cues during material identification of the same set of sounds.

In line with the findings from McAdams et al. (2010), the findings from this dissertation demonstrate that categorization performance (Chapter II) was not necessarily predicted by dissimilarity ratings (Chapter III). We found assimilations of the atypical interactions to typical ones during the categorization task, but the interactions did not seem to cluster in multidimensional space

based on these assimilations. So, listeners prioritized different acoustic cues and processed the sounds differently across these two studies, especially for the unfamiliar, atypical interactions.

On the surface, it may seem that in the comparison of the findings between Chapters III and IV, similarity is now correlated with categorization. However, Chapter IV included a crucial agent that was lacking in Chapter II: supervised learning. Had it not been for supervised learning, categorization performance in Chapter IV would have resembled that of Chapter II. But because of supervised learning, corrective feedback motivated the detection of the appropriate invariants that determined membership to the different excitation, resonator, and interaction categories. This was particularly true for the atypical interactions. So, we infer that before the typical interactions were learned (i.e., at a point when they were novel or unfamiliar), perhaps the relationship between similarity and categorization was not as clear. Following learning, listeners understood that similarities in the sounds' features (be they acoustic or mechanical) can predict categorization. Therefore, we speculate that the correlation between similarity and categorization for unfamiliar and novel stimuli becomes more obvious after supervised learning takes place. That is, after supervised learning, things that belong in the same category will be perceived as more similar, whereas things belonging in different categories will be perceived as more dissimilar (Goldstone et al., 2001; Harnad, 1987a; Pérez-Gay et al., 2014). This result can also explain why the dissimilarity ratings of mostly familiar musical tones have been found to be influenced by their identification (Giordano & McAdams, 2010) or categorical information pertaining to their mechanical components (Siedenburg et al., 2016).

5.2.3 The Role of Attention and Mechanical Plausibility

During each of the discussed experimental tasks, there are intertwining roles of attention and perceived mechanical plausibility. We define attention in terms of what we notice in stimuli under a given context that is informative for us to make predictions or perform certain tasks (Schwartzstein, 2014). Mechanical plausibility is defined in terms of typical versus atypical interactions. Typical interactions (BoSg, BlAc, SkSg, SkPl) are mechanically plausible because they can be physically produced in the real world and their interactions are easy to conceptualize. Atypical interactions are mechanically implausible because it is difficult for listeners to conceptualize their production in the real world. More importantly, they are all physically impossible based on how Modalys (Dudas, 2014) simulates them. We note that two of the atypical interactions, BoPl and SkAc, are more likely to be perceived as mechanically implausible by nonmusicians than musicians. Musicians might be able to conceptualize these interactions, depending on their familiarization with extended playing techniques.

Musicians might have encountered instances where a plate is bowed at its edge or a slap tongue technique is applied to the mouthpiece of a clarinet or saxophone. However, these are not the ways that Modalys simulates these interactions. BoPl is simulated such that the bow passes through a plate that is fixed at its edges. For SkAc, a hammer is applied to the air molecules at a position along the length of the air column. Moreover, categorization performance of BoPl and SkAc in Chapter II confirms that they fall under the class of atypical excitation-resonator interactions as listeners often confused them for a different excitation (as in the case of BoPl) or for different resonators (as in the case of SkAc).

Hierarchical cluster analyses in Chapter II showed that the interactions were grouped differently depending on the categorization task. So, attention to excitations or resonators mediated differential categorization performance and was thus associated with how the interactions were grouped together in the hierarchical cluster analyses. The assimilations of atypical interactions to typical ones can be explained by the interference of and attention to mechanical plausibility. Upon hearing each of the atypical interactions, listeners might have realized they could not identify them. They might have shifted their attention to compare shared features between the atypical interactions and the interactions that are familiar and mechanically plausible.

In dissimilarity perception of Chapter III, performance was guided by listeners' attention to anything that they thought was differentiating the sounds. There were no instructions for them to attend to any specific aspects of the sounds. There was also no interference of mechanical plausibility because listeners did not have to know how the interactions were produced to rate their dissimilarity. Naturally, the interactions were differentiated according to their excitations and resonators based on acoustical cues.

Because the three learning tasks in Chapter IV were based on different source components of the interactions, listeners likely attended to different acoustical features depending on the task. As indicated by the timbre space of Chapter III, when learning was based on excitations, listeners might have attended to the changes in temporal centroid (i.e., a transformational invariant) and ignored the remaining variation (i.e., of structural invariants) to improve categorization performance. However, for the resonator learning task, listeners had to attend to two groups of structural invariants (or at least a more complex subset of acoustic properties) and ignore changes in transformational invariants to reliably sort the sounds into correct resonator categories. This also meant that for the learning task based on interactions, listeners might have focused on the three dimensions of acoustic properties. Not only does the task determine what features listeners might have attended to, but supervised

learning in the form of corrective feedback shapes attention over the course of each task. For example, at the beginning of the training phase of the excitation learning task, listeners might not have known to which aspects of the sounds they should direct their attention. With corrective feedback following each response, they could more accurately decide which transformational invariants (e.g., temporal centroid) to attend to that would facilitate successful categorization. Attention consequently shaped the learning of resonators and interactions in a similar manner. Supervised learning also proved adequate to minimize the interference of perceived mechanical plausibility on categorization performance. Perhaps listeners decided not to focus on whether the sounds could be produced in the physical world, realizing it would not be effective for categorization. Furthermore, Schwartzstein (2014) explains that if attention to appropriate features does not occur, then learning will not occur. So, the participants who were unable to pass the training phase might not have attended to the appropriate features that would have improved categorization and facilitated a learning effect.

5.2.4 The Formation of Mental Models

Mental models are considered internal representations of how systems, much like musical instruments, behave in the world. Becoming more familiar with a musical instrument and the types of sounds it can produce contributes to a better understanding of its capabilities and restrictions. Mental models are not static and can be shaped and updated over time. For example, when a listener hears and/or sees an extended playing technique, such as a slap tongue, being performed on a clarinet for the first time, they will subsequently update their mental model of a clarinet to include the newly encountered timbre. Furthermore, if the listener can see how the sound was produced by a performer, they can begin to conceptualize how they themselves can produce the sound. This is based on the activation of mirror neurons. Mirror neurons fire when a person sees or hears another person doing something, such as playing a musical instrument (Cook et al., 2014). Mirror neurons in turn activate regions of the motor cortex that would be associated with performing the same action. Because the typical interactions resemble the interactions that exist in acoustic musical instruments, it is likely that listeners already have the mental models for them.

We can speculate on the formation of mental models for the atypical interactions, mainly based on the findings of Chapters II and IV. In Chapter III, the most that we can conclude is that mental models of excitations and resonators can be independent, at least implicitly, because listeners seemed to recognize the acoustic similarities among the sounds produced by the same excitations or resonators, regardless of the typicality of the interactions.

Listeners in Chapters II and IV probably did not encounter the sounds produced by the atypical interactions prior to beginning either experiment. To simplify the following explanations, I will focus on the example of the BoAc interaction, but the explanations should apply to the other four atypical interactions. Participants in Chapter II were presented the BoAc sounds and were not given the name of its interaction or the mechanical components that produced them. Hearing the sounds alone might have activated mirror neurons, and listeners began trying to conceptualize how such sounds can be produced. Perhaps listeners thought that the sound best matched those produced by BlAc, a typical interaction for which the mental model exists. So, the BoAc sounds might have been perceived as unique cases of the sounds that can be produced by the BlAc interaction. The listener subsequently updates their mental model of BlAc to include these unique timbres. So, new mental models were not formed for the atypical interactions. Instead, the atypical interactions conformed to the pre-existing mental models of the typical interactions.

The participants in Chapter IV, on the other hand, were presented with the BoAc sound and its category name during the familiarization phase, and through correct categorization and error-correcting feedback during the training phase. Hearing the BoAc sound and associating it with its category name might have activated mirror neurons, and listeners perhaps tried to conceptualize the course of actions required to bow an air column themselves. Encountering more examples of BoAc sounds sharpens this process and initiates the formation of a mental model for this atypical interaction. Of course, for the mental model to be strengthened and exist long-term, listeners would have to encounter many more instances of BoAc sounds (e.g., being played at different pitches and dynamics, and by using different articulations and gestures, etc.). Furthermore, based on the speculation that new mental models were formed for the atypical interactions, we might infer that the mental models for typical interactions were formed in a similar manner at some point early on in human development. For the typical interactions, however, listeners can actually see how their sounds are produced in the physical world. It would be of interest to examine the robustness of mental models for excitation-resonator interactions that can only be experienced digitally and cannot be produced physically.

5.3 Limitations

One of the main limitations in each of the presented experiments concerns the synthesis of the atypical interactions. We used very controlled approaches to simulate the interactions, such that manipulations were applied to only two parameters per interaction. This might be considered

unrealistic because a performer does not simply manipulate two parameters of their instrument across all the notes they can play. It should be noted, however, that changing just one parameter can significantly change the timbre (McAdams & Goodchild, 2017). Moreover, we chose to manipulate parameters that have been commonly manipulated in the physical modeling of acoustic musical instruments: bow speed and bow pressure (Halmrast et al., 2010); breath pressure and embouchure pressure (Dalmont et al., 2005); and a variety of options for percussive sounds, such as, but not limited to mallet force and position of strike (Halmrast et al., 2010).

Another factor concerning the efficacy of Modalys is the simulation of the two sustained excitations, bowing and blowing. We have shown that they are often confused for one another, especially during the categorization task of Chapter II. The physical models of bowing and blowing have been considered to be interchangeable (Ollivier et al., 2004), which might explain the observed confusions. Ollivier et al. (2004) demonstrated that the bowing speed of the bowing excitation can be analogous to the breath pressure of the blowing excitation, and the bow pressure is analogous to the embouchure pressure. If the potential interchangeability of these two excitations plays a primary role in influencing categorization, then we would expect that the atypical interactions produced by a sustained excitation would be categorized as bowing half of the time and as blowing the other half of the time. Interestingly, for some atypical interactions, such as BlSg, listeners more often categorized them as bowed than blown. Because listeners more decisively chose one excitation over the other, we propose that assimilation is primarily at play, whereas the consideration of the interchangeability between bowing and blowing plays a secondary role. We can also argue that this was confirmed by the timbre space generated in Chapter III. Of course, the sustained excitations were perceived as more similar to one another than to the struck sounds; but there appears to be a subtle differentiation between bowed and blown sounds, which indicates that listeners were somewhat able to tell them apart. This implies that the physical models for bowing and blowing might be acoustically different enough in our simulations.

The primary concern for the dissimilarity-rating task in Chapter III was the number of stimuli. With 27 stimuli (i.e., three exemplars of each excitation-resonator interaction), there were too few observations for our statistical analyses. Multiple linear regression can be computed to determine the acoustic correlates of each timbre space dimension. For each regression analysis, the dependent variable was the unique position that each sound occupied on a given dimension, so the number of observations was equal to the number of stimuli. The number of audio descriptors we intended to include as predictors in each model was too large for the number of observations. Such a model would

be prone to overfitting and multicollinearity especially considering that some audio descriptors are naturally highly correlated (Peeters et al., 2011). A statistical method that can account for multicollinearity is Partial Least Squares (PLS) regression, such that sets of highly correlated audio descriptors can be reduced to a smaller number of variables that predict the observations. However, PLS regression requires large sample sizes and consequently a large number of observations (Hair et al., 2011; Peng & Lai, 2012). So, we had to use an exploratory approach with model selection. This involved backward stepwise regression that started with a full model (i.e., including all considered predictors). In each step, a predictor was removed until a model with a smaller number of predictors was found to best explain the data. The Bayesian Information Criterion (BIC) was computed at each step to assess the contribution of a given predictor. The resulting model with the reduced set of predictors has the lowest BIC.

One factor we did not consider in the dissimilarity-rating task (Chapter III) was that there might be asymmetries in the ratings. Siedenburg et al. (2016) conducted a few dissimilarity-rating experiments for three sound sets. One included just recorded tones of musical instruments. Another set comprised transformed sounds which, among other manipulations, included chimeric sounds that combined the time-varying spectrum of one instrument sound with the time-varying spectral envelope of another. The transformed sounds were confirmed to be unfamiliar by listeners. The final sound set comprised recorded and transformed sounds. Siedenburg et al. (2016) found that the combined sound set was prone to asymmetries, particularly when a sound pair comprised one familiar, recorded sound and one unfamiliar, transformed sound. Transformed-recorded pairs generally had higher dissimilarity ratings than recorded-transformed pairs. This is important to consider given that the interactions in the current studies also differ in their familiarity (with atypical interactions being highly unfamiliar). Accordingly, our dissimilarity ratings might have been influenced by asymmetries as well.

In the learning tasks of Chapter IV, we found that listeners could learn to categorize the sounds based on their excitations, resonators, and interactions. The majority of the participants who completed the training phase managed to pass. Moreover, performance between the training and testing phases for passing participants increased from 84.61% to 86.63% for excitation learning, 82.60% to 86.11% for resonator learning, and 75.87% to 84.17% for interaction learning. So, even with the removal of corrective feedback, categorization performance improved at least slightly from training to testing. These learning effects, however, can only be concluded as short-term because we did not conduct a longitudinal study to determine whether learning was maintained over a longer period of time. There is also the concern of whether listeners actually detected invariant features to

reliably determine category membership or if they just relied on their memory of the sounds. We propose that they paid more attention to the invariants because the findings from Chapter III suggested that they are detectable, but the learning paradigm that we used does not allow us to directly determine the strategies for improving categorization performance. A few of the mentioned limitations will be addressed in the next section.

5.4 Future Directions

To address the limitation of the generalizability of our stimuli, a future study can be conducted to manipulate different parameters for each interaction, other than the two parameters we manipulated. We also acknowledge that musical tones in the physical world are not limited to only three types of excitations and three types of resonators. A variety of other excitation and resonator types make up a large quantity of musical tones that were not considered by our study. So, combinations between other excitations and resonators can be simulated. For example, within the impulsive excitations, we can simulate plucking to be compared to, and observe if it is confused for, striking. We can also consider different types of blowing excitations other than those with a single reed mouthpiece. For example, we can simulate air jets, double reeds, and lip valves. Membranes can also be simulated and compared to plates. The same set of experimental tasks can be conducted on an updated stimulus set that includes interactions involving the mentioned excitations and resonators. We can then determine if our findings generalize to other source components that are common to acoustic musical instruments. Not only should other interaction types be included, but more exemplars for each excitation-resonator combination should be added as well. Considering that participants completed the experiments of Chapter II and IV of the current studies within an average of 15 minutes and 40 minutes, respectively, the experimental tasks could include a larger stimulus set. A larger stimulus set with more exemplars for the different types of excitation-resonator interactions would contribute to the robustness of the potential learning effects. Furthermore, the SMACOF algorithm can account for missing ratings, so each participant can still rate the dissimilarity of a subset of pairs if the number of stimuli were to increase. More importantly, increasing the number of stimuli would increase the number of observations, which would benefit the regression analyses aiming to determine the acoustic correlates of each perceptual dimension. This would also allow us to conduct statistical approaches that are more ideal than backward stepwise regression, such as PLS regression or the agnostic approach

implemented by Giordano et al. (2010), which used Principal Component Analysis (PCA) to reduce groups of multicollinear predictors into more general components.

As mentioned, dissimilarity ratings can exhibit asymmetries for pairs of sounds comprising one familiar and one unfamiliar sound (Siedenburg et al., 2016). To determine any effects of asymmetries on the dissimilarity data, we would use the experimental design that Siedenburg et al. (2016) implemented for the dissimilarity ratings of a mixed set of recorded and transformed tones. This would mean that the full matrix of $N \times N$ pairs, including both orders of non-identical pairs would be tested. Then, we could examine if the atypical-typical pairs have higher dissimilarity ratings than pairs presented in the reverse order, which would parallel the findings of Siedenburg et al. (2016). If there are asymmetries in the ratings of atypical and typical interaction pairs, then this would further support the idea that dissimilarity ratings may not be solely based on acoustical information and that they can also be affected by categorical information pertaining to the familiarity of the source.

Given that only a short-term learning effect could be concluded from Chapter IV, it would be interesting to examine if participants can learn the excitation, resonator, or interaction categories over a longer period of time. Longitudinal studies are more prone to attrition, but examining long-term effects of learning would allow us to infer whether new mental models could be formed and maintained for the atypical interactions. Another way to examine the strength of the learning effect would be to adopt an experimental paradigm such as the one used by Pérez-Gay et al. (2017). Participants in their study rated the dissimilarity of visual stimulus pairs before and after a learning task that incorporated supervised learning. The researchers found that successful learners rated stimuli of different categories as more dissimilar after the learning task than before the learning task. To a lesser degree, stimuli belonging in the same category were rated as less dissimilar after the learning task than before learning. This demonstrated that learning generated a perceptual separation of between-category stimuli and a perceptual compression of within-category stimuli. If we were to conduct a similar series of experiments with our sound set and found similar results as Pérez-Gay et al. (2017), then it could be argued that successful learning changes the way the interaction sounds are perceived. Moreover, it would indicate that the learning effect was strong enough to form categorical boundaries between the interactions belonging to different categories.

5.5 Concluding Remarks

Timbre indeed plays a very complex role in sound source recognition. In the body of this thesis, we reported the design and findings of three experimental tasks that investigated the categorization, dissimilarity perception, and learning of sounds that were produced by combining excitations and resonators beyond their typical contexts. In general, the timbres of the atypical interactions were interpreted differently by listeners, depending on the task. Familiarity of the interactions as well as their mechanical plausibility as conveyed by their timbre interfered with categorization in Chapter II. However, the distance perception reported in Chapter III indicated that the excitations and resonators can be differentiated based on acoustic cues, even when listeners did not know the identities of the sounds. These findings also imply that the invariants that differentiate the various excitations, resonators, and interactions from one another were detectable. The invariants of the acoustic cues might have become known to listeners during Chapter IV when they were trained to learn the excitation, resonator, or interaction categories of the sounds. So, mental models can be formed for novel, unfamiliar, and mechanically implausible sound sources based on supervised learning and the detection of invariants that distinguish them from other sound sources for which the mental models already exist. We infer that mental models for the typical interactions—at a point when they were once novel and unfamiliar in early development—were formed in a similar manner. These findings have important implications for our understanding of timbre's role in communicating information about sound source mechanics. Taking these results together, this dissertation highlights timbre as a multidimensional attribute that inevitably contributes to the discernability, identification, and learning of everyday sounds.

References

Agus, T., Suied, C., Thorpe, S., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *Journal of the Acoustical Society of America, 131*(5), 4124–4133. https://doi.org/10.1121/1.3701865

American National Standards Institute (1994). *American national standard acoustical terminology*, *ANSI S1.1-1994*. New York, NY: American National Standards Institute.

Aramaki, M., Besson, M., Kronland-Martinet, R., & Ystad, S. (2009). Timbre perception of sounds from impacted materials: Behavioral, electrophysiological, and acoustical approaches. In S. Ystad, R. Kronland-Martinet & K. Jensen (Eds.), *Computer music modeling and retrieval 2008*, LNCS 5493 (pp. 1–17). Springer. https://doi.org/10.1007/978-3-642-02518-1_1

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Berger, K. W. (1964). Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, *36*(10), 1888–1891. https://doi.org/10.1121/1.1919287

Böttcher, N., Gelineck, S., & Serafin, S. (2007, June). PHYSMISM: A control interface for creative exploration of physical models. In *Proceedings of the 2007 Conference on New Interfaces for Musical Expression* (pp. 31–36). New York, NY. https://doi.org/10.1145/1279740.1279743

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, *118*(1). https://doi.org/10.1121/1.1929229

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319. https://doi.org/10.1007/BF02310791

Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Ystad, S., & Kronland-Martinet, R. (2014). An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling. *Computer Music Journal*, *38*(4), 24–37. https://doi.org/10.1162/COMJ_a_00266

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, *37*(2), 177–192. https://doi.org/10.1017/S0140525X13000903

Coyle, W. L., Guillemain, P., Kergomard, J., & Dalmont, J.-P. (2015). Predicting playing frequencies for clarinets: A comparison between numerical simulations and simplified analytical formulas. *Journal of the Acoustical Society of America*, *138*(5), 2770–2781. https://doi.org/10.1121/1.4932169

Dalmont, J.-P., Gilbert, J., Kergomard, J., & Ollivier, S. (2005). An analytical prediction of the oscillation and extinction thresholds of a clarinet. *Journal of the Acoustical Society of America*, *118*(5), 3294–3305. https://doi.org/10.1121/1.2041207

de Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in *R*. *Journal of Statistical Software*, *31*(3). https://doi.org/10.18637/jss.v031.i03

Dudas, R. (2014). Modalys, Version 3.5.0.rc1 [computer software]. Paris: Institut de recherche et coordination acoustique/musique.

Eckel, G., Iovino, F., & Caussé, R. (1995, July). Sound synthesis by physical modeling with Modalys. In *Proceedings of the International Symposium on Musical Acoustics* (pp. 479–482). Dourdan, France.

Elliott, C. A. (1975). Attacks and releases as factors in instrument identification. *Journal of Research in Music Education*, *23*(1), 35–40. https://doi.org/10.2307/3345201

Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *Journal of the Acoustical Society of America*, *133*(1), 389–404. https://doi.org/10.1121/1.4770244

Ellis, N., Bensoam, J., & Causse, R. (2005). Modalys demonstration. *Proceedings of International Computer Music Conference* (pp. 101–102). Barcelona, Spain: Michigan Publishing.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/10.1037/0096-3445.127.2.107

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*(1), 160–168. https://doi.org/10.3758/BF03196273

Estes, W. K. (1994). Models for category learning. In W. K. Estes (Ed.), *Classification and cognition* (pp. 59–89). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195073355.001.0001

Fabiani, M., & Friberg, A. (2011). Influence of pitch, loudness, and timbre on the perception of instrument dynamics. *Journal of the Acoustical Society of America*, *130*(4), EL193–EL199. https://doi.org/10.1121/1.3633687

Fletcher, N. H. (1999). The nonlinear physics of musical instruments. *Reports on Progress in Physics*, *62*(5), 723–764. https://doi.org/10.1088/0034-4885/62/5/202

Giordano, B. L. (2005). *Sound source perception in impact sounds* [PhD Thesis, Institute de recherche et coordination acoustique/musique].

Giordano, B. L., & McAdams, S. (2006). Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *Journal of the Acoustical Society of America*, *119*(2), 1171–1181. https://doi.org/10.1121/1.2149839

Giordano, B. L., & McAdams, S. (2010). Sound source mechanics and musical timbre perception: Evidence from previous studies. *Music Perception*, *28*(2), 155–168. https://doi.org/10.1525/mp.2010.28.2.155

Giordano, B. L., Rocchesso, D., & McAdams, S. (2010). Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 462–476. https://doi.org/10.1037/a0018388

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157. https://doi.org/10.1016/0010-0277(94)90065-5

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*(2–3), 231–262. https://doi.org/10.1016/S0010-0277(97)00047-4

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2013). Concepts and Categorization. In A. F. Healy, R. W. Proctor, & I. B. Weiner (Eds.), *Handbook of psychology: Experimental psychology* (pp. 607–630). John Wiley & Sons, Inc.

Goldstone, R. L., Lippa, Y., & Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933. https://doi.org/10.1037/a0015781

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, *61*(5), 1270–1277. https://doi.org/10.1121/1.381428

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*(5), 1493–1500. https://doi.org/10.1121/1.381843

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hajda, J. M. (2007). The effect of dynamic acoustical features on musical timbre. In J. W. Beauchamp (Ed.), *Analysis, synthesis, and perception of musical sounds* (pp. 250–271). New York, NY: Springer Science and Business Media. https://doi.org/10.1007/978-0-387-32576-7_7

Halmrast, T., Guettler, K., Bader, R., & Godøy, R. I. (2010). Gesture and timbre. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 183–211). Taylor & Francis.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*(2–3), 137–165. https://doi.org/10.1016/S0010-0277(97)00042-5

Handel, S., & Erickson, M.L. (2001). A rule of thumb: The bandwidth for timbral invariance is one octave. *Music Perception*, *19*(1), 121–126. https://doi.org/10.1525/mp.2001.19.1.121

Handel, S., & Erickson, M. L. (2004). Sound source identification: The possible role of timbre transformations. *Music Perception*, *21*(4), 587–610. https://doi.org/10.1525/mp.2004.21.4.587

Harnad, S. (1987a). Category induction and representation. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 535–565). New York, NY: Cambridge University Press.

Harnad, S. (1987b). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–52). New York, NY: Cambridge University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harnad, S. (2017). To cognize is to categorize: Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 21–54). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00002-6

Hjortkjær, J., & McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *Journal of the Acoustical Society of America*, *140*(1), 409–420. https://doi.org/10.1121/1.4955181

Houben, M. M. J., Kohlrausch, A., & Hermes, D. J. (2004). Perception of the size and speed of rolling balls by sound. *Speech Communication*, *43*(4), 331–345. https://doi.org/10.1016/j.specom.2004.03.004

Huynh, E. Y. (2019). *Bowed plates and blown strings: Odd combinations of excitation methods and resonance structures impact perception* [Master's Thesis, McGill University]. ProQuest Dissertations Publishing. https://escholarship.mcgill.ca/concern/theses/rf55zd06t

Huynh, E. Y., & McAdams, S. (2023a). *Categorization of typical and atypical combinations of excitations and resonators of musical instruments: Assimilation of the unusual to the familiar* [Manuscript submitted for publication].

Huynh, E. Y., & McAdams, S. (2023b). *Implicit differentiation of typically and atypically combined excitations and resonators of musical instruments*. [Manuscript in preparation].

International Organization for Standardization. (2004). Acoustics – Reference zero for the calibration of audiometric equipment – Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones [ISO Standard No. 389-8].

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*(5), 2595–2603. https://doi.org/10.1121/1.407371

Kazazis, S., Depalle, P., & McAdams, S. (2021a). Ordinal scaling of temporal audio descriptors and perceptual significance of attack temporal centroid in timbre spaces. *Journal of the Acoustical Society of America*, *150*(5), 3461–3473. https://doi.org/10.1121/10.0006788

Kazazis, S., Depalle, P., & McAdams, S. (2021b). Ordinal scaling of timbre-related spectral audio descriptors. *Journal of the Acoustical Society of America*, *149*(6), 3785–3796. https://doi.org/10.1121/10.0005058

Kazazis, S., Depalle, P., & McAdams, S. (2022). "The Timbre Toolbox User's Manual", https://github.com/MPCL-McGill/TimbreToolbox-R2021a

Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, *9*(4), 399–410. https://doi.org/10.1162/105474600566907

Krimphoff, J., McAdams, S., and Winsberg, S. (1994). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique [Characterization of the timbre of complex sounds. II. Acoustical analyses and psychophysical quantification]," *Journal de Physique IV*, *4*(C5), 625–628. https://doi.org/10.1051/jp4:19945134

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Vol. 1989, pp. 43–53). Amsterdam: Excerpta Medica.

Krumhansl, C. L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 739–751. https://doi.org/10.1037/0096-1523.18.3.739

Kruschke, J. K. (2005). Category learning. In K. Lamberts & R. L. Goldstone (Eds.), *The handbook of cognition* (pp. 183–201). London: Sage.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, *62*(7), 1426–1439. https://doi.org/10.3758/BF03212144

Lemaitre, G., & Heller, L. M. (2012). Auditory perception of material is fragile while action is strikingly robust. *Journal of the Acoustical Society of America*, *131*(2), 1337–1348. https://doi.org/10.1121/1.3675946

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America*, *90*(6), 3036–3049. https://doi.org/10.1121/1.401778

Lupyan, G. (2006). Labels facilitate learning of novel categories. In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *Proceedings of the 6th International Conference on the Evolution of Language* (pp. 190–197). World Scientific. https://doi.org/10.1142/9789812774262_0025

Lutfi, R. A., & Oh, E. L. (1997). Auditory discrimination of material changes in a struck-clamped bar. *Journal of the Acoustical Society of America*, *102*(6), 3647–3656. https://doi.org/10.1121/1.420151

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, *114*(5), 2946–2957. https://doi.org/10.1121/1.1618239

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology, 11*(2), 64–66. https://doi.org/10.1055/s-0042-1748011

McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 146–196). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198522577.001.0001

McAdams, S. (2013). Musical timbre perception. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 35–67). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-381460-9.00002-X

McAdams, S. (2019). The perceptual representation of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, perception, and cognition* (pp. 23–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_2

McAdams, S., Chaigne, A., & Roussarie, V. (2004). The psychomechanics of simulated sound sources: Material properties of impacted bars. *Journal of the Acoustical Society of America*, *115*(3), 1306–1320. https://doi.org/10.1121/1.1645855

McAdams, S., & Giordano, B. L. (2015). The perception of musical timbre. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed., pp. 72–80). Oxford, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198722946.013.12

McAdams, S., & Goodchild, M. (2017). Musical structure: Sound and timbre. In R. Ashley & R. Timmers (Eds.), *The Routledge companion to music cognition* (1st ed., pp. 129–139). New York, NY: Taylor & Francis.

McAdams, S., & Rodet, X. (1988). The role of FM-induced AM in dynamic spectral profile analysis. In H. Duifhuis, J. W. Horst, & H. P. Wit (Eds.), *Basic issues in hearing* (pp. 359–369). London: Academic Press. https://hal.science/hal-01105536

McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *Journal of the Acoustical Society of America*, *128*(3), 1401–1413. https://doi.org/10.1121/1.3466867

McAdams, S., Thoret, E., Wang, G., & Montrey, M. (2023). Timbral cues for learning to generalize musical instrument identity across pitch register. *Journal of the Acoustical Society of America*, *153*(2), 797–811. https://doi.org/10.1121/10.0017100

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177–192. https://doi.org/10.1007/BF00419633

McIntyre, M. E., Schumacher, R. T., & Woodhouse, J. (1983). On the oscillations of musical instruments. *Journal of the Acoustical Society of America*, *74*(5), 1325–1345. https://doi.org/10.1121/1.390157

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*(4), 279–290. https://doi.org/10.3758/BF03204942

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 282–304. https://doi.org/10.1037/0278-7393.15.2.282

Ollivier, S., Dalmont, J.-P., & Kergomard, J. (2004). Idealized models of reed woodwinds. Part I: Analogy with the bowed string. *Acta Acustica united with Acustica*, *90*(6), 1192–1203.

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport*, *12*(1), 169–174.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, *130*(5), 2902–2916. https://doi.org/10.1121/1.3642604

Peng, D. X., & Lai, F. (2012). Using partial least squares in operations management research: a practical guideline and summary of past research. *Journal of Operations Management*, *30*(6), 467–480. https://doi.org/10.1016/j.jom.2012.06.002

Pérez-Gay, F., Thériault, C., Gregory, M., Rivas, D., Sabri, H., & Harnad, S. (2017). How and why does category learning cause categorical perception? *International Journal of Comparative Psychology*, 30. https://doi.org/10.46867/ijcp.2017.30.01.01

Pollard, H. F., & Jansson, E. V. (1982). A tristimulus method for the specification of musical timbre. *Acta Acustica united with Acustica*, *51*(3), 162–171.

Rizopoulos, D (2022). bootStepAIC: Bootstrap stepAIC. R package version 1.3-0. https://CRAN.R-project.org/package=bootStepAIC

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Lawrence Erlbaum.

Rossing, T. D., & Hanson, R. J. (2010). Bowed strings. In T. D. Rossing (Ed.), *The science of string instruments* (pp. 197–208). Springer New York. https://doi.org/10.1007/978-1-4419-7110-4_12

Ruengvirayudh, P., & Brooks, G. P. (2016). Comparing stepwise regression models to the best-subsets models, or, the art of stepwise. *General Linear Model Journal*, *42*(1), 1–14.

Saldanha, E. L., & Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, *36*(11), 2021–2026. https://doi.org/10.1121/1.1919317

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145

Schwartzstein, J. (2014). Selective attention and learning. *Journal of the European Economic Association*, *12*(6), 1423–1452. https://doi.org/10.1111/jeea.12104

Siedenburg, K., Jones-Mollerup, K., & McAdams, S. (2016). Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in Psychology*, *6*, 1977. https://doi.org/10.3389/fpsyg.2015.01977

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *TRENDS in Cognitive Sciences*, *7*(6), 246–251. https://doi.org/10.1016/S1364-6613(03)00109-8

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Society for Music Perception and Cognition Conference '95*. Berkeley, CA: University of California, Berkeley.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*(2–3), 167–196. https://doi.org/10.1016/S0010-0277(97)00043-7

Strain, G. M. (2017, April 10). *How well do dogs and other animals hear?* Deafness in dogs and cats. https://www .lsu.edu/deafness/HearingRange.html

Srinivasan, A., Sullivan, D., & Fujinaga, I. (2002). Recognition of isolated instrument tones by conservatory students. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, & E. Renwick (Eds.), *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney* (pp. 720–723). Adelaide, Australia: Causal Productions.

Steele, K. M., & Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? *Music Perception*, *23*(3), 215–220. https://doi.org/10.1525/mp.2006.23.3.215

Thayer, J. D. (2002, April). *Stepwise regression as an exploratory data analysis procedure*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 704–712. https://doi.org/10.1037/0096-1523.10.5.704

Wedin, L., & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, *13*(1), 228–240. https://doi.org/10.1111/j.1467-9450.1972.tb00071.x

Zhang, Z. (2016). Variable selection with stepwise and best subsets approaches. *Annals of Translational Medicine*, *4*(7), 136. https://doi.org/10.21037/atm.2016.03.35