Perspectives on Memory for Musical Timbre

Kai Kristof Siedenburg

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

March 2016

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

2016/03/29

bong bing bang bung bäng

für Eliza

Contents

Abstract

This thesis studies musical timbre cognition from the perspective of short-term recognition and dissimilarity rating tasks. Four independent experimental studies provide in-depth investigations into the role of a) timbre similarity and concurrent pitch variability in short-term memory for timbre, b) the impact of sound source categories and familiarity of tones and sequences in timbre recognition and dissimilarity ratings, c) the musical experience of participants and their memory maintenance strategies.

A first theoretical part proposes three conceptual distinctions for the notion of musical timbre and outlines a comprehensive map of previous research on memory for musical timbre. The second part describes experiments on memory for timbre sequences. Using a serial-matching task, the first study shows that musicians do not differ from nonmusicians on sequences with constant pitch, but are better than nonmusicians in recognizing sequences that featured concurrent pitch variability. The perceptual dissimilarity of timbres is shown to be the major determinant of participants' response choices, suggesting parallels of perceptual representation and short-term storage. A musical case study then explores auditory and verbal memory for *tabla*, a pair of drums central to North Indian classical music. Recognition memory of tabla students is compared to naïve controls, using sequences composed of drum sounds, as well as verbal sounds from a dedicated "tabla solfège" (bols) in which syllables stand for specific tabla strokes. The results suggest a partial dissociation of memory for verbal and instrumental sounds.

The third portion of this thesis explores the ways in which sound source categories retrieved from long-term memory affect timbre cognition. Considering timbre dissimilarity ratings for groups of tones from familiar acoustic instruments and unfamiliar digital transformations, the third study reports rating asymmetries that cannot be explained on acoustical grounds. Descriptors related to sound source categories signif-

icantly improve an acoustic model of timbre dissimilarity. The fourth study highlights the finding that musicians and nonmusicians better recognize familiar acoustic timbres than unfamiliar ("non-lexical") transformations of them. Detrimental effects of verbal and visual interference further suggest that short-term memory for timbre relies on attention-based maintenance of the auditory trace.

This research synthesizes and advances the current knowledge on timbre cognition. Several links to hallmark effects of verbal memory are established, including acoustic similarity, sequential chunking, lexicality, and active memory maintenance. The misleading notion that timbre should only be conceived of as a "surface feature" is countered by the demonstration that it can be deeply ingrained in human memory—which explains why it occupies a central role in music listening.

Résumé

Cette thèse aborde la question de la cognition du timbre musical du point de vue de la reconnaissance à court terme et de tâches de jugement de dissemblance. Quatre séries d'expériences indépendantes ont étudié en profondeur le rôle: a) de la similarité du timbre et parallèlement de la variabilité de la hauteur pour la m'emorisation à court terme du timbre, b) de l'impact des catégories de sources sonores et de la familiarité des sons et des séquences de ceux-ci aussi bien sur la reconnaissance du timbre que sur les jugements de dissemblance et c) de l'expertise musicale des auditeurs et de leurs stratégies de maintien des traces mnésiques.

Dans une première partie théorique, trois distinctions concernant la notion de timbre musical sont proposées, et une revue de littérature des recherches préalables sur la mémorisation du timbre musical est également présentée. La deuxième partie présente des expériences sur la mémorisation de séquences de timbre. Par le biais d'une tâche d'appariement de séries, la première étude a mis en évidence que les musiciens ne se différencient pas des non musiciens pour des séquences à hauteur constante. A l'inverse, il s'avère que les musiciens sont meilleurs que les non musiciens dans la reconnaissance de séquences dont la hauteur varie simultanément de note en note. Les résultats ont montré que la dissemblance perceptive entre les timbres est un facteur crucial dans les réponses des auditeurs, suggérant ainsi un parallélisme entre la représentation perceptive et la mémorisation à court terme. Enfin, une étude de cas musicale a exploré la mémorisation auditive et verbale du tabla, instrument composé d'une paire de tambours et jouant un rôle primordial dans la musique classique de l'Inde du Nord. Ainsi, la mémoire de reconnaissance d'étudiants apprenant le tabla a pu être comparée à celle d'auditeurs naïfs. Des séquences composées de sons de tambour, et de sons verbaux issus d'un solfège du tabla (les bols) dans lequel les syllabes représentent des frappes spécifiques ont été utilisées. Les résultats suggèrent finalement une dissociation par-

tielle entre la mémorisation des sons verbaux et instrumentaux.

La troisième partie de la thèse explore les façns dont les catégories de sources sonores issues de la mémoire à long terme affectent la cognition du timbre. En considérant des jugements de dissemblance entre timbres pour des groupes de sons d'instruments acoustiques familiers mais également des modifications numériques du signal acoustique de ces instruments, la troisième étude a mis en évidence des asymétries de jugement ne pouvant être expliquées par des considérations acoustiques. Les descripteurs liés aux catégories de sources sonores ont ainsi permis une amélioration significative d'un modèle acoustique de dissemblance du timbre. Finalement, la quatrième étude a mis en lumière le fait que musiciens et non musiciens reconnaissent mieux les timbres acoustiques qui leurs sont familiers que les transformations non familières (et non « lexicales ») de ces sons. Des effets préjudiciables de l'interférence verbale et visuelle suggèrent en outre que la mémorisation â court terme du timbre est dépendante d'un maintien de la trace mnésique auditive basé sur des processus attentionnels.

Ces recherches synthétisent et améliorent les connaissances actuelles sur la cognition du timbre. Plusieurs liens entre les effets principaux de la mémoire verbale ont été établis, en particulier concernant la similarité acoustique, le morcèlement séquentiel, la lexicalité et le maintien actif de la mémorisation. La notion trompeuse que le timbre devrait être considéré comme un « trait de surface » a été mise à mal par la preuve qu'il peut être enraciné en profondeur dans la mémoire humaine—cela expliquant pourquoi il occupe un rôle central dans l'écoute musicale.

Acknowledgments

This work would not have been possible without the support of many wonderful people. First and foremost, I wish to express my gratefulness to Stephen McAdams, the supervisor of this thesis. He has been an uncompromising researcher, a meticulous reader, and an incredibly patient teacher. His optimism and enthusiasm for science and music served as immense sources of motivation in devising this thesis. It has been a great honor and the greatest fun to learn from him.

Among the other faculty, I would like to thank Philippe Depalle in particular, who was always around for discussing questions on signal processing, and who helped in many other important ways. Ichiro Fujinaga posed an interesting question for my doctoral comprehensive exam, and I am thankful for his help on refining my answer afterwards. Many thanks to Robert Hasegawa, who has always had an open ear, and who provided musical inspirations. I would also like to thank Evan Balaban for getting me to think critically about spectrotemporal modulations. Special thanks go to Petr Janata for warmly welcoming my family and me at UC Davis (during what turned out to be a horrifying Montreal winter), and for inspiring discussions on memory. I would also like to thank Dr. Barbara Tillmann and Prof. Ilja Frissen for serving as thesis examiners.

The Music Perception and Cognition Lab has been a truly inspiring place to work and I wish to thank my fellow labmates for their support and friendship. Special thanks go to Meghan Goodchild, Sven-Amin Lembke, Yinan Tsao, Moe Touizrar, Etienne Thoret, and Chelsea Douglas. I particularly thank Bennett Smith for all kinds of technical assistance and his inexhaustible humor, Cecilia Taher for being such a wonderful and supportive colleague, David Sears for discussions on music, memory, and other fascinating questions I can't recall, and Jason Noble for his collaboration on the CIRMMT Music Cognition Colloquium and the SMPC panel on the Perception and

Cognition of Interesting Music. I wish to credit Rachel Kim, Kiray Jones-Mollerup, Shrinkhala Dawadi and Min-Jee Kim for their hard work of running 165 out of 315 experimental participants.

The Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) has been a great place to combine work, coffee, and daylight. I wish to thank everybody at CIRMMT for their help and I am particularly grateful for the resources provided for the Music Cognition colloquium and the *improv@CIRMMT* series.

I am pleased to acknowledge institutions that have provided financial support for this dissertation project, including the Ministère de l'Éducation, du Loisir et du Sport du Québec, the Auditory Cognitive Neuroscience Training Network supported by the National Science and Engineering Research Council of Canada, Harman International and the Audio Engineering Society Educational Foundation, and the German Academic Exchange Service. Without the generous support of these funding bodies, this work would not have been feasible.

I am grateful to my parents Silke and Heinz Siedenburg for their continual encouragements and support during these many years of *Fortbildung*. I wish to thank my parents-in-law, Gaby and Helmut Latein, for supporting our little family in so many ways. I also would like to thank Sonja and Michel Schriever for their hospitality and the continuous supply with baby stuff.

I thank my flatmates, Ida and Piet. Getting to know these welcoming Canadians has been the most wonderful experience of my life. Living with them has been a constant source of joy and has helped to keep a grounded lifestyle. At last, I cannot even start to express how deeply I am indebted to my wife Eliza Latein, to whom this work is dedicated. Her optimism brought us to Montreal, and her longing for discovery took us beyond. She managed to never loose her understanding for this weird, three-year "timbral mission". All along the way, it was her love, friendship, humor, and selfless support that has kept me swimming.

Contribution of authors

This is a manuscript-based thesis. It contains four core experimental chapters that have been published in peer-reviewed scientific journals, or have been submitted already.

- Chapter 4: Siedenburg, K. and McAdams, S. (submitted 12/2015). Short-term recognition of timbre sequences: Effects of musical training, pitch variability, and timbral similarity. Manuscript submitted to *Music Perception*.
- Chapter 5: Siedenburg, K., Mativetsky, S., and McAdams, S. (submitted 12/2015). Auditory and verbal memory in North Indian tabla drumming. Manuscript submitted to *Psychomusicology*.
- Chapter 6: Siedenburg, K., Jones-Mollerup, K., and McAdams, S. (2016). Acoustic and categorical timbre similarity: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in Psychology*, 6:1977, doi: 10.3389/fpsyg.2015.01977
- Chapter 7: Siedenburg, K. and McAdams, S. (submitted 12/2015, under revision). The role of long-term familiarity and attentional maintenance in auditory short-term memory for timbre. *Memory* (revision).

Stephen McAdams was the supervisor of this thesis and the director of the laboratory in which the research was conducted. He provided guidance in various stages of the research, including the experimental design, analysis methods, and interpretation of results. He also provided funding for the compensation of experimental participants and for the maintenance of technical facilities.

Shawn Mativetsky contributed detailed musical skills and knowledge on tabla (Ch. 5). Individual sounds used in the experiment were recorded from his tabla playing and vo-

calization. He further provided the idiomatic sequences of the bols and strokes, feedback on the resulting sets of stimuli, as well as comments on the tabla-specific parts of the manuscript.

Kiray Jones-Mollerup was a McGill undergraduate research student in Psychology. Under my supervision, she contributed to the experimental design and data analysis for Chapter 6.

As a principal author, I was responsible for devising the experimental paradigms, programming the experimental interfaces, analyzing the data, interpreting the results, and writing the above-listed manuscripts as well as all other parts of this thesis.

List of Figures

List of Tables

List of Acronyms

ANOVA	Analysis of variance
BPM	Beats per minute
c	Bias index
d'	Sensitivity index
ERP	Event related potential
F0	Fundamental frequency
fMRI	Functional magnetic resonance imaging
ISI	Inter-stimulus interval
LMM	Linear mixed model
LTM	Long-term memory
MDS	Multidimensional scaling
MMN	Mismatch negativity
N.S.	Non-significant
PC	Proportion correct
PLSR	Partial least-squares regression
RMS	Root mean square
SDT	Signal detection theory
STM	Short-term memory
WAM	Western art music

Chapter 1

Introduction

The phenomenon of musical timbre has two contradictory faces. From the scholarly perspective, there neither is a commonly accepted music theory of instrumentation, nor a psychological theory of timbre cognition. Yet from the artistic side of the coin, the exploration of sonority has been one of the most important driving forces of musical evolution in the 20th century in both Western classical and popular music, and facets of timbre have always been central to a variety of non-western musical styles. Timbre appears as a black hole or supernova, depending on the observer's viewpoint.

1.1 Ideas and questions

This thesis explores timbre from the perspective of the so-called "higher" cognitive function of memory. Although both timbre and memory are central notions in auditory cognition, memory for timbre as a research topic has gained momentum only recently. More than half of the studies that relate to the issue were published during the last five years. Timbre traditionally refers to the auditory attributes that lend sounds a sense of "color" and enable the identification of sound sources. Major threads of psychophysical work have attempted to explore the nature of its psychological representation and its physical correlates. Starting with the seminal work of von Helmholtz (1885/1954), researchers have traditionally considered spectral factors such as the relative amplitudes of a tone's harmonics to be its key acoustic determinants. Modern approaches have shown that temporal parameters, e.g., the rapidity of a tone's attack, and the spectrotemporal morphology play important roles as well (e.g., Grey, 1975; McAdams,

Winsberg, Donnadieu, De Soete, & Krimphoff, 1995; Elliott, Hamilton, & Theunissen, 2013). (See Siedenburg, Fujinaga, and McAdams (2015) for a recent review of computational approaches.)

Music theoretical reasons for turning towards sonority and timbre may seem elusive, given the pervading "Platonic view" that endorses pitch and duration—precisely encoded in the symbolic code of music notation—and their derivatives of melody, harmony, rhythm, and meter as the primary subject of the musical discourse. But regarding timbre as "secondary" (L. B. Meyer, 1989; Snyder, 2000) is no longer a viable position: Instrumental and electronic sonorities not only play a vital role in the composition of 20th or 21st century "Western art-music" (WAM), which, unfortunately, may appear negligible today in terms of its audience appeal (cf., Fineberg, 2013), but timbre is similarly central in current popular music, for which sound material and qualities are all-important, and where pitch structures sometimes almost constitute a diminutive feature. Let us not forget other musical cultures around the world such as Indian Tabla, Indonesian Gamelan, Japanese Gagaku, and Tibetan Chant which are musical style systems based on timbral contrast (Nattiez, 2007). A century ago, the composer Arnold Schoenberg's famously exclaimed, "Klangfarbenmelodien [tone-color melodies]! How acute the senses that would be able to perceive them! How high the development of spirit that could find pleasure in such subtle things! In such a domain, who dares ask for theory!"(Schoenberg, 1911/1978, p. 422) The classic conjecture is captivating because we are still far from a solid theoretical or cognitive description of timbre's role in music.

But the cognition of what exact kind of musical structures do we seek to understand? Let us turn from the general to the specific, and consider two illustrations (miserably, the only two examples of true musicality to be encountered during the next 200 pages).

In a piece called *Mémoire/Erosion* for French horn and instrumental ensemble (1976), the composer Tristan Murail provided an (almost pedagogically clear) exemplar of the topos of *Klangfarbenmelodie*. Murail's compositional idea was to mimic the effect of an analog re-injection tape-loop using the sound palette of an instrumental ensemble (Murail, 2005). In a re-injection tape-loop, a tape recorder's delayed output is mixed with an independently provided signal or background noise and fed back to be again recorded, which inevitably leads to the degradation of the initial sound trace. The

musical realization of this idea was a sequence of sonorities that contrasted in timbre rather than pitch. In his own words,

> "As in a re-injection loop, the listener will hear each phrase played by the horn repeated after a certain interval of time; it is the other instruments that produce the re-emission. [...] A process of erosion will be played out that, while destroying the original musical structures played by the horn, will gradually reconstitute new structures that, in turn, will be put to the same process of erosion [...]. Several types of erosion are at the heart of Mémoire/Erosion: erosion through timbre, by the wearing out or smoothing of contours, by proliferation, by interference."(Murail, 2005, p. 125)

Figure 1.1 displays the first two pages of the score. The annotation with colored boxes highlights basic grouping structures. One of the first things to note is that the section contains only one pitch, C4. The French horn's first sforzando attack is immediately followed by an array of soft and quickly fading, flutter-echo-like repetitions in the clarinet. The same gesture is restaged in various instrumental colors by being passed on to the pairs of bassoon and flute, clarinet and viola, violoncello and violin. After around 8.5 s, the initial French horn entry repeats in almost identical form (highlighted by horizontal brackets). Although not visible on the first two pages of the score, the process repeats in varied form after around 2:30 minutes into the piece.

In a sense, Murail's *Mémoire/Erosion* suggests the analogy of auditory memory as a type of tape loop. It is a curious coincidence that around the time the piece was written, the seminal Baddeley and Hitch (1974) model of working memory was published, which conceptualized verbal short-term memory via the *phonological loop*, a similar feedback system where phonological information is thought to be continuously re-activated via articulatory rehearsal, and which Baddeley himself called "a tape loop of limited duration"(Baddeley, 1979, p. 356). So how much cognitive reality is there to "Murail's model", and even more so, to the piece? Are the highlighted repetitions, easily visualized on paper, *re*-cognized as such by the listener? How much of timbre's "startling colors" stick in memory, or do we rather memorize the identity of the instruments themselves? Is novel input inevitably overwritten at some point or is there a way to "mentally replay" portions? Do distinct timbres erode less quickly? Why only a single pitch?

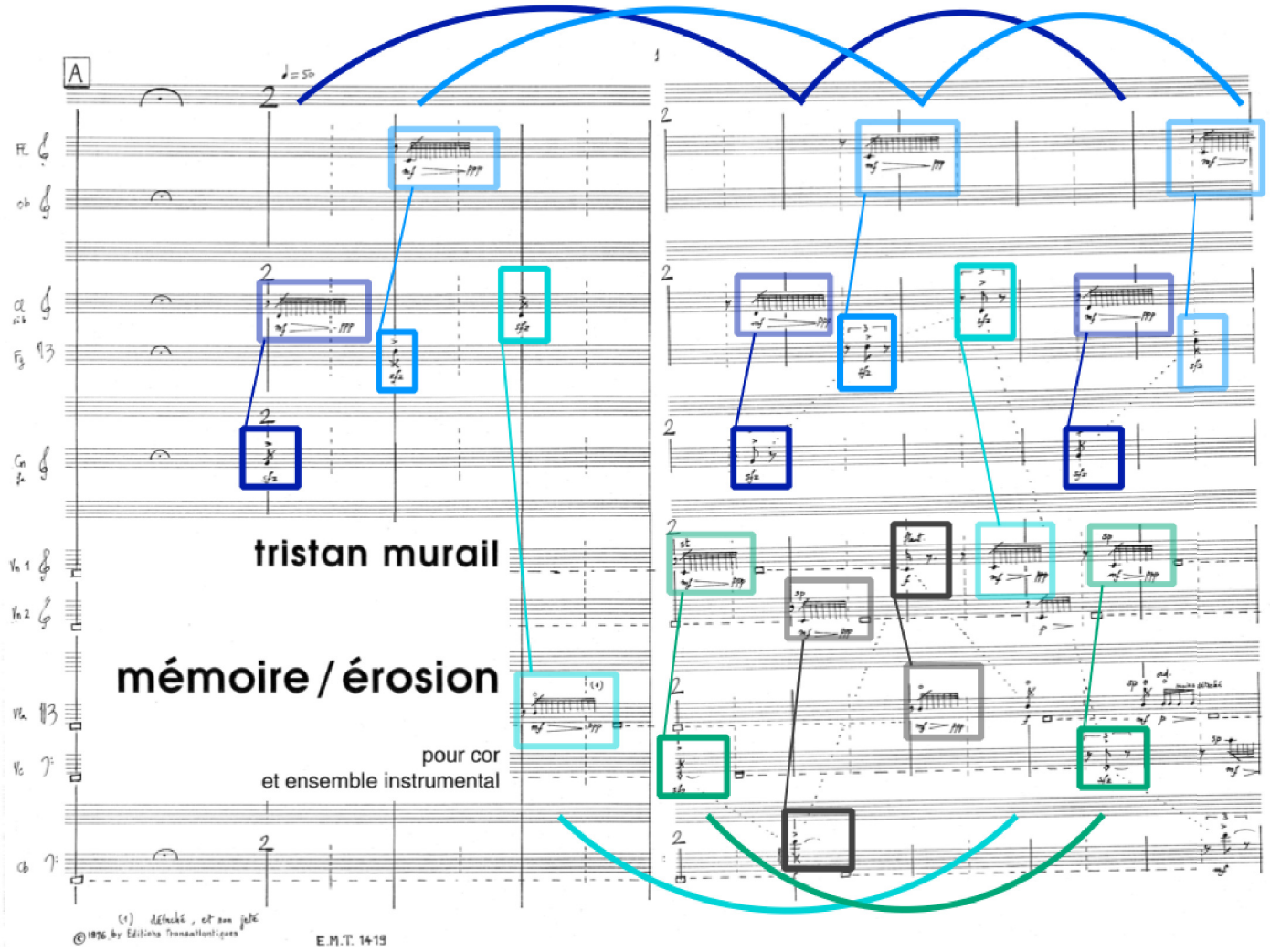

Fig. 1.1 Beginning of Tristan Murail's *Mémoire/Erosion* for French horn and instrumental ensemble. Colored annotation by KS. Score: Editions Transatlantiques, 1976.

Continuing this exercise in analogical reasoning, one could ask what would happen if we fed closely matched verbal and musical material into Baddeley's verbal and Murail's musical loops. Would traces erode at the same pace? This brings us to the second example, and to the musical tradition of another continent. *Tabla* denotes a pair of hand drums of an extremely rich timbral repertoire; this is the most important percussion instrument in North Indian classical music. The centuries old tradition of tabla is taught through an oral tradition. Compositions are learned via the memorization of sequences of *bols*, i.e., solfège-like vocalizations associated with drum strokes, which are also recited in performance.

In tabla solo performance, the verbal recitation of the composition oftentimes precedes the drumming. Figure 1.2 depicts a spectrogram of a selected example.[1] A professional tabla player verbally recited a composition before playing it on the drums. The top depicts the vocal performance, the bottom the corresponding drumming. Notably, there is an impressive coherence of onset timings. Considering that the two signals are generated by different acoustic means, their spectrotemporal morphology appears to feature substantial commonalities. But why does this type of vocalization exist in the first place? Is it abstract poetry or a mnemonic tool? Moreover, could non-experts actually tell the difference between permuted versions of the same sequence?

Questions such as these are inspired by a skepticism about the equation of musical structure and experienced musical form. As formulated by McAdams (1987),

> "Structure is in the world, either notated on paper, stored in computer memory, on magnetic audio tape, impressed on vinyl, or present as vibrations in the air. Form is in the mind and is thus limited by the possibilities of mind."(p. 54)

Differentiating between structure and form is a starting point for acknowledging the intricate memory processes that underlie music listening, and that are by virtue of their ubiquitousness often taken for granted. In order to understand a word or to identify a musical instrument, we integrate information in sensory memory before complete

[1] Retrieved from `http://global.oup.com/us/companion.websites/9780195123753/examples/ch2/svex2.4/` as an audiovisual example of Patel (2008, Ch. 2).

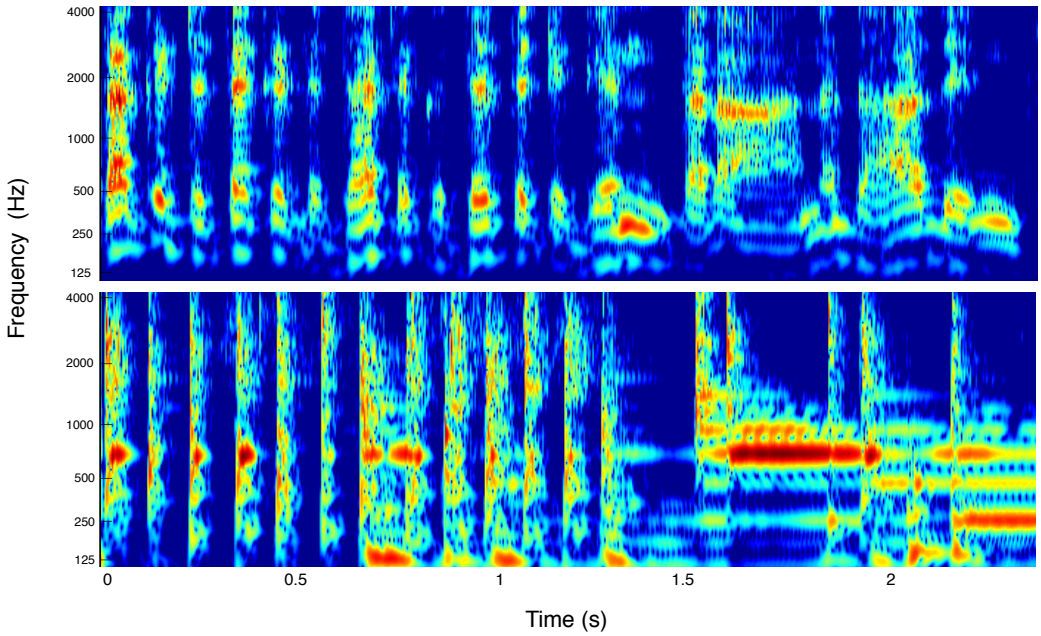

Fig. 1.2 Spectrogram of a recording from a live tabla performance. The composition was first recited verbally using bols (top), and then played on the drum (bottom). Only the first 2.5 seconds of each part are shown.

auditory images of sound events can be formed and matched with templates from long-term memory. For understanding a spoken sentence or recognizing a (timbre-)melody, we need to be able to relate their beginnings and endings in short-term memory. The seemingly most basic mnemonic operation of discriminating two subsequently presented acoustic sequences indeed requires a complex cognitive architecture that keeps track of sound identities and their serial ordering, and concurrently manages perceptual processing, information storage, and matching of representations. That the core part of tests on music processing disorders is constituted by sequence recognition tasks (Peretz, Champod, & Hyde, 2003) illustrates that this type of mnemonic infrastructure is foundational for music listening.

This thesis revolves around the curious mental glue that holds temporally dispersed, fleeting sound events close. More precisely, I investigate acoustic features and psychological mechanisms involved in short-term memory for timbre, and additionally undertake a memory-inspired foray into timbre dissimilarity ratings. The main inno-

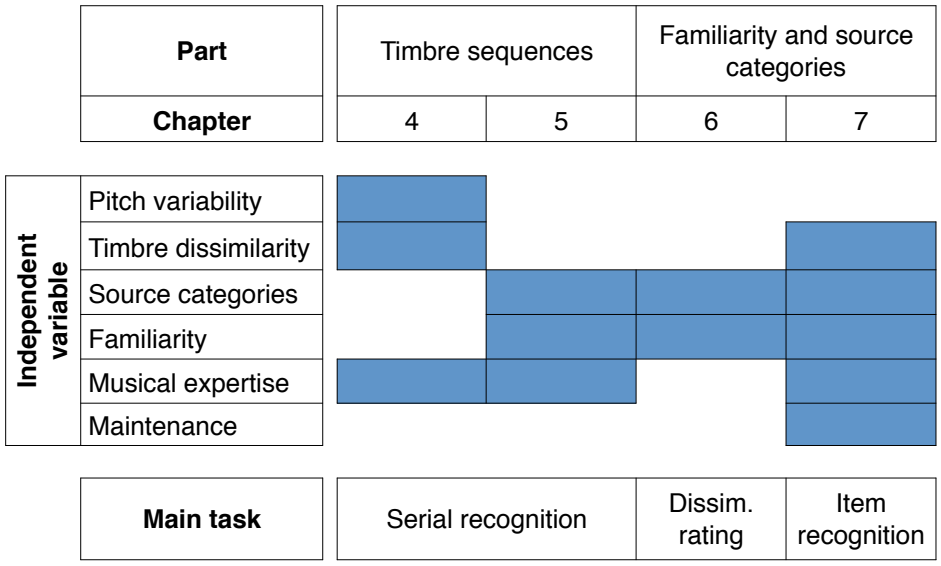

Fig. 1.3 Visualization of the main independent variables and their corresponding chapters, parts, and main experimental tasks.

vation of this thesis is to consider important timbral variables in memory settings, as well as to thoroughly assess effects of listeners' musical expertise. In particular, I will track timbre similarity relations between tones and attempt to account for effects of concurrent variability in pitch. Some familiar timbres may be clearly associable with known sound sources, and I will address whether such categorical affordances affect short-term recognition. Musicians that are familiar with certain sets of sounds may memorize timbre more easily than non-musicians, or could make use of different memory maintenance strategies. As a sideline, I explore how source categories retrieved from long-term memory affect timbre dissimilarity ratings. Fig. 1.3 presents a visualization of these core themes and the chapters in which they will be addressed.

Given timbre's "sisterhood" with speech, this endeavor can also have implications for an understanding of the relations between language and music: after all, speech is the sequencing of vocal timbre, conveying a substantial portion of information via the spectrotemporal shape of auditory units, rather than their pitch or duration. This duality may be particularly well illustrated by an example such as the tabla. Yet, data obtained for pitched instrumental timbres (and variants thereof) may yield similarly valuable boundary conditions on questions regarding the domain specificity of memory models.

1.2 Methods

In the study of complex information-processing systems it is helpful to differentiate three main levels of description (Marr, 2010). On the most global level, the *computational theory* addresses the goal of the computation and its general strategy. The second level of *representations and algorithms* specifies the type of information representations of input and output and the ways these are transformed by algorithms and processes. The third level of *hardware implementation* explicates how representations and algorithms can be implemented physically.

This thesis approaches memory for timbre from a cognitive stance, which means that it seeks to understand the mental representations and algorithms that underlie behavioral memory tasks, corresponding to Marr's second level. Therefore, the cognitive literature will constitute the focus of this work, but neurophysiological studies are taken into account if they contributes to an understanding of the second level (rather than to neurological implementation per se).

Experimentally, the focus will be on participants' behavioral responses as a function of principled stimulus manipulations. For instance, in the classic dissimilarity rating task, two successive tones per trial are presented, and listeners provide a judgment of the respective pairwise similarity on an analog-categorical scale. The employed memory tasks present isochronous sequences of timbres that must either be matched in terms of their serial order, or in terms of the tones' identity, which also is a standard procedure in the current literature.

The dissimilarity data are analyzed in various inferential forms as well as with a regression model that connects acoustic stimulus descriptors, cognitive categories, and behavior. The three short-term memory projects primarily collect binary recognition responses from participants, which are analyzed with signal detection theory (SDT, Macmillan & Creelman, 2005). The benefit of SDT is that it not only yields an unbiased index of discrimination performance (the *sensitivity index d'*), but also a direct measure of response bias, which is a pertinent factor in many recognition tasks and yields an insightful variable in its own right. For the current memory experiments, the Yes/No model is used (Macmillan & Creelman, 2005, Ch. 2), suitable for binary responses. Letting H denote the ratio of hits (e.g., correct responses on match trials) and F that of false alarms (incorrect responses on non-match trials) then the measure of

discrimination sensitivity is given by

$$d' = z(H) - z(F),$$

where z is the inverse cumulative distribution function of the Gaussian distribution. The response bias is measured by the *criterion location* that separates positive from negative responses,

$$c = -\frac{1}{2}\big(z(H) + z(F)\big).$$

Statistical inferences are mostly based on the standard repeated-measures analysis of variance (ANOVA). Although linear mixed models (LMM), taking into account the full structure of the experimental design, have been proposed as an alternative to the classic ANOVA (West, Welch, & Galecki, 2007), we did not use LMMs for two reasons. First, we were hesitant to sacrifice the benefits of the SDT measures which are a valuable interpretative tool for recognition data (Kahana, 2012, Ch. 2), but which require averaging across trials and thus are incompatible with the core of LMMs, namely their by-trial modeling. Second, we found that for the computationally expensive logistic regression models required for the analysis of binary responses, the algorithms often only converged for radically simplified models, which would have further limited the initial benefits of the approach.

1.3 Thesis outline

This dissertation is arranged in four parts. Part I lays out a background for the bridal couple of timbre and memory. Parts II and III constitute the experimental core of the work and deal with memory for timbre sequences and effects of timbre familiarity and source categories on timbre dissimilarity ratings and short-term recognition. Part IV concludes this work.

This is a manuscript-based thesis, which means that Chapters 2–7 attempt to be self-contained, whereas the Conclusion (Ch. 8) integrates findings across chapters. I hope that this modularity, and the fact that the core experimental Chapters 4–7 contain dedicated literature reviews on their own, conveys a sense of theme and variation rather than accumulating boredom (structure $\neq$ form).

Chapter 2, can be read as an introduction to the problem of defining timbre, and

as an overview of the breadth of phenomena associated with the notion. A natural consequence of conceptual breadth is the occurrence of misunderstandings, in particular in interdisciplinary discourses. For that reason, I suggest three conceptual distinctions that may help to contribute to a fresh start. Because one can argue nowhere else as passionately as in the realm of definitions, the chapter is framed as a somewhat polemic essay. In order to lay a groundwork on memory research, Chapter 3 introduces basic concepts and discusses selected key findings in auditory memory. Emphasis will then be placed on the core theme, memory for timbre. I attempt to portray the empirical landscape in comparatively broad strokes, because detailed discussions of particular questions will be given in the following chapters.

Part II comprises two experimental chapters that study short-term memory for timbre sequences. Chapter 4 describes four experiments that investigate the role of concurrent pitch variability in timbre sequence recognition and the impact of musical training, as well as different measures of timbre sequence similarity. As a "real-life" case study, Chapter 5 explores the timbral variety of the tabla. Tabla drumming involves the usage of a verbal solfège system. We compare the recognition performance of tabla students with that of musicians without experience in tabla, and test tabla's drum strokes and solfège vocables in a variety of sequencing conditions. This also allows us to venture into the realm of sequential timbral schemata.

Part III considers the ways in which prior knowledge of instrument categories and their corresponding timbres affect timbre dissimilarity ratings (Ch. 6) and short-term recognition (Ch. 7). At first glance, dissimilarity ratings may not be central to an investigation of memory. As I will argue in Chapter 6, however, not only does memory depend on dissimilarity relations, but dissimilarity relations can be influenced by categorical knowledge about sound sources retrieved from long-term memory. Chapter 6 thus takes a dedicatedly "cognitive" view on timbre similarity by arguing that "it's hard to bypass memory", even in supposedly low-level tasks. The effects of stimulus familiarity and categorical affordances on short-term recognition are scrutinized in Chapter 7. A first experiment tests whether familiar sounds from acoustic instruments are better recognized than unfamiliar synthetic sounds. A second experiment uses a concurrent interference task that draws away participants' attention in order to study the question of maintenance strategies in memory for timbre. Overall, this part attempts to account for the rich array of affordances that timbres from familiar acoustic instruments offer

the listener.

Chapter 8 (Part IV) concludes the thesis. I will summarize the main experimental findings, discuss contributions to the literature, and sketch out a map of processes involved in STM for timbre. The thesis concludes with remarks on the role of timbre in theories of music listening.

Part I

Background

Chapter 2

Three conceptual distinctions for timbre

This chapter features an essay about terminological questions in timbre research. It advocates for a distinction among i) a physical sound and it's timbre, ii) qualitative and source timbre, and iii) different levels of timbral detail. Overall, it may also be read as a conceptual introduction to timbre.

2.1 Introduction

If there is one thing about timbre that researchers in psychoacoustics and music psychology agree on, then it is the claim that it is a poorly understood auditory attribute. One facet of this commonplace conception is that it is not only the complexity of the subject matter that complicates research, but also that timbre is hard to define (cf., Krumhansl, 1989). Perhaps for the lack of a better alternative, one can observe a curious habit in introductory sections of articles on timbre, namely to cite a definition from the American National Standards Institute (ANSI) and to elaborate on its shortcomings. For the sake of completeness (and tradition!) we recall:

> "Timbre. That 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]. NOTE-Timbre depends primarily upon the frequency spectrum, although it also depends upon the sound

> pressure and the temporal characteristics of the sound."(ANSI, 1960/1994, p. 35)

One of the strongest criticisms of this conceptual framing was given by Bregman (1990), commenting,

> "This is, of course, no definition at all. [...] The problem with timbre is that it is the name for an ill-defined wastebasket category. [...] I think the definition [...] should be this: 'We do not know how to define timbre, but it is not loudness and it is not pitch.' [...] What we need is a better vocabulary concerning timbre."(pp. 92-93, but also see, Houtsma, 1997).

In an even more radical spirit, K. D. Martin (1999, p. 43) proposed, "[Timbre] is empty of scientific meaning, and should be expunged from the vocabulary of hearing science." Fifteen years later, while the notion is still part of the terminology, we are far from having reached a clearer taxonomy. One could even ask: Can something useful be done with the wastebasket at last?

The following proposes three conceptual distinctions for timbre. None of these are completely novel; they can be found at various places in the literature to which we will refer. Nonetheless, this is the first attempt of its kind to put these considerations on the same table.

2.2 Sound event vs. timbre

Already the title of Helmholtz's seminal treatise "On the sensations of tone as a physiological basis for the study of music" (von Helmholtz, 1885/1954) distinguishes an external physical sound event (the tone) from its internal perceptual representation (the sensation). If not trivial, the sensation comprises subjective auditory attributes such as pitch, loudness, and timbre, but the physical tone does not. Accordingly, the ANSI definition explicitly addresses sensory attributes. There are, unfortunately, many examples of a different type of usage, where timbre primarily refers to features of physical sound events. These cannot only be found in adjacent academic disciplines such as music theory or music information retrieval, but even in music psychology, where the term is used as a shorthand for a sound event, e.g., a complex tone, the relevant perceptual attribute of which is of timbral nature. This shorthand usage is tempting

but harmful. It encourages the reader to equate the sound event and its timbre, which are in reality connected by a complex sequence of information processing steps, subject to the psychophysics of timbre. This type of usage becomes particularly confusing in conjunction with ecological views of perception, which often appear to circumvent the problem of information transformation by proclaiming a direct correspondence between perception and the world. As noted by Clarke (2005),

> "The amplitude and frequency distribution of the sounds emitted when this piece of hollowed wood is struck are a direct consequence of the physical properties of the wood itself—are an 'imprint' of its physical structure—and an organism does not have to do complex processing to 'decode' the information within the source: it needs to have a perceptual system that will resonate to the information."(p.18)

A crux of the belief that the perceptual system is attuned to the "perceptual invariants" of the environment is, however, that "the detection of physical invariants, like image surfaces, is exactly and precisely an information-processing problem, in modern terminology." (Marr, 2010, p. 30). We need to study the ways in which auditory representations are robust to transformations of the acoustic signal given a specific context, in order to understand the correspondence of tone and sensation (see e.g., Carruthers et al., 2015, for a neuroscientific approach).

One can even observe more hazardous attempts to rephrase timbre as not primarily depending on perception. In a recent ANSI critique from a composer's viewpoint, Roads (2015) states,

> "It [the ANSI definition] describes timbre as a perceptual phenomenon, and not as an attribute of a physical sound. Despite this, everyone has an intuitive sense of timbre as an attribute of a sound like pitch or loudness (e.g., 'the bassoon timbre', or 'Coltrane's saxophone sound'). From a compositional point of view, we are interested in the physical nature of timbre. We want to know how timbre can be made operational, in order to manipulate it for aesthetic purposes."(p. xviii)

Note that this falsely considers all three mentioned auditory attributes to be physical attributes. It is further proposed, paradoxically, that timbre is useful as a musical, but

not as a scientific, notion: "The anachronistic term 'timbre' will likely be superseded by a more precise taxonomy of sound qualities, at least in scientific discourse. In any case, timbral issues are unavoidable in electronic music."(p. xx) In order not to let the indispensable interdisciplinary discourse around timbre disintegrate into terminological incoherence, we should right from the start resist tempting shorthands, and clearly separate physical sound events or tones and their morphologies (as well as their representations via musical scores, sampled time-pressure audio signals, spectrotemporal analyses, etc.) from the resulting auditory sensations. The two distinctions that follow consequently address timbre as a perceptual attribute.

2.3 Qualitative vs. source timbre

The brain is often viewed as a miraculous inference machine (Fuster, 2003). Given sufficient musical experience, an auditory sensory representation may for instance activate violin-specific networks, that could consist of palettes of sensory templates in an auditory lexicon, items in a lexicon of verbal labels, visual templates, or outlines of motor programs associated with the violin (cf., McAdams, 1993). Overall, this process of perceptual inference yields a semantic representation of a sound source that can remain invariant across drastic changes in the acoustic properties of a sound (Handel, 1995). Studies that define timbre as having "a key role" in identification (Patil, Pressnitzer, Shamma, & Elhilali, 2012), or as "foundational" (Elliott et al., 2013) or among the "primary perceptual vehicles" (McAdams, 2013) for sound source identification essentially see timbre as that functional collection of auditory sensory features that *enable* one to infer sound sources via the association of sensory and semantic networks. In contrast, the ANSI definition is not based on an identification task. It addresses an internal representation of auditory nature and does not call for perceptual inference: Two sounds can be declared as dissimilar without bearing semantic associations or without being identified.

Important modern studies on timbre dissimilarity perception do not fully distinguish between qualitative and source timbre in semantic terms (see e.g., Caclin, McAdams, Smith, & Winsberg, 2005; Patil et al., 2012; Elliott et al., 2013), whereas they operationalize timbre in the latter, qualitative sense: Stimuli are equalized in subjective pitch, loudness, and duration, and rated on subjective dissimilarity which

formally does not require any source inference. The latent structure that underlies dissimilarity ratings is then modelled by acoustic properties, implicitly assuming that dissimilarity ratings are solely based on the sensory representation of the sounds' acoustic features and not influenced by semantic categories. It is questionable whether source timbre can be neglected for acoustic stimuli, however, as it can be argued that listeners "can't help" but to integrate semantic information in dissimilarity ratings of Western orchestral instrument tones (Ch. 6, also see, Giordano & McAdams, 2010). In order not to conflate a study of sensory similarity with semantic factors, it is important to take into account the distinction between qualitative and source timbre.

Electronic and digital means of sound production can challenge listeners in their ability to associate timbre and source. As noted by Smalley (1994), "In electroacoustic music, however, sources and causes are many, varied, evident or ambiguous, actual or implied, unknown or unknowable: we can perhaps detect traces of cause or source but realise that neither can exist in reality."(p.37) Such perceptual challenges highlight the interplay between what Gaver (1993) has called *everyday* and *musical* listening, directing attention either towards source-cause identification or the auditory attributes themselves, respectively. From a biological standpoint, so-called *musical listening* may seem obscure because it appears to be bare of evolutionary significance. Although both modes are not mutually exclusive, the conscious sensory introspection of *musical listening* may be the dominant mode of listening in the concert hall (or in a psychoacoustic experiment). As illustrated by Smalley (1994),

> "Although we may recognise immediately that it is water it takes longer to determine what we might refer to as its inner 'timbral' detail because it is not the odd globule but textural behaviour which establishes its imminent identity, for example its resonances, noise content, and pitch-streams or contours. Identifying a crude source-cause is one thing; penetrating its behavioural detail is another."(p. 41)

2.4 Timbre on different scales of detail

When Helmholtz noted "By the quality of a tone [*Klangfarbe*] we mean that peculiarity which distinguishes the musical tone of a violin from that of a flute or that of a clarinet, or that of the human voice, when all these instruments produce the same note at the

same pitch."(von Helmholtz, 1885/1954, p. 10), he (perhaps unwittingly) provided the textbook definition of timbre for the next 150 years. This sentence operationalizes timbre via the perceptual differences based on the distinct acoustics of sound sources such as the flute and clarinet, and, as the ANSI definition, only allows a comparison of timbre across tones with the same pitch, loudness, and duration.

Apart from the cul-de-sac that this deprives any non-pitched sound of its timbre (Bregman, 1990, p. 92), the approach neglects the fact that most pitched musical instruments can give rise to whole palettes of distinct timbral qualities which covary with pitch and loudness (cf., McAdams, 2013, section I.C). Not only do different playing techniques and articulations affect physical and timbral properties of tones (e.g., Barthet, Guillemain, Kronland-Martinet, & Ystad, 2010), but a tone's spectral content also covaries with fundamental frequency (F_0) and playing effort. Low-pitched registers comprise many partial tones, higher tones do not. A *fortissimo* comes along with many pronounced partials (and a correspondingly bright timbre), in a *pianissimo* the amplitudes of higher order partials are attenuated significantly (J. Meyer, 1995). On an even smaller scale that holds excitation-related aspects constant, there may be differences between sounds from exemplars of the same type of sound-producing objects or algorithms (such as a Stradivarius violin and an inexpensive factor-made model)—whether this translates into audible timbral differences or not is subject to the domains of instrument and audio quality (e.g., Fritz, Blackwell, Cross, Woodhouse, & Moore, 2012; Lindau et al., 2014).

The acoustical covariance of F_0 and spectrotemporal envelope shape appears to lead to small but reliable interactions of pitch and timbre. Regarding qualitative timbre, Marozeau, de Cheveigné, McAdams, and Winsberg (2003) collected pairwise timbre dissimilarity ratings for sets of acoustic instrument tones with varying fundamentals (B3, C♯4, and B♭4). When pitch was held fixed within pairs, ratings correlated between $r = .81$ and $r = .88$ across the three levels of F_0. When pitch varied within pairs, ratings were affected most strongly by large pitch differences. Considering source timbre, Handel and Erickson (2001) showed that non-musicians were reliably above chance when instructed to discern whether tones with pitch differences of less than an octave originated from the same musical instrument. Steele and Williams (2006) further showed that musicians even performed well across ranges up to 2.5 octaves. These studies thus suggest systematic, albeit moderate effects of pitch on both qualitative and

source timbre of acoustic instrument tones. The corresponding pitch-timbre "covariance matrices" are likely to be used as a valuable perceptual cue for source identification (Handel & Erickson, 2004), although this research topic has been barely explored. Moreover, even if spectral envelope shape and fundamental frequency are orthogonally varied by means of electronic sound synthesis, they interfere in perceptual processing. Instructed to discriminate timbre, listeners systematically confuse variability in timbre and loudness with that of pitch (Melara & Marks, 1990; Caruso & Balaban, 2014), which tends to be relatively unaffected by musical training (Allen & Oxenham, 2014). Dissimilarity ratings on sounds differing in timbre are similarly affected by variation of fundamental frequency, even when listeners are instructed to ignore pitch (Marozeau & de Cheveigné, 2007).

In sum, it is misleading to suggest that one sound-producing object or instrument yields exactly one timbre. Contrary to parlance of "the timbre of the clarinet", there is no single timbre that *fully* characterizes the clarinet. The timbre of a clarinet tone depends on pitch, playing effort, articulation, fingering, etc. In terms of a biological analogy, a single type of sound-producing object or sound-synthesis algorithm yields a "timbral genus" that may encompass various "timbral species". These species may differ along various parameters, such as playing technique, covariance with pitch and loudness, or expressive intent. Genera group into "families" (e.g., string vs. brass timbres) and at some point into "kingdoms" (timbres related to, say, acoustic vs. electronic means of sound production). Overall, this yields a "hierarchy of embedded distinctions" (Krumhansl, 1989, p. 45), that integrates scales of different timbral detail to which the ANSI definition is agnostic and the textbook definition ignorant.

2.5 Conclusion

By proposing three basic distinctions for the notion of timbre we hope to clear up some confusion around what has been claimed to be the terminological wastebasket of music psychology and psychoacoustics, musical timbre. We proposed to locate timbre on the perceptual side of the "psychophysical divide", i.e., in the mind of the listener instead of in physical properties. We further argued that the notion timbre comprises other components: qualitative and source timbre, large- or small-scale differences (e.g., between "timbral families" or "species"). We do not claim that this is an exhaustive

categorization—more fine-grained taxonomies are necessary depending on the subject matter. In any case, once a few layers of dust are removed, what we had thought of as a wastebasket turns out to be a colorful umbrella(-term) upside down.

Being aware of timbre acting as an umbrella term, one might object, does not solve problems because it does not solve any "real questions". On the contrary, we believe that a sharpened terminology allows us to more flexibly direct our attention towards decisive empirical gaps: From pressure waves in the air to abstract sound source categories in our head, what are the, say, five most important information transformations involved in timbre representation? Do qualitative timbre dissimilarity ratings and source identification rely on the same set of auditory features? Could affordances for source identification facilitate timbre recognition? Does concurrent variability in pitch (according to the pitch-timbre "covariance matrix") aid in sound source identification?

The composer Philippe Manoury (1991) observed that "One of the most striking paradoxes concerning timbre is that when we knew less about it, it didn't pose much of a problem."(p. 293) This can also be put in more optimistic terms: We already know much about timbre. We understand its plentiful, distinct colors are real, and they won't go away. It is time to let inadequate standards rest and start to focus on the specifics.

Chapter 3

A review of research on memory for timbre

This chapter provides a background on basic notions in memory research and discusses ideas and key findings in auditory memory. The main part of this chapter is devoted to a review of experimental work on memory for timbre. I will discuss findings on the role of timbre in memory for melodies, imagery for timbre, and end with a detailed list of previously studied experimental factors on short-term memory. Pertinent research questions will be derived therefrom.

3.1 Introduction

What makes music recognizable? Is it the way that melody and harmony unfold through time? From "spinning the radio dial", we know that we can often classify musical genres and identify individual songs from surprisingly short snippets. Results by Schellenberg, Iverson, and McKinnon (1999) showed that even 100 ms excerpts could be matched to song title and artists with above-chance accuracy, and that time-varying high frequency information (>1 kHz) in particular is important for correct identification. Krumhansl (2010) and Filipic, Tillmann, and Bigand (2010) further specified that emotional content can be consistently judged for excerpts of 250–300 ms length. These findings demonstrate that portions of attributes such as song identity or emotional tone live on fine time scales (e.g., of a granularity of eighth notes at a tempo of 120 beats per minute). What remains when melody, harmony or rhythm are

effectively ruled out as critical features of the memory trace? In fact, "a little piece of timbre", capturing the spectrotemporal configuration of instruments and voices and their emergent musical texture, appears to be enough for the identification of high-level musical features.

This introductory example warrants further exploration of the acoustical features and psychological processes underlying memory for timbre. This review synthesizes research activity in this emerging field. Globally speaking, this pursuit is part of streams of research that discover the realms of (non-symbolic) sensory working memory and long-term memory (e.g., Jolicoeur, Levebre, & Martinez-Trujillo, 2015; Andrillon, Kouider, Agus, & Pressnitzer, 2015), domains that not so long ago may have been hard to find on the psychological map.

At first glance, memory and timbre seem like an odd couple. Musical timbre is a veritable academic niche. The study of human memory, on the contrary, features a proud history and a sheer (if not burdensome) wealth of empirical data. A Google Scholar search of the terms "musical timbre" and "short-term memory" yields differences in search results of three orders of magnitude (around 3,000 vs. 1,000,000). In that sense, one would suspect memory science to be an orderly field of scientific inquiry with commonly agreed-upon theoretical principles. Yet upon closer inspection, one finds that foundational issues have remained controversial. As Surprenant and Neath (2009) remarked,

> "In over 100 years of scientific research on memory, and nearly 50 years after the so-called cognitive revolution, we have nothing that really constitutes a widely accepted and frequently cited law of memory [...]. However, there are a plethora of effects, many of which have extensive literatures and hundreds of published empirical demonstrations."

To take a basic example, how many different kinds of memory are there? Different schools of memory research not only disagree about whether memory should be conceived as unitary or as split into various sub-systems, but also about whether it makes sense to speak about short-term memory as separate from long-term memory (Tulving, 2007; Surprenant & Neath, 2008). As Dudai (2007) formulated, "the most critical concept in the science devoted to its analysis, memory, never had the privilege of sailing

the tranquil waters of consensus. "(p. 11) Note that we encountered comments of a similar flavor in Chapter 2 on timbre.

In order to set the stage, a short glimpse on conceptual issues is provided in Section 3.2, before Section 3.3 outlines recent findings in auditory memory. These are of immediate relevance for timbre because they demonstrate the retention of fine-grained sensory representations over short and long retention intervals, challenging classical accounts of auditory sensory memory. The second part of this chapter, Section 3.4, provides a comprehensive review of studies on the role of timbre in melodies, electrophysiological and behavioral studies of LTM for timbre, as well as studies of STM tasks. Most attention will be paid to STM tasks which were used by a number of recent studies. Because most memory tasks involve many parameters, and given that the phenomenon turns out to be highly context-dependent (Roediger, 2008), the experimental sampling remains sparse and many questions are left open. A subset of them will be discussed in Section 3.5.

3.2 Basic concepts in memory

The notion of memory comprises the tripartition of information i) representation, ii) persistence and maintenance, and iii) reactivation. As Dudai (2007) summarizes, "memory is the retention over time of experience-dependent internal representations, or of the capacity to reactivate or reconstruct such representations."(p. 11) Sensory representations or cognitive states thus stand at the beginning of memory. But it is only their trajectory through time, that lends the phenomenon its full dimensionality. The memory trace, or *engram*, may be recoded, suffer from interference, or be reactivated through maintenance processes. The engram is mutable and elusive, and it is the necessary condition for memory. But a free-standing engram does not constitute memory by itself. Only if the engram interacts with internal or environmental cues, and only if this gives rise to reactivation or retrieval, are the sufficient conditions for the emergence of a memory fulfilled (Moscovitch, 2007).

A commonplace distinction is that between types of memory of different longevity. William James (1890/2004) already thought of primary (conscious, short-lived) and secondary (unconscious, long-lived) memory as independent entities. A more fine-grained distinction became the core of the classic *multistore* or *modal* model, most

prominently elaborated by Atkinson and Shiffrin (1968). It posits three types of stores, namely a sensory *register*, a short-term store, and a long-term store. Sensory information is subject to modality-specific, pre-attentive storage of fast decay (within 2 s), unless there is a subject-controlled "scan" by selective attention, which recodes and transfers portions of the register to the short-term store. This store is thought to retain a categorical, modality-independent code where traces decay within time spans of less than 30 seconds. Their life-span can be lengthened by active rehearsal, which lends them more time to leak into the long-term store.

Beyond its intuitive appeal, among the most persuasive scientific type of evidence for a dissociation of STM and LTM were neuropsychological studies of patients (such as H.M., cf., Scoville & Milner, 1957) who showed normal STM but strongly impaired LTM due to lesions in the temporal lobes and the hippocampus. Influential streams of research subsequently refined the description of the types of representations, information transformations, and maintenance processes within the hypothetical stores. The prominent Baddeley and Hitch (1974) model of working memory decomposed the short-term store into a central executive, a phonological buffer for speech, and a "visuospatial sketchpad" for visual information. The *phonological loop* was conceived as the instantiation of verbal working memory, whereby phonological information circulates between a *phonological buffer* and a *central executive* (the locus of attention). Regarding long-term memory, scholars have parceled out a variety of dissociated memory *systems*, starting with the distinction between *episodic* and *semantic* memory (Tulving, 1972), or the characterization of non-declarative procedural and perceptual-representation-based memory (cf., Squire, 2004).

Instead of conceptualizing memory as a dedicated cognitive faculty, implemented by a multitude of interacting modules (e.g., STM and LTM), an alternative *proceduralist* approach understands memory as an emergent property of the ways in which mental processes operate on perceptual representations or cognitive states (see e.g., Crowder, 1993; Fuster, 2003; Surprenant & Neath, 2008). Famously, Craik and Lockhart (1972) demonstrated that elaborate semantic encoding yields more robust memory traces than superficial perceptual analysis, known as the *levels-of-processing* effect. Contemporary evidence that supports a proceduralist view of STM comes from neuroimaging studies (cf. Jonides et al., 2008; D'Esposito & Postle, 2015). Whereas multicomponent models envision working memory to emerge from information being projected back and forth

between dedicated stores and a central executive system localized in dedicated brain regions (Baddeley, 2003), recent imaging studies indicate that information is stored by the same neural ensembles by which they are perceptually processed. Jonides et al. (2008) conclude that the same neural representations are involved in STM and LTM. Specifically,

> "the same neural representations initially activated during the encoding of a piece of information show sustained activation during STM [...] and are the repository of long-term representations. Because regions of neocortex represent different sorts of information (e.g., verbal, spatial), it is reasonable to expect that STM will have an organization by type of material as well."(p. 201)

Short-term or working memory then emerges as the result of the allocation of attention towards sensory representations or bundles of items in semantic LTM (D'Esposito & Postle, 2015). Representations may be in a potentially capacity-limited focus of attention (Cowan, 2001), and once attention is removed, they transition into a heightened state of activation that is subject to decay and interference.

Another novel type of evidence comes from computational modeling of behavioral data. Addressing the distinction between recognition memory and repetition priming, Berry, Shanks, Speekenbrink, and Henson (2012) compared unitary models, models that embodied multiple memory systems such as explicit and implicit memory, and models that comprised independent recollection and familiarity signals. Surprisingly, overall measures of fit favored a model that relied on a single (signal detection theoretic) signal of memory strength. Although this does not deny that memory can manifest itself in different ways (perhaps 256?, cf., Tulving, 2007), and that it is useful to have a refined vocabulary for these differences, it challenges the assumption that there are strictly independent memory *systems* underlying these manifestations.

Before we proceed, it should be noted that we use STM and LTM as referents to memory function over short- or long *time intervals*, but not to refer to dedicated systems or *stores*. This agnostic usage acknowledges that there may be different time scales of information persistence (e.g., momentary, short, long), as well as different types of formats of information representation (e.g., episodic, semantic, sensory), but does not a priori assume any particular stores or structural configuration of these two "axes". The

modal model, on the contrary, conflates these two facets, by advancing that sensory representations vanish within the moment, and that verbal items constitute the primary currency of the short-term store. Finally note that short-term and *working memory* are often used interchangeably. We consider working memory as a more complex form of cognition that encompasses short-term retention plus the goal-directed manipulation of that information. For most of the tasks we will be concerned with, it thus suffices to speak about STM.

3.3 Key findings in auditory memory

The traditional models of (short-term) memory were developed to account for categorical verbal stimuli (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974). As already mentioned, sensory representations were assumed to have a transient life in the sensory register, before a categorical phonological code could take over. This image of *echoic* memory as attention-independent "echo in the mind" is still the pervasive form of reasoning about auditory STM. It is modeled as a form of leaky integration or temporal smearing (Massaro & Loftus, 1996; Leman, 2000) and is assumed to play a salient role in the processing of musical pitch structures. Nevertheless, it seems intuitively clear that this account must be incomplete. Deutsch (1975) remarked early on,

> "Although we commonly recognize melodies and long works of music by name and can, with musical training, label abstracted tonal relationships, the basic process of music recognition cannot conceivably be verbally mediated. We constantly recognize melodies as familiar without having learned their names. Further, we can accurately identify very short sequences taken from the middle of long works of music. [...] It is clear from such considerations that musical information must be stored in highly specific form for substantial periods of time. [...] It must be concluded that the sensory attributes of a stimulus survive in memory after verbal encoding, and that they continue to be retained in parallel with the verbal attributes."(pp. 3–4)

3.3.1 Auditory sensory memory

Among others, Cowan (1984, 2015) proposed a subdivision of auditory memory. The approach emphasizes parallels with vision, where one finds a seemingly clear structural divide between an automatic sensory storage of almost unlimited capacity and fast decay (< 200 ms) and a more long-lived, attention-dependent short-term memory system of constrained capacity. Cowan's *short auditory store* is hypothesized to be experienced as sensation or sensory afterimage (i.e., is distinct from the sensory type of memory required to integrate and bind perceptual features such as loudness or amplitude modulations over tenths of seconds), contains unanalyzed, pre-categorical content, and has a decay within 200–300 ms. The *long auditory store* is experienced as (short-term) memory, contains partially analyzed or categorized content, and is supposed to decay within 2–20 s. Due to the structural similarity of the long store and categorical STM regarding decay rates and capacity, Cowan (1988) considered the long auditory store to be a special case of STM. This proposal underlined the notion that the faculty of STM may operate on sensory representations, not only on verbal items as assumed in the classic multistore models.

Although Cowan's distinction between a short and automatic and a long and consciously controlled form of auditory memory has intuitive appeal, recent data suggest that it is hard to find clear-cut boundaries. Let us first consider the mismatch-negativity (MMN), an electrophysiological marker of change in a repetitive acoustic pattern, derived from auditory event-related potentials (ERP). During electrophysiological recordings, participants usually attend to a visual stimulus (e.g., a silent movie) while being presented with a sequence of frequent standard sounds and occasional deviants (the "oddballs"). The negative-going wave of the MMN occurs between 100–200 ms after stimulus onset in the difference waveform of the standard's response subtracted from the deviant's response. It has been widely interpreted as an index of a pre-attentive form of auditory sensory memory (e.g., Näätänen, Paavilainen, Rinne, Alho, et al., 2007), which is based on a neural comparison process that detects auditory changes. This perspective appears to be increasingly challenged by a (potentially more parsimonious) account that locates the origins of the MMN in the neural adaptation to the standard, yielding a mismatch because the deviant recruits a fresh population of afferents that are not habituated (May & Tiitinen, 2010). According to this view,

the MMN may better be interpreted as an index of sensory habituation, and therefore stimulus salience (Fishman, 2014). This casts doubt on the traditional interpretation of the MMN as reflecting the operation of a pre-attentive auditory sensory memory, although the debate is not yet resolved (cf. Cowan, 2015).

Behavioral studies highlight difficulties in estimating the exact duration of the shorter type of auditory memory. Testing the discrimination of frequency shifts within non-harmonic tone complexes, Demany, Trost, Serman, and Semal (2008) observed a gradual decay in performance for increasing retention times, not paralleling the steep decline in iconic memory. Importantly, there was no clear sign of differential memory capacity (i.e., a short store of high capacity and a long store of low capacity) within the two-second range of retention times tested. Demany, Semal, Cazalets, and Pressnitzer (2010) more explicitly compared visual and auditory change detection. Whereas visual memory fidelity appeared to decay quickly and drastically within 200 ms, confirming the classical view on iconic memory, there was no such sign for auditory memory which persisted throughout retention times of 500 ms. This indicates that auditory change detection may operate on much longer time scales than visual iconic memory.

3.3.2 Memory for noise

Another type of work that provides an interesting background for timbre uses random waveforms, usually not assumed to possess any global, psychoacoustically well-defined attribute at all. Kaernbach (2004) showed that the periodicity of repeating noise waveforms could be well detected up to at least 10 s of segment length, and single, seamless repetitions of noise waveforms were detected with above chance accuracy up to 2 s. Agus, Thorpe, and Pressnitzer (2010) even demonstrated that there is a form of long-term persistence for features of noise waveforms (for replications of various sorts including neurophysiological data, see, Agus & Pressnitzer, 2013; Luo, Tian, Song, Zhou, & Poeppel, 2013; Kumar et al., 2014; Andrillon et al., 2015). Requiring listeners to detect repetitions of noise segments, it was observed that reoccurring noise stimuli featured far superior hit rates compared to novel noise waveforms. Notably, subjects were not aware that segments reoccurred. This demonstrates that there is implicit, non-declarative long-term auditory memory even for small sensory details.

Overall, these data support the idea that there may be gradual differences in audi-

tory memory fidelity as a function of retention time, but question whether it is possible to clearly discern between one hypothetical auditory store and another solely based on their decay rates or capacities. Listeners not only distinguish subtle repetitions of noise waveforms over short time spans, but also retain implicit, fine-grained long-term memories of noise. These findings cannot be accounted for by the idea of auditory memory as leaky integration, because there would be no discernible auditory (noise or texture) feature left after temporal integration. On the contrary, selective attention appears to play an important role in these results, as it allows for the extraction of potentially idiosyncratic features from otherwise random, featureless waveforms. Accounting for the role of attention seems to be a prerequisite to integrate the various findings on the retention of fine-grained auditory representations reviewed thus far. At some point, it may even be hard to distinguish between process and function, auditory attention and STM. Reflecting a proceduralist position on auditory STM, Alain, Arnott, Gillingham, Leung, and Wong (2015) remarked,

> "Although we can ascribe different perceptual and cognitive operations when we are asked to pay attention to something versus when we are asked to keep something in mind over a period of time, the brain's approach to effecting those behaviors [...] may not follow within the boundaries that our operational definitions provide for what it means to attend or to memorize. Traditionally, selective attention and working memory functions have been studied separately. Yet, increasing direct and indirect evidence suggests that there is considerable overlap in the neural network supporting these two core functions."(p. 224)

Keeping track of the role of attention may be equally central for the study of short-term memory for timbre, as is further pointed out below.

3.4 Memory for timbre

Given that research on memory for timbre is still in its infancy, a general issue underlying many studies is to carve out basic commonalities with other auditory attributes such as pitch, and to consider its general relation to domains such as verbal memory. More specific questions involve the ways in which pitch and timbre interact in memory,

and whether the timbre of the human voice may occupy a privileged mnemonic status.

3.4.1 Timbre in memory for melodies

An important subset of studies highlights the importance of timbre in memory for melodies. Radvansky, Fleming, and Simmons (1995) and Radvansky and Potter (2000) demonstrated that adults' long-term recognition memory for melodies is susceptible to a change in the timbre of the comparison melody, a result that was later verified even for 6-month old infants (Trainor, Wu, & Tsang, 2004). Halpern and Müllensiefen (2008) observed that this effect is unaffected by whether participants' attention was directed towards timbral features via an instrument categorization task or to structural melodic features via a judgment of melody familiarity. Lange and Czernochowski (2013) further showed that a change in timbre affects conscious recollection more strongly than the familiarity-based portion of recognition. Most recently, Schellenberg and Habashi (2015) demonstrated that a change of instrument (from piano to saxophone) impaired melody recognition nearly as strongly as a pitch transposition of six semitones or a tempo shift of 64 BPM. Testing melody recognition after ten minutes, one day, and one week, there was no forgetting but rather a tendency for consolidation of melodic memory. Surprisingly, melody recognition was equally good or better after one week compared to 10 minutes of retention time. Overall these studies suggest that long-term melody recognition does not draw solely from an abstract "lexicon" of melodies to which a perceptual token is matched in the recognition process, but that it relies on rich traces that integrate various perceptual features including timbre. Similar findings have gained prominent status in verbal memory research (e.g., Goldinger, 1996).

Weiss and colleagues added an interesting perspective to this literature by advocating that not all timbres are created equal: Vocal melodies appear to be better recognized than melodies played by musical instruments, although in their experiment the voice timbre was less preferred (Weiss, Trehub, & Schellenberg, 2012). This also held true for nine- to eleven-year-old children (Weiss, Schellenberg, Trehub, & Dawber, 2015) and was independent of musical training (Weiss, Vanzella, Schellenberg, & Trehub, 2015). In a similar vein, Agus, Suied, Thorpe, and Pressnitzer (2012) found that vocal sounds were classified faster as such, compared to acoustic instrument sounds. Suied, Agus, Thorpe, Mesgarani, and Pressnitzer (2014) added that vocal signals re-

quired shorter gated excerpts to be correctly classified.

One interpretation of the vocal advantage suspects that voice sounds attract greater attentional resources and thereby afford a more robust encoding (Weiss, Trehub, Schellenberg, & Habashi, 2015). Although this seems reasonable regarding the biological significance of conspecific vocalizations, it will also be important to address alternative accounts of the phenomenon. For instance, vocal timbres are highly familiar, such that one could expect effects based on familiarity-based processing fluency (Berry et al., 2012). On an even more profane level, it seems crucial to rule out potential confounds due to low-level features, such as loudness. For instance, the aforementioned studies that suggested vocal superiority all normalized root-mean-squared (RMS) signal energy between stimuli, which is an imperfect measure of loudness (e.g., Chalupper, 2008). Critically, Bigand, Delbé, Gérard, and Tillmann (2011) showed that RMS- and peak-amplitude normalization differentially affected the categorization of short spoken voice, classical music, and environmental sounds. Notably, a voice processing advantage only arose with RMS normalization. The need to revisit low-level factors in the interpretation of supposedly "high-level" cognitive effects has recently also been underlined in vision (Firestone & Scholl, 2015).

3.4.2 Neurophysiological perspectives and long-term memory

Addressing pre-attentive auditory sensory memory, the early work of Tervaniemi, Winkler, and Näätänen (1997) showed that changes in timbre elicit MMNs even if tones are as short as 150 ms. Caclin et al. (2006) focused on distinct perceptual dimensions of timbre by synthesizing tones that varied on the dimensions of attack time, spectral centroid, and attenuation of even harmonics. They found that uni-dimensional timbral change elicited MMNs of different latencies, which combined sub-additively. Together with dipole modeling, this result suggested the existence of partially separate MMN generators, i.e., implying the existence of separable, attention-independent processes of timbral change detection.

Other electrophysiological studies have suggested that musicians exhibit specific cortical and subcortical responses to their primary instrument (Pantev, Roberts, Schulz, Engelien, & Ross, 2001; Shahin, Roberts, Chau, Trainor, & Miller, 2008; Strait, Chan, Ashley, & Kraus, 2012), demonstrating that extensive experience with a certain timbre

modulates basic components of auditory processing. By recording MEG, Pantev et al. (2001) observed that professional trumpet players and violinists exhibited stronger N1 ERP components to sounds from their own instrument. The N1 is a negative going peak of the ERP at around 100 ms after stimulus onset that is interpreted as indexing pre-attentive processes related to stimulus detection. This finding suggests that after sufficient musical training, even pre-attentive processing of tones may differ between different groups of musicians. Shahin et al. (2008) further showed in a longitudinal study that gamma-band (25-100Hz) oscillations in EEG-recordings can be enhanced by a year of piano training in children. The same gamma signal differentiated adult musicians from non-musicians in their non-attentive response to different musical timbres. Further research showed that learning not only affects cortical activity, but even modulates "low-level" processing: Strait et al. (2012) showed that filtered brainstem recordings of pianists more closely correlated with the amplitude envelopes of the original piano sounds, compared to recordings of musicians who did not have extensive experience with the piano, but there was no difference between groups for sounds from the tuba and bassoon.

This literature indicates that changes in timbre not only evoke general auditory mechanisms of change detection such as the MMN, but that there may be instrument-specific plasticity that affects the perceptual processing of tones. Nonetheless the reviewed studies did not present behavioral data to anchor their neurophysiological findings. The extent to which the aforementioned results reflect aspects of conscious perceptual experience thus remains unclear.

Collecting both neurophysiological and behavioral data, Halpern, Zatorre, Bouffard, and Johnson (2004) provided evidence for the viability of mental imagery of timbre. They asked musicians to rate perceived dissimilarity of subsequently presented pairs of timbres, while recording brain activity with fMRI. The same procedure was repeated in a condition in which the auditory stimuli were to be imagined. In both perception and imagery conditions, the auditory cortex showed activity with a right-sided asymmetry, and ratings from the two conditions correlated significantly. A different type of behavioral data on auditory imagery for timbre had been provided earlier by Crowder (1989) and Pitt and Crowder (1992). Here, listeners gave pitch-discrimination judgments for pairs of tones differing in pitch and timbre. When tones were identical in timbre, responses were faster and more accurate than when they differed. The same

task was then used in a second experiment, although timbres of the first tone now needed to be imagined and matched in pitch with a pure tone. Overall slower but qualitatively similar results were obtained, suggesting that subjects were able to form accurate mental images of timbre.

Evidence for a very different type of long-term memory for timbre is inspired by the psychological literature on implicit learning of statistical or rule-based auditory regularities (Saffran, 2003; Reber, 1989). Bigand, Perruchet, and Boyer (1998) first demonstrated that implicit learning of timbre sequences is feasible. Tillmann and McAdams (2004) extended these findings by examining the interplay of acoustic similarities and transition rules in the learning of timbre triplets. In an exposure phase, participants listened to a string of tone triplets differing in timbre. In a test phase they were required to distinguish previously encountered triplets from novel triplets. Interestingly, the amount of learning observed was independent of the acoustic structure of the sequence. However, high acoustic similarity within statistical segments and low similarity at segment boundaries supported the perception of timbre grouping and thus fostered overall recognition performance. Overall these results suggest an intricate interplay of potentially automatic acoustic change detection mechanisms, and rule-based long-term traces. An interesting extension of this type of work would be to test whether listeners also generalize the timbral sequencing rules beyond the veridical sequences presented in the learning phase, that is, whether they acquire abstract timbral schemata.

3.4.3 Short-term timbre recognition

Research on short-term memory for timbre is a surprisingly recent endeavor, and most studies have only been published during the last years. Although there is no single guiding question underlying this line of research, all studies use recognition tasks, such that it is worth comparing them at the level of experimental design. Close attention to experimental details is crucial in any case, given that any "law of memory" appears to be affected by contextual variables (Roediger, 2008). In order to provide an overview, Table 3.1 summarizes the specific experimental tasks, independent and dependent variables, groups of participants (whether musicians, nonmusicians, or unspecified), and the number of participants per individual experiment (N). Independent variables are

marked by an asterisk (*) if they yielded significant effects in a majority of experiments within the respective study. Table 3.2 provides the stimulus parameters of list lengths, delay periods, durations of individual tones, inter-stimulus interval (ISI) between tones in the list, whether the source tones were of acoustic or synthetic origin (or from commercial synthesizers), and their (range) of fundamental frequencies. In the remainder, these facets of experimental design are discussed, from which we derive open questions and implications for future work. Secondly, two structural issues are briefly discussed. These address the relation of memory for timbre and pitch, as well as the role of attention in STM for timbre.

Different fields At first glance, the number of participants per experiment could be considered an incidental feature. Nonetheless, the feature reflects the fact that publications on STM for timbre originate from two distinct fields. Studies that feature, say, less than ten participants per experiment are rooted in psychophysics, where it is common practice to extensively train and test a small number of participants, essentially assuming a "standard observer". The results of these studies thus reflect the performance of a small number of participants who usually possess extensive experience on a given experimental task. In music cognition, to the contrary, it is common to test a greater number of participants (>10) with less extensive tests, rather leaning towards the assumption of the "standard stimulus" in psychometrics (Berglund, 2012).

Tasks Apart from two exceptions (Starr & Pitt, 1997; Cousineau et al., 2013), the aforementioned groups also differ in terms of experimental tasks. Psychophysical studies use item-wise same/different discrimination or the classic interpolated tone task which presents additional interfering stimuli between the standard and comparison items (Deutsch, 1970). Studies from the "cognitive camp" usually present standard sequences of tones (so-called memory "lists") with around three to six items, inspired by classic verbal memory tasks (e.g., Sternberg, 1969). One type of matching task then probes memory for serial order by letting participants assess whether a comparison sequence that contains the same items was presented in the same serial order ("serial recognition"). Two studies use slight variations of this task by requiring participants to discriminate (pitch, brightness, loudness) contours (McDermott et al., 2008) or the replacements of a single item in the comparison sequence (Cousineau et al., 2013). A

Table 3.1 Summary of studies on STM for musical timbre. Independent variables (IVs) that yielded significant effects on the listed dependent variable (DV) are indicated with an asterisk (*).

Study	Task	DV	IVs	Training	N
Starr and Pitt (1997)	interp tone	PC	brightness*, pitch proximity, repetition frequency, musical training	nonmus & mus, unspec	55, 24, 28, 43
McDermott et al. (2008)	contour rec	ROC area	material (items differing by F0, spectrum, or intensity), contour streching*	unspec	17, 29, 20, 30
Demany et al. (2008)	s/d discr	d'	delay, no. of components of complex tone	unspec	5,4,3,4
McKeown and Wellsted (2009)	s/d discr	d'	freq of pure tone inducer*	unspec	6, 4, 3, 2
Tillmann et al. (2009)	serial rec	PC	group*, material (sequences of words, pitches, timbres)*	amusic & nonmus	20
Mercer and McKeown (2010)	interp tone	d'	distractor feature overlap*	nonmus & mus	4, 4
McKeown et al. (2011)	s/d discr	d'	delay*, articulatory suppression	unspec	3
Marin et al. (2012)	serial rec	PC	length*, group*	amusic & nonmus	26
Schulze and Tillmann (2013)	serial rec	PC	length, material* (sequences of words, pitches, timbres)*	nonmus & mus	20, 20, 25
Nolden et al. (2013)	serial rec	ERP	length*	unspec	47
Golubock and Janata (2013)	item rec	k	length*, delay*	unspec	52, 36
Cousineau et al. (2013)	seq rec	d'	material* (items differing by pitch, brightness, loudness)	unspec	6
Mercer and McKeown (2014)	s/d discr	d'	delay*, alerts*, inter-trial interval	unspec	6, 4
Soemer and Saito (2015)	item rec	PC	list length*, delay*, interference*	unspec	60, 36

Key. Task column: Interpolated tone paradigm (interp tone), recognition (rec), same/different (s/d), discrimination (discr), sequence (seq). Dependent variables (DV) column: Proportion correct (PC), receiver operating characteristic (ROC), event-related potential (ERP), working memory capacity index (k). Training column: Musicians (mus), nonmusicians (nonmus), unspecified population of participants (unspec). No. of participants per experiment (N).

Table 3.2 Summary of studies on STM for musical timbre (cont'd).

Study	List length	Delay [ms]	Tone dur [ms]	ISI [ms]	Source	F0 [Hz]
Starr and Pitt (1997)	1	5000	300	na	synth	110 (A2) – 880 (A5)
McDermott et al. (2008)	5	300	300	0	synth	100 (∼G2)
Demany et al. (2008)	1	0, 250, 750, 2000	600	na	synth	na
McKeown and Wellsted (2009)	1	350	200	na	synth	na
Tillmann et al. (2009)	5	3000	500	40	comm synth	330 (E4)
Mercer and McKeown (2010)	1	10,000	200	na	synth	130 (C3) – 220 (A3)
McKeown et al. (2011)	1	5, 10, 20, 30 $\times 10^3$	300	na	synth	130 (C3) – 493 (B4)
Marin et al. (2012)	4–8	3000	500	40	rec	311 (D#4)
Schulze and Tillmann (2013)	3–6	3000	500	20	rec	330 (E4)
Nolden et al. (2013)	1–3	2000	200	0	synth	440 (A4)
Golubock and Janata (2013)	2–6	1, 2, 4, 6 $\times 10^3$	250–391	0	(comm) synth	330 (E4)
Cousineau et al. (2013)	1, 2, 4	400	200	0	synth	125 (∼B2)
Mercer and McKeown (2014)	1	2, 32 $\times 10^3$	200	na	synth	146 (D3) – 466 (Bb4)
Soemer and Saito (2015)	2–4	3, 12 $\times 10^3$	500	500	comm synth	131 (C3)

Key. Source column: Digital synthesis (synth), commercial synthesizer (comm synth), recorded acoustic sound (rec).

task that tests memory for item identity also presents a standard sequence of tones, but then requires participants to assess whether a single probe item was part of the previously presented sequence or not ("item recognition").

Musical training It becomes clear from Table 3.2 that only a small portion of studies (3/14) tested between-subjects factors related to musical expertise, two of which compared amusic individuals with non-musician control groups (Tillmann et al., 2009; Marin et al., 2012). Testing both musicians and non-musicians in their first experiment, Starr and Pitt (1997) did not find any significant differences between groups and thus suspended this variable for the remaining three experiments. Schulze and Tillmann (2013) only found a weak correlation between the number of years of participants' musical training and recognition accuracy in one out of three experiments. These results suggest that the role of musical training, investigated intensively for various other aspects of music cognition, is either negligible in STM for timbre, which the current state of the literature advocates, or has not yet been approached from the right angle.

Stimuli selection and timbre familiarity Most studies worked with synthesized tones. Of course, the rationale behind this is to use well-defined stimulus materials and to avoid anything but auditory memory—if sounds can be easily identified, then supposedly auditory tasks may in fact boil down to verbal ones because participants are able to maintain the instrument labels, for instance.

An example of the effect of stimulus selection was given by Golubock and Janata (2013). They observed surprisingly low STM capacity estimates for synthetic tones, but capacity increased once a more variable set of tones was used. This was interpreted as an effect of a higher overall timbral variability in the second set (selected from commercial synthesizers instead of from a custom made synthesis algorithm). The tones were selected to minimize perceptual familiarity (assessed subjectively by the authors), which is the same stimulus selection strategy used by Soemer and Saito (2015).

However, the great portion of subjectivity in the mentioned selection criterion for sounds from commercial synthesizers may demand for further experimental scrutiny. Moreover, it has never been tested whether and how the affordance for sound source

identification affects STM for timbre, as there is no direct comparison of timbre recognition performance for natural acoustic and synthetic sounds. This addresses potential interactions between LTM and STM, because familiar acoustic instrument sounds can be assumed to possess more LTM "baggage". To date, however, it is not clear whether phenomena such as the lexicality effect in verbal memory (words are better recalled than non-lexical pseudo-words) (Thorn, Frankish, & Gathercole, 2008) are singular to the domain of language.

Regarding LTM, there may be another, "horizontal" type of familiarity that may affect timbre sequence recognition. Certain timbre *transitions* may be more familiar than others. Sequential schemata are one of the primary occupations of the field of music cognition, but have only rarely been considered for timbre (Bigand et al., 1998; Tillmann & McAdams, 2004), and not yet in an ecologically realistic scenario. Such familiar transitions may facilitate the chunking of timbre sequences and effectively facilitate short-term retention.

Similarity A factor that has received surprisingly little attention in the timbre literature is similarity-based interference. It has been known for a long time that the acoustic structure of an intervening item markedly affects the strength of interference (Deutsch, 1975): speech, for instance, does not interfere with memory for pitch, in contrast to tones with neighboring fundamental frequencies. Starr and Pitt (1997) used additively synthesized tones and obtained significant interference effects that depended on the similarity in brightness of the interfering tone to the standard. However, most other studies have not attempted to systematically control for this variable (beyond ensuring discriminability). This circumstance is curious given that similarity effects have played out in various different fields of STM research, ranging from auditory or visual (Visscher, Kaplan, Kahana, & Sekuler, 2007) up to the famous phonological similarity effect for linguistic stimuli (Baddeley, 2012; Nimmo & Roodenrys, 2005).

What do these studies more generally imply about the structure of STM for timbre and its maintenance processes? The following summarizes the main points.

Timbre and pitch Considering the relation of timbre and pitch, some authors emphasize commonalities (Starr & Pitt, 1997; McDermott et al., 2008; Cousineau et al.,

2013). Specifically, listeners were able to recognize brightness contours and even familiar pitch melodies "transposed" to the dimension of timbral brightness (McDermott et al., 2008). This suggests that pitch and brightness share aspects of relative perception, likely because both are partially based on a tonotopic neural representation. Similarly exploring differences and commonalities in the processing of auditory sequences differing in pitch, brightness or loudness, Cousineau et al. (2013) did not find qualitative differences between pitch and brightness recognition. Moreover, neuropsychological studies have shown that individuals with music processing deficits not only have difficulty with the recognition of pitch sequences, but the deficit appears to extend to timbre sequences (Tillmann et al., 2009; Marin et al., 2012). Other studies have suggested that timbre is retained according to different principles than pitch (Schulze & Tillmann, 2013), in particular because it cannot rely on subvocal rehearsal processes. At the same time, no systematic mutual interference of pitch and timbre has been observed for short-term memory so far (Semal & Demany, 1991; Starr & Pitt, 1997), although the previous Chapter 2 mentioned the mutual interference of pitch and timbre in perceptual tasks. As a whole, this implies the questionable conclusion that pitch and timbre are retained in structurally similar ways, yet dissociated from each other in memory (but not in perception). More research is required to refine our understanding of the mutual relation of memory for both attributes.

Maintenance The second broad issue concerns the extent to which STM for timbre is an automatic process (in the sense of classic sensory memory) or whether it depends on active maintenance of the engram. Although it was mentioned above that Caclin et al. (2006) demonstrated the existence of an attention-independent MMN for timbre, it is not clear whether this type of automatic change detection reflects sensory memory or rather neural adaption (May & Tiitinen, 2010), and what its exact perceptual correlates are. Nolden et al. (2013) found an ERP component that depended on attention and indexed STM capacity for timbre in similar ways as in studies of pitch memory (Alunni-Menichini et al., 2014). Notably, ERPs differed strongly between a memory condition and a passive listening condition, although stimulation patterns were identical. Golubock and Janata (2013) further demonstrated severe capacity limits of around 1.5 items across a range of retention intervals (1–6 s), which suggests that the underlying form of memory was even more severely limited in capacity than in the verbal

or visual domains, where limitations of around 4 items were reported (Cowan, 2001). Soemer and Saito (2015) argued for the importance of consciously controlled imagery in timbre maintenance, once more indicating the need for attentional resources allocated to sensory modalities. Opposed to articulatory suppression, tapping, and visual imagery, the secondary task of auditory imagery was the only condition that strongly impaired item recognition performance. Similarly, no interference by articulatory suppression was observed by Schulze and Tillmann (2013), using a serial-recognition-like task. Moreover, McKeown et al. (2011) suggested that timbre information can be retained over long retention times (30 s) while participants read aloud. In this case, the results were interpreted as a sign of a kind of "auditory 'sensory' memory that is neither transient nor verbally coded nor attentionally maintained."(p. 1202) It should be noted that the primary interest of the latter study, as well as of related works (McKeown & Wellsted, 2009; Mercer & McKeown, 2010, 2014), is to study the retention of the smallest sensory details. A surprising outcome of these experiments then is that such supposedly transient details are more robustly retained than traditional accounts of auditory sensory memory suggest. In that sense, what may seem like conflicting evidence may turn out to be a different research focus. Whereas these latter studies focus on memory for properties *of* auditory events, the previously reviewed works test memory *for* auditory events as a whole. From a bird's-eye view, however, there remains a divide in the emphasis that is placed on attention.

3.5 Summary and open questions

At the beginning of this chapter, we distinguished two facets of memory. These correspond to the time scales of information persistence required for certain tasks (e.g., short-term, long-term), and the types of information being represented and retained (e.g., sensory, phonetic, semantic, episodic). Traditional accounts of memory conflate both facets by suggesting that sensory memory only persists for a short time. On the contrary, we reviewed studies that demonstrate that auditory memory is more than of echoic nature, quickly decaying and inaccessible to attentional control, but that it retains fine-grained representations over long time spans. Examples included studies on the learning of white noise excerpts, auditory imagery for timbre, and on timbre in memory for melodies. Although there is no single comprehensive framework of audi-

tory memory that incorporates this diverse array of empirical findings, we hypothesize that attention is a key factor to address, in particular for STM.

In the introductory Chapter 1, five variables central to this thesis were presented. To recapitulate, these are: 1) the role of perceptual similarity of the timbre of tones, 2) familiarity and categorical affordances of tones, 3) the role of concurrent pitch variability in memory for timbre, 4) musical training or expertise of the listener, 5) the role of maintenance strategies. The previous section gave reasons for why a better understanding of these factors is important for the development of a more comprehensive account of STM for timbre.

There are various other aspects that have remained unexplored (and will remain so over the course of this thesis). As Table 3.1 reveals, it is clear that no study has systematically manipulated the variables of the duration of individual tones, as well as the ISI. Tone durations are all in the range of 200–600 ms. Tasks that present sequences of tones have used ISIs of 0 or 40 ms, with one exception that used a larger interval of 500 ms (Soemer & Saito, 2015). Both parameters seem to be closely linked to the question of whether sequences are perceived in terms of a succession of items or as a single auditory *Gestalt*. In fact, there is early research on auditory sequence perception that suggests that the duration of items determines whether sequences are compared by a type of global pattern matching or by identification of individual items (Warren, 1974). For that reason, a more detailed examination of these parameters may reveal much about the formation of "auditory objects" in STM.

There is similarly open terrain regarding the properties of LTM for timbre. For instance, no work has used behavioral LTM tasks in order to estimate the magnitude of timbral detail that is preserved over long retention intervals. Intuition suggests that this type of memory can be fine grained in certain cases, although it could be of a rather implicit nature. It would also be interesting to extend the breadth of studies on the role of timbre in memory for melodies and scrutinize the role of different low-level features such as loudness normalizations, as well as similarity relations between timbres of exposure and test melodies. As mentioned, memory for timbral sequencing rules has not been studied comprehensively either, although it may be beneficial to start with an existing musical style that could be modeled by the experiment (the rich world of drum and percussion music may be a good starting place). Finally, it should be obvious by

now that this review did not even try to touch on the vast field of the formation of auditory categories, enabling the identification of sound sources (see, McAdams, 1993, for a review of models).

This review attempted to show that musical timbre, this multidimensional attribute that is so hard to properly translate into symbolic form, can be ingrained in human memory in manifold ways—which may be a key to why it affords a central role in music listening. We have portrayed the memory processes under study as multifaceted and highly interactive, operating from short to long time scales upon auditory sensory formats but being closely linked to other types of codes. The experiments described in the following focus on the many faces of short-term memory for timbre.

Part II

Timbre sequences

Chapter 4

Short-term recognition of timbre sequences

A majority of previous studies on short-term recognition memory for musical timbre have focussed on factors that are generic to memory tasks, such as the length of the presented sequence or the retention time. Variables that are specific to the auditory attribute of timbre have remained unexplored. As the first study of this Part II on memory for timbre sequences, this chapter investigates the role of concurrent pitch variability in timbre sequence recognition and the impact of musical training, as well as different measures of timbre sequence similarity. This chapter thus attempts to contribute to a more comprehensive picture of the relevant factors in short-term memory for timbre.

This chapter is based on the following research article:

> Siedenburg, K. and McAdams, S. (in preparation). Short-term recognition of timbre sequences: Effects of musical training, pitch variability, and timbral similarity. Manuscript prepared for submission to *Music Perception*.

Abstract. Timbre is a basic parameter of audition, but there is a paucity of empirical data on how it is cognitively processed. The present study investigated short-term recognition memory for musical timbre sequences using a serial matching task. Experiment 1 revealed significant effects of sequence length and dissimilarity of items on d' scores, as well as an interaction of musical training and pitch variability: musicians performed better with variable-pitch sequences, but did not differ from nonmusicians with constant-pitch sequences. Exp. 2 yielded a significant effect of pitch variability for musicians when pitch patterns also varied between standard and comparison sequences. Exps. 3 and 4 highlighted the impact of the perceptual dissimilarity of items that were swapped in the sequence on response choice behavior (accounting for around 90% of the variance in response choices across the four experiments), but did not find any effects of timbral heterogeneity in the sequence. The present results extend findings regarding the impact of musical training and pitch variability from the literature on timbre perception to the domain of short-term memory and demonstrate the importance of controlling timbre stimuli for their perceptual similarity relations.

4.1 Introduction

Most research in music cognition has traditionally focussed on the "royal couple" of music theory, that is, pitch and duration. At the same time, the musical relevance of timbre has evolved enormously during the 20th century. There are a variety of musical styles for which sequences of timbres act as the primary conveyors of musical information. Apart from abundant examples in popular or non-western music (Nattiez, 2007), an example from 20th century art music is the so-called *Klangfarbenmelodie* ("timbre melody"), featuring timbral configurations that are sculpted over time (Erickson, 1975). Already in 1911, the composer Schoenberg (1911/1978) had famously conjectured, "Tone-color melodies (Klangfarbenmelodien)! How acute the senses that would be able to perceive them! How high the development of spirit that could find pleasure in such subtle things! In such a domain, who dares ask for theory!"(Schoenberg, 1911/1978, p. 422) Surprisingly coherent with this early note is the fact that despite a long history of research on timbre—at least dating back to von Helmholtz (1863)—the understanding of the cognitive processing of timbral structures in musical contexts is still in its infancy, because empirical data have only begun to emerge in recent years. Here we seek to contribute to this literature by analyzing a process that is foundational for timbre's function in musical contexts: the capacity to recognize timbre sequences from short-term memory (STM). We understand the latter as the cognitive faculty re-

sponsible for the retention of sensory and categorical information over spans of roughly 1–30 seconds (D'Esposito & Postle, 2015). In fact, short-term sequence recognition appears to be essential for the parsing and integration of streams of musical events into phrase structures, and eventually for the experience of musical form (McAdams, 1989).

Timbre specifically denotes the bundle of perceptual attributes that lends tones a sense of "color" or "shape" and identity. It encompasses continuous perceptual qualities of sounds such as brightness, sharpness of attack, spectrotemporal irregularity, roughness and noisiness in addition to auditory features specific to certain instruments. The perceptual structure of timbre has been modeled by multidimensional scaling of pairwise dissimilarity judgments, yielding spatial configurations of timbres (McAdams, 2013). McAdams et al. (1995) found spectral, temporal, and, to a lesser extent, spectrotemporal properties of tones to be the major acoustic correlates of the resulting timbre space.

Outside the lab, timbral contrast does not occur in isolation, but mostly covaries with other parameters, such as pitch. A central question of the current study thus is whether timbre pattern recognition is robust to concurrent variability in pitch. Given that timbre recognition may seem to be a "specialist domain" (at least in Schoenberg's eyes one century ago), we were interested in whether performance would differ across groups of trained musicians and nonmusicians who do not have any experience in playing or analyzing music. Of particular concern was furthermore to take into account timbre's "perceptual topology", that is, to study for the first time how the dissimilarity relations that govern timbre perception affect timbre sequence recognition.

4.1.1 Timbre recognition in the literature

A few studies have started to explore the cognitive underpinnings of short-term memory for timbre, for which serial recognition is a commonly used experimental task: Two sequences that comprise the same items are presented subsequently, and participants are required to tell whether standard and comparison sequence were of the same serial order. Testing timbre processing in amusic participants and normal controls, Marin et al. (2012) obtained generally higher serial recognition scores for shorter sequences (4–8 items), as well as better performance of the control group. Nolden et al. (2013) recorded

EEG during a serial recognition task with electronically synthesized timbres that differed in spectral distribution. In a control condition, participants were asked to ignore the standard sequence and to merely judge a property of the last tone of a comparison sequence. Significant differences in event-related potentials (ERP) between control and memory conditions were found during the retention interval; the higher the memory load, the stronger was the ERP negativity. Only the task differed between conditions, not the sensory stimulation pattern. These findings are coherent with Alunni-Menichini et al. (2014), who demonstrated that the same ERP component robustly indexes STM capacity. This result indicates that the retention of abstract electronic timbres requires a generic, attention-dependent form of STM. Schulze and Tillmann (2013) compared serial recognition for the materials of timbres, words, and pitches with sequences of five and six items. Timbre did not yield an effect of length in forward recognition, whereas words and pitches did. The authors argued that the missing effect of length in forward recognition indicates a domain-specific sensory storage of timbre, contrary to words and pitches, which may engage motor-based rehearsal mechanisms. In order to revisit this aspect, our first two experiments compare sequences that comprise 4, 5, and 6 sounds.

4.1.2 Retention of pitch and timbre and musical training

Most studies on the interaction of pitch and timbre processing are based on pairwise discrimination with only short retention times below 1 s. Providing a groundwork for many later studies on interactions of auditory dimensions, Melara and Marks (1990) used speeded classification of stimuli varying in pitch and timbre with either independent or correlated changes along the two dimensions. Participants were asked to discriminate stimuli only along one dimension. Reaction times were slower when changes in attended and unattended attributes were independent, but faster when both dimensions were correlated. This was interpreted as evidence for integral processing of the two auditory attributes, conceptualized as a cross-talk between "higher-level channels" responsible for the computation of the perceptual attributes pitch and timbre. These findings were replicated for nonmusicians and musicians (Krumhansl & Iverson, 1992; Pitt, 1994), and recently by Caruso and Balaban (2014), showing that the greater a concurrent change in pitch, the harder it was to correctly discriminate timbre. This

result was replicated by Allen and Oxenham (2014) who measured difference limens for musicians and nonmusicians using stimuli with concurrent random variations along the nonattended dimension. Ensuring that the experimental units of timbre and pitch were of the same perceptual magnitude, they found symmetric mutual interference of pitch and timbre in the discrimination task. Musicians yielded higher discrimination overall, but there was no interaction of musicianship and auditory parameter (pitch/timbre).

The sole study that did not find interactions between the two parameters used a classical STM task, an interpolated tone paradigm adapted to timbre with a retention time of 5 s (Starr & Pitt, 1997). Their first experiment provided an effect of timbre similarity without differences between musicians and nonmusicians. They further tested a mixed group of participants for interactions of timbre and pitch in STM, which turned out to be negligible. Nonetheless, it may be hasty, based on a single null result, to infer that the reliable perceptual interaction of timbre and pitch (in discrimination tasks, as reviewed above) is consolidated in STM. Note that Starr and Pitt (1997) did not include a factor of musicianship in their second experiment, which tested the interaction of pitch and timbre. Musicians are well known to have superior memory for pitch (see, e.g., Schulze, Zysset, Mueller, Friederici, & Koelsch, 2011), and therefore musicians might be less likely to confuse variation in pitch and timbre in memory tasks with longer retention times. Given that results concerning the impact of pitch variability on timbre in STM by Starr and Pitt (1997) are incongruent with the literature on perceptual processing (Caruso & Balaban, 2014; Allen & Oxenham, 2014; Melara & Marks, 1990), further research is required to investigate how generic STM for timbre is affected by pitch variability, and whether musical training plays a role in this context.

The latter issue relates to an open question in timbre research, namely whether musical training affects timbre processing. So far, no systematic differences between musicians and nonmusicians have been found in experiments on the perception of timbral dissimilarity (McAdams et al., 1995; Kendall, Carterette, & Hajda, 1999; Lakatos, 2000; Alluri & Toiviainen, 2012). On the other hand, Chartrand and Belin (2006) reported that musicians possess superior discrimination abilities for vocal and instrumental timbres. Furthermore, there is growing neurophysiological evidence that suggests more sensitive timbre processing in musicians (Pantev et al., 2001; Shahin et al., 2008; Strait et al., 2012). Investigating memory for timbre may provide valuable complementary perspectives on this issue. Experiments 1 and 2 thus set out to test

the role of musical training and pitch variability in STM for timbre.

4.1.3 Similarity effects

None of the studies discussed so far has attempted to systematically control perceived similarity of timbre, a surprising circumstance given that similarity effects are pertinent in verbal and visual STM. The only authors who investigated a similarity-specific dimension were Starr and Pitt (1997). They used synthesized harmonic complexes with four partial tones of which the upper three were shifted in rank in order to yield different degrees of brightness. A classical interpolated tone paradigm was employed (cf., Deutsch, 1970), requiring participants to match a standard and a comparison stimulus, separated by a 5 s interval which included distractor tones of varying brightness. They observed that both musicians and nonmusicians performed with greater accuracy when the timbre of the distractor tones was dissimilar to the target timbre, an effect that was robust over distractors with varying pitch. In the current experiment, we attempt to closely track the impact of perceptual similarity relations on serial recognition of timbre.

Addressing the issue of STM capacity for timbre, Golubock and Janata (2013) synthesized electronic sounds whose discriminability was ensured for all three synthesis dimensions, spectral centroid, attack time, spectral flux (i.e., spectral variability over time). They found capacity estimates of one to two items, a surprisingly small number when compared to usual capacity estimates of around four items (Cowan, 2001). A second experiment used stimuli with increased perceptual variability and found significantly greater capacities, suggesting that perceptual variability enhances recognition. Note that variability within the stimulus set is a direct function of pairwise perceptual similarity between items (in particular between the probe and the list); these results thus potentially index an underlying similarity effect. Yet, some authors claim that rather than overall perceptual similarity, overlap of perceptual features and interference determine forgetting for timbre (Mercer & McKeown, 2010). A more thorough assessment of the relation of feature overlap and similarity perception for timbre, especially for timbres that are more complex than the ones used by Mercer and McKeown, would need to be performed to allow for a debate about corresponding memory mechanisms, although it only seems reasonable to assume that feature overlap and similarity

are closely related (Tversky, 1977). Finally, although not strictly working with timbre stimuli, Visscher et al. (2007) demonstrated how similarity effects analogously unfold in auditory and visual STM for moving ripple stimuli (frequency-modulated sinusoid-complexes in audition, Gabor patches in vision). Effects were intriguingly similar across domains when discriminability of stimulus dimensions was controlled on a by-participant basis. They showed effects of probe-to-list similarity, as well as list homogeneity. In conclusion, although the reviewed studies generally suggest that perceived timbral similarity might play a role in serial recognition of timbre, it is as yet unclear how this effect would manifest itself in a sequential context, and what kind of variables could predict response behavior as a function of timbre dissimilarity. To address these questions was the aim of Exps. 1, 3, and 4.

4.1.4 The present study

The first central goal of this study was to investigate the robustness of timbre-sequence recognition to interference by concurrent variability in pitch. In sum, many studies that have used pairwise discrimination found interactive processing and interference. A study using a task that more strongly tapped into STM found non-congruent results (Starr & Pitt, 1997). We were further interested in whether musical training would affect timbre recognition. Secondly, we attempted to account for the variable of timbre similarity. Starr and Pitt (1997) found similarity-based interference in an interpolated tone task. Yet, it is unclear whether similarity plays a role in serial recognition of timbres that are easily discriminated when only the order of items may change. Because Williamson, Baddeley, and Hitch (2010) observed that pitch similarity effects only arose with regards to specific sequence lengths, and because a factor of sequence length would allow a connection with other recent timbre memory studies that used serial recognition (Marin et al., 2012; Nolden et al., 2013; Schulze & Tillmann, 2013), we included sequences of varying length in the first two experiments.

Specifically, Exp. 1 included one between-subjects factor (level of musical training) and three within-subject factors (list length, pitch variability, similarity). Exp. 2 followed up on the aspect of musical training and pitch variability, and Exp. 3 examined similarity and serial position. Finally, Exp. 4 tested the impact of timbral heterogeneity.

4.2 Experiment 1: Group, length, pitch variability, and timbre dissimilarity

The research reported in this manuscript was carried out according to the principles expressed in the Declaration of Helsinki, and the Research Ethics Board II of McGill University has reviewed and certified this study for ethical compliance (certificate # 67-0905).

4.2.1 Method

Participants

Sixty listeners participated in the experiment. These consisted of 30 musicians, recruited from a mailing list of the Schulich School of Music at McGill University, and 30 nonmusicians recruited via web-based, classified advertisements. The musicians had an average age of 23 years (SD = 4.4, range: 19–33), included 19 male participants, and featured an average of 14 years of instrumental training (SD = 5.2) and 4 years (SD = 2.8) of formal music-theoretical instruction or ear-training. Nonmusicians were on average 25 years old (SD = 7.2, range: 19–50), included 25 female participants, had an average of 0.3 years (SD = 0.66) of instrumental instruction and an average of 1.1 (SD = 1.74) years of musical instruction in elementary school and no further formal musical training from there on. All participants reported normal hearing, which was confirmed by conducting a standard pure-tone audiogram measured right before the main experiment (ISO 398-8, 2004; F. N. Martin, Champlin, et al., 2000). They were required to have hearing thresholds of 20 dB HL or better, assessed at octave spaced frequencies from 125 to 8000 Hz.

Stimuli

The same timbre stimuli were used as in McAdams et al. (1995) based on FM-synthesized sounds (Yamaha DX7, Yamaha Corp., Shizuoka, Japan) created by Wessel, Bristow and Settel (1987), to some extent emulating instruments from the classical orchestra. All timbres were synthesized at pitch E♭4 (fundamental frequency of 311 Hz) and had been perceptually normalized with regards to loudness and duration in the original

study. We used this particular set of timbres because it not only had been perceptually normalized in pitch and loudness, but also allowed us to make use of its extant dissimilarity data. These had been collected through pairwise dissimilarity judgments of timbres on a 1-9 rating scale, collected from 98 participants. We were thus able to construct timbre sequences with varying degrees of inter-item similarity. From the 18 timbres, we selected a subset of eight, containing four instruments with impulsive excitation (plucked, struck) and four with continuous excitation (blown, bowed). The selected instruments were electronic emulations of the bassoon, clarinet, guitar, harpsichord, horn, piano, trumpet and vibraphone.

The sounds contained subtle hiss noise, which was removed by using a state-of-the-art audio noise removal algorithm (Siedenburg & Dörfler, 2013) implemented in MATLAB version R2013a (The MathWorks, Inc., Natick, MA). In order to construct sequences with variable pitch height, and given that the original synthesizer was no longer available to us, we created transposed versions of plus/minus one whole tone using the audio-editing software AudioSculpt (IRCAM, Paris, France). Note that timbre dissimilarity relations and timbral identity can be assumed to remain stable for pitch transpositions that are below one octave (Marozeau et al., 2003; Handel & Erickson, 2001; Steele & Williams, 2006). No audible artifacts were introduced by noise removal or transposition. Sounds were then cropped to a duration of 520 ms using a linear fade out from 480-500 ms followed by 20 ms of silence.

Sequences contained 4, 5 or 6 items with 520 ms inter-onset interval (IOI). They had constant or varying pitch. They also had low, medium or high mean transition dissimilarity (MTD) according to the raw dissimilarity data on a scale of 1 to 9 collected in McAdams et al. (1995). MTD measures the mean dissimilarity of the transitions between successive timbres in the sequence. There were 180 sequences in total with 10 sequences per condition. Each sequence was obtained by randomly drawing items without replacement. The selected sets had mean MTDs of 3.3 (low; $SD = 0.19$), 5.1 (medium; $SD = 0.04$) and 7.3 (high; $SD = 0.20$).

Comparison sequences followed the standard after a silent interval of 3 s and were generated either by using the identical sequence or by swapping the last and third-to-last items. In each condition, 50% of the comparison sequences were identical, and 50% were different from the standard sequence.

For any of the sequences of length 4, 5, and 6, melodic templates were generated

and used for all levels of MTD. Melodies were restricted to the tones D♭4, E♭4, and F4 and were created by interleaving two random permutations of that set and truncating according to length. For instance, given the permutations (D♭4, E♭4, F4) and (F4, E♭4, D♭4) and a five item sequence, this would yield the pitches (D♭4, F4, E♭4, E♭4, F4). This ensured that any third-to-last and last tone would have different pitches, i.e. that the same two pitches would not occur in the positions at which the swapped timbres were located. For any given trial, the same pitch pattern was used for standard and comparison sequences.

Apparatus

Experiments took place in a double-walled sound-isolation chamber (IAC Acoustics, Bronx, NY). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronics GmBH, Wedemark, Germany), using a Macintosh computer with digital-to-analog conversion on a Grace Design m904 (Grace Design, Boulder, CO) monitor system. The output level was 67 dB SPL on average (range: 58–75 dB) as measured with a Brüel & Kjær Type 2205 sound-level meter (A-weighting) with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). The experimental interface and data collection were realized with the audio software Max/MSP (Cycling 74, San Francisco, CA).

Procedure

There was one between-subjects factor of musical training and three within-subject factors: pitch variability, sequence length, and mean transition dissimilarity (MTD). These were split into two blocks containing the constant-pitch vs. variable-pitch conditions. The order of the presentation of the blocks was counterbalanced across participants. The order of sequences within blocks was fully randomized.

After having completed the audiogram, participants read through the experimental instructions and completed a set of six training trials, not part of the main experiment. Participants were instructed that if there was a pitch change during the sequence, only the order of the sounds (timbres) might or might not change, but not the order of the pitches and that they could ignore pitch. The first three training trials were from the constant-pitch condition, in order to ensure that participants understood that

they should focus on timbre. The latter three training trials were from the variable-pitch condition. Feedback on response correctness was provided for training trials, and potential questions could be clarified with the experimenter. Participants could listen to the sequences as often as they wished.

In the main experiment, participants listened to the standard sequence and after 3 s of silence to the comparison sequence. They then gave their response by clicking on the appropriate button (*same/different*). They subsequently provided an assessment of their level of confidence, although these ratings will not be taken into account in the following analyses. In contrast to the training, no feedback was provided. Participants could then proceed to the next trial. After finishing the first experimental block, which took around 20 min, they were asked to take a break for 5 min. Having finished both blocks, participants filled out a questionnaire concerning their musical background. Overall, the experiment lasted around 50 min for which participants received monetary compensation.

4.2.2 Results

Discrimination sensitivity and response bias were assessed by calculating d' scores and criterion location c as based upon the yes-no model (Macmillan & Creelman, 2005, Chapter 2). Hits were defined as trials that participants correctly identified as different, false alarms as trials that were incorrectly judged as different. Hit and false alarm rates for cells with the maximum number of hits or false alarms were set to .99 or .01, respectively. Here and in all following analyses, no violations of sphericity were observed (Mauchly's test). We report original p-values in post-hoc comparisons, but compare them against the Bonferroni-corrected α-level.

Sensitivity

d' scores were significantly above chance in all conditions, as validated by one-sample t-tests against d'=0 (all $p < 0.001$). Grand averages for musicians and nonmusicians were $M = 1.67$ and $M = 1.30$, respectively. Figure 4.1 depicts mean sensitivities over all three within-subject factors including the interaction with the between-subjects factor of musical training.

Employing a mixed Group (2) × Pitch Variability (2) × Length (3) × MTD (3)

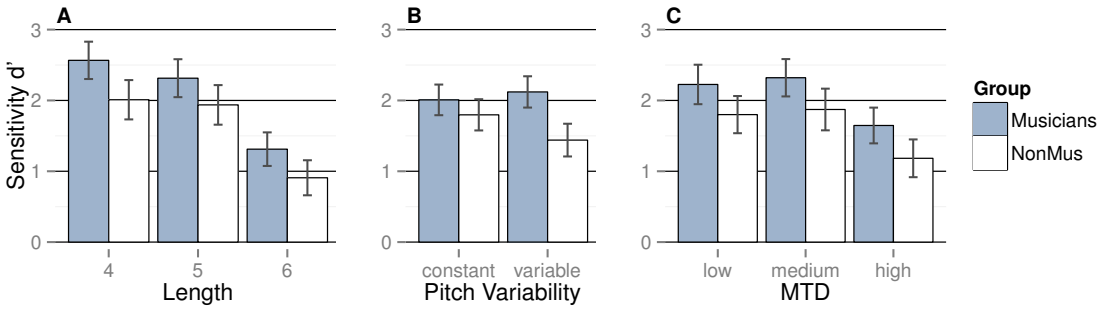

Fig. 4.1 Exp. 1: d' scores for the main within-subject factors of sequence length (A), pitch variability (B) and mean transition dissimilarity (C) for groups of musicians and nonmusicians. Error bars indicate 95% confidence intervals.

ANOVA yielded a weak effect of experimental group, $F(1, 58) = 4.38$, $p = .041, \eta_p^2 = .070$. There was no main effect of pitch variability, $F(1, 58) = 1.54, p = .22$, but effects of sequence length, $F(2, 116) = 56.88$, $p < .001$, $\eta_p^2 = 0.49$, and MTD, $F(2, 116) = 21.90$, $p < .001$, $\eta_p^2 = .27$, were both significant. There was a significant interaction between experimental group and pitch variability, $F(1, 58) = 4.51$, $p = .037, \eta_p^2 = .072$, as well as between sequence length and MTD, $F(4, 232) = 2.65$, $p = .034, \eta_p^2 = .044$.

Post-hoc comparisons revealed that performance was not significantly different for sequences of length 4 and 5, $t(59) = 1.40, p = 0.50$, but scores were higher for lengths 4 and 5 compared to 6, $t(59) > 7.73, p < .001$. However, a monotonic decay of d' scores over sequence length was confirmed by a linear contrast on sequence length (on the full pool of both groups of participants, corresponding to the main effect of length), $\beta = -0.83$, $t(116) = -9.4$, $p < .001$. There was no difference between low and medium MTD, $t(59) < 1$, but both yielded greater sensitivity than high MTD, $t(59) > 4.86, p < .001$.

The interaction of experimental group and pitch variability was due to significantly lower accuracy of nonmusicians compared to musicians in the variable-pitch condition, two-sample $t(58) = 2.73, p = .0084$, but no differences between groups in the constant-pitch condition, two-sample $t(58) = 0.83, p = .41$. Neither musicians, nor nonmusicians differed significantly across pitch conditions, paired $t(29) < 2.34, p > \alpha_{crit} = .0125$).

The interaction between sequence length and MTD was due to a combination of significant differences of medium ($M = 2.8$) and high MTD ($M = 1.8$) at length 4,

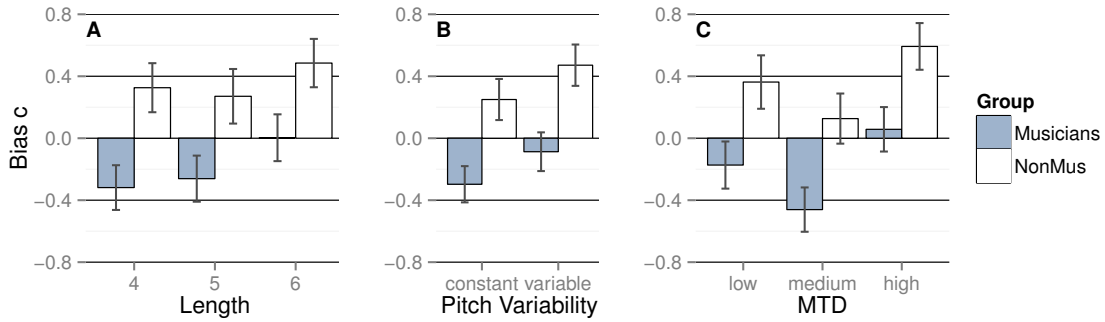

Fig. 4.2 Exp. 1: Estimated bias (criterion location c) for the three within-subject factors of sequence length (A), pitch variability (B) and mean transition dissimilarity (C), for groups of musicians and nonmusicians. Error bars indicate 95% confidence intervals.

paired $t(59) = 4.8, p < .001$, and differences for low ($M = 2.5$) and high MTD ($M = 1.7$) at length 5, paired $t(59) = 4.0, p < .001$, but no significant differences at length 6, all $p > \alpha_{crit} = .0056$).

Response bias

Response bias as estimated by the decision criterion location was lower for musicians ($M = -0.14$), indicating more "different" responses, compared to nonmusicians ($M = 0.29$). This difference is visible in Figure 4.2 depicting bias for both groups over all three within-subject factors. A mixed Group (2) × Pitch Variability (2) × Length (3) × MTD (3) ANOVA confirmed an effect of group, $F(1, 58) = 12.03, p < .001, \eta_p^2 = .17$, as well as significant main effects of pitch, $F(1, 58) = 7.78, p = .007, \eta_p^2 = .12$, length, $F(2, 116) = 7.82, p < .001, \eta_p^2 = .12$, and MTD, $F(2, 116) = 26.41, p < .001, \eta_p^2 = 0.32$. There was a significant interaction between length and MTD, $F(4, 232) = 5.90, p < .001, \eta_p^2 = .092$.

Post-hoc comparisons showed that there were no significant differences in bias for lengths 4 and 5, paired $t(59) < 1$, but biases for lengths 4 and 6, as well as for lengths 5 and 6, were significantly different, paired $t(59) > 3.74, p < .001$. Biases were significantly different between all three pairs of MTD conditions, paired $t(59) > 2.98$, all $p < .004$.

The interaction of length and MTD arose through significant deviations of the medium MTD from the low and high MTD conditions in lengths 5 and 6, paired

$t(59) > 3.8, p < .001$, but otherwise no differences between MTD conditions at any given length, paired $t(59) < 2.6, p > \alpha_{crit} = .0056$).

4.2.3 Discussion

This first experiment revealed that multiple factors play into the serial recognition of timbre. Sensitivity decreased monotonically with sequence length, in line with results from Marin et al. (2012) and Nolden et al. (2013), but different from those of Schulze and Tillmann (2013). More centrally, our results address the role of musical training with regards to concurrent pitch variability in timbre sequence recognition. Because d' scores did not differ significantly between musicians and nonmusicians in the constant-pitch condition, the main effect of musical training can be considered to be driven by the interaction with pitch. In the variable-pitch condition, which required participants to disentangle pitch and timbre in STM, nonmusicians' performance was significantly lower due to an inflated false-negative rate, contrary to musicians who performed similarly in both conditions. This result suggests that there is no difference between groups as far as timbre is concerned in isolation. As soon as concurrent pitch variability enters the picture, however, nonmusicians can no longer disentangle pitch from timbre as well as musicians. This interpretation is in line with parts of the music cognition literature, which has consistently demonstrated effects of formal musical training on the processing of pitch structures, e.g. (Krumhansl, 1990; Patel, 2008), but not for timbre perception (McAdams et al., 1995; Kendall et al., 1999; Lakatos, 2000; Alluri & Toiviainen, 2012). Given that musicians' performance was not inferior in the variable-pitch condition, a corollary question concerns whether musicians' memory for timbre is unaffected by simultaneous variability in pitch, no matter what degree of complexity that variability takes. This question was considered in Experiment 2.

The results regarding timbral mean transition dissimilarity (MTD) seem surprising at first glance. High MTD had significantly lower d' scores than low and medium MTDs. A generic similarity effect, on the contrary, would have yielded the reverse order, featuring high sensitivity for sequences with dissimilar items (cf., Baddeley, 2012; Nimmo & Roodenrys, 2005). MTD yields a global and temporally directed measure of perceptual similarity. Considering a rather local measure instead, namely the timbral dissimilarity between the two items that swapped order (TDS) yields a

different picture. For the set of sequences belonging to low, medium and high MTD, the average TDS values turn out to be 4.7, 5.8, and 3.9, respectively, on a scale of 1 to 9. The rank order of TDS thus is identical to that of the d' scores. The surprisingly low TDS value for the sequences of high MTD can be understood by the structure of the set of timbres: four continuously excited and four impulsively excited timbres were used, and maximizing MTD (for the "high" level) meant leaping between these two clusters. Because timbres within the two clusters were relatively close, and only last and third-to-last timbres were swapped, this meant that the timbres that switched order were on average quite similar in the "high" condition. The perceptual impact of this factor might also depend on the serial positions of swaps that are involved. Experiment 3 studied in greater detail the question of whether TDS is a better-suited predictor for similarity effects in serial recognition of timbre and how the swap used here, involving the last and third-to-last items, compares with other serial positions.

The observed interaction between MTD and length remains somewhat less straightforward to interpret. For length 6, there was no similarity effect, whereas for shorter sequences, there was. Similar results were found by Williamson et al. (2010), using a pitch-sequence reconstruction task. Here the effects of tonal similarity only arose for rather short sequence lengths.

Significant differences in bias estimates between musicians and nonmusicians further characterize the two groups' rating behavior: musicians tend to respond more often with "different", yielding too many false positives, whereas nonmusicians tend more often to respond with false negatives. In what follows, reports of response bias will be omitted for the sake of brevity: the data from Exps. 2–4 exclusively stemmed from musician participants, and the within-subject comparison of response bias is not as informative.

4.3 Experiment 2: Length and pitch variability

In order to further explore the effect of pitch variability on memory for timbre sequences in musicians, we increased the degree of variability in the pitch domain by presenting different pitch sequences for standard and comparison sequences and comparing this situation to a constant-pitch baseline. As in Exp. 1, sequences of length 4, 5, and 6 were presented.

4.3.1 Method

Participants

Twenty-two musicians were recruited over mailing lists of the Schulich School of Music at McGill University. None of them had participated in the previous experiment. The group had a mean age of 26 years (SD = 7.6, range: 19–51), included 8 females, and featured an average of 18 years (SD = 7.8) of instrumental training and 5 years (SD = 3.4) of formal music-theoretical instruction or ear-training. As before, it was confirmed that all participants had normal hearing.

Stimuli and Apparatus

We selected the 30 timbre sequences from the medium MTD condition of Experiment 1, featuring 10 sequences, each with lengths of 4, 5, and 6 items. Each of these sequences was presented once with identical standard and comparison, and once with different standard and comparison. As before, the last and third-to-last items were swapped for the nonidentical condition. These 60 sequences were played at a pitch of E♭4 (fundamental frequency of 311 Hz) for the constant-pitch condition. The sequences of timbres were constructed with pitch progressions in the following way for the variable-pitch condition: We used the same transpositions of pitches as in the first experiment, namely D♭, E♭, F. Again, two random permutations of these three pitches were interleaved. Contrary to Experiment 1, there were now different successions of pitches for the standard and comparison sequences. We did not allow pairs of standard and comparison sequences $P_1^S, \ldots, P_L^S$ and $P_1^C, \ldots, P_L^C$, to have pitch progressions that paralleled the potential swap of last and third-to-last timbres, i.e. we discarded pairs for which both $P_{L-2}^S = P_L^C$ and $P_L^S = P_{L-2}^C$ for any length $L = 4, 5, 6$. Finally, in order to enhance the contrast between standard and comparison, we selected pairs of pitch sequences that had a fairly high edit distance (or "Levenshtein Distance", LD), which measures the minimum number of single-item edits (insertion, deletion, substitution) needed to transform one sequence into another. To transform "123" into "321", for instance, one requires at least two replacements, yielding an LD of two. We selected standard-comparison pairs of six items, whose LD equaled five (n being the maximum LD for sequences of length n), before truncating pitch templates to the appropriate length of four to six items. This means pitch templates of different lengths featured

different LDs overall, but this left the "local" pitch variability fairly constant across length conditions. The apparatus was identical to the one used in Experiment 1.

Procedure

This experiment featured the within-subject factors of length (4, 5, 6) and pitch variability (constant, variable). Each condition contained 20 sequences with 50% identical and 50% nonidentical trials, yielding 120 trials overall. The two levels of the pitch factor were presented in two blocks and their order was counterbalanced across participants.

4.3.2 Results

We analyzed d' scores in a within-subject Pitch Variability (2) × Length (3) ANOVA. Results are presented in Figure 4.3. Mean d' scores for constant-pitch ($M = 2.09$) were greater than for the variable-pitch condition ($M = 1.48$), as confirmed by a significant main effect of pitch, $F(1, 21) = 12.89, p = .002, \eta_p^2 = .38$. There was also a main effect of length, $F(2, 42) = 13.66, p < .001, \eta_p^2 = .39$. There was no interaction between pitch and length. Post-hoc comparisons revealed that the effect of length was due to significant differences of level 6 from levels 4 and 5, paired $t(21) > 3.2, p < .0039$, whereas the difference between lengths 4 and 5 did not reach significance, $t(21) = 1.82, p = .082$.

In order to directly compare the performance of musicians from Exps. 1 and 2, we selected musicians' trials from Exp. 1 for the condition of medium MTD (as presented in Exp. 2), and computed a mixed ANOVA with the factors Length (3) × Pitch Variability (2) × Experiment (2). It revealed significant effects of length (as discussed and analyzed above), and pitch, $F(1, 50) = 4.4$, $p = .041$, $\eta_p^2 = .08$. As expected (due to the different type of pitch variability in the two experiments), the latter effect was driven by an interaction of pitch and experiment, $F(1, 50) = 8.0$, $p < .007$, $\eta_p^2 = .14$. Whereas d' scores in the constant pitch condition of Exp. 2 ($M = 2.5$) did not differ from those in Exp. 1 ($M = 2.3$), two-sample $t(50) = -.82$, $p = .42$, performance was significantly worse in the variable pitch condition of Exp. 2 ($M = 1.3$) compared to Exp. 1 ($M = 2.4$), two-sample $t(50) = 2.7$, $p = .010$.

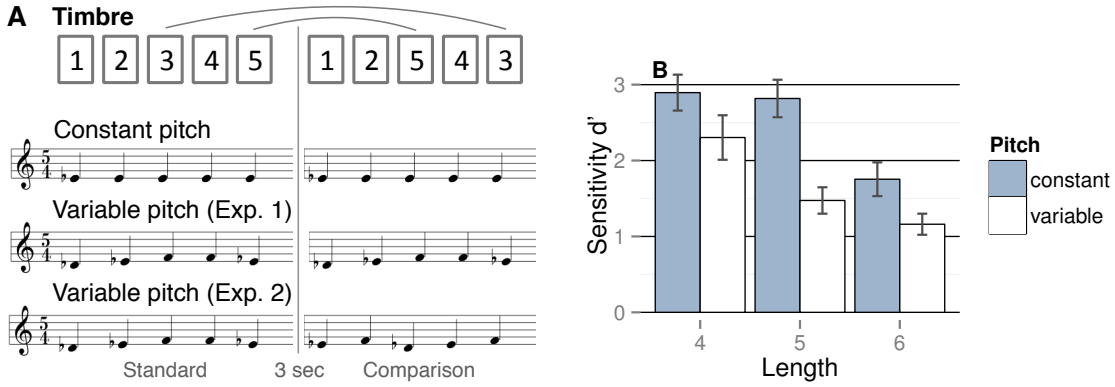

Fig. 4.3 (A) Schematic of pitch variability in Exps. 1 and 2 for an exemplary 5-item sequence. Although pitch sequences are identical for standard and comparison in Exp. 1, they differ in Exp. 2. (B) d' scores for the factors sequence length and pitch variability. Error bars indicate 95% confidence intervals.

4.3.3 Discussion

This experiment replicated the main effect of length from Exp. 1. Sequences of length 6 were significantly harder to retain compared to those of lengths 4 and 5. More importantly, we now obtained a significant main effect of pitch variability for musicians by using an altered variable-pitch condition. With different pitch patterns for standard and comparison sequences, timbral order was harder to match compared to the constant-pitch baseline. Moreover, the post-hoc analysis on the differences across Exps. 1 and 2 showed that the two independent groups of musicians in the two experiments did not differ in the constant pitch condition, whereas the variable pitch condition of Exp. 2 was significantly worse than that of Exp. 1.

These findings imply that even highly trained musicians (music students/professionals with a mean age of 26 years and a mean of 18 years of instrumental instruction) are not immune to cross-channel interference by pitch in STM for timbre, if pitch templates vary across the sequences to be matched (Exp. 2). In the simpler scenario in which the pitch pattern repeats across standard and comparison sequence (Exp. 1), there was no interference, however. For that reason, pitch may be assumed to interfere with timbre pattern recognition if the degree of pitch variability is sufficiently high.

These results may serve as a complement to the state of affairs in perceptual tasks. In Allen and Oxenham (2014), musicians had lower difference limens for pitch than nonmusicians. If variation in the nonattended condition was adjusted as a multiple of the individual threshold, interference from pitch to timbre did not differ across groups. This means that musicians need higher degrees of variability in pitch to exhibit interference with timbre in basic discrimination. Given that nonmusicians showed interference for the "easier" variable pitch condition of Exp. 1, the current data indicate that this circumstance could naturally extend to STM. Musical training allows participants to better disentangle pitch and timbre in STM, but musical training does not fully "immunize" against drastic concurrent variability in pitch.

4.4 Experiment 3: Similarity and position

The third experiment was conceived to specify more precisely the nature of the timbral similarity effect observed in Exp. 1. Therefore, the length of sequences was held fixed at 4 items, and we tested the influence of the perceived similarity of timbres that were swapped. Figure 4.4 illustrates the different similarity measures tested throughout Exps. 1, 3, and 4. We further included a factor of the serial position of swap in order to check whether similarity effects could be specific to certain serial positions.

4.4.1 Method

Participants

Twenty-two musicians (9 female) with an average age of 24 years ($SD = 4.3$, range: 19–36) participated. They had received an average of 15 years ($SD = 4.9$) of instrumental training and 4 years ($SD = 2.9$) of formal music-theoretical instruction including ear training. None of the listeners had participated in either of the two previous experiments. As above it was confirmed that all participants had normal hearing.

Stimuli and apparatus

This experiment used the same timbre stimuli as before, held at constant pitch, and concatenated to sequences of length 4. Four different swap positions were employed:

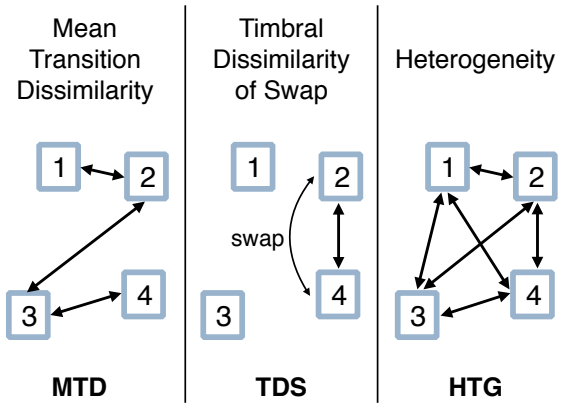

Fig. 4.4 Schematic of the tested types of dissimilarity measures. Numbered boxes represent the timbre sequence in a hypothetical dissimilarity space. Bold arrows indicate the respective pairwise relations taken into account. Exp. 1 tested MTD, Exp. 3 tested TDS, Exp. 4 tested TDS and HTG.

1&2, 2&3, 3&4, 2&4. In terms of timbral similarity, we included each of the 28 possible pairs to be swapped, given our set of eight timbres. The remaining two items per sequence were chosen randomly without replacement from the resulting set of six timbres. Each pair A-B was presented in both orders (e.g. C-A-B-D and C-B-A-D), yielding $2 \times 28 \times 4 = 224$ sequences in total, half of which were presented with identical and half with nonidentical comparison sequences. The apparatus was identical to that of the previous experiments.

Procedure

This experiment featured a within-subject design with factors of swap position (4 levels) and timbral dissimilarity of swapped items (TDS, 2 levels). The TDS factor partitioned the full range of TDS as obtained from the 28 pairings described above into a lower and upper half, such that the factor's first level comprised the 14 swaps of low TDS (×4 positions), and the second comprised the 14 pairs with high TDS values (×4 positions). The 224 trials were presented in fully randomized order, partitioned into two blocks of around 22 min duration each.

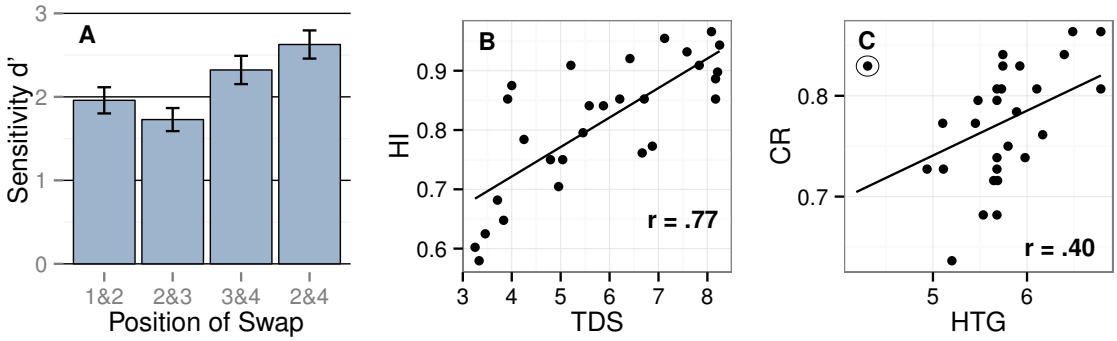

Fig. 4.5 Exp. 3: d' scores for the main effect of swap position (A). Error bars indicate 95% confidence intervals. (B) Hit-rate as a function of timbral dissimilarity of swap (TDS) with data averaged over positions of swap. (C) Correct-rejection-rate as a function of timbral heterogeneity (HTG) with data averaged over positions of swap. The circled point refers to the outlier described in the text.

4.4.2 Results

Figure 4.5 shows d' scores as a function of swap position (A), but also hit rate (B) and correct rejections (C) as a function of TDS and HTG, respectively. Concerning d' scores, a within-subject TDS (2) × Position (4) ANOVA revealed main effects of position, $F(3, 63) = 6.24, p < .001, \eta_p^2 = .23$, and of TDS, $F(1, 21) = 20.88, p < .001, \eta_p^2 = .50$. The interaction of both factors failed to reach significance, $F(3, 63) = 2.67, p = .055, \eta_p^2 = .11$.

Post-hoc comparisons showed that the effect of position was due to significant differences of the swap positions 2&3 and 2&4, $t(21) = -4.97, p < .001$, as well as marginal differences between swaps 2&3 and 3&4, $t(21) = -2.67, p = .011$ ($\alpha_{crit} = .0083$), as well as swaps 1&2 and 2&4, $t(21) = -2.8, p = .014$, whereas for all other pairs $p > .087$.

An exploratory correlation analysis was conducted to determine whether the relation between accuracy and timbral similarity could be modeled parametrically. We therefore zoomed in on the subcomponents of d' scores, i.e., hit-rate (HI) and correct-rejection rate (CR), in order to differentiate between the two different situations of identical and nonidentical sequences. For nonidentical trials, TDS and HI yielded a Pearson correlation of $r(26) = .77, p < .001$, for data averaged across participants

and across the four different swap positions. For all identical trials, TDS is zero by definition and therefore does not yield any useful predictions. However, the sequence heterogeneity (HTG), measuring average pairwise dissimilarities between all items of a sequence, was correlated with CR, $r(26) = .63, p < .001$, disregarding one outlier; the correlation without exclusion was $r(26) = .40, p = .033$. HTG did not correlate with HI, $r(26) = .08, p = .65$. Also note that when the data were only averaged across participants (and not across positions of swap as before), the correlation of TDS and HI remained significant, $r(110) = .30, p = .001$, but vanished for HTG and CR, $r(110) = .02, p > .1$. The section "Response choice and timbre dissimilarity" below analyzes dissimilarity-based variables and their ability to model response behavior in more detail.

4.4.3 Discussion

This experiment demonstrated a main effect of serial position of swap, which was due to increased performance for sequences with swaps occurring at final positions (2&4, 3&4), indexing a recency effect in serial recognition of timbre. We note that the 2&4 condition in this experiment was equivalent to the swaps from Exps. 1 and 2, thus providing some perspective on these previous choices. From a formal standpoint, it could be argued that the constancy of the position of the swap in Exps. 1 and 2 could have allowed participants to optimize their strategy by only focusing on the two relevant positions (once they would have noticed this circumstance over the course of the experiment). Yet, a comparison with the data from Exp. 3, where swaps were equally distributed, renders this stance implausible. There were no significant differences of d' scores from the 2&4 condition in Exp. 3 ($M = 2.6$) compared to the corresponding conditions (constant pitch, length 4) from the two previous experiments, neither for the group of musicians in Exp. 1 ($M = 2.5$), two-sample $t(50) = 0.3$, $p = .75$, nor for the (musician) participants of Exp. 2 ($M = 2.9$), two-sample $t(42) = -0.75$, $p = .46$. Of course, this perspective is also supported by the strong effect of length in both experiments, equally unlikely in conjunction with attention distributed selectively to only one or two serial positions. In the informal post-experimental feedback of Exps. 1 and 2, none of the participants further reported having noticed a constancy in the position of the swap, and comments rather revolved around various different strategies to process

entire sequences.

Regarding the similarity factor of Exp. 3, we tested the measure of timbral dissimilarity of swap (TDS), identified on the basis of Exp. 1 as an alternative to the previously used MTD (measuring all transition dissimilarities). The experiment demonstrated a similarity effect for timbre, paralleling a phonological similarity effect for verbal material in serial recognition (Nimmo & Roodenrys, 2005): highly dissimilar swaps yielded better memory performance than similar ones. More specifically, TDS predicted well the hit-rate (HI) on nonidentical trials. On identical trials, correct-rejection rate (CR) correlated significantly with sequence heterogeneity, when data were averaged over the different swap positions.

Interestingly, the variable of sequence heterogeneity (or its direct inverse: homogeneity) has become relevant in exemplar-based models of item-recognition memory (Kahana & Sekuler, 2002; Nosofsky & Kantner, 2006; Visscher et al., 2007; Viswanathan, Perl, Kahana, Sekuler, et al., 2010). Instead of modeling accuracy, however, these studies attempt to model response tendencies in the form of recognition choice probabilities for highly confusable stimuli. In these studies, the factor of list homogeneity was confirmed to be one of two additive factors, together with the summed similarities of all probe-item dissimilarities. It was shown that on lure trials, where the probe item does not match any list item, an increase in heterogeneity yields a decrease of correct responses independent of list-probe similarity. These findings were confirmed and reinterpreted as an adaptive shift of participants' response criteria by Nosofsky and Kantner (2006): the more homogeneous a list is, the more likely a participant is to respond "old". To the best of our knowledge, no study has yet demonstrated effects of homogeneity or heterogeneity in serial recognition. Experimental task is only one of various aspects in which the current study differs from the cited body of work, however. Another important factor is the confusability of stimuli, which is much greater in the cited studies as compared to the discrete and easily discriminable timbres used here. The data from Exp. 3 only exhibited a correlation between HTG and the probability of "same" responses for identical trials, and only when averaged across different swap positions. We were thus interested in whether sequence heterogeneity would play a significant role when treated as a controlled factor.

4.5 Experiment 4: Facets of similarity

This experiment studied the role of sequence heterogeneity (HTG) and timbral dissimilarity of swapped items (TDS) in a factorial design.

4.5.1 Method

Participants

Twenty-six musicians (11 female) with an average age of 23 years ($SD = 3.1$, range: 19-28) participated. They had received an average of 14 ($SD = 3.8$) years of instrumental training and 4 ($SD = 3.4$) years of formal music-theoretical instruction including ear training. None of the participants had participated in any of the previous experiments. As previously, it was confirmed that participants had normal hearing.

Stimuli and apparatus

The same eight timbres as above were used in sequences of length 4. Half of the comparison sequences were identical. For the other half, items 2 and 3 from the standard sequence swapped order. Items 2 and 3 occurred in both orders in the set of standard sequences (i.e., ABCD and ACBD). The main interest in this experiment was to independently manipulate the two factors of timbral dissimilarity of swap (TDS) and timbral heterogeneity (HTG), each with two levels. Because both variables are correlated (e.g., an increase in TDS implies an increase in HTG), sequences had to be selected carefully in order to guarantee an independent factor design. Figure 4.6 graphs TDS and HTG values of all possible four-item sequences based on the eight timbres in use and the 12 selected sequences per condition. Low TDS sequences ranged between 3.2 and 4.0 dissimilarity units, high TDS sequences between 4.8 and 5.9. Low HTG sequences ranged between 4.0 and 4.9 units, high HTG from 6.1 to 6.8. None of the TDS and HTG distributions from the sub-conditions (e.g., TDS-low x HTG-low) differed significantly on their corresponding factors (i.e., TDS did not differ for TDS-low x HTG-low as compared to TDS-low x HTG-high), as indicated by two-sample t-tests, all $p > .45$. In summary, there were 12 distinct sequences per TDS/HTG condition, each of the factors had two levels, sequences were presented in identical and nonidentical order in the comparison, and items 2 and 3 were present in both orders

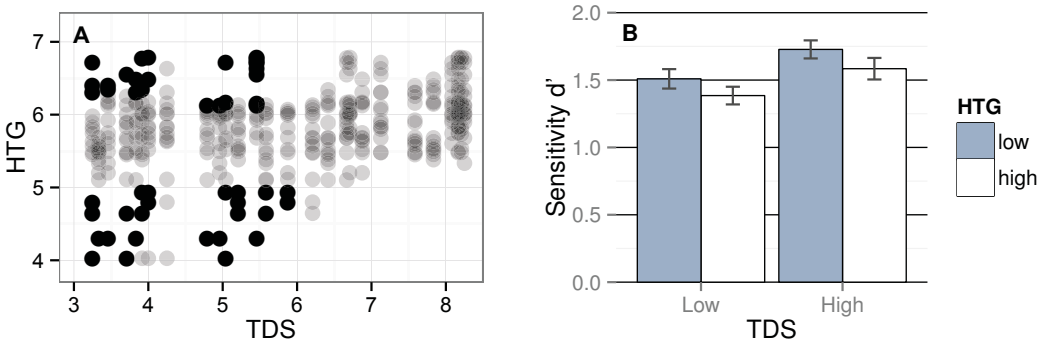

Fig. 4.6 (A) Timbral dissimilarity of swap (TDS) and heterogeneity (HTG) for the 12 sequences per condition selected for this experiment (black dots) and all other possible sequences composed (without replacement) with four out of the eight timbres (gray dots). (B) Sensitivity (d') as a function of timbral dissimilarity of swap (TDS) and heterogeneity (HTG). Errorbars: 95% CI.

in the standard sequence, yielding a total of 192 trials. The apparatus was the same as described above.

Procedure and design

The order of the 192 trials was fully randomized in two experimental blocks of around 20 min duration each. Otherwise, the procedure was identical to that of Exps. 1–3.

4.5.2 Results

Notably, this was the first experiment for which a considerable number of participants (9 out of 26) exhibited chance performance in one of the four experimental conditions. The distribution of mean accuracies for the full experiment followed a bimodal distribution with an average proportion correct of $M = .60$ ($SD = .04$) for the low-performance group and $M = .82$ ($SD = .06$) in the high-performance group, without overlap of distribution, and significantly different as indicated by a two-sample t-test, $t(24) = 10.1, p < .0001$. We did not find biographical factors that accounted for this gap. Yet, the setup of the experiment may have been more difficult than that of Exp. 3, presenting sequences of the same length, but swaps now only occurring at the most difficult position 2&3 in the middle of the sequence, as well as TDS now only being in

the lower half of the TDS range. Average d' sensitivity was 1.55 ($SD = .89$) for the full set of participants, and 2.07 ($SD = .57$) not including chance performers. However, removal of chance performers from the data set did not affect the qualitative pattern of results, which is why results from the full data set are reported in the following.

A repeated-measures TDS (2) × HTG (2) ANOVA indicated that d' scores were affected by TDS, $F(1, 25) = 10.6, p < .004, \eta_p^2 = .25$, but not by HTG, $F(1, 25) = 2.18, p = .15, \eta_p^2 = .004$. There was no interaction, $F(1, 25) < 1$. Note that the same analysis for the dependent variable of proportion of "same" responses (as considered in the original homogeneity studies, cf., Kahana & Sekuler, 2002) did not yield a different picture: there was a significant effect of TDS, $F(2, 50) = 77.1, p < .0001, \eta_p^2 = .75$, but neither an independent effect of HTG, $F(1, 25) = 0.10, p = .75$, nor an interaction of the two variables, $F(2, 50) = 1.78, p = .17$.

4.5.3 Discussion

This experiment confirmed the role of the timbral dissimilarity of the swap (TDS) as an essential variable allowing participants to distinguish between identical and non-identical trials. At the same time, the experiment questioned the impact of sequence heterogeneity (HTG) for the current paradigm, as it did not affect sensitivity or response choices in any systematic way.

4.6 Response choice and timbre dissimilarity

We now attempt to provide a larger context for the somewhat contradictory finding that heterogeneity (HTG) did not yield an effect on choice behavior in Exp. 4, but correlated significantly with the rate of correct rejections in Exp. 3. For that purpose, the choice behavior data from all four experiments are predicted by the dissimilarity-based regressors TDS and HTG. For Exps. 1 and 2, only four-item sequences at constant pitch are considered. To increase the signal-to-noise ratio of the choice data with regards to the factor of interest, i.e., similarity, we averaged data over all identical TDS×HTG conditions, such that there are no two data points that have identical (TDS, HTG) pairs (models for Exps. 1–4 contained 29, 19, 72, 84 points, respectively).

Table 4.1 shows the estimated regression coefficients for all four experiments. All experiments achieve a good fit with roughly 90% explained variance in the choice data.

Table 4.1 Multiple linear regression results for Exps. 1–4 with timbral dissimilarity of swap (TDS) and heterogeneity (HTG) as independent variables. For all four experiments, response choice probability acts as dependent variable. Leftmost column: proportion of variance explained and number of participants.

	Variable	B	SE B	β	p
Exp. 1	Intcpt.	.513	.108	1.96	$< .0001$
($R^2 = .92$)	TDS	.079	.005	.953	$< .0001$
$n = 60$	HTG	-.037	.017	-.016	.058
Exp. 2	Intcpt.	.860	.468	2.61	.084
($R^2 = .90$)	TDS	.092	.008	.940	$< .0001$
$n = 22$	HTG	-.092	.079	-.091	.262
Exp. 3	Intcpt.	.568	.117	1.86	$< .0001$
($R^2 = .90$)	TDS	.095	.004	.958	$< .0001$
$n = 22$	HTG	-.055	.021	-.100	.009
Exp. 4	Intcpt.	.271	.058	1.12	$< .0001$
($R^2 = .89$)	TDS	.101	.004	.940	$< .0001$
$n = 26$	HTG	-.005	.009	-.020	.585

However, only in Exp. 3 is there a significant contribution from HTG, and in Exp. 1 there only is a marginally significant impact. Exps. 2 and 4, however, do not exhibit a significant impact of HTG. At the same time, absolute values of standardized β coefficients are roughly an order of magnitude lower for HTG compared to TDS, again reflecting its significantly inferior predictive power.

Even more important is the fact that stepwise multiple regression did not enter HTG into the model for any of the data from the four experiments (the deviations of the resulting parameter estimates are only marginal as compared to those listed in Table 4.1 and are thus omitted for the sake of brevity). This means that if a parsimonious criterion is taken, the impact of HTG vanishes. The R^2 value in these univariate models is .89 for Exps. 2, 3, and 4, and .91 for Exp. 1, i.e., not more than 1% below the proportion of variances explained by the full models as listed in Table 4.1. A parsimonious account of response behavior thus does not require HTG. Our negative finding is paralleled in recent work by Nosofsky, Little, Donkin, and Fific (2011), where homogeneity did not add explanatory power to an exemplar-based model of response times in item recognition.

There could be a multitude of reasons why studies starting with Kahana and Sekuler (2002) have demonstrated strong effects of heterogeneity in visual and auditory STM, and why this does not extend to the current scenario. These include task differences, because serial recognition probes memory for serial order and not item identity as is the case in item recognition. It has been argued that both types of memory signals rely on different mnemonic mechanisms (Henson, Hartley, Burgess, Hitch, & Flude, 2003), potentially constraining the role of sequence heterogeneity to the item identity case. Further note that for Exp. 3, heterogeneity became relevant for identical trials, whereas in the mentioned item recognition tasks, effects of summed similarity and homogeneity were presented for trials with new probe items. Another factor concerns the confusability of items that are used in typical heterogeneity studies, such as Gabor patches in vision or moving ripples in audition (cf. Visscher et al., 2007) To the contrary, we used clearly distinguishable timbres that varied on multiple perceptual dimensions (as opposed to a single one). This distinctiveness that would set study items apart might be a case in which study list homogeneity does not become relevant.

4.7 General discussion

The current series of experiments attempted to provide a detailed picture of the short-term recognition of musical timbre sequences. Four experiments explored the role of musicianship (Exp. 1), concurrent pitch variability (Exp. 1 & 2), and similarity (Exp. 1, 3, 4).

Exps. 1 and 2 demonstrate that variability in pitch interferes with STM for timbre patterns. This relation, however, is subject to the musical training of participants and the type of variability in pitch. In Exp. 1, we obtained an interaction of the pitch and musicianship factors, such that musicians were only better than nonmusicians on trials with variable pitch. There was no reliable difference between musicians and nonmusicians in the constant-pitch condition. In Exp. 2, we further varied pitch patterns across standard and comparison sequences and showed that musicians show interference in this more complex situation. A post-hoc comparison of Exp. 2 with the corresponding conditions from Exp. 1 confirmed that performance was significantly worse in the variable pitch condition of Exp. 2. The relation between pitch and timbre in STM thus appears to be both a function of musical experience and the type or degree of

variability in the interference domain.

These results complement the perceptual literature, which has not found consistent differences in the timbre dissimilarity judgments of musicians and nonmusicians (McAdams et al., 1995; Lakatos, 2000), but has reported differences in a discrimination task (Chartrand & Belin, 2006) and in neurophysiological responses (Pantev et al., 2001; Shahin et al., 2008; Strait et al., 2012). Using an interpolated tone task, Starr and Pitt (1997) neither observed an effect of interference of pitch, nor an effect of musical training in STM for timbre. Differences between musicians' and nonmusicians' timbre processing may thus be subtle and highly task-dependent. The current data do not show significant differences between musicians and nonmusicians when pitch is constant. Given the previous notes of caution, it seems plausible that this null result remains specific to the current experimental task, serial recognition, and potentially even to the stimuli employed, which were quite different from the timbres (mostly originating from acoustic instruments) that musicians typically deal with on a daily basis.

Let us note that our finding on the interference by pitch variability constitutes a curious perceptual analogy to practices in 20th century music composition, because composers who wished to draw their listeners' attention towards timbral structures have often drastically reduced concurrent complexity in pitch, for instance in "drone"-type pieces. Other well-known classical examples for this include Schoenberg's *Five pieces for orchestra*, op. 16, no. 3 ("Farben"), as well as many works by Giacinto Scelsi, Tristan Murail, and many others (cf. Erickson, 1975; Murail, 2005). One can similarly observe that the focus on sound qualities in popular music has lead to pitch structures that, at times, almost constitute a diminutive feature.

The presented similarity effects extend results from Starr and Pitt (1997) and Visscher et al. (2007) to serial recognition. In short, our observations imply that similarity relations play an integral role in STM for timbre. Exps. 1, 3, and 4 showed that TDS, the timbral dissimilarity of swapped items, is a good predictor of hit rate in serial recognition. Moreover, we tested the "homogeneity-computation hypothesis" (Kahana & Sekuler, 2002), but did not find a strong effect of list homo/heterogeneity. The most powerful predictor of response behavior in our current study was the perceived timbral dissimilarity of items that exchanged order. This single variable predicted the greatest portion of variance of response choices throughout all four experiments. To

the best of our knowledge, this is the first study on auditory STM that has introduced a parsimonious and parametric notion of similarity for serial recognition.

Conceptualizing the matching process, there are two potential strategies that can be roughly distinguished. On the one hand, there may be a tendency to match complete, integrated sequences as *Gestalts*. On the other hand, participants could also match items individually, item by item. Response choices were shown to be well predicted by the summed item-by-item dissimilarities (which in our case were all zero apart from the two items that were swapped, i.e., yielding TDS), but this variable could equally be a strong correlate of any sequence-wise distance measure, and thus cannot distinguish between these two hypotheses. In the current case, strategies appear to depend specifically on characteristics of the listeners (e.g., musical training) as well as the experimental situation (e.g., the presence or absence of pitch variability). A solely item-wise strategy cannot explain the effect of pitch variability for nonmusicians (Exp. 1), because the pitch templates were constant across standard and comparison sequences, and therefore there was no item-wise difference in pitch that could have hampered the computation of timbral difference. For nonmusicians, it thus seems plausible to posit a sequence-wise discrimination process with cross-channel interference from pitch to timbre (Melara & Marks, 1990). In the variable pitch condition of Exp. 1, the (match) result of the pitch-sequence discrimination then biases the discrimination in the target channel of timbre (confirmed by the main effect of pitch variability on bias in Exp. 1). To the contrary, the results obtained for musicians may be better explained by item-wise strategies, which, particularly in the variable pitch condition, may have better allowed timbre to be isolated from pitch on a local level (what Jones & Boltz, 1989, called "analytic attending"). Musicians' sensitivity did not differ between constant and variable pitch conditions in Exp. 1, contrary to Exp. 2, where item-wise differences in pitch may have interfered with local matching of timbre. Needless to say, further experimentation is required to develop the current hypothesis about the matching process' dependency on experimental scenario and musical experience (or "listening skills").

Overall, our findings resonate with what Sekuler and Kahana (2007) dubbed the "stimulus-oriented approach to memory" that emphasizes the interrelatedness of sensory representation and short-term recognition. As the authors note, "But when memory models fail to link their stimulus representations to measures of perceptual simi-

larity, they needlessly limit their ability to account for a variety of important phenomena"(p. 305). Considerations such as these together with the current results imply that it would be hazardous to neglect similarity relations in future studies on short-term recognition of timbre, even if the employed stimuli are easily discriminable. On the contrary, if perceptual similarity is part of the experimental design, similar effects emerge across domains as diverse as musical timbre, musical pitches (Williamson et al., 2010), words and non-words (Nimmo & Roodenrys, 2005), and even auditory ripple noise and visual Gabor patches (Visscher et al., 2007).

Regarding the three variables of interest, pitch variability, musical training, and perceptual similarity, our results can be seen as natural extensions of findings from work on perceptual processing. This characterization favors the view of short-term memory not as a dedicated neural system (Baddeley, 2003), but as an active, top-down-type of trace maintenance that is based on sensory recruitment, i.e., the dedication of attention to sensory representations (Sreenivasan, Curtis, & D'Esposito, 2014; D'Esposito & Postle, 2015). Expressed in other words, our results point towards a sensory-cognitive continuum (cf. Collins, Tillmann, Barrett, Delbé, & Janata, 2014), in which the faculty of nonverbal auditory STM naturally grows on the basis and the properties of sensory representations, rather than being one of many separate "cognitive shoe-boxes" for the retention of modality-specific information.

Chapter 5

Auditory and verbal memory in North Indian tabla drumming

The previous chapter investigated memory for musical timbre in a strictly controlled, but potentially "sterile" experimental setting. This chapter explores a concrete musical scenario in which musical and verbal short-term memory are of immediate relevance in music pedagogy, namely the North Indian *tabla*. The types of sequences encountered in the realm of tabla also yield the incentive to address redundant sequential structures. In that sense, tabla not only allows an exploration of STM for timbre in an ecologically valid setting, but also ventures into the realm of sequential timbral schemata.

This chapter is based on the following research article:

> Siedenburg, K., Mativetsky, S., and McAdams, S. (in preparation). Auditory and Verbal Memory in North Indian Tabla Drumming. Manuscript prepared for submission to *Psychomusicology*.

Abstract. *Tabla* denotes a pair of hand drums that is among the most important instruments in North Indian classical music. Tabla is taught through an oral tradition. Compositions are learned via the memorization of sequences of *bols*, solfège-like vocalizations associated to drum strokes. This study probed short-term serial recognition memory of tabla students and musicians who are naïve to tabla. For investigating the role of familiarity and chunking in the cognitive sequencing of tabla, idiomatic tabla sequences of bols and drum strokes were compared with: i) counterparts reversed in order, ii) sequences with random order and identical item content, and iii) items randomly selected without replacement. A strong main effect of sequence type emerged with monotonically decaying performance (i>ii>iii), underlining the role of chunking in auditory serial recognition. Furthermore, differences between tabla players and musicians only emerged for idiomatic sequences of bols, which constitutes a familiarity effect for verbal, but not for instrumental musical timbres. This is interpreted as a partial dissociation of memory for musical and verbal sounds.

5.1 Introduction

A pertinent theme in the study of auditory cognition addresses the extent to which there is overlap in the cognitive mechanisms involved in the parsing of musical structures and language (see e.g., Patel, 2008; Bigand, Delbé, Poulin-Charronnat, Leman, & Tillmann, 2014; Koelsch, 2009; Williamson et al., 2010). A crucial mechanism for the processing of music and speech is short-term memory (STM). It is responsible for the storage of sensory and categorical information over time spans of roughly 1–30 seconds (Jonides et al., 2008; Baddeley, 2012), and thus allows for the integration of strings of words into phrases and sentences, as well as the detection of repetition and variation in musical phrases. Whether STM is a "blank-slate"-type of buffer or affected by long-term familiarity with stimuli has been intensively discussed in the verbal domain (see e.g., Thorn & Page, 2008; Cowan, 2008; Baddeley, 2012), but research that has compared memory for verbal and musical materials directly has remained scarce.

In the current study, we investigated whether familiarity facilitates auditory short-term serial recognition of sequences comprised of vocal and drum sounds. We explored the example of the *tabla*, a pair of hand drums that is an integral part of North Indian classical music. Its tradition incorporates vocalizations that closely correspond to drum strokes, and thus allows for an ecologically relevant comparison between memory for verbal and instrumental acoustic stimuli. At the same time, tabla music is unknown to many western musicians such that we were able to recruit a truly "naïve" control

group of participants, which we compared to a group of tabla students. The example of tabla therefore seems to be well suited to explore the role of long-term experience in the short-term matching of sound sequences, as well as potential differences between memory for speech and musical sequences.

5.1.1 The North Indian tabla

Tabla is among the most versatile instruments in North Indian music. With its great timbral variety, it can be performed solo, in dance accompaniment (*kathak*) or to accompany melodic instruments such as the sitar, violin or voice. There are six different stylistic schools of playing (*gharanas*), many varied techniques, and a repertoire made up of a multitude of cyclic and cadential compositional forms (Shepherd, 1976). The instrument has traditionally been taught through an oral tradition (Saxena, 2008). Compositions are memorized via a system of *bols* that comprise a solfège system for tabla. The sounds of both drums can be expressed either individually or when two sounds are produced at once, simultaneously. As summarized by Shepherd (1976), "The bol is an aid to memory and not a means of notating tabla strokes exactly. Each stroke on the tabla has one or more corresponding bols. The tabla bol does not require the lips to touch and therefore can be spoken at great speed. In fact the recitation of composition is an art practised in itself."(p. 279)

Tabla vocables have been described as a case of onomatopoeia, or verbal sound symbolism (Patel & Iversen, 2003). It is important to emphasize, however, that the mapping between bols and tabla sounds may vary across schools, and even more importantly, is not one to one. In the *Benares gharana*—the school of tabla from the Varanasi region that we focus on in the current study—the relation is dependent on the musical context. Multiple bols can denote the same stroke, but one and the same bol can also refer to different strokes. Examples are given in Table 5.1.

In order to unambiguously characterize our stimuli in writing, we use the best available approximation in English phonetics of the bols, and a redundant notation for strokes that combines bols and indices that identify the drum and means of sound production. The complete list is given in Tables 5.1 and 5.2. Specifications in superscript refer to the smaller drum (the *dahina*, usually played with the right hand), subscripts refer to the larger drum (the left-hand *baya*). They indicate whether sounds

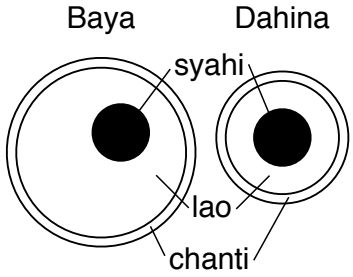

Fig. 5.1 Schematic drawing of the tabla drum surface including the outer ring (chanti), the inner surface head (lao), and the patch (syahi). The Baya is usually played with the left hand, the Dahina with the right hand.

are resonant (o: "open") or damped (x: "closed"). Dahina sounds are further specified by the main point of contact on the drum surface, which is either the rim (c: *chanti*), a circular patch on the drum head (s: *syahi*), or the remaining head (l: *lao*). We also list all alternative bols as they apply for the idiomatic phrases used in the current study. Table 5.2 provides the combination sounds. For a more comprehensive treatment, see Shepherd (1976).

5.1.2 Voice superiority effects

As outlined above, tabla compositions are taught and memorized via the sequencing of verbal material. A natural question regarding this pedagogical practice is whether it exploits processing and memory benefits for conspecific vocalizations. There is a growing number of studies that have not only suggested general processing benefits of vocal compared to instrumental musical timbres (Chartrand & Belin, 2006; Agus et al., 2012; Suied et al., 2014), but also enhanced recognition memory of melodies presented via voice timbres (Weiss et al., 2012; Weiss, Schellenberg, et al., 2015; Weiss, Vanzella, et al., 2015). The experimental task in the latter studies was to recognize melodies from an exposure phase within a set of foils. Testing adults (Weiss et al., 2012), children from different age groups (Weiss, Schellenberg, et al., 2015), and musicians with and without absolute pitch (Weiss, Vanzella, et al., 2015), results converged on a mnemonic advantage for melodies presented via vocal timbres, even if these were rated as least pleasurable among the test timbres (Weiss et al., 2012). Furthermore, mere timbre familiarity seems an unlikely cause of this effect, given that pianists recognized piano melodies not as well as vocal melodies (Weiss, Vanzella, et al., 2015), also see (Halpern

Table 5.1 We use a redundant notation that specifies any stroke by its bol, by whether it is produced on the high-pitched *dahina* (superscript indices) or low-pitched *baya* (subscript), and by whether it is resonant (o: "open") or non-resonant (x: "closed"). Any dahina stroke is further specified by the major point of contact on the drum surface, the rim (c: *chanti*), the head (l: *lao*), or the black patch in the centre of the drum (s: *syahi*), or whether the head and rim is struck by the palm (p). The last column lists alternative bols with an examplary context in brackets.

Drum	Symbol	Playing technique	Alternative bols
Dahina	Na^{oc}	Index finger strikes chanti, while ring finger lightly touches edge of syahi, producing an overtone	
	Tin^{ol}	Index finger strikes lao, middle finger strikes syahi, ring finger lightly touches edge of syahi, producing an overtone	Taa (TaaKaTaaKa)
	Tun^{os}	Index finger strikes center of syahi. NB: in the experiment, this sound is only used in combination (see Table 5.2), but not solo	
	Te^{x1}	Index finger strikes center of syahi	Ra (TiRaKiTa), Taa (KiTaTaaKa)
	Te^{x2}	Middle and ring finger strike center of syahi	Ti, Ta (TiRaKiTa)
	Te^{xp}	Right side of palm strikes syahi and much of lao	
	Re^{xp}	Left side of palm strikes syahi and much of lao	
Baya	Ge_{o}	Index or middle finger strikes lao, wrist on head can induce pitch glide	
	Ke_{x}	Flat hand strikes lao and syahi	Ki, Ka (KiTaTaaKa)

Table 5.2 Combination sounds of both drums. Notation refers to the third column of Table 5.1

Symbol	Dahina	Baya	Alternative bols
Dha_{o}^{oc}	Na^{oc}	Ge_{o}	
Dhin_{o}^{ol}	Tin^{ol}	Ge_{o}	Dha (DhaGeTeTe)
Dhin_{o}^{os}	Tun^{os}	Ge_{o}	
Dhe_{o}^{x2}	Te^{x2}	Ge_{o}	
Dhe_{o}^{xp}	Te^{xp}	Ge_{o}	
Tin_{x}^{os}	Tun^{os}	Ke_{x}	

& Müllensiefen, 2008).

Nonetheless, not all studies that address this issue report unambiguous memory advantages for verbal materials. Schulze and Tillmann (2013) compared the matching of sequences of words (presented vocally) with that of sequences comprised of different musical instrument timbres from western orchestral instruments presented with constant pitch. Across three different tasks (forward serial recognition; backward recognition, requiring participants to match a reversed comparison sequence; and backward recognition with articulatory suppression, additionally requiring participants to speak aloud during the retention interval), they compared performance for words to instrumental timbres. Although performance did not differ in absolute terms across the three experiments, backward recognition memory for words was more strongly affected by articulatory suppression than was the case for timbres. Also comparing memory across domains, Williamson et al. (2010) suggested that STM for auditorily presented letters and tones from a diatonic scale is shaped by similar mechanisms, such as limited capacity and a detrimental effect of perceptual proximity. In their experiment, notably, only non-musicians' performance was reduced by pitch-proximity, but not that of musicians. This furthermore demonstrates that long-term experience and expertise play a role in STM for musical materials.

Given the emphasis on vocalization in tabla pedagogy, we expected to observe generally better short-term recognition memory for vocal sequences, compared to tabla sequences.

5.1.3 Familiarity and chunking

Familiarity with a musical style may be considered to be both founded on knowledge of a style's basic acoustic units with their characteristic acoustic interrelations and on experience with the ways in which units connect over time to build sequences. Working on the schema of the diatonic scale, Schulze, Jay Dowling, and Tillmann (2012) showed that pitch sequences that conformed to diatonic scales were easier to match to comparison sequences than were non-diatonic sequences for both musicians and non-musicians. The magnitude of the effect varied across different sequence lengths and vanished for backward recognition, which indicates that these effects may be highly sensitive to the specific experimental task. In fact, the sensitivity of familiarity with

the experimental task has been a well-known phenomenon for the most prominent familiarity effect in verbal memory, the *lexicality effect*: short-term serial recall of a list of non-words (non-lexical vocables that lack a LTM representation) is worse than that of words, but this difference vanishes for auditory serial recognition (Thorn et al., 2008; Macken, Taylor, & Jones, 2014). Whereas the contrast of diatonic vs. non-diatonic sequences by Schulze et al. (2012) constitutes a familiarity effect based on a cognitive representation of relations among the test items, namely a scale from which tones are drawn, the results of Vuvan, Podolak, and Schmuckler (2014) highlight effects of knowledge about structures unfolding over time, i.e., sequential schemata. Here the authors showed that tonal melodic expectations affected STM by eliciting more false memories for highly expected tones than for less expected ones. Moreover, the strength with which melodic expectations played into STM was a function of the specificity of the elicited melodic expectations.

Idiomatic tabla phrases are usually constructed by a nested succession of groups of items. The eight-item phrase comprised by the bols DhaDhaTe↑Te↓DhaDhaTinNa, for instance, features the sub-element Te↑Te↓ (arrows denoting an upward or downward vocal pitch contour that usually accompanies this pair of bols), which is a frequently occurring chunk in the tabla repertoire, and so is DhaDhaTe↑Te↓ (see Table 5.3 for more examples). It is a classic insight that subdividing memory lists via chunking is a highly effective mnemonic strategy (Miller, 1956). Cowan (2001) defines a chunk as a "collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use."(p. 89) Furthermore, sequences of letters such as IRSCIAFBI are far easier to memorize as the chunks IRS CIA FBI, familiar US federal agencies, than as raw item-by-item successions (Cowan, 2008). Memory over the short term thus is also affected by the ways in which LTM mediates chunking of sequences into meaningful subgroups of items. Regarding tabla, it is therefore likely that participants who are familiar with the tabla repertoire possess LTM representations that allow for the efficient encoding of these basic building blocks or chunks of the tabla repertoire.

In addition to this latter facet of chunking due to long-term associations between items from the list, there is another facet that is based on the immediate structure of auditory stimuli. In the tabla sequence DhaDhaTe↑Te↓DhaDhaTinNa, the subgroups DhaDha and Te↑Te↓ naturally emerge as chunks, because identical items feature strong

associations among each other. For that reason, the immediate and non-immediate repetition of items in a sequence not only reduces memory load in terms of item identity, but also facilitates the chunking and segmentation of sequences (Gobet et al., 2001). Furthermore, idiomatic sequences feature hierarchical structure that arises through nested repetition of chunks, a form of simple temporal recursivity (Jones & Boltz, 1989). The above example can be well subdivided into two phrases, DhaDhaTe↑Te↓, and DhaDhaTinNa, both of which start with the same chunk and comprise two chunks of two items each. It is likely that this structure is exploited in the encoding of sequences, as listeners have been shown to be sensitive to hierarchical structure both in language and music (Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013; Patel, 2008; Gobet et al., 2001; Jones & Boltz, 1989).

In sum, there are at least three potential components that differentiate an idiomatic tabla phrase from a random sequence: i) reliance on idiomatic chunks of items, ii) hierarchical ordering, i.e., nested repetition of chunks, iii) repetition of items. The sequencing factor included conditions that attempted to differentiate between these components. The factor's first condition comprised idiomatic phrases. The second condition presented the idiomatic sequences in backward order; this operation preserved the hierarchical structure inherent to sequences, but was expected to remove familiarity with the sequential patterns for tabla players, turning familiar idiomatic phrases into non-idiomatic phrases. For "naïve" musicians, we did not expect to observe a difference between these two conditions. The third condition contained sequences with the same items as in conditions 1 and 2 but in a randomly scrambled order. This should remove any effect of hierarchical ordering while preserving the same overall redundancy of items within sequences. As a baseline, items were drawn randomly without replacement from the full set of sounds in the fourth condition.

5.1.4 The present experiment

We tested the role of sequential schemata with vocal and percussion sounds from the North Indian tabla in a serial order recognition task. A within-subjects factor varied the structure of the sequences and included four conditions: 1) idiomatic tabla sequences, 2) reversed idiomatic sequences, 3) sequences with random order but the same items as the sequences in 1 or 2, and 4) sequences with items drawn randomly without

replacement. We expected that recognition memory would decay over conditions 2–4 for all participants , but would only be better for condition 1 over condition 2 for tabla players. We expected a main effect of the within-subjects factor of material (table vs. bols) with processing advantages for verbal over drum sounds. Due to the lack of lexicality effects based on familiarity with items in auditory serial recognition (Macken et al., 2014), we did not expect to observe a main effect of group.

5.2 Methods

5.2.1 Participants

Twenty-one tabla players participated in the experiment. They were recruited in a concerted effort among the second author's tabla students. One participant did not have hearing thresholds that fell in the range required for the experiment (see below) and was excluded from the data analysis. For the control group, we tested 23 musicians without experience in tabla. One participant could not complete the experiment due to a fire alarm, the data of two participants were lost due to errors with a computer server.

Tabla players ($N = 20$, 6 female) had a median age of 23 years ($M = 27$, $SD = 11.2$, range: 21–62), had received instruction on a musical instrument for a median of 15 years ($M = 14$, $SD = 4.0$), and a median of 5 years ($M = 5$, $SD = 3.8$) in formal music theoretical training including ear training. At the time of testing, they had played the tabla for a median duration of 11 months ($M = 35$, $SD = 37.7$, range: 3–120) for a median of 5 hours per week ($M = 5$, $SD = 2.3$, range: 3–10). The sample should thus be considered as representing a beginner's level in tabla. Among these, there were eleven participants whose primary instrument was percussion other than tabla, three primary tabla players, three pianists, one guitarist, one flutist, and one saxophonist.

Musicians without experience in tabla ($N = 20$, 9 female) had a median age of 22 years ($M = 23$, $SD = 3.6$, range: 19–36), had received musical instrumental instruction for a median duration of 15 years ($M = 15$, $SD = 4.6$), and possessed a median of 5 years ($M = 6$, $SD = 3.3$) of experience in formal music-theoretical training including ear training. Primary instruments from participants in this group were fairly equally distributed among common western instruments (3 piano, 3 guitar, 3 clarinet, 3 singer,

2 trumpet, 2 cello, 1 violin, 1 viola, 1 acoustic bass, and 1 bassoon).

A standard pure-tone audiogram with octave frequency spacing was measured right before the main experiment (ISO 398-8, 2004 F. N. Martin et al., 2000) in order to confirm that participants had hearing thresholds of 20 dB HL or better in the range of 125–8000 Hz (octave-spaced).

5.2.2 Stimuli

Individual sounds Tabla and bol sounds were recorded in a sound recording studio at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) in Montreal, QC, Canada, using a matched pair of AKG C414 B-XLS large-diaphragm condenser microphones (AKG Acoustics GmbH, Vienna, Austria) with wide-cardioid characteristics and the Reaper digital audio workstation (Cockos Inc., New York, NY, USA). One microphone was positioned with a vertical elevation of around 20 cm above the drum surfaces, and a direct distance of around 30 cm to both drums. Another microphone was placed around 20 cm in front of the player's mouth for recording the vocalizations. Only the respective mono channel (tabla or bols) was used subsequently. We recorded all common tabla sounds and vocalization as played and spoken in isolation by the second author. A realization of every tabla stroke and bol was selected among multiple recordings, followed by manual equalization of the onset lag. In order not to distort natural differences in loudness between stimuli, we normalized the group-averaged root-mean-squared energy (computed for the interval 50–250 ms after stimulus onset of each sound) of all tabla strokes and bols that were used in the experiment. Each sound was then cropped to a duration of 400 ms by applying a 20 ms raised cosine fade-out.

Memory sequences Every standard sequence consisted of eight sounds. Within sequences, there were 10 ms of silence between sounds, yielding an inter-onset-interval of 410 ms. After the standard sequence, there was a delay of 3280 ms, followed by a comparison sequence that was of identical order in half of the trials and of non-identical order in the other half. In the latter case, items 4 and 5 swapped. We distributed sounds in the sequencing conditions 2–4 such that items 4 and 5 were fixed across conditions (as outlined below). This held the perceptual dissimilarity of the swap constant across sequencing conditions, because that dissimilarity had been found

to be a crucial variable in serial recognition of musical timbre in previous experiments, and had explained a major portion of variance in response choices (Ch. 4). This seemed to be the only way not to confound the sequencing factor by differing similarities of swaps, while at the same time retaining a feasible design.

There were four sequencing conditions. Condition 1 ("Idio") contained twelve idiomatic tabla sequences that are commonly used in the repertoire of the *Benares gharana*, the style of the tabla school from Varanasi in North India that the group of tabla participants studied with the second author. They were selected from a larger pool of idiomatic phrases that had been provided by the second author. Selection was based on the criteria that sequences should be isochronous and without gaps (pauses), and that items 4 and 5 should not be identical, but also not very perceptually similar (neither for tabla, nor for bols). Extremely similar bols are Te^{x2} and Te^{x1}, or Te^{xp} and Re^{xp}, for instance. Table 5.3 lists all idiomatic sequences.

As outlined above, the correspondence of bols and strokes is not one-to-one. This resulted in that the number of items per sequence varied across material conditions, although sequences were matched otherwise. We preferred to allow for this variability in order not to render the verbal condition overtly unnatural and irritating to tabla players. This yielded sets of 14 unique tabla stimuli and 16 unique bols from which sequences were constructed. The number of different items per sequence varied between three and six for both material conditions, but with an average number of $M = 4.2$ items per sequence for tabla strokes, and $M = 5.1$ items for bols.

Condition 2 ("Rev") contained all idiomatic sequences in reversed order. This kept the hierarchical structure of repeating subgroups of items intact, but rendered their sequential structure completely unfamiliar to tabla players. Condition 3 ("RO") presented the same sequences as used in conditions 1 and 2, but shuffled their order, apart from items 4 and 5 which were kept at place for reasons explained above. This means that for every sequence used in condition 1 (and 2), there existed an analogue in condition 3 that contained the same items in random order. This preserved the same number of items per sequence, but annihilated their hierarchical structure. Condition 4 ("RI") presented randomly drawn items (without replacement) yielding eight different items per sequence. As before, items 4 and 5 from condition 1 were kept in place.

In order to measure the effects of random ordering and random item contents on memory, and not the peculiarities of one particular realization of a random process,

Table 5.3 Idiomatic tabla and bol sequences. We included up- or downward arrows for the bol *Te*, because *Te*↑ (corresponding to the stroke Te^{x2}) is usually pronounced with an upward and *Te*↓ (Te^{x1}) with a downward pitch countour.

Strokes								Bols							
Dha_o^{oc}	Dha_o^{oc}	Te^{x2}	Te^{x1}	Dha_o^{oc}	Dha_o^{oc}	Tin_x^{os}	Na^{oc}	Dha	Dha	Te↑	Te↓	Dha	Dha	Tin	Na
Te^{x1}	Ke_x	Te^{x2}	Te^{x1}	Ke_x	Te^{x2}	Te^{x1}	Ke_x	Taa	Ka	Ti	Ra	Ki	Ta	Taa	Ka
Dhe_o^{xp}	Re^{xp}	Te^{xp}	Re^{xp}	Ke_x	Te^{x2}	Te^{x1}	Ke_x	Dhe	Re	Dhe	Re	Ki	Ta	Taa	Ka
Ke_x	Te^{x2}	Ke_x	Te^{x2}	Ge_o	Ge_o	Te^{x2}	Te^{x1}	Ka	Ta	Ka	Ta	Ge	Ge	Te↑	Te↓
Ge_o	Ge_o	Te^{x2}	Te^{x1}	Ge_o	Ge_o	Na^{oc}	Na^{oc}	Ge	Ge	Te↑	Te↓	Ge	Ge	Na	Na
Te^{x2}	Te^{x1}	Ke_x	Te^{x2}	Ke_x	Te^{x2}	Te^{x1}	Ke_x	Ti	Ra	Ki	Ta	Ki	Ta	Taa	Ka
Te^{x2}	Te^{x1}	Ke_x	Te^{x2}	Tin^{ol}	Ke_x	Tin^{ol}	Ke_x	Ti	Ra	Ki	Ta	Taa	Ka	Taa	Ka
Te^{x2}	Te^{x1}	Ke_x	Te^{x2}	Tin_x^{os}	Na^{oc}	Tin_x^{os}	Na^{oc}	Ti	Ra	Ki	Ta	Tin	Na	Tin	Na
Dha_o^{oc}	Te^{x1}	Dha_o^{oc}	Dhe_o^{x2}	Te^{x2}	Te^{x1}	Dha_o^{oc}	Te^{x1}	Dha	Ra	Dha	Dhe	Te↑	Te↓	Dha	Ra
Dhin_o^{os}	Ge_o	Te^{x2}	Te^{x1}	Tin^{ol}	Ge_o	Te^{x2}	Te^{x1}	Dha	Ge	Te↑	Te↓	Taa	Ge	Te↑	Te↓
Dha_o^{oc}	Ge_o	Na^{oc}	Ge_o	Dhin_o^{os}	Na^{oc}	Ge_o	Na^{oc}	Dha	Ge	Na	Ge	Dhin	Na	Ge	Na
Dha_o^{oc}	Tin_x^{os}	Na^{oc}	Dha_o^{os}	Tin_x^{os}	Na^{oc}	Dha_o^{oc}	Tin_x^{os}	Dha	Tin	Na	Dha	Tin	Na	Dha	Tin

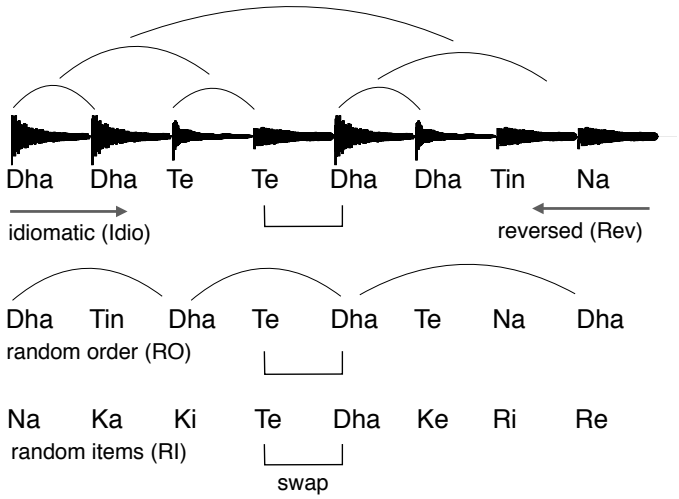

Fig. 5.2 Example of the four sequencing conditions. An idiomatic sequence of bols and the corresponding reversed, random order, and random items conditions. Note that in all conditions, the same items swap positions in non-match trials.

every participant received different randomizations for conditions 3 and 4.

5.2.3 Presentation and apparatus

The average presentation level was $M = 68$ dB SPL ($SD = 4.7$) as measured with a Brüel & Kjær Type 2205 sound-level meter (maximum level, A-weighting) with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). Experiments took place in a double-walled sound-isolation chamber (Industrial Acoustics Company, Bronx, NY, USA). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany), using a Macintosh computer with digital-to-analog conversion on a Grace Design m904 (Grace Digital Audio, San Diego, CA, USA) monitor system. The experimental interface and data collection were conducted by the audio software MaxMSP (Cycling 74, San Francisco, CA, USA).

5.2.4 Procedure and design

Participants were presented four example trials, two of which consisted of tabla trials, two of bols. Correct responses were provided on the screen. In the main experiment,

they were asked to listen to the two sequences and respond to the question "Was the order of the two sequences identical?" by pressing dedicated "Yes" and "No" buttons on the computer keyboard. After participants had provided their response, the message "recording response" was displayed on the screen for 4 s. They could then proceed to the next trial.

Twelve sequences per condition were presented (each as identical and non-identical standard-comparison pairs), yielding 12 × 2 (material: tabla vs. bols) × 4 (sequencing conditions) × 2 (identical/non-identical) = 192 trials in total. The experiment was administered in a split-plot design, containing four blocks of 48 trials each with material counterbalanced across participants (e.g., tabla-bols-tabla-bols). Sequencing conditions were fully randomized within every block. Every block required around 15 min to complete and participants were required to take breaks of 5 min between blocks. After completing the listening experiment, participants filled out a questionnaire on biographical information, musical background, and strategies employed during the experiment. Participants were compensated with CAD 15 for their time.

5.3 Results

Discrimination sensitivity was assessed by calculating d' scores and criterion location c based upon the yes-no model (Macmillan & Creelman, 2005, Ch. 2). Hits were defined as trials that participants correctly identified as non-identical, false alarms as trials that were incorrectly judged as non-identical. No violations of sphericity were observed (Mauchly's test). Figure 5.3 shows sensitivity and bias for all conditions.

A mixed 2 × 2 × 4 ANOVA did not yield main effects of group, $F(1, 38) = 2.32$, $p = .135$, or material, $F(1, 38) = 0.451$, $p = .502$, but a strong main effect of sequencing, $F(3, 114) = 132.1$, $p < .001$, $\eta_p^2 = .776$. The sequencing factor interacted with group, $F(3, 114) = 4.27$, $p = .006, \eta_p^2 = .101$, and material $F(3, 114) = 14.7$, $p < .001$, $\eta_p^2 = .278$. Furthermore, there was a marginally significant three-way interaction between group, material and sequencing, $F(3, 114) = 2.35$, $p = .076$, $\eta_p^2 = .058$.

The main effect of sequencing was qualified by post-hoc contrasts, for which both linear, $\beta = -1.18$, $t(266) = -19.9$, $p < .001$ and quadratic $\beta = -0.25$, $t(266) = -4.3$, $p < .001$ polynomials were significant. Post-hoc comparisons were conducted, here and in the following using the Bonferroni-Holm correction for multiple comparisons with

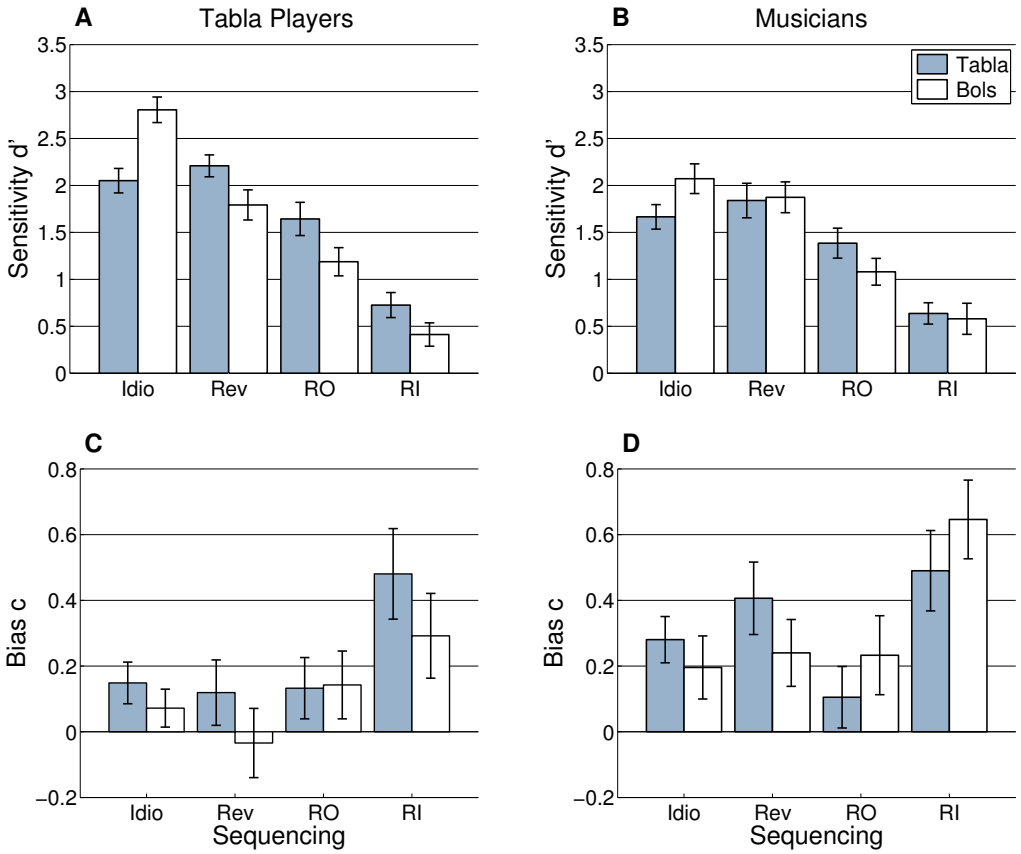

Fig. 5.3 d' scores for tabla players (A) and musicians (B) for idiomatic sequences (Idio), reversed sequences (Rev), random order (RO), and random items (RI). Response bias as given by the criterion location c for tabla players (C) and musicians (D). Error bars display standard errors of the mean.

the critical level $\alpha_{crit} = 0.05/k$, k corresponding to the inverse rank order of the corresponding p-value among all p-values considered for the interpretation of the respective (interaction) effect. Comparisons attested that the interaction between sequencing and group occurred because sensitivity was higher for tabla players than for musicians in the idiomatic condition, $t(38) = -3.63, p = .0008$ ($k = 4$, $\alpha_{crit} = .0125$), but the two groups were equivalent in all other three conditions, all $|t(38)| < 0.92, p > .36$. The interaction between sequencing and material factors reflected the greater sensitivity to bols than to tabla strokes in the idiomatic condition, $t(39) = -4.65, p < .0001$ ($k = 4$, $\alpha_{crit} = .0125$), paired with a greater sensitivity for tabla compared to bols in the random order condition, $t(39) = 3.71, p = .0006$ ($k = 3$, $\alpha_{crit} = .0167$). No other

significant differences between materials for the remaining two sequencing condition were found, $t(39) < 1.73, p > .09$.

Both two-way interactions, however, appear to be partially driven by the far superior performance of tabla players for idiomatic sequences of bols, giving rise to the (marginally significant) three-way interaction: Tabla players were better than musicians for idiomatic bols, $t(38) = 3.51$, $p = .0012$ ($k = 14$, $\alpha_{crit} = .0036$), but not for idiomatic tabla, $t(38) = 2.08, p = .043$ ($k = 13$, $\alpha_{crit} = .0038$), and there were no other significant differences between tabla players and musicians elsewhere $t(38) < 1.69$, $p > .098$. Furthermore, tabla players were better on idiomatic bols than on idiomatic tabla, $t(19) = -4.46, p = .0003$ ($k = 16$, $\alpha_{crit} = .0031$), but worse for bols compared to tabla on random order sequences, $t(19) = -4.04, p = .0006$ ($k = 15$, $\alpha_{crit} = .0033$). Other differences between tabla and bols for tabla players did not survive correction for multiple comparisons, all $t(19) < 2.38$, N.S.. Similarly, none of the differences between tabla and bols sequences were significant for musicians, all $|t(19)| < 2.27$, N.S..

For tabla players, we did not find significant positive correlations of d' scores and the number of month of experience with tabla in any of the experimental conditions with corrected α-levels, $|r(18)| < .48$, N.S..

Response bias as measured according to the criterion location yielded a main effect of sequencing, $F(3, 114) = 18.3, p < .001, \eta_p^2 = .325$. There was no significant main effect of material, $F(1, 38) = 1.9, p = .174$, nor of group, $F(1, 38) = 1.82, p = .184$. There was no significant interaction, all $p > .10$. Post-hoc t-tests confirmed that the main effect of sequencing on response bias was due to significantly higher bias for the RI condition compared to any other condition, all $t(39) > 4.9, p < .0001$ ($k = 6$, $\alpha_{crit} = .0083$, Bonferroni-Holm corrected α-level), but no differences otherwise, all $p > .57$. This means that in the RI condition, both groups missed more non-identical trials.

In summary, we observed a strong main effect of sequencing condition on sensitivity. As expected, worst performance occurred for sequences without repeating items (RI), followed by randomly structured sequences with a smaller number of items (RO). Further, sequences that contained hierarchical groupings (Idio and Rev) were memorized most easily. We interpret the interactions between sequencing, material, and group to be driven by the superior performance of tabla players on idiomatic bols. Notably,

there was no difference between idiomatic and reversed tabla sequences, neither for musicians without experience in tabla (as we expected), nor for tabla players (contrary to our hypothesis). Considering bol vocalizations on the contrary, tabla players were better for idiomatic sequences compared to reversed sequences, but this difference was not statistically significant for musicians without experience in tabla.

5.4 Discussion

The current findings suggest a dissociation of recognition memory for verbal and musical material. We only observed a facilitated recognition memory for verbal material on idiomatic sequences, but not for any other sequence types. This advantage was significant for tabla players who had on average 11 month of experience. It should be remarked that both groups were closely matched in terms of their training in Western music, close to a professional level for many participants, whereas tabla students were mostly at a beginner's level on the tabla. Within that sample of tabla students, surprisingly, performance (d' scores) in none of the experimental conditions correlated significantly with the reported number of months of practice on the instrument and the associated vocalizations. These circumstances suggest that the observed facilitation of verbal memory may be due to learning that proceeds comparatively quickly on the timescale of months.

Two potential limitations of the current experimental design deserve further notice. First, given that we worked with natural tabla and bol sounds, we did not determine which exact auditory features were decisive for distinguishing items. The most likely candidate attribute is timbre, denoting that bundle of spectrotemporal auditory features that conveys the identity of a sound source and that for many instruments covaries with pitch and loudness (McAdams, 2013). Items within the sets of bols and strokes were not equalized in relative pitch or loudness, however, and it is therefore hard to assess the overall impact of these two attributes. Pitch is an especially salient feature for tabla strokes, because the resonant sounds of the dahina and baya are spaced at around an octave and thus can be easily distinguished. Pitch differences in bols were not as salient, but here participants may have also relied on pitch contour. Notwithstanding, timbre was the only cue that distinguished all items from each other, not only subsets of items, for both sets of bols and strokes.

Second, in order not to confound the sequencing factor by varying the timbral dissimilarities of the swaps, we kept the positions of the swapped items constant (items 4 and 5). The relevance of this approach was justified by a post-hoc analysis that compared performance on similar and dissimilar swaps of strokes. Similar strokes were here defined, for the lack of direct dissimilarity ratings, as swaps that comprised either both resonant or non-resonant (dampened) strokes, and dissimilar swaps comprised mixed pairs. The hit rate for dissimilar swaps (including 7/12 pairs) was $M = .70$ ($SD = 0.15$), and for similar pairs (5/12 pairs) it was $M = .61$ ($SD = 0.17$). Comparing participant-wise hit rates for these two types of swaps confirmed a highly significant advantage for dissimilar pairs, $t(39) = 4.45, p < .001$. This design may have allowed for the possibility that participants selectively attended to items 4 and 5 without attempting to memorize the full sequence. It seems unlikely that this was the case, however. Participants were asked to describe the strategies they had employed in a questionnaire after completing the experiment. None of the participants mentioned focusing on a particular serial position. On the contrary, musicians reported a variety of different strategies, such as trying to memorize the full sequential pattern, chunking the sequence into smaller subgroups, or focusing on damped-resonant patterns. Tabla players additionally mentioned using idiomatic patterns from the repertoire as a memory aid, associating movements on the instrument with the sequence, or making use of verbal rehearsal. Furthermore, in case a majority of participants would have only focused on these two serial positions, performance would have likely shown ceiling effects across all sequencing conditions, which is clearly not the case. The variety of strategies just mentioned therefore suggests that participants selectively attended to the variety of perceptual-motor affordances of items (cf., Macken et al., 2014), rather than to a limited number of serial positions.

The only notable difference in response bias was between RI sequences and all other sequencing conditions, a circumstance that may reflect the categorical difference in the construction of these sequences. RI was the only sequencing condition that did not contain repetitions of items. Here participants adopted a less "critical" criterion (i.e., more often considering sequences as identical). Sequences in the RO condition contained repetitions of items, thus reducing the memory load in terms of item identity. The Rev and Idio conditions, moreover, contained repeating subgroups of items,

therefore potentially allowing participants to hierarchically represent sequences by considering successions of subgroups. Both groups of participants showed a linear decay of sensitivity across these three conditions (Rev > RO > RI), reflecting a mnemonic hierarchy of sequential structure. Although the material factor interacted with that of sequencing, there was no main effect of material, contrary to our hypotheses. Congruent results have been observed for the serial recognition of words and timbres by Schulze and Tillmann (2013). Using three different matching tasks, they did not find differences in performance between sequences of words and timbres. However, they only used non-structured sequences, comparable to the current RI condition.

The current results can be viewed both from encoding and maintenance perspectives. Recall that the idiomatic phrases were composed by subgroups of items that feature a high frequency of co-occurrence, and thus are likely to be encoded as chunks by tabla players. Take the phrase TiRaKiTa as an example, a frequently occurring "word" of the tabla repertoire. Instead of memorizing four items, participants who are familiar with the style only need to retain one chunk that represents the phrase. Although familiarity-based chunking in immediate memory is usually considered as a domain-general mechanism (Cowan, 2001; Gobet et al., 2001), tabla strokes were curiously not processed in the same way. Otherwise they would have yielded the same boost of recognition performance for idiomatic sequences, and there would not have been an interaction between material, sequencing, and group. Verbal material appears to be particularly suited for chunking, because language learning requires the acquisition of an enormous amount of vocabulary based on the hierarchical representation of chunks of phonological sequences (e.g., Pinker, 1994; Patel, 2008). It thus seems natural that familiar chunks of verbal material are represented most efficiently. Memory for sequences of musical timbres, on the other hand, may be veridical and of high fidelity in immediate recognition, but less apt for hierarchical abstraction via familiarity-based chunking.

Considering subvocal rehearsal as a potential maintenance mechanism, it is clear that all experimental conditions afforded for (sub)vocal rehearsal by tabla players in principle (e.g., Baddeley, 2012). In fact, in the course of learning the instrument, students must acquire strong associations of bols and tabla sounds in order to be able to memorize and recite bol sequences. Bols are, of course, particularly suited for sub-

vocal rehearsal because they already are in a verbal format. The current three-way interaction could then be interpreted as caused by facilitated rehearsal of idiomatically structured verbal sequences. In particular, tabla players are trained in the rapid recitation of idiomatic sequences of bols, an ability that may aid subvocal rehearsal. An effect of articulatory fluency on serial recall was observed by Woodward, Macken, and Jones (2008). Here participants were trained to quickly articulate lists of (non-)words, and in consequence showed better recall performance. The advantage was independent of item-wise articulation fluency, but only occurred when lists were composed of items that co-occurred in the training phase. As noted, however, results from serial recall do not directly apply to those from serial recognition, in particular when dealing with familiarity effects (Thorn et al., 2008). Although recognition performance for sequences of timbres and words did not differ in the data of Schulze and Tillmann (2013), backward serial recognition of words with articulatory suppression was significantly worse than without, but not so for timbres. This finding may suggest that articulatory rehearsal processes are preferentially used in verbal memory. Nonetheless, this evidence remains ambiguous for an account solely based on articulatory rehearsal, given that maintenance strategies have been shown to be dependent on the concurrent load in verbal working memory tasks (Camos, Mora, & Oberauer, 2011): in tasks with high concurrent load, subvocal rehearsal was shown to be the predominant strategy of young adults. On the other hand, maintenance strategies such as attentional refreshing were preferentially used in low-load tasks. Due to no interference whatsoever and a relatively short retention interval (3280 ms), the current serial recognition task certainly constitutes a low-load task, in contrast to backward serial recognition from Schulze and Tillmann (2013) in which participants needed to reverse the order of the standard sequences. The role of subvocal rehearsal in the current results thus remains questionable. Considering both potential encoding and maintenance processes may nonetheless be the most realistic vantage point for ecologically relevant scenarios, using stimuli that feature a broad array of perceptual-motor affordances.

In conclusion, this study explored the cognitive sequencing of verbal and drum sounds in the example of the North Indian tabla. Although tabla constitutes a musical tradition that at first glance treats language and music in surprisingly similar ways, human memory appears to feature subtle but important differences in dealing with

these two classes of acoustic stimuli. We observed a dissociation between memory for verbal and instrumental sounds, that arises in conjunction with familiarity with the musical style. Tabla students better recognized familiar sequences of verbal material compared to sounds of tabla strokes and a control group of music students without experience in tabla.

This dissociation may be one of the driving forces behind traditional tabla pedagogy that requires students to memorize compositions by listening to their teachers' vocalizations, rather than their tabla drumming. Such a claim remains speculative insofar as the role of multimodal cues in short-term memory should not be underestimated (Quak, London, & Talsma, 2015); tabla students' memory may well benefit from *watching* their teachers play. Nonetheless, when memory for auditory stimuli are considered as such, and participants are familiar with the basic building blocks of the repertoire, this study has shown clear mnemonic advantages for tabla's vocal solfège, underlining the cognitive utility of this centuries-old pedagogical tradition.

Part III

Source categories and familiarity

Chapter 6

Acoustic and categorical dissimilarity of musical timbre

Part III considers the ways in which timbre familiarity and prior knowledge of instrument categories affect timbre cognition. The current chapter takes a dedicatedly "cognitive" view on timbre dissimilarity and argues that long-term familiarity and knowledge about instrument categories affect even such a supposedly low-level task as dissimilarity rating. The main pieces of evidence come from rating asymmetries and a regression model. In that sense, the chapter advocates for a more comprehensive view of timbre dissimilarity.

This chapter is based on the following research article:

Siedenburg, K., Jones-Mollerup, K., and McAdams, S. (under revision). Acoustic and categorical timbre similarity: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in Psychology*.

Abstract. We investigate the role of acoustic and categorical information in timbre dissimilarity ratings. Using a Gammatone-filterbank-based sound transformation, we created tones that were rated as less familiar than recorded tones from orchestral instruments and that were harder to associate with an unambiguous sound source (Exp. 1). A subset of transformed tones, a set of orchestral recordings, and a mixed set were then rated on pairwise dissimilarity (Exp. 2A). We observed that recorded instrument timbres clustered into subsets that distinguished timbres according to acoustic and categorical properties. For the subset of cross-category comparisons in the mixed set, we observed asymmetries in the distribution of ratings, as well as a stark decay of inter-rater agreement. These effects were replicated in a within-subjects design (Exp. 2B) and cannot be explained by acoustic factors alone. We finally introduced a novel model of timbre dissimilarity based on partial least-squares regression that compared the contributions of both acoustic and categorical timbre descriptors. The best model fit ($R^2 = .88$) was achieved when both types of descriptors were taken into account. This provides evidence for an interplay of acoustic and categorical information in timbre dissimilarity perception.

6.1 Introduction

Timbre is often considered as one of the "last frontiers" in auditory science. Leaving aside the general agreement that a definition by negation (ANSI, 1960/1994) is unsatisfactory (Bregman, 1990; Krumhansl, 1989; Hajda, Kendall, Carterette, & Harshberger, 1997), the notion is usually understood in a twofold manner. Timbre first denotes that auditory attribute that lends sounds a sense of "color". This quality emerges from a number of acoustic cues—perceptually integrated into a *timbral Gestalt*—the most important of which include the spectral envelope shape, attack sharpness, spectrotemporal variation or modulation, roughness, noisiness, in addition to features that may be idiosyncratic to certain instruments (McAdams, 2013). For acoustic instruments, this bundle of features usually covaries with pitch register and dynamics or playing effort (Handel & Erickson, 2001). At the same time, timbre allows for the categorization of sound sources (Pressnitzer, Agus, & Suied, 2013; McAdams, 1993) and for the inference of the mechanics of sound-producing objects and events (Giordano & McAdams, 2010). This gives rise to a cognitive representation of a sound in terms of its source-cause properties that can remain invariant across drastic changes in the acoustic signal (Handel, 1995).

Most cornerstones of the perceptual representation of musical timbre are based on dissimilarity ratings: Two tones are presented in succession per experimental trial, and

listeners rate their degree of dissimilarity, such that the task does not require any verbal labeling of sounds. Starting with the early work of Plomp (1970), Wessel (1973), and Grey (1975), multidimensional scaling (MDS, see Kruskal, 1964; Winsberg & De Soete, 1993) has been the most important tool for the analysis of the resulting dissimilarity data. Its basic idea is to yield a spatial configuration of the rated stimuli, the *timbre space*, in which spatial distance corresponds to rated dissimilarity. The space is spanned by the rating data's latent dimensions which can be interpreted psychophysically by correlation with continuous acoustic descriptors. For example, McAdams et al. (1995) presented a three-dimensional solution including values for dimensions or features specific to each sound, as well as weights on shared dimensions and specificities for latent classes of subjects. The first spatial dimension correlated with (log-) attack time (AT), the second with the spectral center of gravity (SCG), the third with spectral variation over time ("spectral flux"). SCG and AT have been confirmed to be perceptually salient in a number of studies (Lakatos, 2000; Halpern et al., 2004; Caclin et al., 2005). Recently, Elliott et al. (2013) used high-dimensional modulation spectra that represent a signal's joint spectro-temporal variability, followed by methods of dimensionality reduction in order to provide an acoustic basis for the five-dimensional MDS space they had obtained. They observed that the approach has similar predictive power compared to an acoustic description based on scalar audio descriptors (including measures such as spectral and temporal center of gravity).

Two implicit assumptions of this approach deserve further notice. First, dissimilarity ratings are symmetric (none of the mentioned studies tested this empirically). Second, dissimilarity ratings are based on the sounds' acoustic properties and are not related to source categories or semantic associations (none of the above-mentioned studies reported having specifically instructed participants to rate acoustic quality). The goal of the present paper is to demonstrate that there are cases under which these subtle but important assumptions can fail.

In order to provide some background, we first review previous work on timbre similarity and categorization. We then outline a related controversy on the continuous or categorical nature of psychological similarity, exploring more deeply the conditions under which asymmetric similarities are likely to occur. Note that this work concerns the role of familiar instrument categories in dissimilarity ratings; we will not address *categorical perception* of timbre in the sense of differential inter- and intra-category

discriminability of stimuli (see, Donnadieu, 2008).

6.1.1 Sound source categories and similarity

Regarding the inference of material and excitation properties, listeners have been shown to reliably infer geometry and material properties such as damping of sounding objects (McAdams, Chaigne, & Roussarie, 2004; Giordano & McAdams, 2006; McAdams, Roussarie, Chaigne, & Giordano, 2010; Giordano, Rocchesso, & McAdams, 2010; Giordano & McAdams, 2010). However, acoustic cues used for dissimilarity rating and categorization partially differed (McAdams et al., 2010). Research with vocal and instrumental timbres has demonstrated that neither solely spectral, nor solely temporal cues are sufficient to account for timbre categorization (Agus et al., 2012). Curiously, Suied et al. (2014) highlighted in a subsequent study that acoustic cues for timbre categorization may reside on very short time-scales, i.e., likely in the spectral domain. Using gated vocal and instrumental sounds, listeners could reliably categorize sounds of gate durations as short as around 8 ms. Taken together, these diverse findings suggest that the perceptual system might exploit sensory cues in an opportunistic fashion. Rather than always using the same fixed set of acoustic cues, only the most informative cues are employed with respect to the scenario of a particular perceptual task (also see McAdams et al., 2010; Suied et al., 2014).

Coming back to the similarity rating task, Lakatos (2000) used a set of harmonic instrumental sounds, percussive sounds, and a mixed set to explore MDS and clustering solutions of dissimilarity ratings. As acoustic complexity of sounds increased, in particular for the set of percussive sounds, listeners' responses were interpreted to rely more on categorical representations. Accordingly, Lemaitre, Houix, Misdariis, and Susini (2010) proposed to distinguish between acoustical sound similarity (cognitively represented by auditory sensory representations), causal similarity (represented via the shared and distinct features of the perceptually inferred source-cause mechanisms), and semantic similarity (related to associated meaning or knowledge about the underlying sound event; see Slevc & Patel, 2011, for a more general discussion of semantics in music). Halpern et al. (2004) compared musicians' dissimilarity of heard and imagined musical instrument tones while recording functional magnetic resonance imaging (fMRI). Both conditions presented instrument names visually, and the "heard"

condition also presented the instrument's sound. Auditory cortex was active during perception and imagery and behavioral ratings of perceived and imagined dissimilarity correlated significantly (r = .84). Note that the fMRI data are the only suggestive piece of evidence that there was indeed sensory imagery for timbre, as the correlation could well have been explained by participants comparing non-auditory features of instruments in both tasks, i.e., relying on causal or semantic similarity. In a similar vein, Iverson and Krumhansl (1993) had already found similar MDS solutions for sets of orchestral instrument sounds for which either full tones, only attack portions (80ms) or only remainders were presented. Giordano and McAdams (2010) presented a meta-analysis of studies on instrument identification and dissimilarity judgments. Instruments were more often confused in identification and rated as more similar when they were members of the same family or were generated by the same manner of excitation (impulsive, sustained), underlining the strong correspondence between continuous sensory and categorical types of timbre similarity.

There still remains the question of whether these links between acoustics and source category are of an intrinsic correlational nature, based on the partial coincidence of acoustic similarity and categories of source mechanics (instruments that feature similar source mechanics will likely feature similar acoustic qualities), or because listeners give significant weight to the causal similarity of stimuli. Most timbre dissimilarity studies have used tones from western orchestral instruments or their synthetic emulations. These are stimuli with which western listeners, whether musicians or non-musicians, inevitably have a lifelong listening experience, and thus can be assumed to possess long-term mental categories (cf., Agus et al., 2010). For unaltered instrumental tones, it thus seems hard to experimentally disentangle acoustic and categorical factors. An example of such a dissociation was nonetheless given by Giordano, McDonnell, and McAdams (2010), albeit not for timbre specifically. These authors outlined how processing strategies may differ across sound categories: sounds from non-living objects are sorted mainly based on acoustic criteria, whereas the evaluation of living sounds is biased towards semantic information that is partially independent of acoustic cues. The interplay of affordances for source identification and listening experience was further studied by Lemaitre et al. (2010). They observed that sounds with low *causal uncertainty* (measuring the amount of reported alternative causes for a sound) tended to be classified on the basis of their *causal similarities* (i.e., based on source-cause prop-

erties), whereas sounds with high causal uncertainty were rather grouped on the basis of acoustic cues. Moreover, so-called "expert listeners" (i.e., musicians, sound artists, sound engineers, etc.) tended to rely more heavily on acoustic cues than non-experts when categorizing sounds with low causal uncertainty.

6.1.2 Similarity and categorization

The previous observations on timbre are surrounded by a long-lasting debate on the nature of perceptual dissimilarity. One basic question is whether similarity is best described by continuous multidimensional spaces or via set-theoretic models based on categorical stimulus features (cf. Tversky, 1977; Shepard, 1987; Ashby, 1992; Tenenbaum, Griffiths, et al., 2001; Goldstone, de Leeuw, & Landy, 2015).

Classic work in cognitive psychology shows that for complex, semantically loaded stimuli, geometric reasoning about psychological similarity may be inadequate. In a pioneering paper, Rosch (1975) presented asymmetric data of psychological similarity. Subsequently, Tversky (1977) developed a similarity model based on categorical *features*, binary attributes that a stimulus may or may not possess. Tversky also attacked the symmetry assumption: "Similarity judgments can be regarded as extensions of similarity statements, that is, statements of the form 'a is like b'. Such a statement is directional [...]. We tend to select the more salient stimulus, or the prototype, as a referent, and the less salient stimulus, or the variant, as a subject. [...] We say 'North Korea is like Red China' rather than 'Red China is like North Korea'." Tversky (1977, p. 328) He provided a variety of asymmetric empirical data in which the similarity of a prototypical stimulus to a variant was smaller than the reverse.

Shepard (1987) commented that the observed problems of spatial models might only concern stimuli with highly separable perceptual dimensions that do not interfere with each other in perceptual processing. Nonetheless, the results by Melara, Marks, and Lesko (1992) seem to render this hypothesis unlikely. Their subjects rated the pairwise similarity of sets of stimuli with varying separability of perceptual dimensions. A perceptually separable audio-visual condition presented stimuli varying in pitch accompanied by visually presented crosses with varying positions. A perceptually integral condition presented auditory stimuli varying in pitch and loudness. For both conditions, a first group of subjects was instructed to judge similarity on the

basis of the overall *Gestalt*, another to attend to all perceptual dimensions separately. Data from the latter group were best fitted by a cityblock metric (additive sum of the individual dimensions), whereas dissimilarities from the group that applied a holistic strategy were best approximated by a Euclidean metric (a nonlinear combination of the dimensions). The malleability of ratings, easily modified by instructions, therefore led the authors to conclude that direct similarity ratings involve an interplay of *optional* and *mandatory* perceptual processes. Mandatory processes refer to hard-wired perceptual processes where the weighting of stimulus dimensions is thought not to be under direct control of subjects (Shepard, 1987). Optional processes were interpreted to give subjects a choice of what stimulus facets to attend to and rate. Importantly, Melara et al. (1992) observed both kinds of processes for all stimulus sets they tested, even those classically considered as integral.

Perceptual dimensions of timbre have been described as interactive (Caclin, Giard, Smith, & McAdams, 2007). The above considerations thus suggest that optional processes are likely to be at play, particularly so if sounds can be easily identified or possess heterogeneous semantic affordances. On the other hand, if participants exclusively relied on a stimulus's sensory representation, rating asymmetries should not occur.

6.1.3 The present study

For circumventing the co-occurrence of acoustic similarity and source categories, we chose to compare musicians' dissimilarity ratings of familiar acoustic and unfamiliar synthetic tones specifically generated for the study. We first created timbral transformations that partially preserved the acoustic properties of a set of recorded orchestral instruments (similar to Z. M. Smith, Delgutte, & Oxenham, 2002) and let musicians identify and rate the subjective familiarity of the sounds (Exp. 1). The 14 transformations rated as most unfamiliar were then selected for comparison with the 14 recorded acoustic instrumental tones. In Exp. 2A, we then collected dissimilarity ratings for the set of recorded tones, transformed tones, and a mixed set (methodically similar to Lakatos, 2000). We were interested in observing the relation of instrument categories and acoustic similarity in the clustering of the dissimilarity data, as well as potential category-based asymmetries in dissimilarity ratings. We hypothesized that if asymme-

tries would occur, they would most likely be found between recorded acoustic tones and synthetic transformations, i.e., in the mixed set. Such mixed pairs feature a particularly strong categorical dissimilarity, because one sound is acoustic and the other synthetic, and because there is a gap in familiarity between these two classes of sounds (experimentally controlled by virtue of Exp. 1). We finally conducted an exploratory regression analysis that enabled us to trace out the role of categorical factors for the set of recordings and the mixed set in more detail.

6.2 Experiment 1: Identification and familiarity

This experiment was conducted in order to provide a basis for the selection of unfamiliar stimuli without readily available source-cause associations for Exp. 2.

6.2.1 Method

Participants

There were 15 participants (nine male, six female) with ages between 18 and 36 ($M = 22.2$, $SD = 4.6$). They had a mean of 9.4 years of musical instruction ($SD = 3.5$) and a mean of 5 years experience playing in ensembles ($SD = 2.9$). Two reported possessing absolute pitch. Participants were compensated for their time.

Stimuli and presentation

Stimuli consisted of 14 recordings of single tones from common musical instruments and 70 tones that were derived by digital transformation of the 14 acoustic tones. The recorded timbres consisted of the bass clarinet (BCL), bassoon (BSN), flute (FLT), harpsichord (HCD), horn (HRN), harp (HRP), marimba (MBA), piano (PNO), trumpet (TRP), bowed violoncello (VCE), violoncello pizzicato (VCP), vibraphone (VIB), bowed violin (VLI), and violin pizzicato (VLP), all played at mezzo-forte without vibrato. Piano and harpsichord samples were taken from Logic Professional 7; all other samples came from the Vienna Symphonic Library (http://vsl.co.at, last accessed April 12, 2014). The audio sampling rate used throughout this study was 44.1 kHz. Sounds had a fundamental frequency of 311 Hz (E♭4), and only left channels were used. According to the VSL, the samples were played as 8th-notes at 120 beats per minute,

i.e. of 250 ms "musical duration". However, actual durations varied and were slightly longer than 500 ms for all sounds, such that we used barely noticeable fade-outs of 20 ms duration (raised-cosine windows), in order to obtain stimuli of uniform duration (500 ms). Peak amplitude was normalized across all sounds. This set of 14 timbres is hereafter referred to as "recordings".

A second set of timbres was generated digitally. The goal was to obtain stimuli for which associations of an underlying source were not readily available and that possessed a reduced degree of perceptual familiarity. At the same time, these stimuli should not differ too strongly in their overall acoustic variability compared to the set of original recordings. We thus decided to digitally transform the spectro-temporal envelopes and acoustic fine structures of the recordings, a procedure that was demonstrated to yield altered ("chimæric") perceptual properties for speech signals (Z. M. Smith et al., 2002). Any novel sound was derived from a source signal ("chimæra-source" or "c-source"), the spectrotemporal fine structure of which was amplitude modulated by the spectrotemporal envelope of a second signal that acted as a time-varying filter ("c-filter"). These abbreviations will be used in the rest of this paper in order not to confuse this specific approach with the general technique of source-filter synthesis. More specifically, chimaeras were generated in MATLAB version R2013a (The MathWorks, Inc., Natick, MA). Sound signals were decomposed by a linear 24-band Gammatone-filterbank (Patterson et al., 1992) as implemented in the MIRtoolbox (Lartillot & Toiviainen, 2007). Amplitude-envelopes were extracted for every filterband of both c-sources and c-filters, using low-pass filtering and half-wave rectification (Lartillot & Toiviainen, 2007). For every band, the c-filter's envelope values were then imposed onto the c-source by normalizing the c-source's filterband envelopes, followed by point-wise multiplication with the c-filter's time-varying envelope magnitudes. The resulting signal hence possessed the spectrotemporal envelope of the c-filter and the fine structure of the c-source (cf. Z. M. Smith et al., 2002).

We chose to use three different types of sounds to act both as c-sources and c-filters. The first type consisted of the fourteen recordings mentioned above. Sounds of the second type (conceived to further decrease perceptual familiarity) were generated in four steps: We i) decomposed the acoustic sounds into twenty-four Gammatone-filterbands, ii) randomly selected four sounds from the fourteen, iii) allocated their filterbands such that each of the four sounds contributed to the new sound with six

different bands, and iv) added all twenty-four distinct bands. This process is called "filterband scrambling" (FBS) hereafter. Six such sounds were selected, denoted as FBS 1–6 below. Among these, FBS 1&2 possessed a slow attack, FBS 3&4 a sharp attack, and FBS 5&6 attacks in between the two extremes. The third type of sounds simply consisted of a zero-phase harmonic tone complex with a fundamental frequency of 311 Hz. Note that on their own, type one should be highly familiar to participants, and type two should be less familiar. Despite its artificiality, the harmonic tone complex may be familiar due to its status in electronic music. If taken as c-filter, the harmonic tone complex has a neutral effect due to its flat spectral envelope, i.e., coincides with no envelope filtering at all. Using sounds of type one as c-filter should affect familiarity of recordings acting as c-sources, as spectrotemporal envelope properties are substantially altered. This provided 21 (14+6+1) distinct sounds in total. Any possible combination of c-sources and c-filters was then used to generate 441 (21×21) chimaeric signals, 70 of which were pre-selected manually for the experiment. The selection was subject to the constraint that every c-source and c-filter signal was required to be selected at least once; for recordings acting as c-filters, each c-filter was selected at least twice. Additionally, the selection favored timbres that seemed unfamiliar to the experimenters, but did not contain too much narrowband noise (an artifact that was introduced in some transformations by boosting the amplitude of filterbands with low energy). All sounds were normalized in peak amplitude.

Procedure

The research reported in this manuscript was carried out according to the principles expressed in the Declaration of Helsinki and the Research Ethics Board II of McGill University has reviewed and approved this study (certificate # 67-0905).

Participants first completed a standard pure-tone audiogram to ensure normal hearing with hearing thresholds of 20 dB HL or better with octave spacing in the range of 250–8000 Hz (ISO 398-8, 2004; Martin & Champlin, 2000). In every trial of the experiment, a single stimulus from the 70 transformations and 14 recordings was presented to participants. They were asked to choose an identifier from a list of eight possible options. The list consisted of six musical instrument names. For recorded timbres, it contained the correct label and five randomly chosen labels from the remaining set.

For transformations, it involved the two labels of the timbres that had been involved as c-source and c-filter, plus four labels chosen randomly from the remaining set. For instance, if a transformation was derived from a piano as a c-source, whose time-varying spectral envelope was exchanged with that of a violin, then both instrument names, piano and violin, would be part of the list. The list further contained the two options "unidentifiable" and "identifiable but not contained in list". If the participant selected the latter option, a dialogue box appeared prompting them to enter an appropriate identifier in the text box on screen. They could then continue, whereupon they heard the sound a second time and were presented with two analog-categorical scales on which they had to rate familiarity (1-highly unfamiliar, 5-highly familiar) and artificiality (1-very natural, 5-very artificial). Sounds were presented in randomized order. Three example trials preceded the 84 experimental trials. The full experiment took around 45 minutes.

Experiments took place in a double-walled sound-isolation chamber (IAC Acoustics, Bronx, NY). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany), using a Macintosh computer with digital-to-analog conversion on a Grace Design m904 (Grace Digital Audio, San Diego, CA) monitor system. The experimental interface and data collection were programmed in the Max/MSP audio software environment (Cycling 74, San Francisco, CA). The average presentation level was 78 dB SPL (range=75–82 dB SPL) as measured with a Brüel & Kjær Type 2205 sound-level meter (A-weighting) with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark).

6.2.2 Results

By construction, correct responses for the identification task only existed for the recordings. Here, correct identification rates ranged from .46 (BCL and BSN) to 1.0 (TRP). The mean identification rate for all 14 recordings was .73 ($SD = 0.180$) with chance baseline equal to $1/8 = .125$. The bass-clarinet (BCL) was the only recording for which an alternative category, "unidentifiable", was selected most often (.53).

From the remaining 70 transformations, 29 were most often identified as other musical instruments (i.e., the category that was selected by the majority of subjects) with

average selection rates of .47 ($SD = .12$). From these 29 transformations, the category chosen most often for 23 sounds was an instrument that acted either as c-source or c-filter in its generation. Only six transformations failed to be related to their source or filter by a majority of participants; these were the transformations BCL-VLP ($\rightarrow$ heard as VIB; BCL denoting c-source, VLP c-filter), VIB-BCL ($\rightarrow$ MBA), VLI-BSN ($\rightarrow$ HRN), VCP-FBS ($\rightarrow$ MBA), FBS-FBS ($\rightarrow$ VIB) and FBS ($\rightarrow$ MBA). Thirteen transformations were most often selected as "unidentifiable" with stimulus-wise mean selection rates of .55 ($SD = .21$). Twenty-eight transformations were selected as "identifiable, but not in the list" with mean selection rates of .55 ($SD = .16$). If subjects had selected the latter category, they were asked to briefly describe what they had heard in a written response. Three different types of responses appeared most often here: 41% of these responses mentioned single orchestral instruments; 37% mentioned a mix of multiple instruments (e.g. "piano and trombone in unison"); 16% mentioned electronic means of sound synthesis; 6% were hard to categorize (e.g., participant 7: "Ahh yes patch 87: plucking a frog.").

Pearson correlations between the proportion of "unidentifiable" votes per stimulus and mean familiarity ratings were strong and negatively associated, $r(82) = -.88, p < .001$, as was the correlation between familiarity and artificiality, $r(82) = -0.86, p < .001$. The harmonic tone complex without filtering obtained maximal artificiality ratings ($M = 4.95$, $SD = 0.10$) and medium familiarity ($M = 3.37$, $SD = 1.34$) and was an obvious outlier in the latter correlation; removing this datum increased the correlation to $r(81) = -0.89$.

Mean familiarity ratings as a function of c-source and c-filter are displayed in Figure 6.1. Given that the pre-selection of stimuli attempted to select unfamiliar timbres, a causal interpretation of effects of c-source and c-filter on familiarity would not be appropriate. It should be remarked, however, that the highest familiarity ratings were as expected obtained by the non-filtered recordings. At the same time, the filterbank scrambled signals (FBS) acting as c-filters achieved the lowest average familiarity ratings for all c-sources.

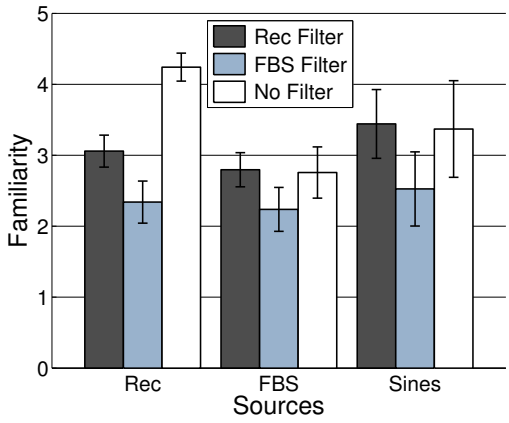

Fig. 6.1 Experiment 1: Mean familiarity of signals generated by nine different combinations of c-source (x-axis) and c-filter (color-coded), see text for a description of c-source and c-filter. Error bars represent 95% confidence intervals.

6.2.3 Discussion

The identification scores for the 14 recorded timbres yielded correct choices by the majority of subjects for all instruments except the bass-clarinet, for which "unidentifiable" took the lead. Apart from this one exception which also possessed the lowest familiarity ratings among the 14 unaltered timbres (familiarity and rates of "unidentifiable" choices were strongly correlated), results indicated that musicians were able to identify acoustic timbres of less than 500 ms duration from a single presentation. Yet, the current data exhibit considerable variance in the percentage of correct identifications across different instruments (ranging from 46% to 100%), a finding that parallels the divergent estimates of identification accuracy in the literature (Srinivasan, Sullivan, & Fujinaga, 2002). From the 70 transformations, the majority vote identified 29 as alternative instruments that were provided in the list of options. Among these 29, around 80% were correctly identified as instruments that had either acted as c-source or c-filter in the synthesis process. This underlines musicians' abilities to identify sound source properties and mechanics (Giordano & McAdams, 2010), even in situations where these are severely altered.

Familiarity ratings and the proportion of "unidentifiable" votes were strongly correlated. Familiarity and artificiality ratings shared around 77% of mutual variance if one outlier was removed. The most likely factor that may have caused this strong correla-

tion could be the digital transformation used in the production of stimuli. The more impact the transformation had on the original signal structure, the less familiar the resulting timbres appeared to be. However, the plain harmonic tone complex received the highest artificiality ratings, while far from being rated as least familiar. The fact that this signal did not follow the overall trend suggests that not any digitally synthesized tone obtains low familiarity ratings, which justifies our usage of a somewhat elaborate signal transformation. Not surprisingly, the highest familiarity ratings were obtained for the unaltered recordings.

6.3 Experiment 2: Timbre dissimilarity of acoustic recordings and synthetic transformations

Exp. 1 suggested that overtly simple means of sound synthesis may fail to create tones that are unfamiliar to musicians, but confirmed that familiarity and source identifiability were highly related in the presented set of recordings and transformations. In order to study the role of sound categories and familiarity in dissimilarity perception, we selected 14 transformations rated as least familiar in Exp. 1 and used them together with a set of recordings in a dissimilarity rating task for musicians. Due to the strong correlation of familiarity and identifiability, the selected transformations consequently only scarcely afforded unambiguous identification of source-cause categories. We were interested in the ways in which the dissimilarity structures would be affected by categorical properties of tones, such as instrument families within the set of recordings, and whether asymmetries would occur between synthetic and acoustic tones.

Specifically, we collected dissimilarity ratings for the set of recordings (Set 1), transformations (Set 2) and a mixed set (Set 3). In Exp. 2A, the order of the presentation of tones within a pair was counterbalanced across (musician) participants. Using a within-subjects design, Exp. 2B was conducted in order to confirm rating asymmetries in Set 3 from Exp. 2A; a new group of musicians rated both orders of presentations of only the mixed set of tones (Set 3). For the sake of brevity, methods and results of both experiments will be described in the same section below.

6.3.1 Method

Participants

Experiment 2A Twenty-four musicians (11 male, 13 female) with ages between 18 and 36 years (mean age=24.1, $SD = 5.3$) took part. Participants had a mean of 12.8 years of musical instruction ($SD = 6.4$) and a mean of 7.3 years experience playing in ensembles ($SD = 4.6$). One participant reported possessing absolute pitch.

Experiment 2B Twenty-four musicians (10 male, 14 female) with ages between 18 and 28 years (mean age=22.5, $SD = 2.7$) participated. They had a mean of 11.1 years of musical instruction ($SD = 3.7$) and a mean of 6.3 years experience playing in ensembles ($SD = 3.6$). Seven participants reported possessing absolute pitch. All participants (Exps. 2A and 2B) were compensated for their time.

Stimuli and presentation

Experiment 2A In every trial, pairs of timbres of 500-ms duration each were presented with a 300-ms inter-stimulus interval. Stimuli consisted of the 14 acoustic recordings (Set 1) and 14 transformed sounds (Set 2) that had obtained the lowest familiarity ratings in Exp. 1. A mixed set contained the seven most familiar recordings and the seven least familiar transformations (Set 3). All stimuli had a 311 Hz fundamental frequency. Tables 7.2 list all stimulus names, labels, and their mean familiarity ratings from Exp. 1 for recorded and transformed stimuli, respectively. Stimuli included in Set 3 are indicated with asterisks.

Six expert listeners equalized the perceived loudness of sounds against a reference sound (marimba), using a protocol designed in PsiExp (Smith, 1995) for the music-programming environment Pure Data (http://puredata.info, last accessed April 12, 2014). Stimuli were presented through a Grace m904 amplifier, and listeners used a slider on the computer screen to adjust the amplitude-multiplier of the test sound until it matched the loudness of the reference sound. Loudness was then normalized on the basis of the median loudness adjustments. Both for loudness equalization and the main experiment, the same apparatus was used as in Exp. 1. The average presentation level after loudness normalization was 66 dB SPL (range=58–71 dB SPL).

Table 6.1 List of recordings and transformations used in Exps. 2A and 2B with mean familiarity ratings (Fam.). Labels with asterisks (*) indicate timbres that were also used in Set 3.

	Set 1 (Recordings)			Set 2 (Transformations)			
#	Instrument	Label	Fam.	Source	Filter	Label	Fam.
1	Bass Clarinet	BCL*	4.3	Bass Clarinet	FBS2	BCL-FBS2*	1.6
2	Bassoon	BSN	3.1	Bassoon	Harpsichord	BSN-HRP*	1.9
3	Flute	FLT	4.1	FBS1	Violoncello	FBS1-VCE*	1.8
4	Harpsichord	HCD*	4.5	FBS2	Violoncello	FBS2-VCE	2.1
5	Horn	HRN	4.2	FBS3	FBS2	FBS3-FBS2	2.1
6	Harp	HRP	4.1	FBS6	Trumpet	FBS6-TRP*	1.9
7	Marimba	MBA*	4.6	Flute	FBS1	FLT-FBS1	2.1
8	Piano	PNO	4.3	Harp	FBS3	HRP-FBS3*	1.7
9	Trumpet	TRP*	4.8	Harpsichord	FBS4	HRP-FBS4	2.3
10	Violoncello	VCE*	4.7	Horn	FBS6	HRN-FBS6*	2.0
11	Violonc. Pizz.	VCP*	4.5	Marimba	Harpsichord	MBA-HRP	2.0
12	Vibraphone	VIB	4.3	Trumpet	FBS5	TRP-FBS5	2.3
13	Violin	VLI	3.4	Violin	Piano	VLP-PNO	2.4
14	Violin Pizz.	VLP*	4.4	Violoncello	Vibraphone	VCE-VBS*	2.0

Experiment 2B Only the mixed set (Set 3) was used. Otherwise, stimuli were identical to Exp. 2A.

Procedure

Experiment 2A Normal hearing was ensured as in Exp. 1. Subjects were asked to rate the dissimilarity of two successively presented sounds on an analog-categorical scale (a continuous rating scale with marks between 1-identical and 9-very dissimilar at the extremes) by answering the question "How dissimilar are these two sounds?" They were able to hear the pair as many times as desired by pressing a play button, but were encouraged to move at a reasonable pace. Four example trials were given. Before the start of each experimental session, participants heard all sounds from the respective set in random order. The overall experiment consisted of one session per set. The mixed set came last for all participants. For all three sets, each pair was presented once in one order. The order of presentation (AB vs. BA for timbres A and B) was counterbalanced across subjects. Pairs of identical timbres were included, yielding 105 comparisons per set. There was a ten-minute break between each set. The

full experiment took 1.5–2 hours to complete.

Experiment 2B In contrast to Exp. 2A, the full 14x14 matrix of pairwise comparisons including both orders of pairs was presented to every subject. This was administered in a single session with 196 trials in fully randomized order, lasting on average less than 40 minutes.

6.3.2 Results

Average dissimilarity ratings for Sets 1–3 from Exps. 2A and 2B are displayed in Figure 6.2. In Exp. 2A, mean ratings were $M = 5.5$ ($SD = 1.9$) for Set 1, $M = 4.9$ ($SD = 1.7$) for Set 2, and $M = 5.4$ ($SD = 1.7$) for Set 3. Mean ratings in Exp. 2B for Set 3 were $M = 5.7$ ($SD = 1.7$). Ratings for Set 3 from Exp. 2A and 2B were highly correlated, $r(194) = .94, p < .001$.

Dissimilarity clusters Hierarchical cluster analyses were computed on the basis of dissimilarity data averaged over the directionality of the comparison (symmetry being a condition of the clustering algorithm). This approach admittedly can only serve as a rough approximation for the subset of recordings-transformations from Set 3, as indicated by the analyses on asymmetries below. Figure 6.3 shows the corresponding clustering trees, using the complete-linkage method. The latter is based on a function that iteratively computes the distance of the two elements (one in each cluster) that are the farthest away from each other. Thresholds for overall grouping (indicated by color-coding in the figure) was 70% of maximal linkage (the default value of the Matlab dendrogram.m function that was used). Sets 1, 2, and 3 yielded 4, 3, and 5 clusters, respectively. The cophenetic correlation coefficients (the linear correlations between the tree solutions and the original dissimilarities) were .80, .86, and .65 for Sets 1–3, respectively, indicating the worst fit for Set 3.

More specifically, the clustering solution for Set 1 partially corresponded to the well-known families of musical instruments: wind instruments clustered together (turquoise), similarly to bowed string instruments (VLI and VCE, violet). The top cluster (green) corresponds to impulsively excited instruments, and there is one cluster with two very bright and impulsive instruments (VIB and HCD, red).

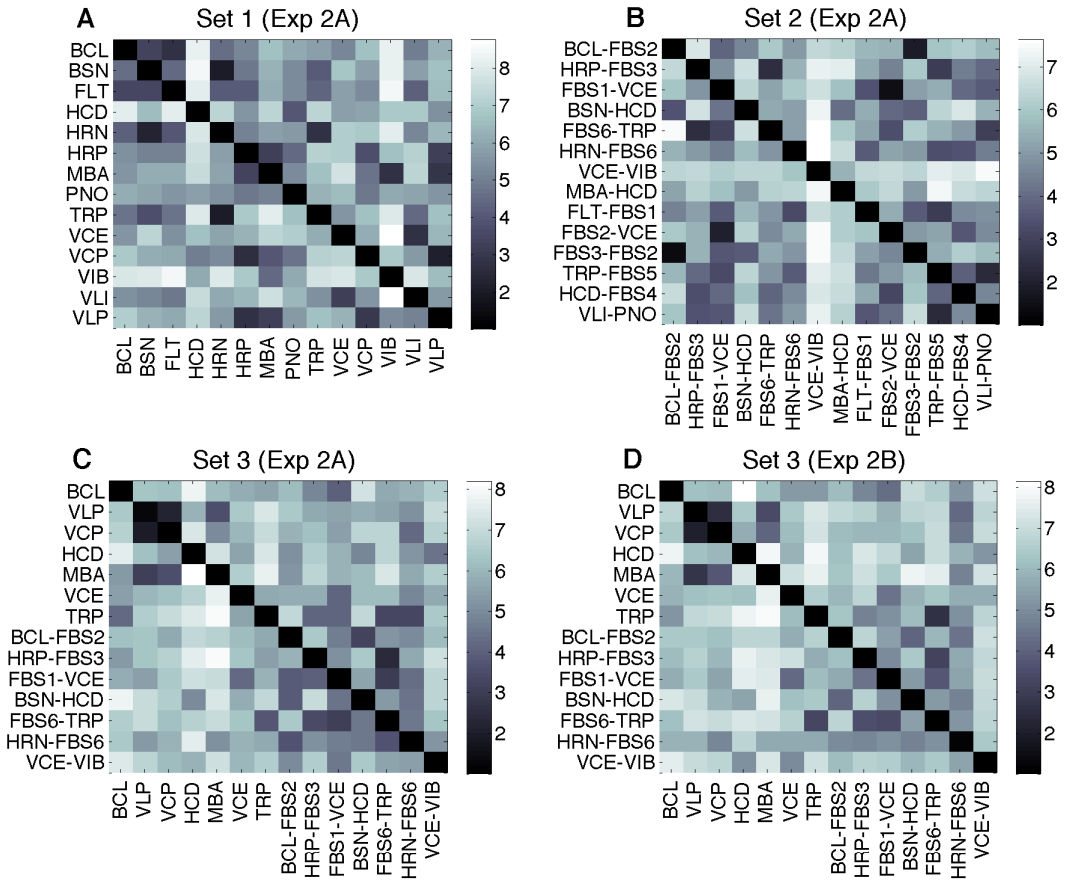

Fig. 6.2 Mean dissimilarity ratings for Exp. 2A, Set 1 (A), Set 2 (B), Set 3 (C), and Exp. 2B, Set 3 (D). Rows determine the first stimulus, columns the second.

The tree for Set 2 is harder to interpret, due to the lack of definite source categories. Here, only three clusters emerged, one of which contained nine of the 14 timbres. It is further to be noted that the identity of the timbres corresponding to c-sources or c-filters did not seem to play out as a definite predictor for clustering. For instance, although the timbres MBA-HCD and BSN-HCD contain the same c-filter, they turned out to be maximally far apart in the tree. On the other hand, the timbres FBS1-VCE and FBS2-VCE were very close in the tree.

Set 3 yielded a solution with five clusters, from which two were mixed clusters (containing both recordings and transformations), two contained recordings only, and one contained only transformations. From bottom to top, the first cluster (in violet) retained impulsively excited timbres from their cluster in Set 1. The second cluster

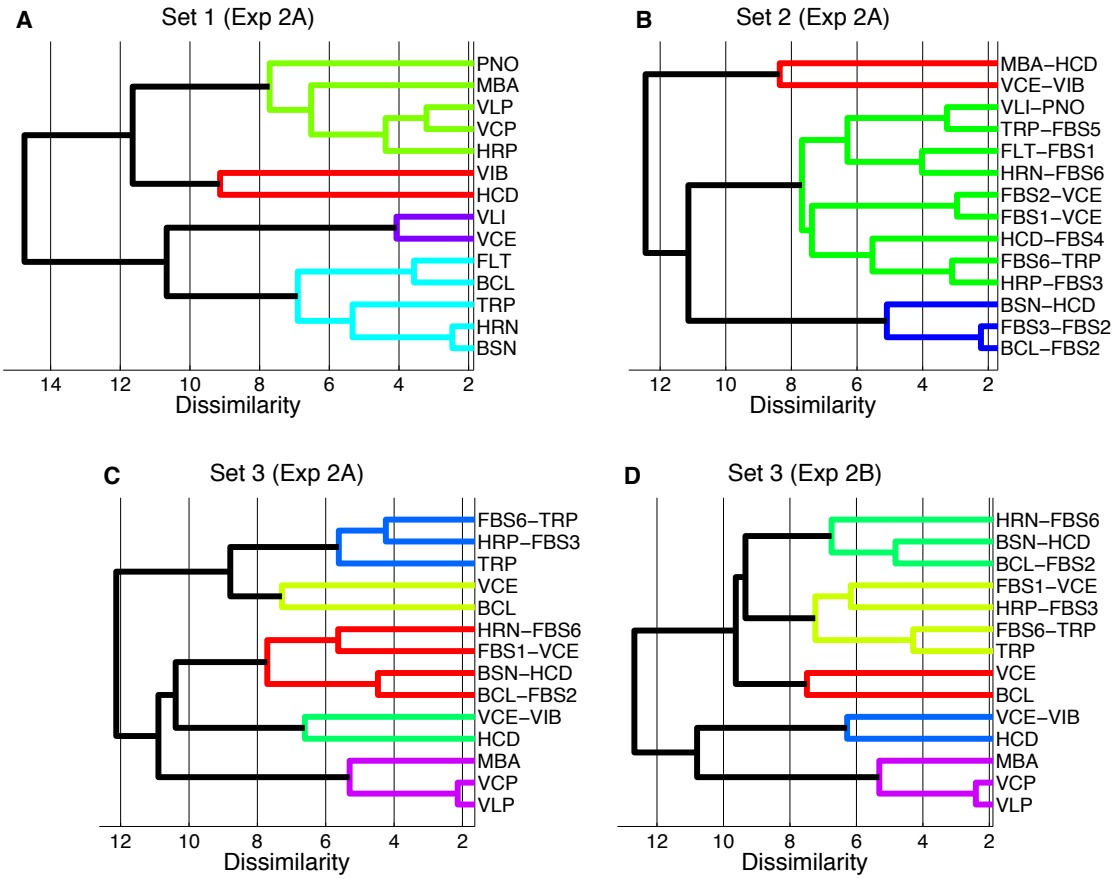

Fig. 6.3 Hierarchical clustering of mean dissimilarity ratings from Exp. 2A, Sets 1 (A), Set 2 (B), and Set 3 (C), as well as Exp. 2B, Set 3 (D), using the complete-linkage method. Color-coded groups are specified by a 70% linkage cutoff.

(green) joined the brightest recordings (HCD) and transformations (VCE-VIB). The largest cluster of this set (red) contained four transformations, two each stemming from relatively close clusters in Set 2. The cluster consisting only of VCE and BCL (bright green) again joined relatively proximal timbres from Set 1. Finally, the last cluster (blue) connected two very similar timbres from Set 2 with a single recording (TRP). The clusters obtained from Exp. 2B (Set 3) were identical apart from the timbre FBS1-VCE.

Asymmetry Difference matrices for dissimilarity ratings from Exp. 2A were obtained by excluding identical pairs, i.e., the comparisons (A,A), (B,B), etc., and sub-

tracting mean dissimilarity ratings for pairs with reversed order, i.e., dissim(A,B) – dissim(B,A). Specifically, we subtracted the upper from the lower triangular entries of the initial dissimilarity matrices. The values of the resulting triangular difference matrices should be centered at zero, if dissimilarity ratings were symmetric. Shapiro-Wilk tests did not indicate deviations from normality for any of these four difference matrices, all $p > .49$. Set 3 contained three types of pairs that were analyzed in their own right: recordings-recordings (RR), transformations-transformations (TT), and recordings-transformations (RT).

Figure 6.4 (panel A) depicts means and confidence intervals of the corresponding differences data (lower minus upper triangular matrices). The positive mean for the subset of mixed pairs from Set 3 ("S3-RT") indicates that dissimilarity ratings tended to be greater for transformations followed by recordings (lower triangular matrix) than for recordings followed by transformations (upper triangular matrix). No other (sub)set featured such an asymmetry: after correction for multiple comparisons (using the Bonferroni method, n = 6 comparisons, i.e., $\alpha_{crit} = .0083$), two-sided single-sample t-tests against a mean of zero for difference matrices yielded non-significant results for all sets apart from the subset of mixed (RT) pairs (Set 1: $t(90) = 1.24, p = .26$, Set 2: $t(90) = -0.18, p = .85$, Set 3: $t(90) = 2.12, p = .037$, Set 3-RR: $t(20) = -0.87, p = .49$, Set 3-TT: $t(20) = -2.0, p = .04$, Set 3-RT: $t(48) = 5.3, p < .001$). This means that only the ratings of mixed pairs exhibited reliable asymmetries. This pattern of results was replicated in Exp. 2B (Set 3: $t(90) = 1.8, p = .073$, Set 3-RR: $t(20) = -0.70, p = .49$, Set 3-TT: $t(20) = -2.2, p = .08$, but Set 3-RT: $t(48) = 4.3, p < .001$).

Inter-rater agreement We assessed inter-rater agreement by calculating inter-rater correlations (IRC) for ratings from Sets 1, 2, and 3, as well as the RR, TT, and RT subsets of Set 3. For any such (sub)set of comparisons and N subjects, we obtained the IRC by computing the mean (Fisher-transformed) Pearson correlation coefficients between all N(N-1)/2 pairs of subjects. Mean (back-transformed) IRCs are displayed in Figure 6.4 (panel B) with 95% confidence intervals as obtained by bootstrapping (Efron & Tibshirani, 1994). Every bootstrap sample drew 28 comparisons with replacement (the cardinality of the smallest subsets of comparisons, RR and TT, such that comparison of IRC across different (sub-)sets is not confounded by a difference in variable size); we used 1,000 random drawings and the percentile method to obtain

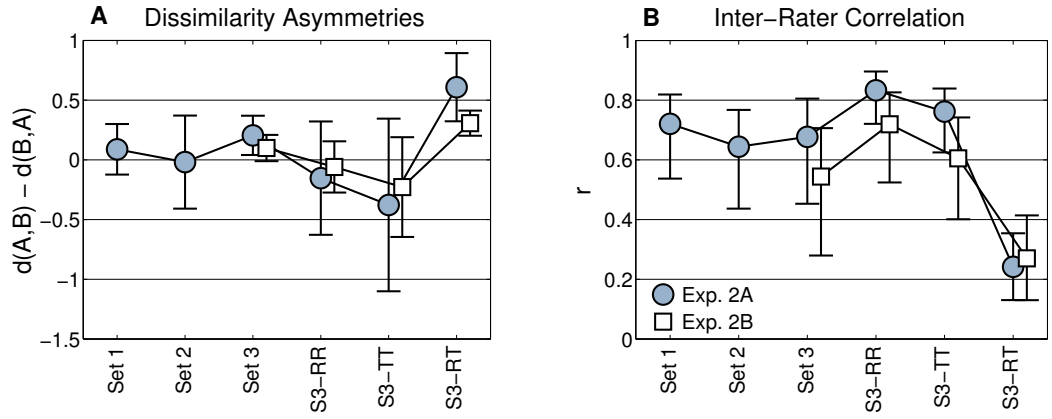

Fig. 6.4 Exps. 2A and 2B. (A) Mean rating asymmetries across the three sets, and the subsets of Set 3 with the pairs recording-recording (RR), transformation-transformation (TT), recording-transformation (RT). Errorbars indicate 95% confidence intervals. (B) Inter-rater agreement as measured by mean Pearson correlation coefficients. Errorbars indicate 95% confidence intervals obtained by bootstrapping.

confidence intervals (Efron & Tibshirani, 1994). Most obviously, mean IRCs for the first five sets are in the range of 0.6–0.8 for Exp. 2A and somewhat lower for Exp. 2B, but not significantly so. However, for the comparison of mixed pairs (RT), the IRC decreases to around .3 in both Exp. 2A and 2B. In this last subset, the IRC in Exp. 2A is significantly smaller than for any other (sub)set in Exp. 2A. In Exp. 2B, there is a significant difference between the dissimilarity ratings for RT and RR pairs.

6.3.3 Discussion

The clustering solution for Set 1 could be interpreted as featuring two distinct facets of timbre, namely instrument categories (or families) and continuous acoustic aspects such as sound brightness. Two of the four clusters were constituted by instruments with impulsive excitation, the other two subsumed continuously excited instruments. The two impulsive clusters differentiated themselves by spectral qualities rather than instrumental families, because the two very bright timbres, vibraphone and harpsichord, were part of one cluster. The two clusters of continuously excited instruments split into woodwinds and string instruments. This interpretation of the clustering solution suggests that multiple acoustic and categorical factors may affect musicians' dissimilarity

ratings of western orchestral instruments. The last section of this manuscript attempts to model these intertwining and correlated aspects in more quantitative detail. The clustering solutions of the mixed Set 3 (Exps. 2A & 2B) exhibited two clusters of recorded tones, one cluster for transformed tones, and two mixed clusters. This means that both category membership (recordings, transformations) and acoustic similarity (e.g., brightness) appeared to act as differentiating features.

Ratings in Exp. 2A were symmetric for pairs within each set of recordings or transformations. As expected, asymmetries occurred for cross-category comparisons involving recorded and transformed tones. Pairs in which the acoustic recording was followed by the synthetic transformation generally exhibited lower dissimilarity ratings than the reverse order. This effect was replicated in a within-subjects design with a different group of musicians in Exp. 2B, and no such effect occurred for any other (sub)set. This finding suggests that sound category membership may exert an effect on dissimilarity ratings, as no simple acoustic factor can plausibly account for this effect of directionality.

To our knowledge, this is the first systematic report of asymmetries in timbre dissimilarity ratings. This effect occurred although subjects were not instructed to treat one sound as a referent and one as a subject of the comparison. Neither did we implement a directed dissimilarity rating ("How different is A to B?"), but an undirected one ("How different are A and B?"). If one assumes that auditory presentation is analogous to language, i.e., places the comparison's subject before the referent, then the direction of observed asymmetries would be opposed to what was observed for the similarities of stimuli such as countries, figures, letters, morse-code signals and integers by Tversky (1977) and Rosch (1975)—we saw that the transformation-recording pairs were generally rated less similar compared to the reverse order. The only auditory stimuli discussed by Tversky (1977) were morse-code signals, where it was assumed that longer signals act as referents and where the reported asymmetries yielded higher similarity for short-long pairs than the reverse. On that basis, it was concluded that the directionality of comparisons must be identical in the auditory domain, such that the referent follows the subject. For spectrally rich timbral stimuli, the opposite could be true, however, as the presentation of a stimulus affects processing of any stimulus presented shortly after, due to automatic stimulus-specific neural adaptation as part

of sensory memory (Demany & Semal, 2007; McKeown & Wellsted, 2009). From that perspective, the second timbre is interpreted "in light of" the first, meaning the first would act as a referent. What further complicates the issue is that asymmetries only occurred systematically for cross-category comparisons, which may suggest that categorical representations independent of sensory memory are driving the effect. Note that this also leaves open the question of whether the current effect is of a perceptual nature or due to a shift in judgment strategies, commonly found in "top-down effects" (Storrs, 2015; Firestone & Scholl, 2015).

It was finally shown that inter-rater correlations (IRC) in Exps. 2A and 2B are relatively high for all pairs of timbres, except the cross-category comparisons of Set 3. This indicates that in this type of comparison, raters lost a common frame of reference. We interpret this as an index of optional processes in dissimilarity ratings (Melara et al., 1992). In the within-set comparisons of Set 1 and Set 2, comparisons may have been driven to a larger extent by sets of acoustic or categorical features similarly weighted across subjects.

Because the reduced IRCs and the rating asymmetries occurred conjointly, one may argue that one effect drove the other. It seems unlikely, though, that asymmetries were simply a coincidental artifact of a reduced IRC, given that they were reproduced in Exp. 2B in an altered design. Further research is required to better understand subjective rating behavior for timbres that have very different source origins and categorical affordances.

6.4 Dissimilarity models and analyses

The above findings on cross-category comparisons provide evidence for that categorical information may play a role in timbre dissimilarity ratings as these results seem unlikely to be explained on acoustic grounds alone. At the same time, they are based on a rather pathological comparison, namely that of familiar instrumental recordings and unfamiliar digital transformations. The question therefore becomes whether similar processes take place in the perhaps more "standard" scenario of comparing sounds from acoustic instruments. For the latter, instrument category and acoustic qualities of course coincide to a large extent (Giordano & McAdams, 2010), although not completely. Take the difference between the piano and the harpsichord or the vibraphone and marimba;

the members of both pairs may feature quite different acoustic qualities although they belong to the same instrument family: keyboard and mallet instruments, respectively. Using an exploratory regression analysis, we thus set out to quantify which types of stimulus representation, acoustic or categorical, musicians took into account in their timbre dissimilarity ratings.

In the following, we first present a latent-variable-based model of acoustic timbre dissimilarity (partial least-squares regression, PLSR), well suited to deal with collinear predictors. We then add categorical predictors to the model, which solely take into account instrument families, excitation mechanisms and types of acoustic resonators. We finally demonstrate that the highest correlations are obtained by taking into account both classes of predictors, acoustic and categorical. Note that the acoustic model will be treated in a "black-box" approach—the aim of this section is not to pin down the most parsimonious acoustic description of timbre, but for the sake of argument it must suffice to provide a robust, although potentially over-complete, acoustic model and to show that the model fit still improves with the inclusion of categorical variables.

6.4.1 Approach

We used the TimbreToolbox (Peeters et al., 2011), a large set of audio descriptors that describes the acoustic structure of audio signals with a focus on timbral qualities. We selected 34 out of its 164 descriptors, derived from measures of the temporal and spectral envelopes of the signal. The temporal envelope is computed by the Hilbert transform. Temporal descriptors focus on attack (McAdams et al., 1995) and decay properties of tones and measures of energy modulation (Elliott et al., 2013). Spectral descriptors are computed from an ERB-spaced Gammatone filterbank decomposition of the signal. They are measured for each 25-ms time frame and are summarized via the median and interquartile range as measures of central tendency and variability, respectively. Spectral descriptors include the first four moments of the spectral distribution, such as the spectral centroid that has been shown to correlate with perceived brightness (McAdams et al., 1995). Additional descriptors of the spectral distribution such as spectral slope or rolloff are included, but also measures of spectrotemporal variation, relevant to capture the perceptual dimension of *spectral flux* (McAdams et al., 1995). A full list of the descriptors is given in Table 6.2.

Table 6.2 List of acoustic descriptors from the TimbreToolbox (Peeters et al., 2011). For spectral descriptors and the RMS envelope, medians (med) and interquartile range (IQR) summarize the time-varying descriptors computed over time frames of 25 ms. Square brackets provide descriptor units (a: audio signal amplitude, F: ERB-rate units). Temporal descriptors are computed from the signal energy (temporal) envelope, spectral (and spectro-temporal) descriptors from the ERB gammatone filterbank representation.

Temporal	Spectral
1)Attack duration [s]	13) Centroid (med) [F]
2) Decay duration [s]	14) Centroid (IQR) [F]
3) Release [s]	15) Spread (med) [F]
4) Log-attack Time [log(s)]	16) Spread (IQR) [F]
5) Attack slope [a/s]	17) Skew (med) [-]
6) Decrease slope [log(a)/s]	18) Skew (IQR) [-]
7) Temporal centroid [s]	19) Kurtosis (med) -
8) Effective duration [s]	20) Kurtosis (IQR) [-]
9) Frequency of energy modulation [Hz]	21) Slope (med) [F^{-1}]
10) Amplitude of energy modulation [a]	22) Slope (IQR) [F^{-1}]
11) RMS envelope (med) [a]	23) Decrease (med) [-]
12) RMS envelope (IQR) [a]	24) Decrease (IQR) [-]
	25) Rolloff (med) [F]
	26) Rolloff (IQR) [F]
	27) Spectro-temporal variation (med) [-]
	28) Spectro-temporal variation (IQR) [-]
	29) Frame energy (med) [a^2]
	30) Frame energy (IQR) [a^2]
	31) Flatness (med) [-]
	32) Flatness (IQR) [-]
	33) Crest (med) [-]
	34) Crest (IQR) [-]

The TimbreToolbox provided the $n = 34$ scalar descriptors for all 14 sounds. In order to obtain a predictor of acoustic dissimilarity, we computed the absolute difference of descriptor values (deltas) for each pair of sounds, yielding $m = 105$ comparisons. The final design matrix X $(m \times n)$ thus concatenated descriptor deltas as column vectors. The dependent variable y $(m \times 1)$ contained the 105 mean dissimilarity ratings for the respective set (averaged over the order of presentation).

In order to handle collinearity of predictors (Peeters et al., 2011), we used partial least-squares regression (PLSR) (Geladi & Kowalski, 1986; Wold, Sjöström, & Eriksson, 2001). PLSR is a regression technique that projects the predicted and observed variables onto respective sets of latent variables, such that the sets' mutual covariance is maximized. More precisely, given a dependent variable y and an design matrix X, PLSR generates a latent decomposition such that $X = TP' + E$ and $y = Wq' + F$ with loadings matrices P $(n \times k)$ and q $(1 \times k)$, and components ("scores") T $(m \times k)$ and W $(m \times k)$ plus error terms E and F. The decomposition maximizes the covariance of T and W, which yields latent variables that are optimized to capture the linear relation between observations and predictions. For that reason, PLSR also differs from principal component analysis (PCA) followed by multivariate linear regression (MLR), which does not specifically adapt the latent decomposition to the dependent variable of interest. The regression coefficients for the original design matrix can be obtained by $\beta = W(P'W)^{-1}q$ (cf., Mehmood, Liland, Snipen, & Sæbø, 2012), which yields a link to the original MLR design via $y = X\beta + F$. In order to prevent overfitting of the response variable, the model complexity k is usually selected via cross-validation (Wold et al., 2001). Here we use PLSR as implemented in the plsregress.m function provided by MATLAB version R2013a (The MathWorks, Inc., Natick, MA), which applies the SIMPLS algorithm (De Jong, 1993). The significance of the individual coefficients β_i $(i = 1, ..., n)$ was estimated by bootstrapping 95% confidence intervals (percentile method) for the set of $\beta = (\beta_i)_i$ coefficients (Mehmood et al., 2012); if intervals overlapped with zero, a variable's contribution was considered to be not significant. All variables were z-normalized before entering the model.

Table 6.3 Variance explained (R^2) for timbre dissimilarity models and their generalization performance across sets and experiments. Models fitted to the four data sets (rows) from Exps. 2A, 2B, cross-validated on the same four sets (columns). Numbers in parentheses indicate performance of the reduced model for which non-significant coefficients (estimated by bootstrapping) were omitted.

	Data (y, X)			
Model (β)	Set 1 (E2A)	Set 2 (E2A)	Set 3 (E2A)	Set 3 (E2B)
Set 1 (E2A)	.79 (.78)	.57 (.57)	.72 (.72)	.72 (.70)
Set 2 (E2A)	.58 (.60)	.84 (.84)	.68 (.69)	.60 (.61)
Set 3 (E2A)	.74 (.73)	.67 (.66)	.83 (.84)	.81 (.81)
Set 3 (E2B)	.75 (.72)	.58 (.59)	.82 (.83)	.83 (.83)

6.4.2 Acoustic model

We opted to use a model with $k = 2$ components, which exhibited minimal 6-fold cross-validation error in the response variable compared to all other choices of k. This solution explained 47% of variance in the design matrix X. In order to first validate the general approach, we fitted models to all four dissimilarity data sets from Exps. 2A and 2B. These were tested on every other set. This evaluated the model not only on one fairly homogeneous set of sounds (as would be the case for conducting regular cross-validation on Set 1), but also allowed us to observe effects of model generalization to completely novel sets of sounds (Training: Set 1, Test: Set 3), sets in which half of the sounds are new (e.g., Train: Set 2, Test: Set 3), as well as same sets of sound but with the dependent variable stemming from a different set of participants (e.g., Train: Set 3, Exp. 2A, Test: Set 3, Exp. 2B).

Table 6.3 provides the proportions of explained variance (R^2) in y. Values for each model, tested on the data to which it was fitted (i.e., the table's diagonal), range between .79 (Set 1, Exp. 2A) and .84 (Set 2, Exp. 2A). Numbers in brackets correspond to the model variant in which non-significant variables were omitted. The fact that R^2 values only differ marginally between the full models and those with omitted variables indicates that these variables indeed had negligible effects on explaining the response variable. Models generalized fairly well, in particular when only the participants changed (i.e., for Exps. 2A and 2B for Set 3), but also when only half of the sounds were novel to the model. The worst generalization was for the models fitted

to Set 1 or 2, evaluated on Sets 2 and 1, respectively, yielding a little less than 60% of explained variance. Overall, this demonstrates that this approach is quite robust as it explains the largest proportion of variance in the rating data on acoustic properties alone, even for models whose training sets differed from the test sets.

6.4.3 Including categorical variables

Figure 6.5 (panel A) displays the predicted and observed dissimilarities for the acoustic model introduced above. Although there is generally a good fit, the plot highlights two outliers (annotated as 1 and 2 in the plot). Point 1 stems from the marimba-vibraphone pair for which the acoustic model overestimated the dissimilarity rating, and point 2 from the harp-trumpet pair, for which ratings were underestimated on acoustic grounds alone. This again suggests that listeners not only based their ratings on acoustic information, but also took into account categorical information such as instrument families: Because the marimba and the vibraphone are both percussion instruments, they were rated as more similar than would be predicted given their acoustic differences. The reverse may have been at play for the harp and the trumpet, members of the string and brass families, respectively. In order to provide a quantitative footing for this intuition, we considered four additional categorical predictors of dissimilarity related to the mechanics of instruments and their families. These categories were not based on continuous acoustic descriptions of the audio signal, but may have been inferred perceptually and therefore influenced the dissimilarity ratings.

Table 6.4 lists all 14 instruments and their class memberships (cf., Lakatos, 2000). Here we considered categories based on two types of differences in instrument excitation (impulsive, continuous; pluck, struck, bowed, blown), resonator type (string, air column, bar), and common instrument families in the western orchestra (woodwinds, brass, strings, keyboards, percussion). For all of the four category types, dissimilarity between instruments was treated as a binary code (Giordano, McAdams, Zatorre, Kriegeskorte, & Belin, 2012), i.e., given a 0 if members from a pair shared the same category and a 1 otherwise. The question was whether taking these variables into account would improve the model fit (given that mere overfitting was controlled for by using PLSR).

In order to take examples from the opposite ends of the scale, let us start with

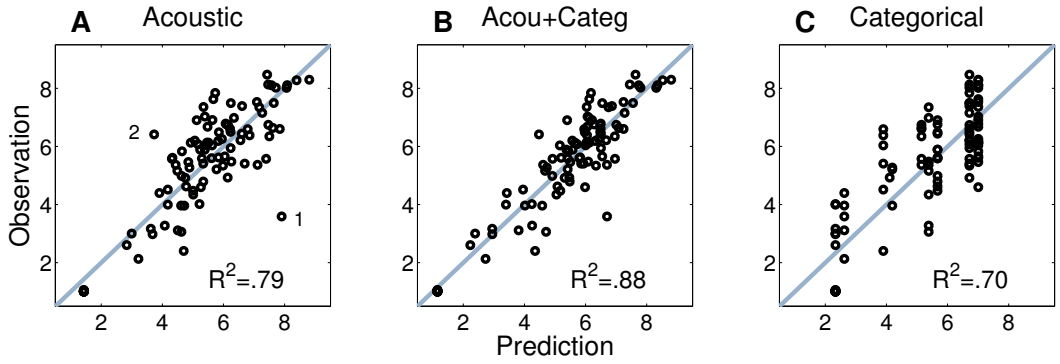

Fig. 6.5 Mean pairwise dissimilarity ratings for Set 1 (observations; y axis) and predictions based upon acoustic descriptors (A), audio and categorical predictors combined (B), and category membership of the instruments (C). Data points 1 and 2 in the left panel are discussed in the text.

the dissimilarity of the marimba and the vibraphone. The above categorical variables would yield a zero contribution to the overall dissimilarity of this pair, because both instruments fall into the same categories for all four variables. The harp and the trumpet, on the contrary, do not share any category. By including these categorical variables in the regression model, the predicted dissimilarity of this pair would thus increase by the sum of the four variables' regression coefficients.

Categorical descriptor 1 (family) correlated significantly ($p < .05$) with all (!) of the other 34 acoustic descriptors with median correlations of med r(103)=.28. Excitation 1 (impulsive, continuous) correlated with 18 (med $r(103) = .19$), excitation 2 (pluck, struck, bowed, blown) with 33 (med $r(103) = .29$), and resonator type with eight acoustic descriptors (med $r(103) = .11$).

Figure 6.5 (panel B) displays predicted and observed values for the model including the full set of acoustic and categorical variables, significantly improving the model fit by 10% as compared to the solely acoustic model (Fisher's $z = -2.22, p = 0.026$, two-tailed), and also visibly improving the fit for the two outliers discussed above. Notably, all categorical descriptors yield significant contributions as their (bootstrapped) confidence intervals do not overlap with zero, as highlighted in Figure 6.6 (white diamonds), which depicts the estimated coefficients (standardized β) for the full model. For the spectral descriptors, the majority of the inter-quartile-range descriptors appear to not provide an important contribution, whereas all but one of the median descriptors do contribute significantly. Similarly, all temporal descriptors contribute significantly.

Table 6.4 Instrumental categories based upon excitation and resonator. Membership to instrument families is indicated by superscript numerals: (1) woodwinds, (2) brass, (3) keyboards, (4) string, (5) percussion.

		Resonator		
Excitation		String	Air Column	Bar
Continuous	blown		BCL[1], BSN[1], FLT[1], HRN[2], TRP[2]	
	bowed	VLI[4], VCE[4]		
Impulsive	struck	PNO[3]		VBS[5], MBA[5]
	pluck	VLP[4], VCP[4], HCD[3], HRP[4]		

Contributions from all four categorical descriptors are significant, although differences in resonator type (encoded by the rightmost variable) are not as strongly taken into account. Moreover, the four categorical descriptors on their own (Figure 6.5, right), already explain 70% of the variance in the ratings (which is not significantly different from the fit of the solely acoustic model, Fisher's $z = 1.41, p = .16$, two-tailed). For this exclusively categorical model, resonator type was the only variable that failed to make a significant contribution as indicated by bootstrapped confidence intervals (not presented here).

We finally considered whether these findings would generalize to Set 3. Dissimilarity ratings for Set 3 were averaged over the order of presentation, as well as across Exps. 2A and 2B. We included the same four categorical predictors as above (although they only applied to the subset of 21 pairs among the seven acoustic recordings part of Set 3) and further added a binary variable that encoded across-category comparisons (indexing rec-trans or trans-rec pairs as 1, and all other pairs as 0). Because categorical descriptors were here construed to encode the dissimilarity based on shared features, instrument categories could not be taken into account for mixed pairs (because they are undefined for transformations). This means that for the subset of 21 pairs from the recordings, the same predictors were considered as in Set 1, but among the 21 pairs of transformations or the 49 mixed pairs, there weren't any categorical dissimilarities

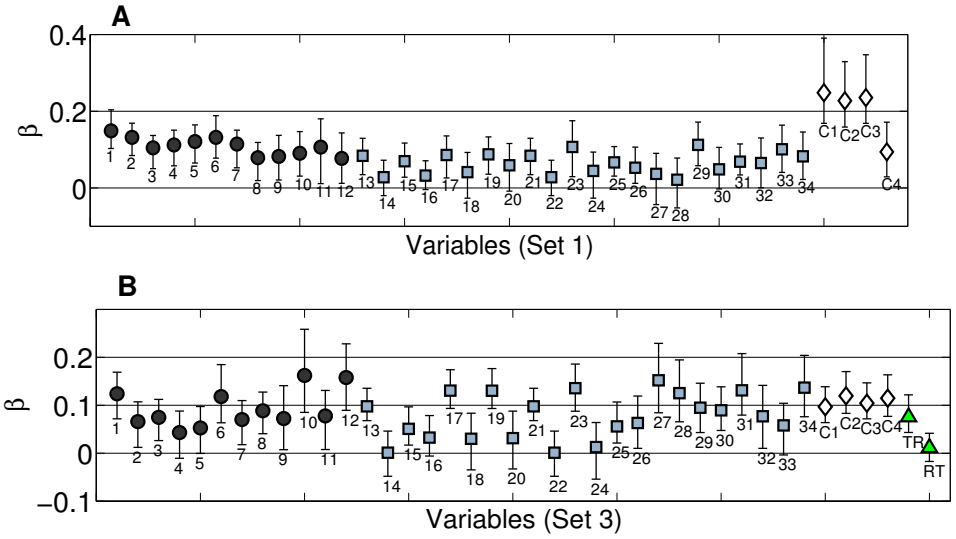

Fig. 6.6 Bootstrapped regression coefficients (standardized) for complete models (acoustic+categorical descriptors) of Set 1 (A) and Set 3 (B, depicts the model that predicts both orders of presentation). (Black) circles correspond to temporal envelope descriptors, (blue) squares to spectral descriptors, (white) diamonds to the four categorical descriptors (within recordings), (green) triangles (Set 3) to across sound category (rec-trans) comparisons. Enumerations of variables corresponds to Table 6.2. Categorical variables correspond to C1) instrument family, C2) excitation 1 (impulsive, continuous), C3) excitation 2 (struck, pluck, bowed, blown), and C4) resonator type (string, air column, bar). RT encode recording-transformation pairs, TR the reverse. Error bars correspond to bootstrapped 95% confidence intervals.

contributing to the regression model. In comparison to Set 1, categorical dissimilarity was therefore encoded quite coarsely. Nonetheless, all five categorical variables contributed significantly as indicated by bootstrapped confidence intervals that did not overlap with zero. The model fit increased from $R^2 = .83$ for the solely acoustic model to $R^2 = .86$ for the complete model, although that increase was not significant (Fisher's $z = 0.66$, $p = .51$, two-tailed).

In a last step, we considered the same model without averaging rating data over the order of presentation (i.e., yielding a model with 182 data points instead of 91 as above). In order to control for asymmetries in ratings of mixed pairs (see, Sec. 6.3.2), we used two independent binary variables, one indexing the order recording-transformation (i.e., yielding 1 for rec-trans pairs, and 0 otherwise), the other encoding the reverse order

(i.e., yielding 1 for trans-rec pairs, 0: otherwise), in addition to the other four categorical variables that only applied to pairs among recordings. Again, the inclusion of the categorical variables improved from $R^2 = .73$ to $R^2 = .77$ (although insignificantly, Fisher's $z = .78$, $p = .42$, two-tailed), and regression coefficients of all four categorical variables specific to the recordings were significantly different from zero, as indicated by bootstrapping. However, only the variable encoding the mixed pair with the order trans-rec had significant positive weight; the variable encoding the reversed order was deemed insignificant by bootstrapping. Figure 6.6 (bottom) displays the corresponding model coefficients.

Note that in contrast to Set 1, where categorical variables alone already explained 77% of variance in the ratings, the solely categorical model achieved a fit of $R^2 = .41$ and $R^2 = .31$ for Set 3 (averaged and not averaged across orders of presentation, respectively). This reflects the above mentioned coarseness of the encoding of the categorical dissimilarity for Set 3.

6.4.4 Discussion

This section described a novel model of timbre dissimilarity using partial least-squares regression. Scalar descriptors of the acoustic signal provided good predictions of timbre dissimilarity ratings, which generalized to other sets of sounds. By a post-hoc inclusion of a set of categorical predictors that described an instrument's family membership and facts about source and excitation mechanisms in Set 1, correlations with the observed timbre dissimilarities could be improved by around ten percentage points of the explained variance with significantly better fit compared to the solely acoustic or categorical model. Notably, these categories alone predicted around 70% of the rating variance in Set 1.

The model for Set 3 improved by 3–4 percentage points when categorical variables were added. Importantly, the model qualified the asymmetries discussed above by suggesting that only when the transformation precedes the recording does categorical information seem to strongly affect ratings, but this does not hold for the reversed order. The smaller increase in fit achieved by categorical variables for Set 3 compared to Set 1 may be attributed to the circumstance that the fine grained categorization by the four within-recordings variables only encompassed a quarter of all comparisons in

Set 3, thus effectively reducing their predictive power when quantified on the basis of the full set.

Overall, we interpret these results as evidence that timbre dissimilarity ratings are informed by both continuously varying "low-level" acoustic properties, transformed into an auditory sensory representation available to the listener, as well as more "cognitive" categorical and semantic information from long-term memory inferred from the sensory representation. The regression analysis thus plausibly extends the above hypothesis on category effects in timbre dissimilarity ratings to within-set comparisons for well-known acoustic timbres from the western orchestra that can be easily associated with instrument categories. In effect, optional processes in dissimilarity ratings (Melara et al., 1992) may not only be at play in the "pathological" situation of comparing sounds with very different origins (instrumental vs. synthetic), but in any dissimilarity rating of stimuli that evokes source categories. More generally, this interpretation resonates with Tversky's comments on similarity as a complex concept. "Similarity has two faces: causal and derivative. It serves as a basis for the classification of objects, but it is also influenced by the adopted classification."(Tversky, 1977, p. 344)

It could be argued that the categorical descriptors only described acoustic and sensory aspects in a more precise way than the acoustic descriptors. However, their rough binary nature (e.g. describing attack quality by simply two categories) together with the comparatively good fit that the exclusively acoustic model achieved renders that hypothesis unlikely. This relates to the discussed experimental obstacle in this domain, namely the inherent coupling of acoustics and categories, that allows listeners to infer categories in the first place: a majority of acoustic variables correlated significantly with any of the categorical ones, making it impossible to fully disentangle sensory and cognitive aspects for natural acoustic stimuli that listeners are familiar with, i.e., for which they possess categories. However, there are exceptions to this coupling, as illustrated by the example of the marimba-vibraphone pair (within instrumental family), whose dissimilarity was overestimated on acoustic grounds, or the trumpet-harp comparison (across family) whose dissimilarity was underestimated in the solely acoustic model. A natural follow-up question then would be whether the suggested effects are under intentional subjective control, that is, whether instructing and training participants to base their ratings solely upon acoustic properties would diminish the observed effects.

Broadening the view, the current findings feature certain parallels with aspects of the literature on speech perception. For example, Zarate, Tian, Woods, and Poeppel (2015) suggested that acoustical, as well as pre-lexical phonological information, contribute to speaker identification (also see, Remez, Fellowes, & Nagel, 2007; Obleser & Eisner, 2009). Identification performance was above chance for non-speech vocalizations, demonstrating the importance of solely acoustic information, but native English speakers' accuracy improved with increasing phonological familiarity of speech tokens (Mandarin, German, Pseudo-English, English). Again, there seems to be an interplay of basic acoustic factors and higher-level properties of speech signals that listeners need to be familiar with in order for it to become useful to them. From an even broader perspective, related observations have been made in computational music classification. For example, McKay and Fujinaga (2008) showed that combining variables extracted from the audio signal with non-acoustic types of information (e.g., symbolic MIDI data) markedly increased genre classification accuracy, again underlining the value of combining acoustic and categorical types of information representations.

6.5 Conclusion

This paper explored the role of acoustic and categorical information in timbre dissimilarity ratings. Exp. 1 provided data on the identifiability and familiarity of sounds. By means of filterbank-based sound analysis-synthesis, we created transformed tones that were generally rated as less familiar than recorded acoustic tones. We selected a subset of stimuli from the least familiar transformations that were subsequently rated on pairwise dissimilarity in Exps. 2A and 2B., along with a set of recorded acoustic tones and a mixed set. We observed that the dissimilarity data of the recorded instrument timbres clustered into subsets that distinguished timbres according to acoustic and categorical properties, such as brightness and instrument family, respectively. For the subset of cross-category comparisons in Set 3 that involved both recordings and transformations, we observed asymmetries in the distribution of ratings, as well as a stark decay of inter-rater agreement. Subsequently, these effects were replicated in a more robust within-subjects design in Exp. 2B and cannot be explained by merely acoustic factors. Note that within-set dissimilarities did not show asymmetric tendencies. In a last section we explored a novel model of timbre dissimilarity that compared

the contributions of both acoustic and categorical features. The strongest correlation with the observed dissimilarities was achieved when both kinds of timbre descriptors were taken into account.

In the introduction, musical timbre was defined as a seemingly hybrid concept that encompasses both sensory and categorical components. Subsuming both facets under one term does not, consequently, constitute a lack of definitional precision, but acknowledges the multifaceted nature of information representation in the human mind. To borrow from Fuster (2003),

> "Every percept has two components intertwined, the sensory-induced *re-cognition* of a category of cognitive information in memory and the categorization of new sensory impressions in the light of that retrieved memory. Perception can thus be viewed as the interpretation of new experiences based on assumptions from prior experience."(p. 84)

Our data on the interaction of acoustic and categorical facets in timbre dissimilarity suggest that the percept of timbre is a superb example of this duality. Timbre perception naturally associates a sensory representation of an acoustic waveform to hierarchically ordered categories of sound production stored in long-term memory. The listening brain represents, simultaneously, "the sound" and "the idea" of a musical instrument. Future research on timbre perception should attempt to distinguish and further disentangle these levels of representation.

Chapter 7

Familiarity and attentional maintenance in memory for timbre

By comparing the recognition of timbres from familiar acoustic instruments and unfamiliar synthetic transformations, this chapter extrapolates the previous chapter on timbre dissimilarity ratings towards short-term memory. Beyond systematically addressing for the first time the role of stimulus material in short-term recognition of timbre, the chapter further explores memory maintenance strategies and factors of musical training. Overall, this work attempts to account for the rich array of affordances that familiar timbres offer the listener.

This chapter is based on the following research article:

> Siedenburg, K. and McAdams, S. (in preparation). The role of long-term familiarity and attentional maintenance in auditory short-term memory for timbre. Manuscript prepared for submission to *Memory*.

Abstract. We study short-term recognition of timbre using familiar recorded tones from acoustic instruments and unfamiliar transformed tones that do not readily evoke sound-source categories. Participants indicated whether the timbre of a probe sound matched with one of three previously presented sounds (item recognition). In Exp. 1, musicians better recognized familiar acoustic compared to unfamiliar synthetic sounds, and this advantage was particularly large in the medial serial position. There was a strong correlation between correct rejection rate and the mean perceptual dissimilarity of the probe to the tones from the sequence. Exp. 2 compared musicians' and non-musicians' performance with concurrent articulatory suppression, visual interference, and with a silent control condition. Both suppression tasks disrupted performance of a similar margin, regardless of musical training of participants or type of sounds. Our results suggest that familiarity with sound source categories and attention play important roles in short-term memory for timbre, which rules out accounts solely based on sensory persistence.

7.1 Introduction

Timbre is a major component of audition, but many facets of its cognitive processing have only been investigated sporadically. Timbre refers to the auditory attributes that lend sounds a sense of "color" or "shape" and enable the inference of sound sources. The percept emerges from acoustic cues such as the spectral envelope distribution, attack sharpness, spectrotemporal variation or modulation, roughness, and noisiness, in addition to features that may be idiosyncratic to certain instruments (McAdams, 2013). Whereas music cognition has studied non-verbal auditory schemata acquired through long-term experience in the domain of pitch, harmony, and rhythm (see e.g., Krumhansl, 1990; Huron, 2006), timbre has traditionally been treated as a primarily sensory phenomenon that resides "in the moment". Accordingly, timbre processing should not be subject to long-term familiarization. Neurophysiological studies on timbre processing have started to provide evidence for the contrary position (cf., Pantev et al., 2001; Shahin et al., 2008; Strait et al., 2012), although it is unclear whether these results reflect conscious perceptual experience. In the present study, we investigate the role of long-term timbre familiarity in a behavioral short-term memory (STM) task.

A related experimental cornerstone in the verbal domain is the *lexicality effect*: Memory for item identity is generally better for words than for closely matched *pseudowords* (Thorn et al., 2008), defined as vocables that respect phonotactic constraints of a language but are meaningless, i.e., not part of the dictionary. Similar effects

have also been shown for variables such as word frequency and imaginability (Thorn et al., 2008). Whether caused by greater activation strength, facilitated rehearsal, or more robust memory retrieval (cf., Thorn, Gathercole, & Frankish, 2002; Macken et al., 2014), these effects underline the importance of long-term knowledge in verbal short-term remembering.

In simplest terms, words reference things or activities in the world. Timbre has similar properties, in the sense that familiar timbres from acoustic instruments can be perceived as referents to sound sources (e.g., a violin) and the cause or activity that set them into vibration (e.g., plucking), likely by virtue of a learned, long-term association (McAdams, 1993). Comparing short-term memory for unfamiliar tones with hidden underlying source/causes to familiar tones from acoustic instruments may therefore create a scenario that is analogous to experiments that give rise to the verbal lexicality effect. A challenge lies in the selection of unfamiliar sounds. A simple idea would be to use abstract digitally synthesized sounds, created by additive synthesis of sinusoidal components, for instance. One problem of such an approach is that the overall acoustic variability (or complexity) of a stimulus set appears to affect short-term memory. Golubock and Janata (2013) observed severe capacity limits of short-term memory for the timbre of tones created by additive synthesis, but less so when a more variable set of tones, selected from commercial synthesizers, was used.

One piece of the problem is to define what it means to retain similar degrees of "acoustic variability". Digitally synthesized tones usually vary on a small number of dimensions, whereas natural sounds vary in manifold ways. A perspective that has proven to be of relevance in a variety of settings is the distinction between an acoustic signal's temporal fine structure and time-varying envelope (e.g., Moore, 2015). For instance, Z. M. Smith et al. (2002) superimposed the envelope of one type of speech signal onto the fine structure of another and obtained "chimæric" perceptual properties (also see, Agus et al., 2012). Here, we decided to use this approach as a starting point and thereby generated transformed tones that have similar physical properties (i.e., regarding temporal fine structure and envelope) compared to a set of tones recorded from acoustic musical instruments. At the same time, they have drastically reduced arrays of identifiable source categories and were rated as perceptually less familiar than the original recordings. On the one hand, one might suspect potential differences in memory performance for such "referential" (familiar) and "non-referential" (unfamiliar)

timbres to emerge from encoding, where familiar timbres may be assumed to more strongly activate semantic long-term memory representations than unfamiliar timbres, affording a level-of-processing phenomenon (Craik & Lockhart, 1972). On the other hand, differential maintenance strategies could also be a factor. The extent to which sound source categories give rise to verbal rehearsal in STM for timbre has in fact been discussed by a number of recent studies (McKeown et al., 2011; Schulze & Tillmann, 2013; Soemer & Saito, 2015), and the issue relates to the general question of whether there is active maintenance in STM for timbre. Having participants discriminate small changes in spectral aspects of timbre, McKeown et al. (2011) showed that sensitivity was above chance even for extended retention intervals of 5–30 s. Notably, this effect was robust to an articulatory suppression task in which participants were required to read aloud during the retention time. The authors interpreted these results as evidence for a type of sensory persistence that is "neither transient nor verbally coded nor attentionally maintained."(p. 1202) Nonetheless, they also emphasized that there may be various other forms of memory for timbre. Schulze and Tillmann (2013) compared the serial recognition of timbres, pitches, and words in various experimental variants, using sampled acoustic instrument tones and recorded verbalizations. They found that the retention of timbre, contrary to that of pitches and words, did not suffer from concurrent articulatory suppression. On the basis of these results, they suspected that working memory for timbre is not subject to active rehearsal but instead relies on the passively stored sensory trace.

Other studies have underlined the necessity of attentional maintenance. Nolden et al. (2013) recorded EEG during a serial order recognition task with synthesized timbres differing in spectral envelope. In a control condition, participants received the same stimuli but were asked to ignore the standard and to judge a property of the last tone of the comparison sequence. Significant differences in event-related potentials (ERP) were found during the retention interval; the higher the memory load, the stronger the ERP negativity. These findings cohere with Alunni-Menichini et al. (2014), demonstrating that the same ERP component robustly indexes STM capacity, providing evidence for an attention-dependent form of STM. Most recently, Soemer and Saito (2015) observed that short-term item recognition of timbre was only inconsistently disrupted by articulatory suppression, but was more strongly impaired by a concurrent auditory imagery task. This was interpreted as evidence for that memory for timbre can be

an active process that deteriorates when attentional resources are removed. These findings further suggested that a process similar to what has been called *attentional refreshing* (Camos, Lagner, & Barrouillet, 2009) may be a viable candidate mechanism for maintenance of timbre. Refreshing emerges through the reactivation of a target's mental representation by means of attentional focusing (Cowan, 1988; Johnson, 1992). The target briefly reenters conscious awareness, whereby its representation is kept in an active state. The process has been shown to be independent of subvocalization-based rehearsal (Camos et al., 2009) and is preferentially employed in verbal working memory tasks with low concurrent processing load (Camos et al., 2011). Nonetheless, these considerations all emerged in the domain of verbal working memory, and currently it is unclear whether similar processes play a role in memory for timbre.

In order to explore the role of sound source categories, long-term familiarity, and maintenance strategies in short-term memory for timbre, we compare recognition of acoustic musical instrument sounds and their digital transformations. Exp. 1 tests effects of timbre familiarity and list-probe delay, as well as effects of serial position and list-probe dissimilarity. Exp. 2 uses a subset of trials from Exp. 1 and exposes a group of musicians and non-musicians to articulatory suppression and a visual distractor task, and also includes a silent control condition.

7.2 Experiment 1: Material and delay

We explored the effect of long-term timbre familiarity and delay interval on musicians' short-term item recognition performance. Because we expected the timbral memory traces of unfamiliar transformations to be more transient, we hypothesized that a potential familiarity advantage would even be greater at 6 s compared to 2 s of delay.

7.2.1 Methods

The research reported in this manuscript was carried out according to the principles expressed in the Declaration of Helsinki and the Research Ethics Board II of McGill University has reviewed and certified this study for ethical compliance (certificate #67-0905).

Participants

Thirty musicians (22 female) participated in the experiment for monetary compensation. They were recruited from a mailing list of the Schulich School of Music at McGill University and had an average age of 21 years (SD = 3.7, range: 18–29). They had 10 years (SD = 3.8) of instruction on at least one musical instrument and had received 5 years (SD = 3.6) of formal music-theoretical training. Participants reported normal hearing, which was confirmed in a standard pure-tone audiogram measured before the main experiment (ISO 398-8, 2004; F. N. Martin et al., 2000) and had hearing thresholds of 20 dB HL or better for octave-spaced frequencies from 125 Hz to 8000 Hz.

Stimuli

Recorded and transformed sounds A material factor contained two conditions with different types of sounds, familiar acoustic recordings, and unfamiliar synthetic transformations. The first set consisted of 14 recordings of single tones from common musical instruments, all played at mezzo-forte without vibrato.

Piano and harpsichord samples were taken from Logic Professional 7 (Apple Computer, Cupertino, CA), all others were drawn from the Vienna Symphonic Library (http://vsl.co.at, last accessed April 12, 2014); see Table 7.2 for a complete list. The audio sampling rate was 44.1 kHz with 16-bit amplitude resolution. Sounds had a fundamental frequency of 311 Hz (E♭4), and only the left channel of the stereo sound file was used. According to VSL, the samples were played as 8th-notes at 120 beats per minute, i.e., of 250 ms "musical duration". Nonetheless, actual durations were all slightly longer than 500 ms. We therefore applied barely noticeable fade-outs of 20 ms duration (raised-cosines), in order to obtain uniform stimulus durations of 500 ms.

A set of 70 unfamiliar sounds was generated digitally in order to obscure associations with an underlying source. We digitally transformed the spectro-temporal envelopes and acoustic fine structures of the recordings, a procedure that was demonstrated to yield altered, "chimæric" perceptual properties for speech signals (Z. M. Smith et al., 2002). Each novel sound was derived from a source signal, the spectrotemporal envelope of which was shaped by the spectrotemporal envelope of a second signal that acted as a time-varying filter. The resulting signal possessed the spectrotemporal envelope of the one filtering signal and the fine structure of the source signal. More details on the

sound synthesis, familiarity and dissimilarity ratings can be found in the Appendix.

Using the 14 recorded acoustic tones and the 70 resulting transformations, 15 musicians rated perceptual familiarity and identified sounds by selecting one out of eight options (including six instrument names and the labels "unidentifiable", and "identifiable, but not in the list"). The 14 transformations that had received the smallest mean familiarity ratings were selected for the main experiment. Mean familiarity of the 14 original recordings ($M = 4.2$, range: 3.1–4.8) was significantly higher than that of the 14 selected transformations ($M = 2.0$, range: 1.6–2.4) as indicated by an independent-samples t-test, $t(26) = 15.5, p < .001$. The mean proportion of "unidentifiable" ratings selected for the 14 recordings and the 14 selected transformations was $M = 0.04$ ($SD = 0.06$) and $M = 0.52$ ($SD = 0.11$), respectively, which constituted a significant difference (independent-samples t-test, $t(26) = 13.8, p < .001$). Pearson correlations between the proportion of "unidentifiable" votes per stimulus and mean familiarity ratings were strong and negatively associated, $r(82) = -.88, p < .001$. Table 7.2 lists the stimuli used for the current memory experiments.

Perceived loudness was matched on the basis of six expert listeners' adjustments. Subsequently, 24 musicians rated pairwise dissimilarity for both sets of sounds on an analog-categorical scale (1-identical, 9-very dissimilar).

Memory sequences We used an item-recognition task for the main experiment. Every trial featured a "study list", that is, a sequence of three distinct sounds of 500 ms duration each, which were concatenated with an inter-stimulus interval of 100 ms. The list was followed by a delay of 2 or 6 s before a probe tone was presented.

Fourteen study lists were generated by drawing sounds (nos. 1–14) randomly without replacement under the constraint that every tone occurred equally often (i.e., 3 times) in the 14 lists. Note that the underlying list structure was identical for both material conditions (i.e., recordings and transformations); only the individual sounds that represented the numbering scheme differed. Per material condition, every list was paired with two matching and two non-matching probes. Matching probes were taken from all three serial positions, such that there were overall 8, 10, and 10 probes from the first, second, and third serial position, respectively. New probes were selected among the remaining $14 - 3 = 11$ sounds from the set of recordings or transformations such that for every list there was one probe that was dissimilar (i.e., with a list-probe

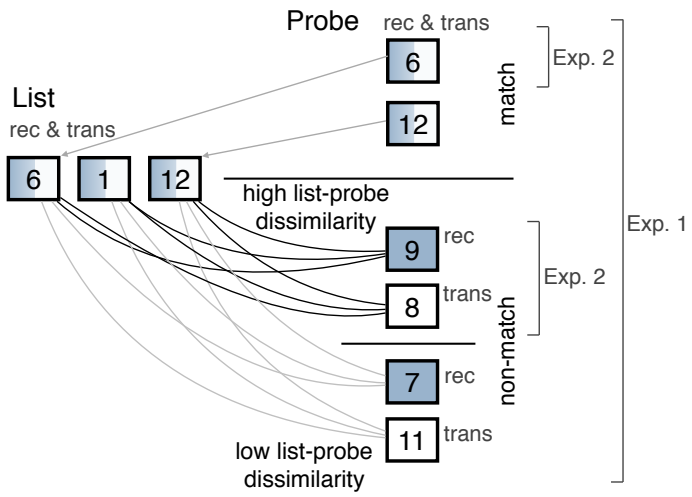

Fig. 7.1 Illustration of the construction of list-probe sequences. Digits refer to individual sounds (#1–14), blue boxes to recordings, white boxes to transformations, half blue/half white boxes to numbers that are instantiated by both materials. Per list, there were two matching probes, equally selected from all three serial positions across the different trials (see Table 7.1). Non-matching probes were selected such that both materials' lists had a probe with high, and another with low list-probe dissimilarity (and the distribution of dissimilarities did not differ across material). Exp. 2 only used a subset of trials.

dissimilarity above the median) and another that was similar (i.e., a below-median dissimilarity). The fact that timbre dissimilarity relations are different between recordings and transformations required us to use differently numbered non-matching probes in the two material conditions. Figure 7.1 illustrates this graphically.

List-probe dissimilarity has been proven to be important in various short-term item recognition studies (see e.g., Visscher et al., 2007). In our case, the resulting distribution of dissimilarities did not differ between recording ($M = 5.2$, $SD = 1.09$) and transformation trials ($M = 5.2$, $SD = 0.90$), neither in terms of means, $t(54) < 1$ (two-sample t-test), nor in terms of shape, $D = 0.18, p = .72$ (two-sample Kolmogorov-Smirnov test). The complete list of memory sequences is given in Table 7.1. Overall, there were 14 × 2 (new probes: low and high dissim.) × 2 (old probes) × 2 (delay: 2 s and 6 s) = 112 trials per material condition.

Table 7.1 List of memory sequences. Digits 1–14 refer to the materials of recordings (recs) and transformations (trans) as provided in Table 7.2. Lists and matching probes rely on the same numbering structure for both materials. Non-matching probes are selected differently across materials, in order to obtain a similar distribution of list-probe dissimilarities across material conditions. Non-match probes in the A columns feature high list-probe dissimilarity, and the B columns contain low-dissimilarity probes. Exp. 1 uses all trials as indicated (i.e., 14 lists × (2 match probes+ 2 non-match probes) = 56 trials per material condition), presented at 2 and 6 s retention intervals. In Exp. 2, only the probes listed in the columns A are used.

Lists			Probes					
recs & trans			recs & trans match		recs non-match		trans non-match	
			A	B	A	B	A	B
11	12	6	11	12	1	7	8	1
11	4	3	11	4	13	6	2	10
10	7	4	10	7	5	14	12	11
2	1	9	2	1	11	5	8	12
5	14	13	14	13	1	9	8	12
1	5	11	5	11	7	2	13	4
2	6	8	6	8	13	5	4	13
8	14	2	14	2	1	6	11	12
10	13	3	13	3	14	1	8	14
9	4	7	7	9	12	5	5	3
10	13	7	7	10	5	14	4	2
5	3	9	9	5	11	2	4	10
8	12	14	14	8	2	7	1	2
6	1	12	12	6	9	7	8	11

Presentation and apparatus

The average presentation level after loudness-normalization was 66 dB SPL (range: 58–71 dB SPL) as measured with a Brüel & Kjær Type 2205 sound-level meter (A-weighting) with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). Experiments took place in a double-walled sound-isolation chamber (Industrial Acoustics Company, Bronx, NY). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronics GmbH, Wedemark, Germany), using a Macintosh computer (Apple Computer, Cupertino, CA) with digital-to-analog conversion on a Grace Design m904 (Grace Digital Audio, San Diego, CA) monitor system. The experimental interface and data collection were conducted with the audio software Max/MSP (Cycling 74, San Francisco, CA).

Procedure and design

In the item recognition task, participants were asked to respond to the question "Did the final sound exactly match any previous sound from the sequence?" by pressing a button on a response box corresponding to "Yes" or "No". If participants responded "Yes", they were asked to indicate the serial position of the match by pressing the corresponding number on the computer keyboard. We only consider the data from the first binary task for the current analyses.

Trials were presented in four blocks, with two containing recordings and two transformations. They were interleaved (e.g., rec, trans, rec, trans) with order counterbalanced across subjects. Within each material condition, the order of trials was fully randomized. Every block required around 15 min to complete, and participants took a mandatory break of 5 min between blocks. In order to get used to the recognition task, participants received four example trials from the recordings for which correct responses were provided. After completion of the experiment, participants filled out a questionnaire about biographical information and about the experiment itself.

Data analysis

We measured sensitivity with d' scores and response bias with the criterion location c, as provided by the Yes/No model (Macmillan & Creelman, 2005, Ch. 1–2). Hits were defined as a correctly recognized match trial (i.e.,"old"), false alarms as incorrectly

identified non-match trials ("new"). The sensitivity d' thus indicates how well participants discriminate between old and new trials. The criterion c describes whether participants are biased toward responding "non-match" ($c > 0$) or "match" ($c < 0$). We did not consider individual responses that were faster than 200 ms or slower than 4000 ms (less than 5% of overall responses). We did not analyze response times in the full factorial designs because instead of reflecting memory fidelity, response times may have been confounded by the factors of delay in Exp. 1 and suppression in Exp. 2. The following set of analyses considers the variables of material, delay, serial position, and list-probe dissimilarity, as well as potential effects of online familiarization. ANOVAs are conducted for the dependent variables of i) sensitivity and ii) bias as a function of material and delay. The factor of position could not be included in this analysis, because it is only defined on match trials, whereas the signal detection theoretic variables require match and non-match trials to be combined. We thus computed another ANOVA for an analysis of iii) hit rate as a function of material, delay, and position. For non-match trials, we analyzed iv) correlations between list-probe dissimilarities and correct-rejection rates. In order to assess potential effects of online familiarization, we finally computed two ANOVAs on v) sensitivity and vi) bias as a function of experimental block[1] (1st vs. 2nd) and material. Because multiple null hypotheses tests (such as the five ANOVAs just mentioned) inflate experiment-wise Type I error rates, we used the adjusted significance level of $\alpha = .01$ for the main analyses[2].

7.2.2 Results

i) A repeated-measures ANOVA on d' scores yielded effects of material, $F(1, 29) = 11.1$, $p = .002$, $\eta_p^2 = .276$, and delay, $F(1, 29) = 30.3$, $p < .001$, $\eta_p^2 = .511$, but no significant interaction.

ii) The criterion location c was significantly affected by material, $F(1, 29) = 12.3$, $p = .002$, $\eta_p^2 = .297$, and delay, $F(1, 29) = 100$, $p < .001$, $\eta_p^2 = .776$, but the interaction

[1] We did not analyze material, delay, and block conjointly because in our randomization scheme, each subject was presented with a varying number of trials for a given block×material×delay condition, which would have rendered the calculation of signal detection theoretic measures problematic (Macmillan & Creelman, 2005, pp. 8–9).

[2] See for instance the statistical guidelines of the *Psychonomic Bulletin & Review* for corresponding recommendations: `http://www.springer.com/psychology/cognitive+psychology/journal/13423`

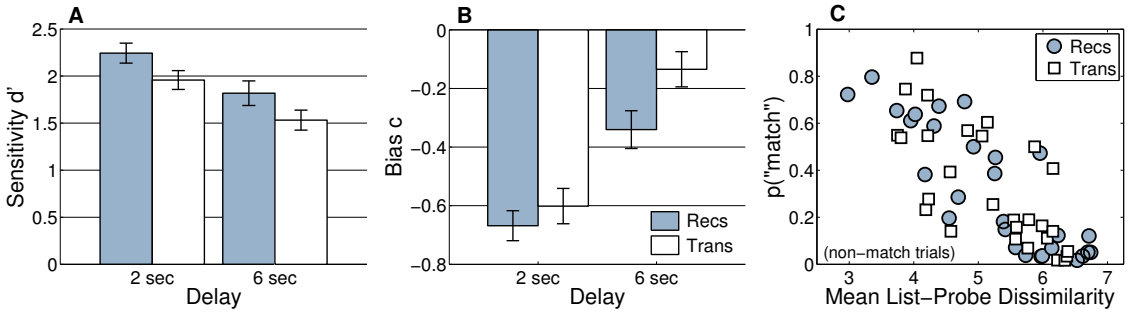

Fig. 7.2 Exp. 1: d' scores (A), response biases (B), and proportion of match responses as a function of mean dissimilarity of list and probe items for non-match trials (C). Error bars depict standard errors of the mean.

of both factors failed to reach significance ($\alpha = .01$), $F(1, 29) = 4.66$, $p = .039$, $\eta_p^2 = .139$. Figure 7.2 depicts sensitivity and criterion locations for delay and material conditions.

iii) Considering effects of serial position, a repeated measures ANOVA on hit rates with the factors position, material, and delay yielded an effect of position, $F(2, 58) = 13.4$, $p < .001$, $\eta_p^2 = .316$, and of material, $F(1, 29) = 17.9$, $p < .001$, $\eta_p^2 = .382$, as well as a significant interaction between the two, $F(2, 58) = 12.6$, $p < .001$, $\eta_p^2 = .304$. The main effect of position stemmed from significantly lower performance in the second position compared to the first and third positions, paired $t(29) > 2.9$, $p < .007$, but only a marginal difference between first and third positions, $t(29) = -2.3$, $p = .028$ (n = 3 comparisons, Bonferroni-corrected $\alpha_{crit} = .0167$). The interaction of position and material was due to higher sensitivity for recordings in the second position, paired $t(29) = 5.2$, $p < .001$, see Figure 7.3, but no differences between recordings and transformations in the other two serial positions, $p > .040$ (n = 3 comparisons, Bonferroni-corrected $\alpha_{crit} = .0167$).

There also was an effect of delay, $F(1, 29) = 52.4$, $p < .001$, $\eta_p^2 = .644$, and an interaction of delay and position, $F(2, 58) = 4.2$, $p = .002$, $\eta_p^2 = .127$, as visible in Figure 7.3 (A and B). The latter was due to the fact that in addition to the main effect of position (featuring lowest performance in the second serial position) hit rates were particularly low in this serial position with 6 s of delay ($M = .75$, $SD = .20$, compared to $M = .90$, $SD = .12$ for 2 s), as confirmed by post-hoc contrasts, $\beta = .074$, $t(325) = 5.14$, $p < .001$.

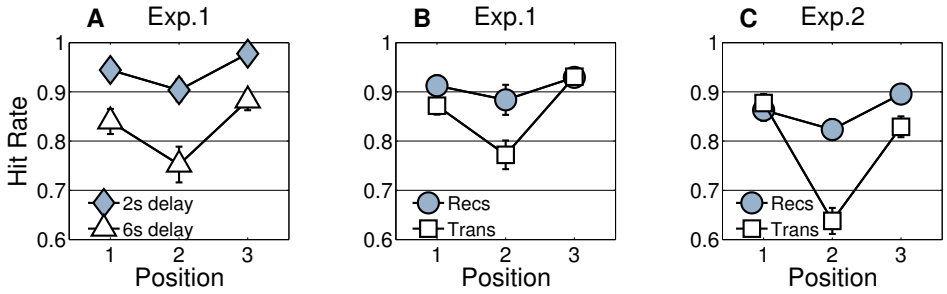

Fig. 7.3 Hit rates as a function of serial position depicted for delay conditions in Exp. 1 (A), and material conditions in Exp. 1 (B) and Exp. 2 (C). Error bars show standard error of the mean.

iv) Figure 7.2 (C) shows the strong relation between mean list-probe dissimilarity and response choice with significant correlations for recordings, $r(26) = .85, p < .001$, and transformations, $r(26) = .72, p < .001$. The figure also demonstrates that responses are strongly biased, because trials with the lowest dissimilarity ratings had correct-rejection rates of less than 50%. This strong bias warrants the usage of the signal detection theoretic measures. As with the unbiased d' measure, performance on the lower half of list-probe dissimilarities ranged above chance with $M = 1.4$ ($SD = 0.65$), and $M = 1.3$ ($SD = 0.48$) for recordings and transformations, respectively. For the other half of trials with high dissimilarities, sensitivity was at $M = 2.9$ ($SD = 0.69$) and $M = 2.3$ ($SD = 0.60$) for the two respective material conditions. There was no significant correlation of list heterogeneity and correct rejection rate or hit rate, $r(27) < .40$, $p > .12$.

v) Finally, we addressed potential effects of online familiarization in a dedicated repeated measures ANOVA with the factors of material and experimental block. If there was online familiarization with the initially unfamiliar transformations, one would expect an interaction between the two variables. Besides the main effect of material on d' scores already analyzed above, there was neither an effect of block, $F(1, 29) = 0.4$, $p = .53$, nor an interaction, $F(1, 29) = 1.2$, $p = .27$. vi) The criterion location c was affected by material (analyzed above), but not significantly affected by experimental block, $F(1, 29) = 0.17$, $p = .68$, and the interaction of material and block failed to reach significance ($\alpha = .01$), $F(1, 29) = 5.4$, $p = .026$, $\eta_p^2 = .158$.

7.2.3 Summary and discussion

We compared short-term item recognition of musicians for a set of familiar orchestral tones and a set of unfamiliar synthetic tones. The stimulus sets were tightly matched in terms of physical properties such as spectrotemporal envelope profiles, and so were the resulting sets of memory lists and probes which were almost identical in structure and did not differ with regards to list-probe similarity (cf., Visscher et al., 2007). The main effect of material on sensitivity was coherent with our hypotheses. Familiar timbres that musicians can associate with well-known instrument categories are better recognized than are unfamiliar timbres. The main effects of delay on sensitivity and bias seem intuitive and are coherent with results from Golubock and Janata (2013).

We had expected an even larger difference in sensitivity across material conditions at 6 s of list-probe delay where we thought the multiple affordances for encoding and maintenance of familiar timbres would lead to more robust recognition. This was not the case, although there was a tendency for an interaction effect on response bias: participants judged more transformation trials than recording trials as "new", and this was particularly so for the longer delay condition. That is, rather than affecting memory fidelity as such, the interplay of material and retention time only tended to weakly affect response behavior.

Considering the serial position data, there was not only a main effect of position on hit rate, but transformations were even less well recognized when they were presented in the medial position of the sequence (according to the main effect of position, the medial position was less salient than the first and last serial positions). The same interaction was obtained for the delay factor, with the 6 s delay being worst in the medial position.

There was a strong correlation between correct rejections and dissimilarity: the more dissimilar the probe was to the elements of the list, the more likely it was to be recognized as new. Note that we did not find any significant effect of list homogeneity (pairwise similarity of a study list) on correct rejections or on hit rates. This contrasts with the findings from Visscher et al. (2007), but also raises the question of whether the increase in capacity across the experiments that Golubock and Janata (2013) interpreted as resulting from a global increase in perceptual heterogeneity of the stimulus set might be more precisely described as a trial-wise similarity effect.

An intricate question is whether the initially unfamiliar transformations become more familiar over the course of the experiment. Other studies (e.g., Golubock & Janata, 2013; Soemer & Saito, 2015) have selected large numbers of supposedly unfamiliar stimuli by relying on the subjective familiarity judgments of the authors alone, as well as audio-descriptor-based models of timbre dissimilarity (which have only been perceptually validated to a limited extent). We chose a "closed set" design that repeats items, because we wanted to thoroughly control the items' perceptual familiarity and identifiability as well as perceptual dissimilarities between target list items and probe items on the basis of experimental data (as reported in the stimulus section above). Because the number of pairwise dissimilarity ratings grows quadratically with set size, we thus needed to settle on two relatively small sets of tones. Every sound, whether as part of a sequence or as probe, appeared on average around 32 times over the course of the entire experiment. In that sense, the current design may conflate the aspects of familiarity and source identification, which theoretically may have different dynamics: The transformed sounds do not readily evoke sound source categories, and this is unlikely to change with repeated listening (because there aren't any). On the contrary, it could be that listeners became progressively more familiar with the transformations, supporting processing fluency.

Our data, however, do not feature significant effects of online familiarization, as would have been indicated via a material×block interaction for sensitivity or bias. Although there was a tendency of an interaction effect for the latter variable, participants did not manage to adapt their strategy for the transformed sounds in a way that optimized sensitivity. For that reason, we conclude that the current data are not substantially affected by online familiarization.

Turning towards the underpinnings of the observed effect of material, a crude explanation could posit that musicians verbally labeled recordings but not transformations and subsequently rehearsed the label. In Exp. 2, we set out to test whether disruption of maintenance may have differential effects on our sounds that featured explicit differences in affordances for verbal labeling, as well as groups of participants with different levels of musical expertise.

7.3 Experiment 2: Material, suppression, and group

The experiment contained a between-subjects factor that compared a group of non-musicians with a group of musicians, and besides the material factor, a novel suppression factor with the conditions of articulatory suppression, visual suppression, and a silent control condition. For non-musicians, we expected a diminished advantage of recordings over transformations (as expressed in a material×group interaction). We were also curious whether the interruption of verbal labeling via articulatory suppression would specifically interfere with performance on the acoustic recordings.

7.3.1 Methods

Participants

Forty-eight listeners participated in the experiment for monetary compensation. A group of 24 musicians (13 female) was recruited from a mailing list of the Schulich School of Music at McGill University. They had mean ages of 23 years ($SD = 4.2$, range: 18–34), had received 15 years ($SD = 4.5$) of instruction on at least one musical instrument (including the voice) and had received 6 years ($SD = 4.3$) of formal music theoretical instruction. None of them had participated in Exp. 1. The group of 24 "non-musicians" was recruited via classified advertisements on a McGill University webpage. They had a mean age of 28 years (median: 23.5, $SD = 11.6$, range: 19-67), 0.4 years ($SD = 0.91$) of instruction on a musical instrument, and no formal music theoretical training beyond elementary school. Normal hearing was confirmed as in Exp. 1.

Stimuli

Memory sequences We used the memory lists from Exp. 1 but only in conjunction with the group of non-match probes that possessed high list-probe dissimilarity, plus one of the two subsets of old probes (see Tab. 7.1). This yielded 14 × 2 (match, non-match) = 28 trials per material condition. Every trial was presented in each of the three suppression conditions.

Suppression conditions There was a silent condition, a visual distractor task, and an articulatory suppression condition. In the visual task, a sequence of 4 × 4 grids of

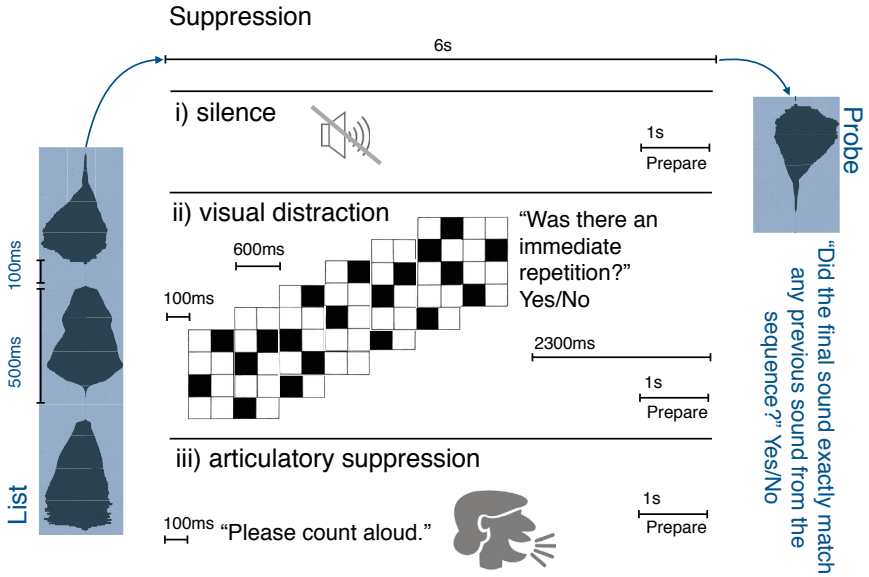

Fig. 7.4 Sketch of the three different suppression conditions in Exp. 2.

filled black and white squares appeared on the screen, similar to the method used by Pechmann and Mohr (1992) and Schendel and Palmer (2007). Participants were asked to indicate, using the same yes/no buttons on the response box, whether there was a direct repetition of a grid in the sequence or not. The visual sequence appeared 100 ms after the offset of the study list, and contained 6 grids, each of which was presented for 600 ms. The grids were created randomly such that 5 of the 16 squares were always filled (Pechmann & Mohr, 1992). The grids occupied a 10×10 cm area on the computer screen. In 50% of the visual suppression trials, there was a direct repetition of a visual grid, distributed across the serial positions of the visual sequence. After the end of the visual sequence, subjects thus had at least 2300 ms to respond to the visual task and prepare for the auditory task. One second before the onset of the probe stimuli, the screen into which the grids were embedded disappeared, signalling participants to get ready to respond to the probe. Figure 7.4 illustrates the task demands of the three suppression conditions.

In the articulatory suppression task, a screen appeared 100 ms after offset of the study list that asked participants to count aloud into a microphone, starting at one. The screen disappeared 1 s before the onset of the probe, which indicated to participants

to stop counting and prepare for the auditory task.

Presentation and apparatus

Presentation and apparatus were identical to those in Exp. 1.

Procedure and design

Participants completed the audiogram and read through the experimental instructions. They were then introduced to the basic item recognition task that was used in all three suppression conditions. For that purpose, two example trials without suppression were presented for which the experimenter provided correct responses. Each suppression condition was then presented block-wise and was preceded by six training trials that let participants familiarize with the respective task. During training, participants could clarify questions with the experimenter. All training trials used sounds from the recordings.

In sum, we considered one between-subjects factor (musicians, non-musicians) and two global within-subject factors, suppression (silence, visual, articulatory), and material (recordings vs. transformations). The serial position factor was nested within the subset of matching probes. The six possible orders of presenting the three suppression blocks were counterbalanced across participants (i.e., participants 1&7, 2&8, etc. received the same order of suppression blocks). The material condition was presented block-wise and was nested within the suppression conditions, with order counterbalanced orthogonally to the suppression factor (i.e., participants 1&3, 2&4, etc. received the same succession of material conditions). A questionnaire was administered after the experiment.

Data analysis

To ensure visual distraction, only trials with correct responses to the visual task were taken into account (on average 93%, $SD = 6$). In the articulatory suppression interval, participants' vocalizations were recorded such that we could verify aurally that they counted aloud in all test trials of the articulatory suppression condition. ANOVAs were computed for the variables of i) sensitivity and ii) bias as a function of suppression,

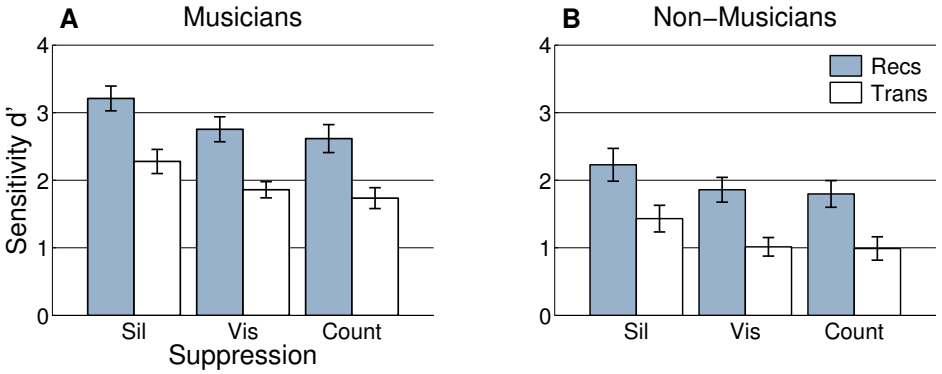

Fig. 7.5 Exp. 2: d' scores for musicians (A) and non-musicians (B) in the suppression conditions of silence, visual suppression, and articulatory suppression (counting).

material, and musical training, iii) hit rate as a function of these latter three independent variables and serial position. The robustness of effects found in analysis (iii) was confirmed by a cross-experiment ANOVA with the variables of material, position, and experiment (iv). For non-match trials, we considered v) correlations between list-probe similarities and correct-rejection rate. Otherwise, data analysis was identical to Exp. 1.

7.3.2 Results

i) A mixed ANOVA indicated that all three factors of group, material, and suppression affected memory fidelity significantly. Figure 7.5 shows the corresponding d' scores. Musicians had higher sensitivity than non-musicians, $F(1, 46) = 25.6$, $p < .001$, $\eta_p^2 = .357$, and recordings were easier to recognize than transformations, $F(1, 46) = 65.0$, $p < .001$, $\eta_p^2 = .586$. There was a main effect of suppression, $F(2, 92) = 13.8$, $p < .001$, $\eta_p^2 = .231$, because the silence condition was both easier than the visual condition, paired $t(47) = 4.01$, $p < .001$, and easier than articulatory suppression, paired $t(47) = 4.96, p < .001$, but there was no difference between visual or articulatory suppression, paired $t(47) = -0.88, p = .383$. There was no interaction.

ii) Response bias was not affected by material, $F(1, 46) < 1$, but was by group, $F(1, 46) = 16.4$, $p < .001$, $\eta_p^2 = .262$, and weakly by suppression condition, $F(2, 92) = 4.68$, $p < .001$, $\eta_p^2 = .092$ (Fig. 7.6). The latter effect arose through significant differences between the silence and counting condition, paired $t(47) = 3.19$, $p = .008$, but no differences otherwise, $p > \alpha_{crit} = .078$.

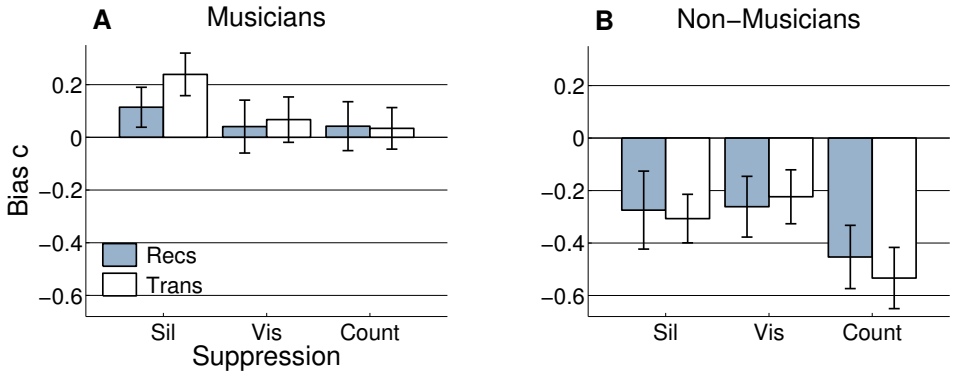

Fig. 7.6 Exp. 2: Response bias as measured with criterion location c for musicians (A) and non-musicians (B) in the suppression conditions of silence, visual suppression, and articulatory suppression (counting).

iii) Regarding effects of serial position, a mixed ANOVA on hit rates did not yield main effects of group, $F(1, 46) < 1$, or suppression, $F(2, 92) = 1.99$, $p = .142$, but did reveal significant effects of position, $F(2, 92) = 44.6$, $p < .001$, $\eta_p^2 = .492$, and material, $F(1, 46) = 32.0$, $p < .001$, $\eta_p^2 = .410$. The effect of position was due to inferior performance in the second position compared to the first and third, paired $t(47) > 8.0$, $p < .001$, but no differences between first and third position, $t < 1$. There was a strong interaction of position and material, $F(2, 92) = 29.1$, $p < .001$, $\eta_p^2 = .387$, that was due to no differences between recordings and transformations in the first serial position, paired $t(47) = -.687$, $p = .49$, but significant differences in the latter two positions, $t(47) > 3.7$, $p < .001$. There was no other significant interaction. Figure 7.3 (C) displays the corresponding hit rates.

iv) The robustness of the position-related effects was confirmed by a post-hoc, cross-experiment ANOVA on hit rate as a function of serial position, material, and experiment, using the subset of musicians from Exp. 2 in the silent suppression condition, and musicians from Exp. 1 in the 6 s delay condition. There were significant main effects of material $F(1, 52) = 43.5$, $p < .001$, $\eta_p^2 = .45$, and position, $F(2, 104) = 31.7$, $p < .001$, $\eta_p^2 = .38$, as well as the interaction of material and position, $F(2, 104) = 23.3$, $p < .001$, $\eta_p^2 = .31$. This interaction arose through significant differences between recordings and transformations in the medial position, paired $t(53) = 7.7$, $p < .001$, but no other significant differences, $p > .06$. Furthermore, there was no significant main effect of experiment, $F(1, 52) = 0.9$, $p = .34$. Although the comparison of panels B) and C)

in Figure 7.3 may suggest a differential effect of position in Exps. 1 and 2 (i.e., an experiment×position interaction), and even differential interactions of material and position across experiments (i.e., a three-way interaction), both two- and three-way interactions failed to fulfill the strict significance level of $\alpha = .01$ and, more importantly, had comparatively small effect sizes, $F(2, 104) < 3.7$, $p > .028$, $\eta_p^2 < .066$.

v) Considering non-match trials, correct rejection rates neither correlated significantly with list-probe dissimilarities for recording trials in any of the three suppression conditions, $r(13) < .511$, $p > .011$, nor was this the case for transformations, $r(13) < .30, p > .29$. The lack of a correlation in Exp. 2 may have been due to its smaller range of dissimilarities (rec: 5.4–6.7, trans: 5.5–6.4) compared to Exp. 1 (rec: 2.8–6.7, trans: 3.7–6.4), where significant correlations were obtained for both groups of sounds.

7.3.3 Summary and discussion

Exp. 2 reproduced the main effect of material on sensitivity, but not on response bias. The interaction of serial position and material from Exp. 1 was also replicated, see Figure 7.3 (panels B and C). A cross-experiment ANOVA further confirmed that this effect was robust across experiments, even though Exp. 1 presented a larger set of stimuli than Exp. 2, and both experiments featured different contextual variables, such as delay in Exp. 1 and suppression in Exp. 2. The position×material interaction suggests that unfamiliar matching probes are particularly difficult to recognize when they are "in the shade" of the medial serial position. Regarding the between-subjects factor of musical training, we observed that musicians featured higher sensitivity and less bias than non-musicians. Note that this is not due to a different approach to the speed-accuracy trade-off, as musicians were also overall faster with a grand average response time of $M = 1358$ ms ($SD = 306$) compared to $M = 1710$ ms ($SD = 337$) for non-musicians, two-sample $t(46) = -3.8, p < .001$.

Contrary to our hypotheses, sensitivity was not affected by an interaction of material and group. This may be surprising at first glance, because one can assume that musicians are more familiar with orchestral instrument sounds (Douglas, 2015) and therefore the difference in their encoding and maintenance of familiar acoustic and unfamiliar synthetic sounds should be particularly large. Nonetheless, considering un-

familiar sounds as a neutral baseline across groups may have been a flawed assumption because musicians possess better auditory skills (Kraus & Chandrasekaran, 2010; Patel, 2012) and may be generally more experienced in memorizing and categorizing sounds, even if novel.

The main effect of suppression was due to reduced performance in both suppression tasks relative to the control condition, and the advantage of recordings persisted throughout all suppression conditions. One could argue that if maintenance of familiar acoustic sounds mainly relied on verbal labeling and subvocal rehearsal, this should lead to strong interference effects by articulatory suppression. If maintenance relied on visual imagery, performance should be strongly disrupted by the visual task. Attentional refreshing, on the contrary, should be moderately disrupted by both types of suppression because articulatory suppression interferes with the very auditory trace to be refreshed, and the visual distractor task reduces those attentional resources that refreshing requires. Refreshing therefore seems to be best supported by the current results.

The finding that articulatory suppression significantly impaired timbre recognition (Exp. 2) is novel and does not cohere with a number of studies (McKeown et al., 2011; Schulze & Tillmann, 2013; Soemer & Saito, 2015). Discerning potential differences with previous studies, it should be first noted that McKeown et al. (2011) used a drastically different experimental scenario. Their task was to discriminate subtle changes in spectral intensity. They tested three participants (two of which were co-authors), and participants underwent daily training for one up to two months with a test phase that lasted for around 10h over 20 days. It thus seems hard to exclude the possibility that their finding—reading aloud does not impair timbre discrimination over long retention intervals—reflects rather specific training effects. Schulze and Tillmann (2013) did not find effects of articulatory suppression in a backward serial recognition task, requiring subjects to match the order of a mentally reversed timbre sequence to a comparison. It is questionable whether participants indeed used refreshing-based strategies in the first place, given that this task certainly constitutes a high working memory load scenario, shown to lead to other maintenance approaches (Camos et al., 2009, 2011). In an experimental design that was relatively close to the current study, Soemer and Saito (2015) only observed a detrimental effect of articulatory suppression in the 2-item list condition (always presented first), but not for lists of length 3 or 4. These results are

particularly hard to reconcile with the current data.

7.4 Questionnaire data

In a post-experiment questionnaire, participants were asked to specify aspects that had made the experimental task difficult. Several free-form responses illustrated the importance of familiarity. A musician noted "The task got easier as I started to have schemas to relate the sound information."(#1, Exp. 2). Another musician (#15, Exp. 1) described memorization as an act of constructing images: "With the synthetic sounds, it was hard to create mental images of what I was listening to, which I found helpful in the acoustic sequences." Others even rephrased the main hypothesis under study (of course not mentioned until the debriefing): "In the synthetic sounds, I didn't really have a point of reference from past experience, and so it was difficult for me to remember the sounds, whereas the acoustic sounds I could easily register in my mind. However, I did notice I had a harder time remembering string sounds than I did with winds [...] which makes sense since I work more closely with winds on a regular basis."(#28, Exp. 1).

We further asked subjects *What kind of strategies did you use to accomplish the task?* The five response options, specific to what was referred to as *acoustic* and *synthetic* sounds, consisted of strategies based on approaches similar to refreshing (*imitating the sounds in my head*), overt vocal imitation (*imitating the sounds out loud*), visual association (*imagining pictures of the instruments*), sensorimotor-association (*imagining playing the instruments*), and verbal labeling and rehearsal (*repeating the names of the instruments*). Figure 7.7 depicts the response choices. Notably, the category covering attentional refreshing was selected most frequently across the two experiments and three groups of participants. Overt imitation was selected least frequently by both groups. Furthermore, musicians from both experiments selected the visual, sensorimotor, and verbal categories much more frequently for recordings compared to transformations. This was not the case for non-musicians, who did not select these latter categories very often.

These reports yield an interesting complementary perspective, although they remain very coarse. The questionnaire only required binary category selection and no judgment on degrees of importance. This does not take into account the fact that some strategies may have simply been used more frequently than others (not to speak

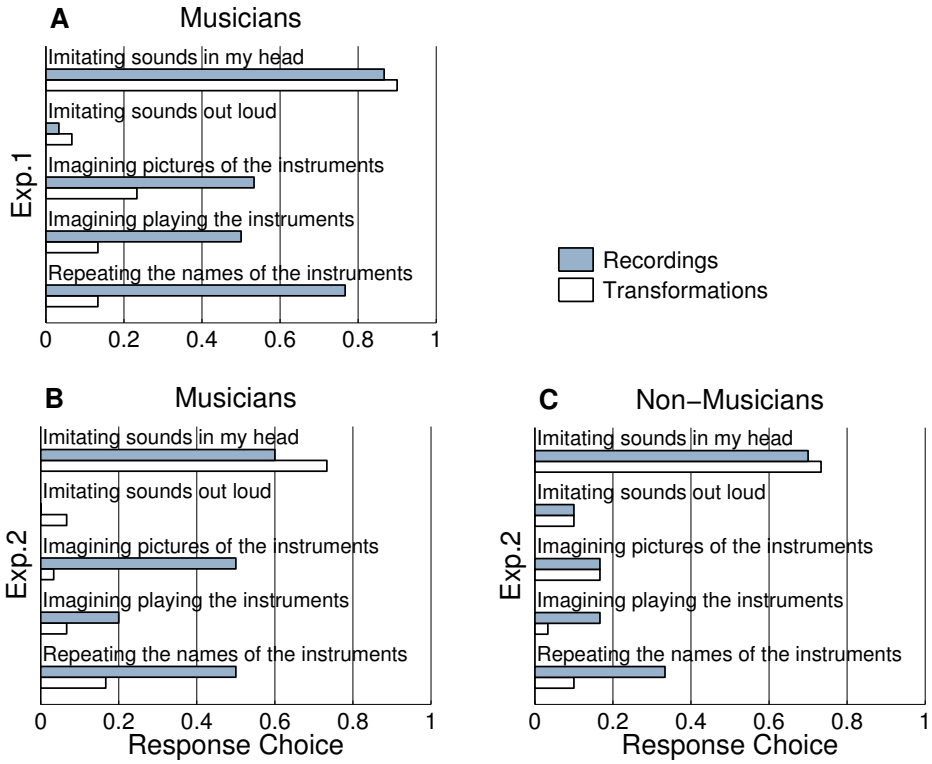

Fig. 7.7 Self reports from post-experiment questionnaires for (A) musicians from Exp. 1, (B) musicians from Exp. 2, (C) non-musicians from Exp. 2. Participants indicated which of the five strategies they had used to accomplish the memory task. Proportion of response choices displayed on the x-axis.

of potential acquiescence biases, letting musicians select categories such as "imagining playing the instrument", for instance, because they are recruited as *musician* participants). Another problem is that responses were not specific to the three suppression conditions in Exp. 2. It is highly likely that participants did not simply endure the interference given by a suppression task, but optimized their maintenance strategy, if possible (Camos et al., 2009). When attentional resources were diminished by the visual task, for instance, some musicians may have indeed relied on verbal labeling and rehearsal, which is less affected by concurrent attentional load (Camos et al., 2011). Nonetheless, the finding that the category that most closely corresponded to refreshing was by far selected most frequently across listeners and material conditions coheres with our interpretation of its role as the most important maintenance strategy for timbre.

7.5 General discussion

We observed a robust recognition advantage for timbres from acoustic instruments compared to timbres from digital transformations. Across material conditions, stimuli were otherwise similar in terms of spectrotemporal envelope properties, temporal fine structure, list-probe dissimilarities, and loudness. Other timbre processing advantages, also independent of musical training, have been reported in the literature. In a series of studies, Weiss and colleagues showed that melodies presented with a vocal timbre were better recognized than the same melodies played by musical instruments (Weiss et al., 2012; Weiss, Vanzella, et al., 2015), and this holds true for musicians and non-musicians. Agus et al. (2012) found that vocal timbres were more rapidly recognized than instrumental timbres. However, the exact mechanisms underlying these vocal processing advantages are not yet understood (cf. Bigand et al., 2011).

The intriguing repetition priming results of Agus et al. (2010) showed that after only a few exposures, participants implicitly learned features of white noise clips, which led to enhanced processing fluency in the detection of clip repetitions. In the current data, response sensitivity did not improve across the two blocks of Exp. 1, questioning the idea that a similar form of processing fluency based on mere exposure could be the sole locus of the current advantage for timbres from acoustic instruments. At the same time, the results from Exp. 2 cast doubts on explanations primarily based on maintenance, for instance via verbal labeling and rehearsal, because both sound types were equally affected by articulatory and visual suppression. Keeping in mind that recognition degraded more strongly for transformations compared to recordings in the least salient medial position of the list, these findings may point towards a different form of encoding for timbres from familiar acoustic instruments. In fact, the self-reports highlighted the idea that in contrast to abstract synthetic sounds, musicians perceived timbres from acoustic instruments as constituting rich arrays of perceptual affordances. Acoustic sounds that possess a type of long-term familiarity and therewith relate to auditory knowledge schemes, may activate not only auditory sensory representations, but to some extent also semantic, visual, and even sensorimotor networks. In consequence, familiar timbres possess more affordances for "deep" encoding. As noted by Craik,

> "Deep processing can be carried out on any type of material: the general principle is that the new information is related conceptually to relevant pre-

> existing schematic knowledge. Thus familiar odors, pictures, melodies and actions are all well remembered if relating to existing bases of meaning at the time of encoding. On the other hand, stimuli that lack an appropriate schematic knowledge base [...], are extremely difficult to remember."(Craik, 2007, p. 131)

Although level-of-processing effects (Craik & Lockhart, 1972) have traditionally been sought in the domain of long-term memory, Rose, Buchsbaum, and Craik (2014) have recently shown that there can be effects of encoding depth (shallow vs. deep, i.e., based on orthographic/phonemic vs. semantic perceptual analysis) on working memory when participants use attentional refreshing.

Beyond advocating a level-of-processing view of timbre, the current study contributes to an emerging picture of attentional refreshing as a primary maintenance strategy for short-term timbre recognition. Refreshing is defined as the attentional reactivation of an item's representation and therefore relies on domain-general attention as well as the fidelity of the respective sensory representation. Should the integrity of either component be disrupted, such as by removal of attention (as in the visual task) or by auditory interference (as in articulatory suppression), the process may be assumed to become prone to errors. Beyond strong evidence for the necessity of attention in timbre recognition tasks from previous studies (Nolden et al., 2013; Golubock & Janata, 2013; Soemer & Saito, 2015), support for attentional refreshing comes from Exp. 2 which found equal disruption of recognition performance by articulatory suppression and visual distractor tasks. The detrimental effect of an attention-demanding visual distractor task, which does not interfere with the auditory trace, appears to be inexplicable with a passive account solely based on sensory decay. The similarity effect on correct rejection rates in Exp. 1 additionally underlines the importance of the distinctiveness of the sounds' auditory sensory representations for response choices and, together with the lack of an interaction of suppression and material, rules out an account of maintenance that solely relies on verbal labelling and rehearsal. The generality of refreshing is supported by the fact that suppression effects occurred regardless of whether familiar recordings or chimæric transformations were used.

By and large, our results suggest that timbre (re)cognition is a multifaceted and active process. It not only functions on the basis of the persistence of sensory features,

but evolves through the interplay of different representational formats, i.e., sensory and sound-source-specific information, attention, and long-term memory. The more a timbre affords multilayered and deep encoding, the more robust becomes its recognition. Short-term memory for timbre should then be seen not as a mere "echo" in the mind of a listener, but rather as a flexible "workspace," which revolves around auditory sensory representations and which trades with a plurality of other mental currencies.

7.6 Appendix: Transformation and selection of sounds

Sound synthesis Transformations were derived from a source signal, the temporal fine structure of which was shaped by the spectrotemporal envelope of a second signal that acted as a time-varying filter. We used MATLAB version R2013a (The MathWorks, Inc., Natick, MA) and a linear 24-band Gammatone-filterbank decomposition (Patterson et al., 1992) as implemented in the MIRtoolbox (Lartillot & Toiviainen, 2007). The temporal fine structure was extracted for every filterband upon which the envelope from the filtering signal was subsequently imposed. The resulting signal thus possessed the spectrotemporal envelope of the filtering signal and the temporal fine structure of the source signal. See (see Z. M. Smith et al., 2002, or Ch. 6 for more details on the transformation process).

Familiarity and identification judgments Among the resulting 441 transformations, we selected 70 to be rated in a dedicated experiment on perceptual familiarity and other variables. The selection was subject to the constraint that every source and filter signal was required to be selected at least once; for recordings acting as filters, each filter was selected at least twice. Additionally, the selection favored timbres that seemed unfamiliar to the experimenters, but did not contain too much narrowband noise (an artifact that was introduced in some transformations by boosting the amplitude of filterbands with low energy). All sounds were normalized in peak amplitude. An experiment assessed perceptual familiarity and source identification of the resulting 70 transformed tones and 14 original recorded acoustic tones. Fifteen musicians participated. In every trial of the experiment, a single stimulus from the 84 tones was presented to participants. They were asked to choose an identifier from a list of eight possible options. The list consisted of six musical instrument names. For recorded timbres, it contained the correct label and five randomly chosen labels from the remaining set. For transformations, it involved the two labels of the timbres that had been involved as source and filter, plus four labels chosen randomly from the remaining set. For instance, if a transformation was derived from a piano as a source, whose time-varying spectral envelope was exchanged with that of a violin, then both instrument names, piano and violin, would be part of the list. The list further contained the two options “unidentifiable” and “identifiable but not contained in list”. If the participant

Table 7.2 List of tones used in Exps. 1 and 2 with mean familiarity ratings. FBS: filterbank scrambling (see text).

	Set 1 (Recordings)			Set 2 (Transformations)			
#	Instrument	Label	Famil.	Source	Filter	Label	Famil.
1	Bass Clarinet	BCL	4.3	Bass Clarinet	FBS2	BCL-FBS2	1.6
2	Bassoon	BSN	3.1	Bassoon	Harpsichord	BSN-HRP	1.9
3	Flute	FLT	4.1	FBS1	Violoncello	FBS1-VCE	1.8
4	Harpsichord	HCD	4.5	FBS2	Violoncello	FBS2-VCE	2.1
5	Horn	HRN	4.2	FBS3	FBS2	FBS3-FBS2	2.1
6	Harp	HRP	4.1	FBS6	Trumpet	FBS6-TRP	1.9
7	Marimba	MBA	4.6	Flute	FBS1	FLT-FBS1	2.1
8	Piano	PNO	4.3	Harp	FBS3	HRP-FBS3	1.7
9	Trumpet	TRP	4.8	Harpsichord	FBS4	HRP-FBS4	2.3
10	Violoncello	VCE	4.7	Horn	FBS6	HRN-FBS6	2.0
11	Violonc. Pizz.	VCP	4.5	Marimba	Harpsichord	MBA-HRP	2.0
12	Vibraphone	VIB	4.3	Trumpet	FBS5	TRP-FBS5	2.3
13	Violin	VLI	3.4	Violin	Piano	VLP-PNO	2.4
14	Violin Pizz.	VLP	4.4	Violoncello	Vibraphone	VCE-VBS	2.0

selected the latter option, a dialogue box appeared prompting them to enter an appropriate identifier in the text box on the screen. They could then continue, whereupon they heard the sound a second time and were presented with two analog-categorical scales on which they had to rate familiarity (1-highly unfamiliar, 5-highly familiar) and artificiality (1-very natural, 5-very artificial).

The 14 transformations that had received the smallest mean familiarity ratings were selected for use in the main experiment, see Table 7.2. The mean familiarity of the 14 recordings ($M = 4.2$, range: 3.1–4.8) was significantly higher than that of the 14 selected transformations ($M = 2.0$, range: 1.6–2.4), as indicated by a two-sided, independent-samples t-test, $t(26) = 15.5, p < .001$.

The mean proportion of "unidentifiable" ratings of recordings and transformations was $M = 0.04$ ($SD = 0.06$) and $M = 0.52$ ($SD = 0.11$), respectively, which constituted a significant difference (independent-samples t-test, $t(26) = 13.8, p < .001$).

Pearson correlations between the proportion of "unidentifiable" votes per stimulus and mean familiarity ratings were strong and negatively associated, $r(82) = -.88, p < .001$, as was the correlation between familiarity and artificiality, $r(82) = -0.86, p < .001$.

Dissimilarity ratings Subsequently, six expert musician listeners equalized perceived loudness of familiar recordings and unfamiliar transformations against a reference sound (marimba) by adjusting the amplitude of the test sound until it matched the loudness of the reference sound. The levels were then set to the median of the loudness adjustments.

In order to be able to control for perceptual similarity among timbres, 24 musicians rated pairwise dissimilarity for both sets of sounds. Sets were presented separately, and the order of sets was counterbalanced across participants. The 105 pairs of stimuli (14 identical, 91 non-identical) were presented at a 300-ms inter-stimulus-interval and participants provided dissimilarity ratings on an analog-categorical scale (1-identical, 9-very dissimilar). The order of stimulus presentation (AB vs. BA) was counterbalanced across participants. See Siedenburg, Jones-Mollerup, McAdams (in prep., Ch. 6) for more details on individual sounds and their familiarity and dissimilarity relations.

Part IV

Conclusion

Chapter 8

Facets of memory for musical timbre

This final chapter comprises a summary, two coordinate transformations, and a vista point. The first section summarizes the most important points from Chapters 2–7 (in "by-experiment coordinates"). The second section transforms this description into a depiction of the current contributions to the literature and yields an account of the experimental factors that affect short-term memory for timbre (in "by-factor coordinates"). The third section transforms the empirical findings into a more abstract account in order to identify a few cognitive processes that are central to memory for timbre (yielding "by-process" coordinates). A final vista discusses implications for theories of music listening.

8.1 Summary

This thesis studied musical timbre cognition and memory for timbre from the perspective of short-term recognition and dissimilarity rating tasks. Featuring four independent chapters, which framed nine separate listening experiments, this thesis provided detailed investigations into the role of a) timbre similarity and concurrent pitch variability in short-term memory for timbre, b) the impact of sound source categories and familiarity of tones and sequences in timbre recognition and dissimilarity ratings, and c) the musical experience of participants and their memory maintenance strategies. Several links to hallmark effects of verbal memory were established, including acoustic similarity, sequential chunking, lexicality, and active memory maintenance. The findings portray the memory processes under study as multifaceted and highly interac-

tive, operating from short to long time scales, and extending from sensory to semantic representations. The main results will now be summarized.

Part I

The first theoretical part provided a background on the notion of timbre and on previous work on memory for timbre. The essay in Chapter 2 advocated for a distinction between i) a sound event and its perceived timbre, ii) qualitative and source timbre, and iii) different scales of timbral detail. It was proposed that these three conceptual axes constitute an adequate taxonomy for timbre.

Chapter 3 outlined basic concepts in memory research and discussed recent findings in auditory memory that bear relevance for accounts of timbre cognition. Subsequently, a comprehensive map of previous research on memory for musical timbre was described, and the main research questions were derived. These address how timbre cognition is affected by the factors of i) timbre dissimilarity, ii) sound source categories and the familiarity of tones, iii) concurrent variability in pitch, iv) the musical training or expertise of the listener, and v) attentional maintenance.

Part II

The second part described experiments on short-term recognition memory for timbre sequences. Using a serial-order recognition task, Exp. 1 showed that musicians did not differ from nonmusicians on sequences with constant pitch, but were better than nonmusicians in matching sequences that featured concurrent pitch variability (identical for standard and comparison sequences). Exp. 2 yielded a significant effect of pitch variability for musicians when pitch templates differed across standard and comparison sequences. Exps. 3 and 4 highlighted how response-choice behavior is affected by the perceptual dissimilarity of items that swapped positions (accounting for around 90% of the variance of responses across the four experiments), but did not find any effects of timbral heterogeneity of the sequence. These results demonstrate the importance of controlling stimuli for their perceptual similarity relations in memory studies, and highlight strong commonalities of principles found in the domains of timbre perception and short-term memory.

As a musical case study, Chapter 5 explored auditory and verbal memory for North

Indian *tabla*, testing tabla students and musicians naïve to tabla. Whereas the other experimental chapters studied the processing of timbral contrast arising from different musical instruments (or emulations and transformations thereof), this chapter zoomed into the sound world of the voice and the tabla, and studied timbre sequences composed on the level of timbral "species" (as opposed to timbral contrast arising from the comparison of sounds from different instruments or instrument families, see the third distinction of Ch. 2). For investigating the role of familiarity and chunking in the cognitive sequencing of tabla, idiomatic tabla sequences of (verbal) bols and (instrumental) drum strokes were compared with: i) counterparts reversed in order, ii) sequences with random order and identical item content, and iii) randomly selected items without replacement. A strong main effect of sequence type emerged that featured monotonically decaying performance (i>ii>iii), underlining the importance of chunking in auditory serial recognition. Furthermore, differences between tabla players and musicians primarily emerged for idiomatic sequences of bols, suggesting a familiarity effect for verbal material, but not for instrumental musical timbres. This result points towards a partial dissociation of memory for musical and verbal sounds.

Part III

The third experimental portion of this thesis explored the ways in which sound source categories of familiar acoustic tones affect timbre dissimilarity ratings and short-term item recognition. Chapter 6 may be seen as an empirical elaboration of the distinction between timbre as auditory quality and timbre as a cue for source recognition highlighted in Chapter 2. Considering timbre dissimilarity ratings for groups of tones from familiar acoustic instruments and unfamiliar digital transformations, rating asymmetries were observed that cannot be explained on acoustical grounds alone (Exp. 2A), and were replicated in an altered design (Exp. 2B). Correspondingly, descriptors related to sound source categories significantly improved an acoustic model of timbre dissimilarity. This yielded evidence that timbre dissimilarity of familiar acoustic tones draws upon both sensory and categorical factors. A novel model of timbre dissimilarity was introduced in order to compare the contributions of acoustic and categorical timbre descriptors. Using partial least-squares regression, the best model fit ($R^2 = .88$) was achieved when both types of descriptors were taken into account.

Continuing with the same set of familiar acoustic tones and unfamiliar synthetic transformations, Chapter 7 explored the role of source categories and maintenance strategies in memory for timbre, comparing musicians and non-musicians in a short-term item-recognition task. In Exp. 1, musicians better recognized acoustic recordings compared to synthetic transformations, and this effect of material was particularly large in the medial serial position. There was a strong correlation of correct rejection rates and the mean perceptual dissimilarity of the probe to the tones from the sequence, which extends the findings on similarity from Chapter 4 to item recognition. Exp. 2 showed that musicians recognized timbres better than non-musicians, regardless of the concurrent suppression task (i.e., articulatory suppression, visual interference, or a silent interval), or the material type (familiar recordings, unfamiliar transformations). These results were interpreted as evidence for the importance of attention-based maintenance of timbre in STM.

8.2 Factors that affect timbre cognition

I will now provide a state of affairs of factors that affect STM for timbre and to a lesser extent, timbre dissimilarity ratings, and discuss the current contributions to the literature. I will start with an overview, before individual variables are discussed in depth.

8.2.1 Overview

Factors influencing timbre dissimilarity ratings There exists a broad literature on timbre dissimilarity perception that cannot be comprehensively summarized here. See McAdams (2013) for an overview, Siedenburg et al. (2015) for a recent review of acoustic variables, and Donnadieu (2008) and Giordano and McAdams (2010) for categorical factors. The innovative contributions of this thesis include:

- the description of systematic rating asymmetries for across-category comparisons (of acoustic and synthetic sounds); and
- a predictive model of acoustic and categorical timbre dissimilarity that generalizes across different sets of sounds.

Factors influencing short-term memory for timbre Most studies on STM for timbre include factors of retention time and sequence length (see Table 3.1), such that there is a relatively good empirical basis for these generic memory factors. The focus of this thesis, on the contrary, was the study of timbre-specific factors in memory tasks, as well as the between-subject factor of musical training. The following list contains the respective experimental variables, the chapters where these variables were addressed, and references to studies that have investigated these factors previously. Positive findings are listed first, null results second (in italics).

- Retention time: Ch. 7; Demany et al. (2008); McKeown et al. (2011); Golubock and Janata (2013); Mercer and McKeown (2014); Soemer and Saito (2015).
- Sequence length: Ch. 4; Marin et al. (2012); Nolden et al. (2013); Golubock and Janata (2013), but also see, *Schulze and Tillmann (2013)*.
- Similarity: Ch. 4; Ch. 7.
- Sound source categories and familiarity: Ch. 7.
- Stimulus type: Ch. 5; Ch. 7; Schulze and Tillmann (2013); Golubock and Janata (2013).
- Sequential structure: Ch. 5.
- Serial position: Ch. 4; Ch. 7.
- Concurrent variability in pitch: Ch. 4; but also see, *Starr and Pitt (1997)*.
- Concurrent suppression: Ch. 7; Soemer and Saito (2015); but also see, *McKeown et al. (2011); Schulze and Tillmann (2013)*.
- Musical training: Ch. 4; Ch. 5; Ch. 7; but also see, *Starr and Pitt (1997)*.

8.2.2 Discussion

Retention time Although it appears to be hard to draw an exact line between early "sensory" types and more long-lasting forms of memory for timbre (see the discussion

in Ch. 3, and Demany et al., 2008, 2010), there is convergent evidence for the deterioration of timbre in STM in the absence of interference (McKeown et al., 2011; Golubock & Janata, 2013; Mercer & McKeown, 2014; Soemer & Saito, 2015). The rate of this process appears to be moderate. For instance, Soemer and Saito (2015) observed a decrease of ten percentage points in item-recognition accuracy for an increase of retention time from 3 to 12 s. Our results (Ch. 7) compared 2 and 6 s of delay and yielded a decrease of d' scores from around 2.1 to 1.6. Golubock and Janata (2013) estimated working memory capacities of 1.7 to 1.3 items for 1 and 6 s delay, respectively. McKeown et al. (2011) even noticed above chance accuracy for the retention of small spectral details after a delay of 30 s. Overall, these results imply that timbre degrades at a moderate pace, and generally that STM for timbre is more long lasting than traditional accounts of memory would have suggested (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974).

Sequence length Two previous studies which have used item recognition found effects of sequence length (Golubock & Janata, 2013; Soemer & Saito, 2015). Using serial recognition, we obtained an effect of length (Ch. 4), similar to the results from Marin et al. (2012) and Nolden et al. (2013), but contrary to Schulze and Tillmann (2013) who observed effects of length for verbal material and items differing in pitch, but not for timbre. There are a couple of differences in design between these experiments, such that it is hard to specifically pinpoint the experimental sources for this divide.

Similarity We showed that trial-wise similarity relations are a major determinant of response choices in short-term recognition of timbre. In serial recognition, the dissimilarity of the items that swapped order was a powerful predictor and has to our knowledge not been considered elsewhere in the literature on short-term memory for serial order. In the item-recognition task of Chapter 7, the mean dissimilarity of the probe to the items of the list correlated significantly with response choice, extending results from (Visscher et al., 2007). Nonetheless, the correlation was weaker than in serial recognition and only significant when the range of dissimilarities was large enough. Overall, these findings also cohere with the *relative distinctiveness principle* (Surprenant & Neath, 2009), which derives memory accuracy as a function of an item's distinctiveness relative to a background, such that distinct items can be assumed to

be generally better recognized. Contrary to reports originating from the literature on visual STM (Kahana & Sekuler, 2002; Visscher et al., 2007), we did not find that sequence heterogeneity plays a significant role, neither in serial nor in item recognition.

Starr and Pitt (1997) were the only other authors that investigated effects of variation within a timbre-specific dimension, namely the brightness of digitally synthesized tones and its effect in an interpolated-tone task. The closer in brightness the interfering tones were to the target, the more detrimental their effect. Otherwise, studies that presented sequences of items (i.e., using serial or item recognition) ensured (in different ways) that items were discriminable. The similarity structure within trials had not been considered previously. Chapters 4 and 7 show that it plays a substantial role in participants' response choices. An interesting future project would be to specify whether there is a differential memory capacity for sets of stimuli that only vary with respect to subcomponents of timbral dissimilarity, such as spectral or temporal envelope properties.

Source categories We showed that sound source categories of familiar acoustic tones play a role in both dissimilarity rating experiments and short-term recognition. In the former, robust rating asymmetries arose for comparisons of tones from the set of familiar acoustic recordings and unfamiliar synthetic transformations, and a regression model suggested that musicians' ratings of acoustic pairs were affected by dissimilarities of source categories. These results inform the broad range of work on timbre dissimilarity that has worked with both digitally synthesized and recorded acoustic tones but mostly sought to explain responses solely in terms of acoustic models.

An open question is whether this finding is specific to musicians who may be suspected to more readily infer source categories than nonmusicians. Working with environmental sounds, Lemaitre et al. (2010) found that *expert listeners* (i.e., musicians or researchers working on sound) more strongly relied on acoustic properties than nonexperts, who tended to use categorical similarities, if available. Future work should therefore determine whether types of listening expertise affect tendencies to hear timbre through the ears of "musical" or "everyday" listening (Gaver, 1993), i.e., with a focus on the qualia arising from acoustic properties or source categories, respectively.

In an item-recognition task, we found that the timbre of familiar acoustic sounds was better recognized than that of unfamiliar transformed sounds. This was interpreted

as a levels-of-processing effect based on the rich array of affordances (sensory, semantic, motor, visual, etc.) that acoustic sounds from familiar instruments offer the listener.

Stimulus types Let us now more generally discuss whether different types of stimuli affect STM for timbre. Schulze and Tillmann (2013) compared verbal material (words) with tones differing in pitch and timbre in a serial recognition task. Although the word and timbre conditions did not differ in absolute terms, it was shown that recognition was differentially affected by task manipulations (backward recognition and backward recognition paired with articulatory suppression). Similar results were found in Chapter 5 for tabla, where verbal bols and drum strokes only differed for students of tabla acquainted with the bol "language", and even for them only for idiomatic phrases. This suggests that there may be no general mnemonic advantage for vocal timbre in serial recognition, but that specific task demands may give rise to vocal superiority. These include situations of high concurrent processing load such as in the backward suppression task from Schulze and Tillmann (2013), or when there is long-term knowledge about sequential structure as in the case of idiomatic tabla phrases.

Golubock and Janata (2013) are the only other authors who (indirectly) tested effects of stimulus type. Their first experiment used digitally synthesized sound that varied on the three dimensions of attack time, spectral centroid, and spectral flux. In the second experiment, a more variable set of tones was compiled from a number of commercial synthesizers, and this set yielded higher working memory capacity estimates in a post-hoc comparison. Although the authors attempted to exclude obviously familiar sounds from the second set, the degree to which the selected sounds were more familiar and elicited source categories remains unclear. In light of the previous discussion on similarity and source categories, it seems reasonable to infer that this comparison conflated aspects of sensory similarity (what Golubock & Janata, 2013, called "perceptual variability") and the affordance for source identification (which they suspected to "more likely activate LTM representations of additional semantic properties such as word labels or other associated concepts", p. 407). Chapter 7 attempted to disentangle both aspects to a greater extent by using two sets of tones that explicitly differed with regards to source identifiability but had similar acoustic properties and were equated in terms of their intrinsic perceptual dissimilarity relations.

It would be fair to object that this only does half the job, because it was already

highlighted that perceptual dissimilarity ratings are to some degree also determined by source categories. Nonetheless, recall that we observed a dissociation of these variables in Exp. 1 of Chapter 7: Contrary to non-match trials where list-probe dissimilarity correlated significantly with performance for both types of materials (familiar recordings vs. unfamiliar transformations), we did not find that similarity played a role for match trials (e.g., as a function of list heterogeneity). At the same time, there was a significant effect of material for match trials. Both variables, sensory dissimilarity and source categories, can therefore be assumed to have conjointly affected these results.

As pointed out in Chapter 6, it may be impossible to fully disentangle these components on the stimulus side. Future work on the role of sound categories and contributions of LTM to short-term recognition may therefore consider learning paradigms such that participants associate artificially synthesized sounds with potential source/cause categories, before comparing them with naïve controls in a recognition task. This would control the contribution of acoustic dissimilarity by operationalizing stimulus familiarity as a between-subjects variable, as has been done in the implicit learning of timbre sequences (Tillmann & McAdams, 2004).

Sequential structure The tabla project (Ch. 5) showed that sequential structure has a strong effect on serial recognition of timbre sequences. Idiomatic sequences of tabla strokes and their reversed versions were recognized best, followed by their counterparts with randomly shuffled order, followed by fully random sequences without repetitions of items. The latter advantage was interpreted as a sign of facilitated chunking due to the repetition of items. Because we tested serial-order recognition, we assume that the advantage of redundancy primarily goes back to chunking and not a reduced load in terms of item identity. The advantage of reversed sequences over randomly shuffled ones was suspected to be related to the hierarchical structure inherent in the idiomatic sequences or their reversed versions. They not only contained item repetitions, but repeating chunks of items, such that sequences could be encoded hierarchically. From an attentional perspective, listeners could attune to more global attentional cycles (Jones & Boltz, 1989). Notably, effects of familiarity with idiomatic sequences (which could have only occurred for the group of tabla students, but not for naïve controls), only occurred for the vocal sounds but not for the drum sounds. This result suggests partially dissociated mechanisms in the cognitive sequencing of verbal

and musical stimuli.

Serial position We observed effects of serial position both in item recognition (Ch. 7) as well as in serial recognition (Ch. 4) . However, because these experiments worked with short sequences (3–4 items), serial position curves were not very detailed. To the best of our knowledge, no other effects of serial position have been reported in the literature.

Concurrent variability in pitch As outlined in Chapters 2 and 4, the perceptual literature has by now converged towards the position that pitch and timbre mutually interfere in discrimination tasks (see, Allen & Oxenham, 2014, for the most recent manifestation thereof). In STM tasks, on the contrary, two studies did not find interference. Semal and Demany (1991) tested STM for pitch with an interpolated tone task and showed that performance was independent of the spectral proximity of the interpolating tone to the target. Using the same task for timbre, Starr and Pitt (1997) did not find interference from changes in pitch on the recognition of the target timbre, suggesting the independence of timbre from pitch.

The review on the role of timbre in memory for melodies (Ch. 3) outlined a different picture, and highlighted the finding that a change of timbre robustly impairs long-term melody recognition. This indicates that melodies can be retained as holistic auditory images, rather than as abstract entries in a melodic lexicon. The current findings regarding the role of concurrent pitch variability (Ch. 4) put forward a similar conclusion, and indeed almost seamlessly extrapolate the perceptual literature towards short-term memory. We observed that even highly trained musicians are not immune to cross-channel interference, if there is substantial variability in pitch. In the findings of Allen and Oxenham (2014), musicians had lower difference limens for pitch than nonmusicians. But if variation in the non-attended condition was adjusted as a multiple of the individual threshold, interference from pitch to timbre was equally high across groups. In other words: musicians need more substantial variability in pitch to exhibit interference on timbre in basic discrimination *and* in short-term recognition.

Concurrent suppression Chapter 7 observed detrimental performance in visual and articulatory suppression tasks, irrespective of stimulus type (familiar recordings

vs. unfamiliar transformations). Self reports of participants confirmed this interpretation in favor of attentional refreshing as the central maintenance mechanism in STM for timbre, also coherent with Soemer and Saito (2015) and Nolden et al. (2013). Other studies did not find effects of articulatory suppression (McKeown et al., 2011; Schulze & Tillmann, 2013), although these studies used markedly different tasks. For instance, Schulze and Tillmann (2013) only employed suppression in a backward serial recognition task (obliging the participants to mentally reverse sequences). The design used by McKeown et al. (2011) differed even more drastically, testing memory for the most subtle kind of timbral detail (also see Ch. 3). In any case, the unanimous point of agreement among all studies is that maintenance of timbre is not primarily guided by identification of verbal labels and subsequent verbal rehearsal, as traditional models of memory would have suggested.

Musical expertise Only one previous experiment explicitly tested effects of expertise on memory for timbre as a between-subjects factor, and it obtained non-significant differences between groups (Starr & Pitt, 1997). On the contrary, our results revealed effects of musical expertise in three experiments. In item recognition (Ch. 7), there was a main effect of group and no interaction with other variables: musicians better recognized the timbre of probes than did non-musicians, regardless of stimulus material or suppression condition. In a serial-recognition task with emulated orchestral tones (Ch. 4), there was only a weak main effect of group, but a strong interaction of the factor of group and pitch variability, as discussed above: Musicians were more robust than nonmusicians, albeit not fully immune to concurrent variability in pitch. The tabla project, also using serial recognition, compared tabla students at a beginners level with a matched group of musicians who were naïve to tabla. The project thus specifically tested for the role of experience with a musical style system. Surprisingly, our results suggested that the experience of tabla players with the stylistic conventions and with the sound material only facilitated performance in the condition of idiomatic bols. In summary, item recognition yielded a main effect of group, whereas in serial recognition the role of expertise only seemed to play out in more delicate interactions.

One interpretation would be that item recognition requires more "analytic" forms of listening than basic serial recognition. In item recognition, the probe must be matched with the elements of the list. One model of item recognition assumes that

subjects sum up the similarity of the probe to all elements of the list (Kahana, 2012, Ch. 2). If it exceeds a criterion level, the probe is considered part of the list (which coheres with the role of similarity in Ch. 7, Exp. 1). This requires a comparison of a context-less probe tone with elements of a sequence of tones; these elements must therefore be "analytically" selected from the sequence before being matched to the probe. Describing the dynamics of attentional cycles, Jones and Boltz (1989) contrasted *analytic attending*, focused on low levels of the temporal hierarchy of a sequence, with the more global focus of *future-oriented attending*. Accordingly, listeners differentially *attune* to time scales of auditory sequences, depending on their goals and the perceptual affordances of the stimuli.

In serial recognition, on the contrary, the whole standard sequence can be matched with the comparison, thus affording a holistic or unitary type of matching process (cf., Warren, 1974). Note that the dissimilarity-based predictor of response choice (the dissimilarity of the swap) is coherent with this hypothesis, because the error of global matching can be assumed to correspond to the item-wise dissimilarities of two sequences, which again correlates with our measure of the dissimilarity of the swap. It might be the case that such a holistic matching process (which can be thought of as one act of matching, instead of multiple acts as in item recognition) may be less reliant on skills in what was tentatively called "analytic listening" above. For that reason, it is only when additional conditions are introduced, such as the challenge of the concurrent variability in pitch, or the affordance of verbal material to be efficiently abstracted into words spanning multiple items, that types of expertise become relevant in serial recognition. In Chapter 4, we noted that musicians may have circumvented a global matching strategy in the conditions of concurrent variability in pitch.

This interpretation is in line with findings on the task dependency of the lexicality effect (which, in a sense, could be viewed as a type of within-subject operationalization of linguistic expertise): Auditorily presented words are better recalled than pseudo-words, but there is no such effect for auditory serial recognition (Gathercole, Pickering, Hall, & Peaker, 2001). Macken et al. (2014) provided evidence that this lack of an effect in serial recognition could be based on different strategies in the two tasks. Whereas in serial recall subjects encode lists segmentally (i.e., item by item), serial recognition affords global matching of the full sequences. This result would support the idea that auditory serial recognition is less strongly affected by expertise (tentatively interpreted

as comprising familiarity with the stimulus materials) than item recognition, where an isolated probe stimulus must be matched.

An important caveat must be added. In the tabla project, we noticed an effect of experience (i.e., tabla players better recognized idiomatic bols than reversed sequences of bols), which contradicts the above argument at first glance. What differentiates the situation in tabla from that of serial recognition of words and pseudo-words is that by virtue of experience, idiomatic tabla sequences can be abstracted into far fewer items (because idiomatic strings of bols constitute "tabla words"). This means that tabla players could encode idiomatic bols as a succession of far fewer items than naïve controls. In tests of the lexicality effect, however, there is no way to combine items into higher-order chunks.

8.3 Processes and principles in memory for timbre

This section characterizes four processes that may be seen as "cognitive undercurrents" for the empirical situations described above. These include heterogeneous representation, chunking, attentional refreshing, and similarity-based matching. The goal is to provide a tentative outline of how a few general ideas cover the life-cycle of memory representation, maintenance, and reactivation in the realm of timbre. Characteristic traits or principles of memory for timbre are discussed subsequently.

8.3.1 Processes

Part III suggested that timbre cognition transcends the domain of sensory representations. Although this does not attempt to question that the main currency of timbre cognition is of auditory sensory nature, acknowledging its inferential and associative tendencies is important for the development of a comprehensive account. The perception of a violin tone is more than the construction of an auditory image. It includes the activation of nodes in a hierarchically ordered, multimodal network. To recall from Chapter 6, the listening brain represents, by virtue of its faculty for massive parallel processing, "the sound" and the many "ideas" of a musical instrument. Fuster (2003) calls this "an associative conglomerate of sensory and semantic features at many levels of the cognitive hierarchy of perceptual knowledge."(p. 124)

It is self evident that the separate layers spanned by such associative networks encode multidimensional attributes by themselves. The sensory representation of timbre, the multidimensional auditory attribute per se, is composed of bundles of auditory attributes that encode spectral, temporal, and spectrotemporal stimulus features. Assuming that long-term learning drives this associative, multilayered representation, it is clear that certain classes of sounds lack definite content at certain layers, and that there are inter-individual differences. This association of different formats of representations via long-term learning may also be interpreted as a form of *long-term heterogeneous association*.

Whereas timbre affords for a heterogeneous, multilayered representation in principle, only a few bundles of features may be functional in STM. We thus need to assume a form of selective attention that weights individual nodes in the network, or bundles of features, depending on task demands. These weights may be assumed to be modulated by endogenous and exogenous attentional factors and therewith change dynamically over time.

Variability in pitch is an example of an exogenous factor that affects the processing of timbre (Ch. 4). Melara and Marks (1990) conceptualized these forms of interference as crosstalk between different auditory "channels". For interacting perceptual dimensions, the output of one channel is weighted by the output of another, thereby modeling failures of selective attention in the form of interference. Nonetheless, attentional channel selection by itself is assumed to be robust, meaning that subjects do not mistakenly attend to the irrelevant attribute. Rather the output of the target channel is automatically modulated by the unattended channel, most likely due to crosstalk at sensory levels of processing.

To some degree, the resulting type of selective attention correlates with the modes of auditory attending that have been called *musical* and *environmental* listening (Gaver, 1993). Let us consider Murail's *Mémoire/Erosion* as an example (see Ch. 1). When a musician listens to the piece for the first time, she may instantly identify the French horn, but not the following clarinet for its reduced intensity and uncommonly staggered articulation, leaving her in doubt about the source. In the introductory section of the piece, a first impulse could then be to scan the musical scene for *what* is there, that is, to focus on instrument identity rather than auditory quality. A longer stretch into the piece, when instrument identities are already established, the focus may "dive in" and

yield greater weights for sensory representations.

A second important associative process is directed in time and binds items within a sequence. The data from the tabla project suggested that short-term recognition of timbre strongly benefits from different types of sequential chunking. To recall, Cowan (2001) defines a chunk as a "collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use."(p. 89) In that sense, chunking relies on intra-segmental association and inter-segmental dissociation of serially ordered items. We assumed that three types of chunking played a role in Chapter 5. The most elementary type was based on redundancy of items within an otherwise randomly shuffled sequence and the resulting affordances for structural grouping in short-term memory. The second referred to the hierarchical structuring of sequences, which could be grouped into repeating chunks of items. This simple type of temporal recursion may yield *attunement shifts* such that listeners attentional cycles evolve on more global time scales (Jones & Boltz, 1989). Such structures then may give rise to future-oriented listening, or, in a stronger form, dynamic expectations that can arise within the context of a single musical phrase (see e.g., Huron, 2006). The third type is based on long-term learning of idiomatic chunks (although we only observed advantages for verbal materials). It can be assumed to draw from a lexicon of temporal patterns (McAdams, 1989) and may contribute to long-term schematic expectations (e.g., Huron, 2006). For the sake of contrast with the *heterogeneous* type of association of different representational contents described above, these various facets of chunking could also be summarized as *temporal association*.

Chunking does of course not work without a basic form of temporal integration that registers the appearance of chunks. It is this "primitive" kind of STM that is studied in most memory tasks that circumvent sequential structure and thus present trials with unique items per sequence (e.g., see Chs. 4 and 7). In fact, only if experimental procedures ensure that the presented items cannot be grouped into higher-order chunks, is it possible to observe "pure" STM capacity limits (Cowan, 2001). Therefore, it is natural to think of temporal association as a process of second order that may enhance the longevity of STM, but does not constitute its basis. Yet it powerfully alters the very representation that is to be maintained by structuring the trace along a hierarchy of time scales (if this sounds abstract, think about how to memorize, ABCXYZABCQ).

The current findings on maintenance processes (Ch. 7) and the discussed behavioral and neurophysiological evidence suggest that auditory refreshing—a domain-specific type of attentional maintenance that reactivates current content in STM (as opposed to the notion of *imagery* which is more frequently used in conjunction with LTM)—is the primary maintenance mechanism in STM for timbre. It is defined as the temporary reactivation of a trace by means of allocation of attention. As previously, this process must be of second order, because there needs to be a trace (i.e., memory persistence) to which attention can be directed in the first place.

This active process, grounded in sensory representations, relates to the notion of *perceptual simulation*, defined as a recreation of schematic facets of perceptual experience. Theories of perceptual symbol systems indeed advocate that cognition is grounded in perceptual simulation (Barsalou, 1999, 2008), contrary to classic theories of cognition, which assert that perception leads to a transduction of sensory states into configurations of amodal symbols. Theories of perceptual symbol systems posit that perceptual learning abstracts schemata or perceptual symbols from sensory states. These schematic representations are subsequently simulated or "restaged". The concept of a chair thus is not assumed to activate neighboring nodes in a semantic network or a list of features inherent to the concept of chair, but it runs a (not necessarily conscious) perceptual simulation that generates a schematic sensory image of a chair upon which further mental processing can take place. This idea is coherent with the previously discussed concept of heterogeneous representation because activation spreads in both bottom-up and top-down directions. In the absence of direct stimulation, perceptual simulation then corresponds to a systems-level description of the top-down activation of sensory representations.

One could question the extent to which attentional reactivation of previously encountered items in the presence of concurrent interference may be an artifact of experimental settings that require participants to accomplish a memory task. In music listening, there (usually) is no recognition test at the end of the phrase. In that sense, the cited studies and the current results should be understood as pointing towards mental capacities: Listeners are able to refresh or simulate past (or future) events with some accuracy even in the midst of concurrent perceptual processing. Such an active type of memory may be used in music listening as an "auditory workspace" that allows listeners to actively explore musical scenes, create relations between past and present

auditory events, and seek their individual pathways through the musical landscape. As noted by McAdams (1984), "Musical listening (as well as viewing visual arts or reading poetry) is, and must be considered seriously by any artist as, a creative act on the part of the participant. [...] Perceiving a work of art can involve conscious and willful acts of composition."(p. 213) STM provides parts of the cognitive infrastructure for this creative endeavor.

If the life cycle of memory comprises the stages of encoding, maintenance, and reactivation, then the matching process corresponds to the last component of the cycle. It therefore depends on the properties of the previously described processes of heterogeneous representation, chunking, and attentional maintenance.

The similarity effects discussed above suggested that a similarity-based matching mechanism could underlie serial and item recognition. In abstract terms, the properties of the memory trace, including the context of encoding and the various levels of heterogeneous and temporal association and attentional weights, could be modeled (at least in principle) by a fairly high-dimensional vector. The matching process could then be conceived of as a summed similarity computation in item recognition (Kahana, 2012), or a global or item-wise matching in serial recognition that is well approximated by the dissimilarity of the swap.

Matching strategies further vary as a function of stimulus affordances, concomitant variation in other auditory attributes, as well as listeners' "auditory skills". If we interpret the similarity rating task as a particular type of matching for a moment, Chapter 6 provided evidence that heterogeneous types of information representations are integrated into such judgments. We interpreted the results from Chapter 4 as pointing towards partially different strategies for musicians and nonmusicians. We suspected that in the face of strong concurrent variability in pitch, musicians could rely more on "analytic" item-by-item based encoding, i.e., focus on local levels of the attentional hierarchy. In the tabla project (Ch. 5), to the contrary, tabla players' knowledge of the bol "language" enabled them to encode and match higher-level chunks that frequently co-occur in the repertoire. In that sense, musical expertise may be suspected to not necessarily favor one level of the attentional hierarchy over the other, but rather to more flexibly select the most reliable level.

8.3.2 Principles

Let us note that temporal association and heterogeneous representation not only share the fundamental mnemonic principle of association, but also share similarities in function. Both generate the representational bases for further processing. Whereas heterogeneous association results in the parallel processing of layers of different information types, which vary in their degree of abstraction, temporal association abstracts from the granularity of the item level and creates novel representations of a coarser temporal resolution, i.e., a map of the event structure of the immediate past.

Different classes of auditory stimuli feature different predispositions for heterogeneous and temporal association. For instance, the representational layers of speech are highly permeable. Listeners proficient with a language do not pay attention to the auditory sensory facets of speech signals, but mostly to its heterogeneous semantic associations and implications. The temporal scales of language are readily associable, in the sense that multiple words can easily be abstracted into single semantic items (think of the "ice cream factory owner"). Part III attempted to create differential predispositions for heterogeneous association in the realm of timbre, and Chapter 5 specifically explored the effects of differential affordances for temporal association.

Summarizing the above discussion of general processes in memory for timbre, both heterogeneous representation and temporal association contribute to information representation, and matching could be assumed to operate via a similarity computation on such representations. Maintenance is naturally attributed to the workings of attentional selection and perceptual simulation. Overall, this description resonates with a dynamic procedural approach that conceives of memory as an emergent property of the ways in which processes like the ones discussed interact with specialized perceptual representations. To summarize this point by means of analogy,

> "There are no dedicated RAM circuits in the primate brain. Rather, the operation of holding information in working memory occurs within the same circuits that process that information in nonmnenomic contexts."(D'Esposito & Postle, 2015).

In contrast to memory for pitch and verbal material, the ways to rehearse timbre by means of motor-based re-encoding (i.e., subvocal singing or articulation, cf., Williamson et al., 2010; Schulze & Koelsch, 2012) seem to be more severely limited. Listeners

may mimic brightness contours, for example, but it may be generally harder to vocally reproduce the articulate spectrotemporal morphology of a complex sound event than to reproduce pitch or to subvocalize words. Nonetheless, several links to important effects of verbal memory were established, including acoustic similarity, sequential chunking, lexicality, and active memory maintenance, suggesting that these variables tap into domain-general principles of memory.

This aspect provokes us to draw a link to a pertinent theme of music cognition, namely the overlap of the cognitive mechanisms involved in music and language processing. Patel (2003, 2008) proposed that cognitive mechanisms for the parsing and integration of linguistic and musical events operate on specialized representations (stored in long-term memory). Whereas these representations are distinct, the interpretative processes (roughly considered as part of working memory) that integrate and interpret auditory sequences of sounds share cognitive resources.

Williamson et al. (2010) argued in favor of considering STM as one component of this question. The authors interpreted their results on STM for pitch as evidence for a pitch store that interacts with the maintenance process of rehearsal. Although we do not share the view of memory as composed by various distinct buffers, we do agree that STM is a worthy contributor to questions on the parallels of music and language as cognitive systems. Overall, our description coheres with Patel's hypothesis in the sense that it fosters a view in terms of perceptual representations that interact with long-term memory and of integrative processes that are at the basis of auditory sequencing. The data from the tabla project (Ch. 5) further supported this view in that differences between verbal bols and instrumental drum strokes occurred for tabla players in the idiomatic sequencing condition. This was interpreted as evidence for the availability of long-term memory representations for verbal materials, "tabla words", which could have reduced the memory load once sequential integration had represented them as chunks. In that sense, verbal and acoustic instrumental stimuli can be thought of as activating distinct representational systems, but otherwise relying on the same STM processes.

The encounter with Murail's *Mémoire/Erosion* and Baddeley's phonological (tape) loop in Chapter 1 raised the question of whether a re-injection tape loop may be a good metaphor for short-term memory for timbre. Such tape loops usually begin with

the transduction of sound pressure waves into electric current, followed by an amplifier that boosts these currents, and the coil of the tape head that creates a magnetic field in order to align the magnetic material in the tape proportional to the original signal. Unfortunately for the sake of the metaphor, the above conclusions would imply that the magnetic tape only minimally contributes to "the tape loop of STM".[1] On the contrary, we must infer that the magnetic field generated by the coil is strong enough to keep objects of considerable complexity in the air. For a moment.

This "lack of tape" metaphorically characterizes an implicit tenet of a proceduralist approach to memory. Our very object of study, short-term memory, somehow dissolved into a set of processes a priori unrelated to short-term memory, such as attention operating on perceptual representations, sequential chunking, or long-term associations. Beyond a primitive form of information persistence, short-term memory as a dedicated cognitive system with its own representations and algorithms (Marr, 2010) is obsolete from this perspective. It may thus be appropriate to conclude by rewriting an esteemed closing statement on musical timbre (Risset & Wessel, 1999, pp. 150–151). Two key notions will be replaced by suitable counterparts.

> "The role of ~~*timbre*~~ *memory* has extended to that of central subject of ~~*the music*~~ *cognition*. Then, paradoxically, the very notion of ~~*timbre*~~ *memory*, this catchall, multidimensional attribute with a poorly defined identity, gets blurred, diffuse, and vanishes into the ~~*music*~~ *mind* itself."

8.4 Remarks on timbre in theories of music listening

This final section, a type of coda, attempts to embed the themes of this thesis into questions on the role of timbre in theories of music listening. I will use the position outlined by Patel (2008) as a starting point. For a summary of compositional and music-theoretical perspectives, see Tan (2015). I will start with specific points that

[1] Otherwise, memory for timbre would be like managing analog tape loops and stacks of punch cards in parallel (heterogeneous association of continuously valued and hierarchically encoded, compressed categorical representations). The machinery would necessitate an uncountable number of controls to allow the listener to focus on certain frequency/modulation channels (selective attention). It would incorporate not only one but multiple loops running in parallel at different speeds (chunking and matching at different temporal scales). This metaphor can be rejected for a lack of parsimony.

complement Patel's position, and work my way up (or down?) to more fundamental aspects.

The basic question addressed by Patel (2008, Ch. 2) is why timbral contrast rarely serves as a basis for musical sound systems. One the one hand, he acknowledged that timbre is of immense aesthetic importance, as illustrated by the disparate aesthetic appeal of a Jazz ballad played expertly on a saxophone, compared to the same piece played on an electronic saxophone synthesizer. On the other hand, he emphasized that musical styles based on note-to-note timbral contrast, as found in tabla, are rarely encountered in Western music. For that reason, and given Schoenberg's call for music that features tone colors "whose relations with one another work with a kind of logic entirely equivalent to that logic which satisfies us in the melody of pitches"(Schoenberg, 1911/1978, p. 421), Patel called it a "curious fact that *Klangfarbenmelodie* has not become a common feature on the musical landscape in Western culture"(p. 34). He concluded that the difficulty of organizing timbre in terms of intervals or scales likely posits the main cognitive obstacle for timbral contrast to serve as a commonly-used structural basis of Western music. This alluded to McAdams (1989), outlining the cognitive prerequisites for an auditory dimension to contribute to the experience of musical form. McAdams argued that a *form-bearing dimension* would likely need to correlate with attributes that affect auditory grouping, would need to have the potential for integration into a hierarchical event structure, and would require the affordance for categorization and the encoding of inter-category relations, which could give rise to abstract musical knowledge structures. Wessel (1979) and McAdams and Cunible (1992) provided evidence that musicians and composers can perceive timbre intervals in principle (thus demonstrating the encoding of inter-categorical relations), but the results of McAdams and Cunible (1992) also suggested a lack of generalizability of perceived intervals across timbre type. Patel (2008) interpreted these results as raising the question whether timbre could be structured with enough uniformity in order to yield a category system that is shared by composers and listeners, as in the case of pitch or relative duration.

Before addressing more fundamental concerns, three immediate comments shall complement this perspective. It should first be noted that different territories of the musical landscape are governed by drastically different musical laws. *Klangfarben-melodie* has become a well-known formula in "Western art music" (WAM), clearly

explored in Schoenberg's 1909 example, *Farben* (Colors, Op. 16, no. 3), directly followed by Webern's *Five pieces for orchestra* (op. 10), and many other examples (see Erickson, 1975), not to forget *Mémoire/Erosion*. For a recent example in popular music, see `https://www.youtube.com/watch?v=Y8w2oVZ6Sio`. Attempting to argue by quantity would miss the point, though. What makes 20th/21st century WAM unique is that much fewer stylistic features are shared between different pieces and composers than in the common practice period. Common traits are subtle in general, and exemplars of *Klangfarbenmelodie* don't share many structural features in particular (apart from being a timbre melody to start with). In fact, one can assume that the more popular a compositional idea becomes, the less likely it is to be used in obvious ways—contemporary composers of WAM would do anything but agree upon an "industry standard". It therefore seems illusory to continue to call for a "genuine timbral syntax"(Lerdahl, 1987) that could become the basis of a novel common practice in WAM, and that could imprint itself in the form of strong cognitive schemata in the mind of a significant number of listeners.

Secondly, other territories of the landscape of 20th/21st century Western music adhere to different laws. The most obvious example of orderly "item-to-item" timbral contrasts that are based on a commonly agreed-upon structure is that of drum and percussion tracks in popular music (in the broadest sense). 150 years ago, it may have been commonplace to consider non-harmonic percussion sounds only "applicable for marches and other boisterous music"(von Helmholtz, 1885/1954, p. 119), but I suspect that contemporary listeners do enjoy drums and percussion, given that they are endured in a majority of genres in popular music. Obviously, they possess a constitutive role in musical textures and contribute to the perception of grouping and sectional boundaries. From music-theoretical viewpoints, they are, however, often left unanalyzed (perhaps because they don't occur in the piano reduction?), and are hardly researched in music cognition. Given the large and detailed body of work on musical expectancy, it may be time to address the question of how the statistical regularities of drum tracks generate timbral schemata and expectations and affect the ways in which musical textures are perceived.

Thirdly, another more subtle type of timbral contrast is all around us (but we don't see it, because it's not in the score). At first glance, most traditional musical instruments vary along the dimensions of pitch, duration, and dynamics (i.e., playing

effort). As outlined in Chapter 2, however, timbre coherently covaries with pitch and dynamics. Although we don't think about a piano sonata in terms of timbral contrast, subtle forms of within-instrument-based timbral contrast are an integral part of natural forms of musical articulation, even for the piano, the most "platonic" instrument of all. I hypothesize that major portions of what lets playing effort contribute to musical expressiveness (e.g., see Bhatara, Tirovolas, Duan, Levy, & Levitin, 2011) is not due to contrasts in perceived loudness, but to a large part based on the resulting fine grained timbral contrast and the resulting acoustic articulation (cf., Lembke, 2014, Ch. 4).

Moreover, most of what lends extreme pitch registers their distinctiveness (and thus their structural and expressive value) may be due not to pitch height, but to timbral distinctiveness (think about the drastic timbral differences between low, middle and high piano tones). In fact, it may be the lack of this subtle layer of timbral (co)articulation that lets poorly synthesized emulations appear unpleasing (see Patel's saxophone example above). The fact that such emulations can be readily identified as synthetic and inauthentic, adds another source of aesthetic dissatisfaction. Similar observations have been made for the clarinet (Barthet, Kronland-Martinet, & Ystad, 2008).

The critical reader may be inclined to dismiss the last point, because the articulatory structure of a musical realization does not appear to touch the point in question, that is, timbral contrast as a structural basis of the musical discourse. This requires us to better specify what is meant by the notion of "structural basis".

Addressing music from the common-practice period, Patel (2008) argued for an abstract similarity of the "syntactic architecture" of musical sequences and that of language. These similarities include multiple levels of organization, hierarchical structuring, grammatical categories that can be filled by different physical entities, as well as differences between structure and elaboration. Specifying roles in the communicative chain of music, Palmer (1996) similarly emphasized the importance of abstraction: "The listener's and performer's experience of a musical piece can be described as a conceptual structure, an abstract message that specifies the relevant musical relationships in a piece."(p. 25) In fact, the divide between music's abstract structural essence and its concrete incarnation, the elaborative "musical surface", is a common vantage point not only for Schenkerian methods of music analysis, but also for large parts of music

theory in general, and was influentially re-phrased in cognitive terms in the *Generative Theory of Tonal Music* (GTTM) by Lerdahl and Jackendoff (1983).

In the dichotomy of structure and elaboration, timbre is commonly construed as part of the non-structural, that is, as a "surface feature". As described by Dolan (2013),

> "[Timbre] is the concept to which we must turn to describe the immediacies of how sounds strike our ears, how they affect us. It is the word we need when we want to discuss sound in terms of its particularities and peculiarities. To put it another way, to talk about timbre is to value sound as sound, and not as a sonic manifestation of abstract principles."(p. 87)

Nevertheless, Schoenberg's call for a *logic* of tone colors has found a strong resonance in theoretical writings on timbre. For instance, in order to lend timbre structural "depth", Lerdahl (1987) proposed a way to organize the parameter according to syntactic rules. Drawing from the methods of GTTM, he suggested transposing pitch structures to the realm of timbre. He remarked, "There is now such an infinity of timbral possibilities that the need for some kind of selection and organization has become acute. [...] The time is ripe to develop a genuine timbral syntax. But according to what principles?"(p. 136) He argued that structuring compositions along *timbral hierarchies* could be a way to give timbre structural function and thereby let it flourish. Hierarchies could be organized along timbral dimensions such as brightness (assuming that bright sounds are more tense and unstable than dull sounds), according to vibrato frequencies, or other timbral facets. This presumes the usage of discretized sound material ordered along various scales of stability. Nattiez (2009) commented on proposals such as these,

> "But we have to ask ourselves, from both an aesthetic and a critical point of view, whether in seeking to treat timbre as a phenomenon contributing to a syntax, which is to say trying to confine it to the properties that a tone has in a scale system, especially the tonal and twelve-tone systems, we are not in danger of denying it one of its basic characteristics, multidimensionality, which explains not only the fascination of composers of electroacoustic music with the endless facets of sounds, but also the enormous richness in our listening to some of their works."(p. 13)

He thus underlined the danger of attempting to "domesticate" the complexity of the timbral facets of sound. In a similar vein, Tristan Murail (2005) emphasized the intricacy and non-linearity of the sound world of contemporary music:

> "The new materials that offer themselves to the composer [...] are often complex sounds, intermediate sounds, hybrids, sounds that possess new dimensions (transitions, development over time), sounds that are neither harmonic complexes nor timbres but something between the two. [...] There is no precise line between pitch and noise, rhythm and frequency; harmony and sound color are continuous phenomena."(pp. 123–124)

Even Murail, however, far from "domesticating" or discretizing his sound palette, appeared to seek for something "deeper than just sound". His writing reflects the theoretical yearning for a timbral *logic* that would transcend the musical moment and contribute to the construction of musical form. He further noted,

> "The entire range of complex sounds can be integrated functionally within a musical logic, rather than used as a startling daub of color, or only for expressive ends, for their anomalous or paroxysmal qualities."(p. 135)

One of the great "alchemists" of musical color seems to be theoretically dissatisfied by their "merely expressive ends".

Given that both composers and music psychologists emphasized abstract syntactic systems, is it adequate to assume that music's structural "architecture" is foundational to listening? Do abstract relations most significantly contribute to the shaping of the musical present and the experience of large-scale musical form?

Taking a conceptual stance, it should not be forgotten that there is a basic terminological inadequacy with the previously encountered notions of the syntactic "architecture" and the "surface vs. structure" dichotomy. These notions are metaphors, mapping time to space in a twisted way, such that their descriptive reach is finite. Composers, music theorists, music producers, all have the chance to continually revisit particular components of pieces of music, and thus may conceptualize music in a mode somewhat independent of time. In listening, however, there is no such thing that would qualify as a "musical surface". There is no outside part or uppermost layer of music.

It is memory over various time scales that gives rise to musical form, and we are only beginning to understand the relative mnemonic salience of different musical features. Although spatial metaphors permeate language, they should be used cautiously in reasoning about music, which is temporal in essence. In other words, "Le Corbusier said, architecture magnifies Space. Today, as in the past, music transfigures Time."(Grisey, 2000, p. 3)

Arguing from the vantage point of musical performance, Cook (2013) described structural ontologies of music as being under the shadow of *Plato's curse*. He characterized that curse as an ideology[2] inherent to predominant discourses in musicology, music theory, and even music cognition. The basic credo is to seek musical substance in the abstract, platonic realm of musical notation as opposed to the temporal and bodily processes of music performance and reception. One facet of the curse is that it yields a spatial conception of musical time. Another facet is that it assumes that listeners primarily seek a structural understanding of musical pieces. With Cook's sarcasm,

> "The performer's role is at best to transcribe the work from the domain of the abstract to that of the concrete , and at worst to deviate from it. [...] It is the performer's obligation to represent the composer's work to the listener, just as it is the listener's obligation to strive towards an adequate understanding of the work itself."(p. 13)

It was mentioned above that the bracketing of perceptual representations also has a long history in cognitive science. Traditional models seek cognitive operations in the realm of amodal symbols, transduced at some point in the processing chain, despite a persistent lack of empirical support for amodal symbols (E. E. Smith & Kosslyn, 2013). Long-term memory was similarly not conceived of as comprising rich sensory representations. Theories of perceptual symbol systems have pledged to revise these traditional accounts and conceive of cognition as operating on concrete perceptual schemata (Barsalou, 1999).

But also the empirical situation in music cognition suggests that there are limits to purely structural perspectives. Reviewing studies on the cognitive processing of short-term and long-term musical structure, Tillmann and Bigand (2004) addressed

[2] In his words, "It presents itself not as an assumption at all but just as the way things are." (Cook, 2013, p. 17)

the divide between listeners' sensitivity for local ordering paired with their apparent insensitivity for hierarchical structure on more global time scales (i.e., beyond the phrase level). The authors argued that there may not be an overall psychological utility for global order processing, when individual musical scenes exhibit rich affordances already.

> "One way to understand this paradox [...] is to consider that local, small-scale musical units are so rich for aesthetic experience that processing larger musical units may fulfill no crucial need." [...] Time may tend to be processed moment by moment when such moments lead to an extremely rich aesthetic, intellectual, spiritual, or emotional experience."(p. 219)

These conclusions emphasize the importance of phrase-level processing and the multiple perceptual affordances of short time scales. The short-term coherence of the musical discourse seems to be more important than long-term relations. STM is part of the cognitive infrastructure for the apprehension of the individual musical moment. In fact, it is the very precondition for moments to encompass more than a glimpse. The flexibility and active nature of this auditory "workspace" allows us to explore musical scenes, compare one sound event to another, look for individual pathways through a musical texture, and in a sense, prolong the musical moment that may (or may not) mold into form at some point.

Turning from the pertinent time scales of the "musical message" to its content, the previous reviews and discussion (in particular Ch. 3) have emphasized the point that musical memory is concrete. Memory for seemingly abstract entities such as melodies (as specified by pitch interval structure) are affected by their instrumental materialization. Further, musical recordings can be identified on the basis of very short excerpts, which rules out most auditory attributes apart from timbre as the critical feature of the memory trace. The literature further suggests that the concrete acoustic properties of musical events critically contribute to memory for musical textures in non-tonal 20th century WAM (Krumhansl, 1991; McAdams, Vieillard, Houix, & Reynolds, 2004; Poulin-Charronnat et al., 2004). Curiously, the elaborative and seemingly inessential "surface" provides the mnemonically most salient features.

If we assume that listeners extract a "message" from a musical realization, this message would preferentially encode musical relationships over short time scales, and

it would be *concrete* in the sense that the material face of music would be an important part of it. In other words, the many accounts of music listening that build on abstraction seem to underestimate the importance of concrete musical features and their sensory representation. Beyond the realm of structural inference, it is ephemeral *presence* that we seek in works of art (cf., Gumbrecht, 2004). Music draws us into the moment and its spectrotemporal essences. The startling colors and buzzing textures of rich acoustic scenes lend music its tangibility and afford active auditory exploration. This is where we enter the realm of aesthetic experience.

References

Agus, T. R., & Pressnitzer, D. (2013). The detection of repetitions in noise before and after perceptual learning. *The Journal of the Acoustical Society of America*, *134*(1), 464–473.

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133.

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*, 610–618.

Alain, C., Arnott, S. R., Gillingham, S., Leung, A. W., & Wong, J. (2015). The interplay between auditory attention and working memory. In P. Jolicoeur, C. Levebre, & J. Martinez-Trujillo (Eds.), *Mechanisms of sensory working memory/ Attention and perfomance XXV* (pp. 215–228). London, UK: Academic Press.

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *Journal of the Acoustical Society of America*, *135*(3), 1371–1379.

Alluri, V., & Toiviainen, P. (2012). Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. *Music Perception*, *29*(3), 297–310.

Alunni-Menichini, K., Guimond, S., Bermudez, P., Nolden, S., Lefebvre, C., & Jolicoeur, P. (2014). Saturation of auditory short-term memory causes a plateau in the sustained anterior negativity event-related potential. *Brain Research*, *1592*, 55–64.

Andrillon, T., Kouider, S., Agus, T., & Pressnitzer, D. (2015). Perceptual learning of acoustic noise generates memory-evoked potentials. *Current Biology*, *25*, 1–7.

ANSI. (1960/1994). *Psychoacoustic terminology: Timbre.*

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.),

Multidimensional models of perception and cognition (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Baddeley, A. D. (1979). Working memory and reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of visible language* (pp. 355–370). Heidelberg, Germany: Springer.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47–89). New York, NY: Academic Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Barthet, M., Guillemain, P., Kronland-Martinet, R., & Ystad, S. (2010). From clarinet control to timbre perception. *Acta Acustica united with Acustica*, *96*(4), 678–689.

Barthet, M., Kronland-Martinet, R., & Ystad, S. (2008). Improving musical expressiveness by time-varying brightness shaping. In R. Kronland-Martinet, S. Ystad, & K. Jensen (Eds.), *Computer music modeling and retrieval. Sense of sounds* (pp. 313–336). Berlin, Germany: Springer.

Berglund, B. (2012). Measurement in psychology. In B. Berglund, G. B. Rossi, J. T. Townsend, & L. R. Pendrill (Eds.), *Measurement with persons: Theory, methods, and implementation areas* (pp. 27–50). New York, NY: Psychology Press.

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review*, *119*(1), 40–79.

Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B., & Levitin, D. J. (2011). Per-

ception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 921–934.

Bigand, E., Delbé, C., Gérard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes? *PloS ONE*, *6*(10), e27024.

Bigand, E., Delbé, C., Poulin-Charronnat, B., Leman, M., & Tillmann, B. (2014). Empirical evidence for musical syntax processing? computer simulations reveal the contribution of auditory short-term memory. *Frontiers in Systems Neuroscience*, *8*, doi: 10.3389/fnsys.2014.00094.

Bigand, E., Perruchet, P., & Boyer, M. (1998). Implicit learning of an artificial grammar of musical timbres. *Current Psychology of Cognition*, *17*(3), 577–600.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., & McAdams, S. (2006). Seperate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, *18*(12), 1959–1972.

Caclin, A., Giard, M.-H., Smith, B. K., & McAdams, S. (2007). Interactive processing of timbre dimensions: A garner interference study. *Brain Research*, *1138*, 159–170.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, *118*(1), 471–482.

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469.

Camos, V., Mora, G., & Oberauer, K. (2011). Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Memory & Cognition*, *39*(2), 231–244.

Carruthers, I. M., Laplagne, D. A., Jaegle, A., Briguglio, J., Mwilambwe-Tshilobo, L., Natan, R. G., & Geffen, M. N. (2015). Emergence of invariant representation of vocalizations in the auditory cortex. *Journal of Neurophysiology*, *DOI: 10.1152/jn.00095.2015*(in press).

Caruso, V. C., & Balaban, E. (2014). Pitch and timbre interfere when both are

parametrically varied. *PLoS ONE*, *9*(1), e87065.

Chalupper, J. (2008). Calculation of loudness for normal and hearing-impaired listeners. In D. Havelock, S. Kuwano, & M. Vorländer (Eds.), *Handbook of signal processing* (Vol. 1, pp. 251–262). Heidelberg, Germany: Springer.

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters*, *405*(3), 164–167.

Clarke, E. F. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford, UK: Oxford University Press.

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C., & Janata, P. (2014). A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological Review*, *121*(1), 33–65.

Cook, N. (2013). *Beyond the score: Music as performance*. Oxford, UK: Oxford University Press.

Cousineau, M., Carcagno, S., Demany, L., & Pressnitzer, D. (2013). What is a melody? on the relationship between pitch and brightness of timbre. *Frontiers in systems neuroscience*, *7*, doi: 10.3389/fnsys.2013.00127.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*(2), 341–370.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338.

Cowan, N. (2015). Sensational memorability: Working memory for things we see, hear, feel, or somehow sense sensational memorability: Working memory for things we see, hear, feel, or somehow sense. In P. Jolicoeur, C. Levebre, & J. Martinez-Trujillo (Eds.), *Mechanisms of sensory working memory/ Attention and perfomance XXV* (pp. 5–22). London, UK: Academic Press.

Craik, F. I. (2007). Encoding: A cognitive perspective. In H. L. Roediger III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 129–136). Oxford, UK: Oxford Univ Press.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Crowder, R. G. (1989). Imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 472–478.

Crowder, R. G. (1993). Auditory memory. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 113–143). Oxford, UK: Oxford University Press.

De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263.

Demany, L., & Semal, C. (2007). The role of memory in auditory perception. In W. A. Yost & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 77–113). New York, NY: Springer.

Demany, L., Semal, C., Cazalets, J.-R., & Pressnitzer, D. (2010). Fundamental differences in change detection between vision and audition. *Experimental Brain Research*, *203*(2), 261–270.

Demany, L., Trost, W., Serman, M., & Semal, C. (2008). Auditory change detection: simple sounds are not memorized better than complex sounds. *Psychological Science*, *19*(1), 85–91.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*(28), 1–28.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science*, *168*(3939), 1604–1605.

Deutsch, D. (1975). Auditory memory. *Canadian Journal of Psychology*, *29*, 87–105.

Dolan, E. (2013). *The orchestral revolution: Haydn and the technologies of timbre.* Cambridge, MA: Cambridge University Press.

Donnadieu, S. (2008). Mental representation of the timbre of complex sounds. In J. W. Beauchamp (Ed.), *Analysis, synthesis, and perception of musical sounds* (pp. 272–319). New York, NY: Springer.

Douglas, C. (2015). *Perceived affect of musical instrument sounds.* Unpublished master's thesis, McGill University.

Dudai, Y. (2007). Memory. In H. L. Roediger III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 129–135). Oxford, UK: Oxford Univ Press.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Elliott, T., Hamilton, L., & Theunissen, F. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *The Journal of the Acoustical Society of America*, *133*(1), 389–404.

Erickson, R. (1975). *Sound structure in music*. Berkeley, CA: Univ of California Press.

Filipic, S., Tillmann, B., & Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. *Psychonomic Bulletin & Review*, *17*(3), 335–341.

Fineberg, J. (2013). *Classical music, why bother? Hearing the world of contemporary culture through a composer's ears*. London, UK: Routledge.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, in press.

Fishman, Y. I. (2014). The mechanisms and meaning of the mismatch negativity. *Brain Topography*, *27*(4), 500–526.

Fritz, C., Blackwell, A. F., Cross, I., Woodhouse, J., & Moore, B. C. (2012). Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. *The Journal of the Acoustical Society of America*, *131*(1), 783–794.

Fuster, J. M. (2003). *Cortex and mind*. Oxford, UK: Oxford Univ Press.

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology: Section A*, *54*(1), 1–30.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, *5*(1), 1–29.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1–17.

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

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

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2012). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral*

Cortex, *23*(9), 2025–2037.

Giordano, B. L., McDonnell, J., & McAdams, S. (2010). Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and Cognition*, *73*(1), 7–19.

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.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24–29.

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 399–412.

Grey, J. M. (1975). *An exploration of musical timbre*. Unpublished doctoral dissertation, CCRMA, Stanford University.

Grisey, G. (2000). Did you say spectral? *Contemporary Music Review*, *19*(3), 1–3.

Gumbrecht, H. U. (2004). *Production of presence: What meaning cannot convey*. Palo Alto, CA: Stanford University Press.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In *Perception and cognition of music* (pp. 253–306). New York, NY: Psychology Press.

Halpern, A. R., & Müllensiefen, D. (2008). Effects of timbre and tempo change on memory for music. *The Quarterly Journal of Experimental Psychology*, *61*(9), 1371–1384.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–1292.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. Moore (Ed.), *Hearing* (Vol. 2, pp. 425–461). San Diego, CA: Academic Press.

Handel, S., & Erickson, M. L. (2001). A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception*, *19*(1), 121–126.

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

Henson, R., Hartley, T., Burgess, N., Hitch, G., & Flude, B. (2003). Selective interference with verbal short-term memory for serial order information: A new paradigm and tests of a timing-signal hypothesis. *Quarterly Journal of Experimental Psychology: Section A*, *56*(8), 1307–1334.

Houtsma, A. J. (1997). Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, *26*(2), 104–115.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

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

James, W. (1890/2004). *The principles of psychology* (Vol. 1). Retrieved Nov 9 2015, from `http://psychclassics.yorku.ca/James/Principles/`

Johnson, M. K. (1992). MEM: Mechanisms of recollection. *Journal of Cognitive Neuroscience*, *4*(3), 268–280.

Jolicoeur, P., Levebre, C., & Martinez-Trujillo, J. (2015). *Mechanisms of sensory working memory/ Attention and perfomance XXV*. London, UK: Academic Press.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193–224.

Kaernbach, C. (2004). The memory of noise. *Experimental Psychology*, *51*(4), 240–248.

Kahana, M. J. (2012). *Foundations of human memory*. Oxford, UK: Oxford University Press.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, *42*(18), 2177–2192.

Kendall, R. A., Carterette, E. C., & Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, *16*(3), 327–363.

Koelsch, S. (2009). Music-syntactic processing and auditory memory: Similarities and differences between ERAN and MMN. *Psychophysiology*, *46*(1), 179–190.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605.

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. 846, pp. 43–53). Amsterdam, The Netherlands: Excerpta Medica.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford, UK: Oxford University Press.

Krumhansl, C. L. (1991). Memory for musical surface. *Memory & Cognition*, *19*(4), 401–411.

Krumhansl, C. L. (2010). Plink: "thin slices" of music. *Music Perception*, *27*(5), 337–354.

Krumhansl, C. L., & Iverson, P. (1992). Peceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 739–751.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kumar, S., Bonnici, H. M., Teki, S., Agus, T. R., Pressnitzer, D., Maguire, E. A., & Griffiths, T. D. (2014). Representations of specific acoustic patterns in the auditory cortex and hippocampus. *Proceedings of the Royal Society B: Biological Sciences*, *281*, doi: 10.1098/rspb.2014.1000.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, *62*(7), 1426–1439.

Lange, K., & Czernochowski, D. (2013). Does this sound familiar? Effects of timbre change on episodic retrieval of novel melodies. *Acta Psychologica*, *143*(1), 136–145.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx), 10–15 September, Bordeaux, France* (pp. 237–244).

Lemaitre, G., Houix, O., Misdariis, N., & Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, *16*(1), 16–32.

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, *17*(4), 481–509.

Lembke, S.-A. (2014). *When timbre blends musically: perception and acoustics underlying orchestration and performance.* Unpublished doctoral dissertation, McGill University.

Lerdahl, F. (1987). Timbral hierarchies. *Contemporary Music Review*, *2*(1), 135–160.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge, MA: MIT Press.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., & Weinzierl, S. (2014). A spatial audio quality inventory (SAQI). *Acta Acustica united with Acustica*, *100*(5), 984–994.

Luo, H., Tian, X., Song, K., Zhou, K., & Poeppel, D. (2013). Neural response phase tracks how listeners learn new acoustic representations. *Current Biology*, *23*(11), 968–974.

Macken, B., Taylor, J. C., & Jones, D. M. (2014). Language and short-term memory: The role of perceptual-motor affordance. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *40*(5), 1257–1270.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Manoury, P. (1991). Les limites de la notion de 'timbre'. In J.-B. Barriere (Ed.), *Le timbre: Métaphore pour la composition* (pp. 293–299). Christian Bourgois, Paris.

Marin, M. M., Gingras, B., & Stewart, L. (2012). Perception of musical timbre in congenital amusia: Categorization, discrimination and short-term memory. *Neuropsychologia*, *50*(3), 367–378.

Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *The Journal of the Acoustical Society of America*, *121*(1), 383–387.

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

Marr, D. (2010). *Vision: A computational approach*. Cambridge, MA: MIT Press.

Martin, F. N., Champlin, C. A., et al. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, *11*(2), 64–66.

Martin, K. D. (1999). *Sound-source recognition: A theory and computational model*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Massaro, D. W., & Loftus, G. R. (1996). Sensory and perceptual storage: Data and theory. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 67–99). San Diego, CA: Academic Press.

May, P. J., & Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology*, *47*(1), 66–122.

McAdams, S. (1984). *Spectral fusion, spectral parsing, and the formation of auditory images*. Unpublished doctoral dissertation, Stanford University.

McAdams, S. (1987). Music: A science of the mind? *Contemporary Music Review*, *2*(1), 1–61.

McAdams, S. (1989). Psychological contraints on form-bearing dimensions in music. *Contemporary Music Review*, *4*, 181–198.

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. 146–198). Oxford, UK: Oxford University Press.

McAdams, S. (2013). Musical timbre perception. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 35–67). San Diego, CA: Academic Press.

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

McAdams, S., & Cunible, J.-C. (1992). Perception of timbral analogies. *Philosophical Transactions of the Royal Society: Biological Sciences*, *336*, 383–389.

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

McAdams, S., Vieillard, S., Houix, O., & Reynolds, R. (2004). Perception of musi-

cal similarity among contemporary thematic materials in two instrumentations. *Music Perception*, *22*(2), 207–237.

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*(3), 177–192.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychological Science*, *19*(12), 1263–1271.

McKay, C., & Fujinaga, I. (2008). Combining features extracted from audio, symbolic and cultural sources. In J. P. Bello, E. Chew, & D. Turnbull (Eds.), *Proceedings of the 2008 International Society for Music Information Retrieval Conference, Philadelphia, USA, Sep 14-18, 2008* (pp. 597–602).

McKeown, D., Mills, R., & Mercer, T. (2011). Comparisons of complex sounds across extended retention intervals survives reading aloud. *Perception*, *40*(10), 1193–1205.

McKeown, D., & Wellsted, D. (2009). Auditory memory for timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 855–875.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69.

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics*, *48*(2), 169–178.

Melara, R. D., Marks, L. E., & Lesko, K. E. (1992). Optional processes in similarity judgments. *Perception & Psychophysics*, *51*(2), 123–133.

Mercer, T., & McKeown, D. (2010). Updating and feature overwriting in short-term memory for timbre. *Attention, Perception, & Psychophysics*, *72*(8), 2289–2303.

Mercer, T., & McKeown, D. (2014). Decay uncovered in nonverbal short-term memory. *Psychonomic bulletin & review*, *21*(1), 128–135.

Meyer, J. (1995). *Akustik und musikalische Aufführungspraxis: Leitfaden für Akustiker, Tonmeister, Musiker, Instrumentenbauer und Architekten*. Bergkirchen, Germany: Bochinsky.

Meyer, L. B. (1989). *Style and music: Theory, history, and ideology*. Chicago, IL: University of Chicago Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on

our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Moore, B. C. (2015). *Auditory processing of temporal fine structure: Effects of age and hearing loss*. Singapoore, China: World Scientific.

Moscovitch, M. (2007). Why the engram is elusive. In H. L. Roediger III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 17–21). Oxford, UK: Oxford Univ Press.

Murail, T. (2005). The revolution of complex sounds. *Contemporary Music Review*, *24*(2/3), 121–135.

Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., et al. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544–2590.

Nattiez, J.-J. (2007). Le timbre est-il un paramètre secondaire? [Is timbre a secondary parameter?]. *Cahiers de la Société Québécoise de Recherche en Musique*, *9*(1–2), 13–24.

Nattiez, J.-J. (2009). *Is timbre a secondary parameter?* (Unpublished manuscript)

Nimmo, L. M., & Roodenrys, S. (2005). The phonological similarity effect in serial recognition. *Memory*, *13*(7), 773–784.

Nolden, S., Bermudez, P., Alunni-Menichini, K., Lefebvre, C., Grimault, S., & Jolicoeur, P. (2013). Electrophysiological correlates of the retention of tones differing in timbre in auditory short-term memory. *Neuropsychologia*, *51*(13), 2740–2746.

Nosofsky, R. M., & Kantner, J. (2006). Exemplar similarity, study list homogeneity, and short-term perceptual recognition. *Memory & Cognition*, *34*(1), 112–124.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in cognitive sciences*, *13*(1), 14–19.

Palmer, C. (1996). On the assignment of structure in music performance. *Music Perception*, *14*(1), 23–56.

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

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*,

6(7), 674–681.

Patel, A. D. (2008). *Music, language, and the brain.* Oxford, UK: Oxford University Press.

Patel, A. D. (2012). The OPERA hypothesis: Assumptions and clarifications. *Annals of the New York Academy of Sciences*, *1252*(1), 124–128.

Patel, A. D., & Iversen, J. R. (2003). Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: An empirical study of sound symbolism. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS), 3-9 Aug 2003, Barcelona* (pp. 925–928).

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in our ears: The biological bases of musical timbre perception. *PLOS Computational Biology*, *8*(11), e1002759.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Auditory Physiology and Perception*, *83*, 429–446.

Pechmann, T., & Mohr, G. (1992). Interference in memory for tonal pitch: Implications for a working-memory model. *Memory & Cognition*, *20*(3), 314–320.

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

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders. *Annals of the New York Academy of Sciences*, *999*(1), 58–75.

Pinker, S. (1994). *The language instinct: The new science of language and mind.* London, UK: Penguin.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 976–986.

Pitt, M. A., & Crowder, R. G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 728–738.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 397–414). Leiden, The Netherlands: Suithoff.

Poulin-Charronnat, B., Bigand, E., Lalitte, P., Madurell, F., Vieillard, S., & McAdams, S. (2004). Effects of a change in instrumentation on the recognition of musical materials. *Music Perception*, *22*(2), 239–263.

Pressnitzer, D., Agus, T. R., & Suied, C. (2013). Acoustic timbre recognition. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience: Springer Reference* (pp. 1–6). Heidelberg, Germany: Springer.

Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in Human Neuroscience*, *9*, doi: 10.3389/fnhum.2015.00197.

Radvansky, G. A., Fleming, K. J., & Simmons, J. A. (1995). Timbre reliance in nonmusicians' and musicians' memory for melodies. *Music Perception*, *13*(2), 127–140.

Radvansky, G. A., & Potter, J. K. (2000). Source cuing: Memory for melodies. *Memory & Cognition*, *28*(5), 693–699.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235.

Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *The Journal of the Acoustical Society of America*, *122*(6), 3688–3696.

Risset, J.-C., & Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 113–169). San Diego, CA, Academic Press.

Roads, C. (2015). *Composing electronic music: A new aesthetic*. Oxford, UK: Oxford University Press.

Roediger, H. L., III. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532–547.

Rose, N. S., Buchsbaum, B. R., & Craik, F. I. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, *42*(5), 689–700.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Saxena, S. K. (2008). *The art of tabla rhythm: Essentials, tradition and creativity*. New Dehli, India: D.K. Printworld Ltd.

Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory & Cognition*, *43*(7), 1021–1031.

Schellenberg, E. G., Iverson, P., & McKinnon, M. C. (1999). Name that tune: Identifying popular recordings from brief excerpts. *Psychonomic Bulletin & Review*, *6*(4), 641–646.

Schendel, Z. A., & Palmer, C. (2007). Suppression effects on musical and verbal memory. *Memory & Cognition*, *35*(4), 640–650.

Schoenberg, A. (1911/1978). *Theory of harmony [Harmonielehre]*. Berkeley, CA: University of California Press (R. E. Carter, Trans. from original German edition, 1911).

Schulze, K., Jay Dowling, W., & Tillmann, B. (2012). Working memory for tonal and atonal sequences during a forward and a backward recognition task. *Music Perception*, *29*(3), 255–267.

Schulze, K., & Koelsch, S. (2012). Working memory for speech and music. *Annals of the New York Academy of Sciences*, *1252*(1), 229–236.

Schulze, K., & Tillmann, B. (2013). Working memory for pitch, timbre, and words. *Memory*, *21*(3), 377–395.

Schulze, K., Zysset, S., Mueller, K., Friederici, A. D., & Koelsch, S. (2011). Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Human Brain Mapping*, *32*(5), 771–783.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21.

Sekuler, R., & Kahana, M. J. (2007). A stimulus-oriented approach to memory. *Current Directions in Psychological Science*, *16*(6), 305–310.

Semal, C., & Demany, L. (1991). Dissociation of pitch from timbre in auditory short-term memory. *The Journal of the Acoustical Society of America*, *89*, 2404–2410.

Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *Neuroimage*, *41*(1), 113–122.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepherd, F. A. (1976). *Tabla and the Benares Gharana*. Unpublished doctoral dissertation, Wesleyan University.

Siedenburg, K., & Dörfler, M. (2013). Persistent time-frequency shrinkage for audio denoising. *Journal of the Audio Engineering Society (AES)*, *61*(1/2), 29–38.

Siedenburg, K., Fujinaga, I., & McAdams, S. (2015). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Under review*.

Slevc, L. R., & Patel, A. D. (2011). Meaning in music and language: Three key differences. Comment on "towards a neural basis of processing musical semantics" by Stefan Koelsch. *Physics of Life Reviews*, *8*(2), 110–111.

Smalley, D. (1994). Defining timbre—refining timbre. *Contemporary Music Review*, *10*(2), 35–48.

Smith, E. E., & Kosslyn, S. M. (2013). *Cognitive psychology: Mind and brain*. New York, NY: Pearson Higher Ed.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Snyder, B. (2000). *Music and memory: An introduction*. Cambridge, MA: MIT Press.

Soemer, A., & Saito, S. (2015). Maintenance of auditory-nonverbal information in working memory. *Psychonomic Bulletin & Review*, *published online*, doi: 10.3758/s13423-015-0854-z.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89.

Srinivasan, A., Sullivan, D., & Fujinaga, I. (2002). Recognition of isolated instrument tones by conservatory students. In *Proceedings of the 2002 International Conference on Music Perception and Cognition, Sydney, July 17–21, 2002* (pp. 17–21).

Starr, G. E., & Pitt, M. A. (1997). Interference effects in short-term memory for timbre. *The Journal of the Acoustical Society of America*, *102*(1), 486–494.

Steele, K. M., & Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? *Music Perception*, *23*(3), 215–220.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Storrs, K. R. (2015). Are high-level aftereffects perceptual? *Frontiers in Psychology*, *6*, doi: dx.doi.org/10.3389/fpsyg.2015.00157.

Strait, D. L., Chan, K., Ashley, R., & Kraus, N. (2012). Specialization among the specialized: Auditory brainstem function is tuned in to timbre. *Cortex*, *48*(3), 360–362.

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. *Journal of the Acoustical Society of America*, *135*(3), 1380–1391.

Surprenant, A., & Neath, I. (2008). The nine lives of short-term memory. In A. S. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16–43). New York, NY: Psychology Press.

Surprenant, A., & Neath, I. (2009). *Principles of memory*. New York, NY: Psychology Press.

Tan, A. (2015). *Ksana: Compositional control of spectral fusion as a parameter of timbre functionality*. Unpublished doctoral dissertation, McGill University.

Tenenbaum, J. B., Griffiths, T. L., et al. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Tervaniemi, M., Winkler, I., & Näätänen, R. (1997). Pre-attentive categorization of sounds by timbre as revealed by event-related potentials. *NeuroReport*, *8*(11), 2571–2574.

Thorn, A. S., Frankish, C. R., & Gathercole, S. E. (2008). The influence of long-term knowledge on short-term memory: Evidence for multiple mechanisms. In A. S. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 198–219). New York, NY: Psychology Press.

Thorn, A. S., Gathercole, S. E., & Frankish, C. R. (2002). Language familiarity effects in short-term memory: The role of output delay and long-term knowledge. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(4), 1363–1383.

Thorn, A. S., & Page, M. (2008). *Interactions between short-term and long-term memory in the verbal domain*. New York, NY: Psychology Press.

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism*, *62*(2), 211–222.

Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences:

Statistical regularities confronted with acoustical (dis)similarities. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*(5), 1131–1142.

Tillmann, B., Schulze, K., & Foxton, J. M. (2009). Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds. *Brain and Cognition*, *71*(3), 259–264.

Trainor, L. J., Wu, L., & Tsang, C. D. (2004). Long-term memory for music: Infants remember tempo and timbre. *Developmental Science*, *7*(3), 289–296.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–402). London: Academic Press.

Tulving, E. (2007). Are there 256 different kinds of memory? In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of henry l. roediger* (pp. 39–52). New York, NY: Psychology Press Festschrift Series.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biology*, *5*(3), e56.

Viswanathan, S., Perl, D. R., Kahana, M. J., Sekuler, R., et al. (2010). Homogeneity computation: How interitem similarity in visual short-term memory alters recognition. *Psychonomic Bulletin & Review*, *17*(1), 59–65.

von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig, Germany: Friedr. Vieweg & Sohn.

von Helmholtz, H. (1885/1954). *On the sensations of tone as a physiological basis for the theory of music* (trans. by A. J. Ellis of 4th German ed., 1877, republ. 1954 ed.). New York, NY: Dover.

Vuvan, D. T., Podolak, O. M., & Schmuckler, M. A. (2014). Memory for musical tones: The impact of tonality and the creation of false memories. *Frontiers in Psychology*, *5*(582). doi: 10.3389/fpsyg.2014.00582

Warren, R. M. (1974). Auditory temporal discrimination by trained listeners. *Cognitive Psychology*, *6*(2), 237–256.

Weiss, M. W., Schellenberg, E. G., Trehub, S. E., & Dawber, E. J. (2015). Enhanced processing of vocal melodies in childhood. *Developmental Psychology*, *51*(3), 370–377.

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way

she sings enhanced memory for vocal melodies. *Psychological Science*, *23*(10), 1074–1078.

Weiss, M. W., Trehub, S. E., Schellenberg, E. G., & Habashi, P. (2015). Eyes wide open for vocal melodies. In J. Grahn (Ed.), *Proceedings of the Meeting of the Society for Music Perception and Cognition, Nashville, TN, Aug 1–5, 2015* (p. 24).

Weiss, M. W., Vanzella, P., Schellenberg, E. G., & Trehub, S. E. (2015). Pianists exhibit enhanced memory for vocal melodies but not piano melodies. *The Quarterly Journal of Experimental Psychology*, *68*(5), 866–877.

Wessel, D. L. (1973). Psychoacoustics and music: A report from Michigan State University. *PACE: Bulletin of the Computer Arts Society*, *30*, 1–2.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*(2), 45–52.

West, B., Welch, K., & Galecki, A. (2007). *Linear mixed models*. Boca Raton, FL: Chapman Hall.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*(2), 163–175.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, *58*(2), 315–330.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130.

Woodward, A. J., Macken, W. J., & Jones, D. M. (2008). Linguistic familiarity in short-term memory: A role for (co-) articulatory fluency? *Journal of Memory and Language*, *58*(1), 48–65.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*, 11475.